Section 3 Comparative Genomics and Phylogenetics

At the end of this section you should be able to:

- Describe what is meant by DNA sequencing.
- Explain what is meant by Bioinformatics and Comparative Genetics.
- Carry out a task using Bioinformatics
- Complete your own Phylogenetic tree using on-line software.

3.1 DNA Sequencing

So what is the next step after Gel Electrophoresis?
Well, if your samples are of good quality with a sufficient concentration of DNA, the samples can undergo DNA sequencing, a technique by which the exact order of nucleotides within a DNA molecule can be determined. The process has come a long way since the initial manual 2-D chromatography methods were developed and today, automated procedures are used which have the advantage of speeding up turnaround times. When you send your sample of DNA away for analysis, the sequencing result that you get back may look something like this:

ATGACCAACATTCGAAAATCTCACCCCTTAATAAAAATTATTAACAGCTCATTTATTGACCTCCCTGCCCCATCAAACATTTCATCTTGATGAAACTTTGGATCTCTCCTAG
GAATTTGCTTAGCACTACAAATTTTAACAGGACTATTTCTAGCTATACACTACACATCAGACACCACAACAGCTTTTAACTCTGTCACCCATATTTGCCGAGATGTAAACTAT
GGTTGAGTTCTACGCTACTTGCATGCAAATGGAGCCTCCATATTTTTTATCTGCCTATATCTCCATGTAGGACGGGGCCTTTACTATGGGTCCTATATATATACAGAAACCTG
AAATATCGGAGTTATTCTATTATTTGCTGTAATAGCAACAGCCTTTATAGGATATGTACTTCCATGAGGACAAATGTCTTTCTGAGGAGCAACAGTAATTACCAACCTGCTCT
CTGCAATTCCGTACATTGGAACAGACCTTGTAGAATGAATCTGAGGCGGCTTCTCTGTTGACAAAGCTACTTTGACCCGATTCTTTGCCTTTCACTTTTTACTCCCATTTA
TTATTGCAGCCATAGTCATAGTCCACCTCCTATTTCTTCACGAAACCGGATCCAATAACCCAACAGGAATCCCCTCCAACGCTGATATAATCCCCTTCCACCCCTACTATAC
AATTAAAGACATTCTTGGCCTGCTATTAATAATTACAGTCCTACTCATACTAGTACTATTCTCCCCCGACCTGCTAGGAGACCCTGACAACTACACACCAGCGAACCCACTAA
ACACCCCTCCCCATATCAAACCGGAATGGTACTTTTTATTCGCATATGCAATTCTACGATCAATTCCAAACAAACTAGGAGGAGTGTTAGCCCTAGTACTATCAATCCTTATT
CTAATTATCATTCCCCTACTCCACACCTCCAAACAACGCAGCATAACTTTTCGTCCCTTAAGCCAGTGCCTATTCTGACTATTAACAGCAGATCTATTCACTCTAACATGAAT
CGGAGGACAGCCCGTCGAACATCCATATGTCATCATTGGCCAACTAGCATCAATTCTTTATTTTTCTATTATCATTATCCTAATACCACTTATTAGCCTGATAGAGAACCACCT
ACTAAAATGAAGA

So what do all these letters mean and what can we do with this information?

All of the genes that have been sequenced are sent to a central database which acts like a huge library. Therefore, when you get your sequencing result like the one above, you can compare it against known results in the database and look for similarities and differences.

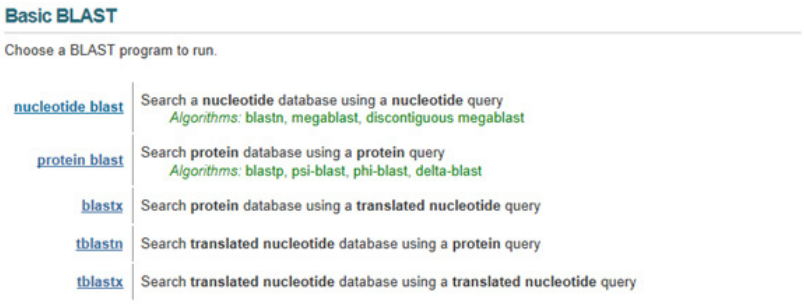
To do this manually would take a very long time and a lot of patience. Thankfully, we can use Bioinformatics. By using software known as BLAST (Basic Local Alignment Search Tool), the unknown or query sequence can be compared and contrasted against a database of known sequences in a matter of seconds.

Please note that the following tasks do not require any software to be downloaded onto your computer. All of the programmes are available online. To find out what species of bat the above unknown sequence (in blue) belongs to, you can use BLAST as follows:

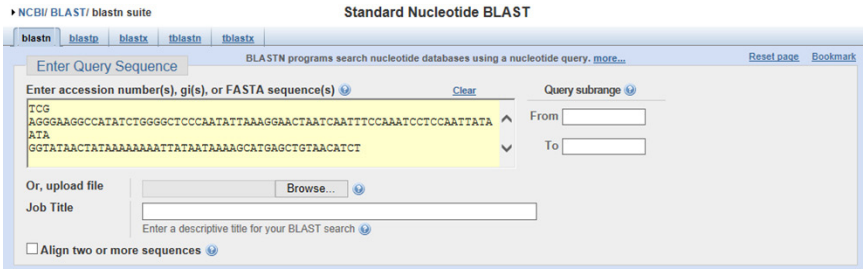
Log on to: http://blast.ncbi.nlm.nih.gov/Blast.cgi

(cont...)

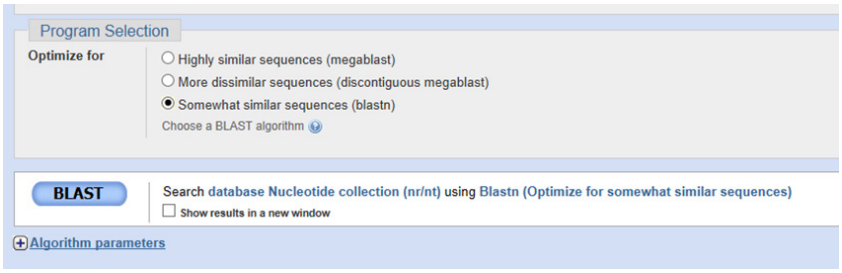
Scroll down the page and click on nucleotide blast.



Copy and paste the sequencing results (provided by your teacher) into the space provided in 'Enter Query Sequence'

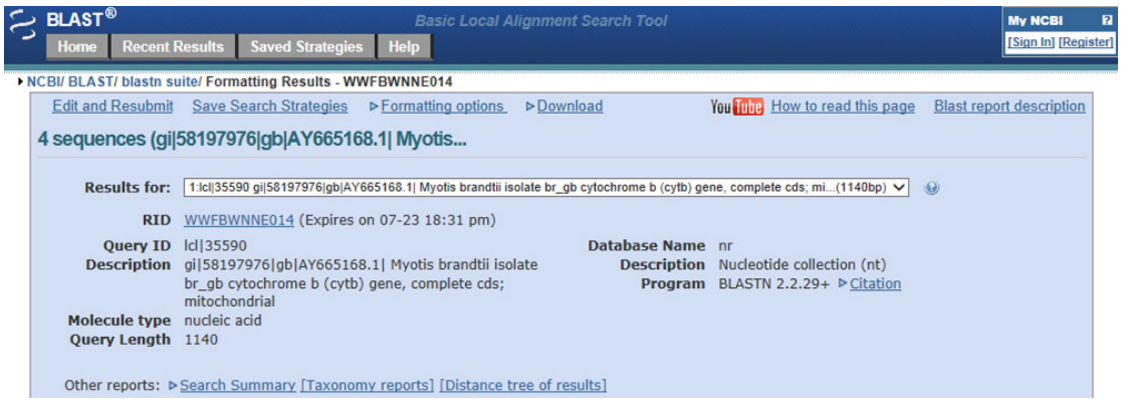


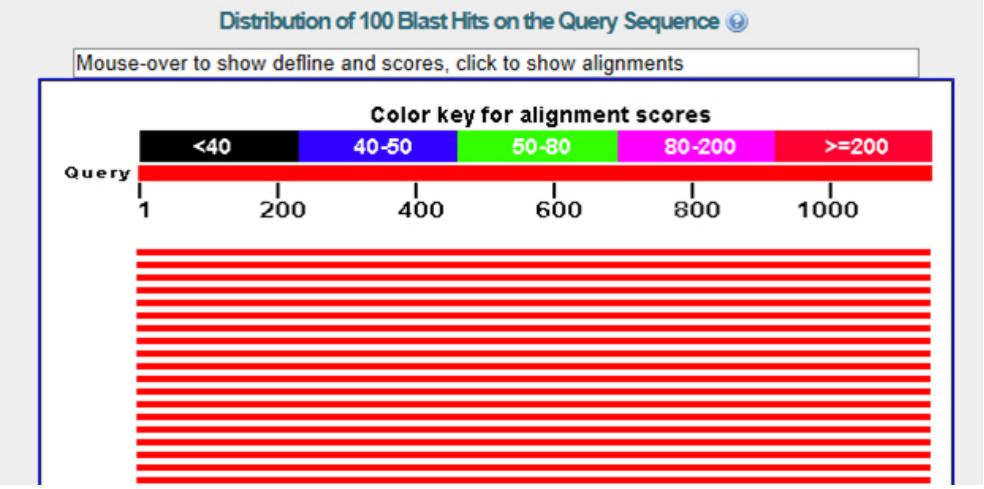
Scroll down the page and click 'Somewhat similar sequences'. Then click BLAST.



You'll need to wait for a few minutes while the programme is running and it is important not to refresh the page or hit the back arrow during this time.

The results will look something like this:





3 1 2

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Myotis brandtii cytochrome b (cytb) gene, complete cds; mitochondrial gene for mitochondrial produ	2057	2057	100%	0.0	100%	AF376844.1
☐	Myotis brandtii isolate 21 cytochrome b (cytb) gene, complete cds; mitochondrial	2051	2051	100%	0.0	99%	AY665139.1
☐	Myotis brandtii mitochondrial cytb gene for cytochrome b, specimen voucher NMP PB 916 (Nat. Mus	2042	2042	100%	0.0	99%	AM261886.1
☐	Myotis keaysi voucher JAG 286 cytochrome b (cytb) gene, complete cds; mitochondrial	1492	1492	100%	0.0	89%	JX130525.1
☐	Myotis keaysi voucher TK 13526 cytochrome b (cytb) gene, complete cds; mitochondrial	1483	1483	100%	0.0	89%	JX130449.1
☐	Myotis keaysi cytochrome b (cytb) gene, complete cds; mitochondrial gene for mitochondrial produc	1480	1480	100%	0.0	89%	AF376852.1
☐	Myotis mystacinus mitochondrial cytb gene for cytochrome b, complete cds, specimen voucher: OC	1471	1471	100%	0.0	89%	AB106605.1
☐	Myotis sp. KK0005 mitochondrial cytb gene for cytochrome b, complete cds, specimen voucher: KK	1462	1462	100%	0.0	88%	AB106609.1
☐	Myotis cf. albescens voucher TK 101479 cytochrome b (cytb) gene, complete cds; mitochondrial	1452	1452	100%	0.0	88%	JX130480.1
☐	Myotis cf. albescens voucher TK 101723 cytochrome b (cytb) gene, complete cds; mitochondrial	1452	1452	100%	0.0	88%	JX130479.1
☐	Myotis atacamensis mitochondrial cytb gene for cytochrome b, specimen voucher MVZ 168933 (Mu	1447	1447	100%	0.0	88%	AM261882.1
☐	Myotis cf. yumanensis voucher TK 48533 cytochrome b (cytb) gene, complete cds; mitochondrial	1443	1443	100%	0.0	88%	JX130483.1

So what do these results mean?

There are three important parameters that you need to look at:

1. The E value: This is a mathematical calculation and gives the probability of the result being a false match. What you are looking for here is a result of zero.
2. Ident: This is the percentage of nucleotides matching between your query sequence and the results in the database. Here you are looking for 100% for an exact match to an unkown species. If you do not know what your gene does, even Ident of 70% or above can still help you figure out your gene's function.
3. Query cover: This tells you how much of the sequence length has matched you query. Again, ideally you are looking for 100%.

Note: The accession number is the unique identifier assigned to each gene in the database.

FUN FACT: There are nine species of bat resident in Ireland. They are all insectivorous.

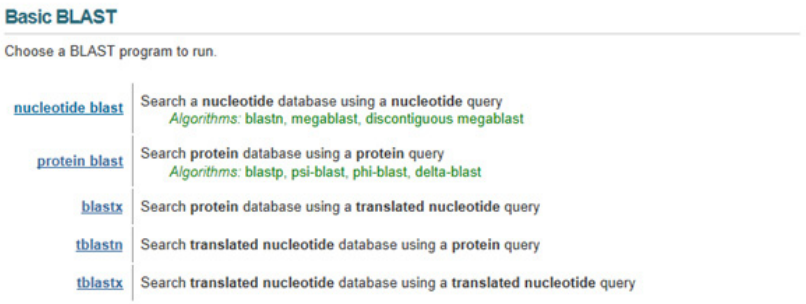
So from the above table it can be deduced that our unknown bat gene has been identified as belonging to the *Myotis brandtii* species.

STUDENT TASK:

While the bats native to Ireland eat only insects, some bats in other parts of the world eat fruit or other animals such as frogs. It is possible to extract DNA from bat faeces (e.g. poop, droppings) to help figure out what the bat's diet consists of. Your task is to use the software tools to analyse bat faeces. By doing this you should be able to find out what these bats ate for dinner. Good Luck!

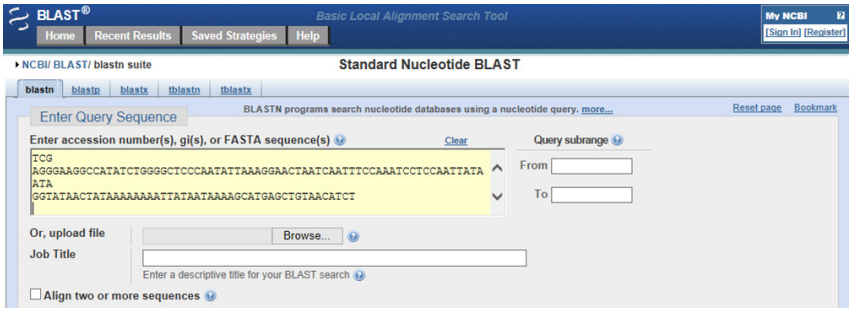
A sample of bat droppings which can be used for analysis of bat diet

Let's get started - log on to:
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Scroll down the page and click on nucleotide blast.

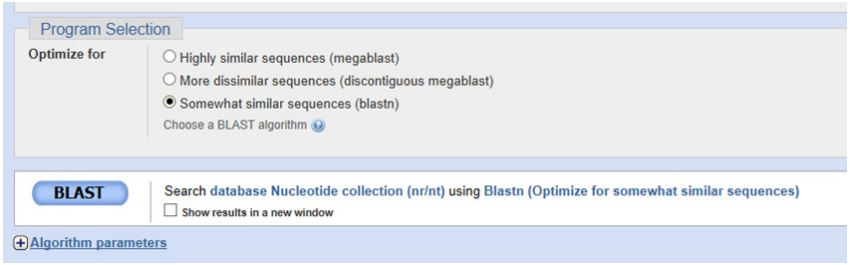


Cut and paste the **gene for unknown bat food 1** into the space under 'Enter Query Sequence' (your teacher will provide this).

STUDENT TASK (cont...):

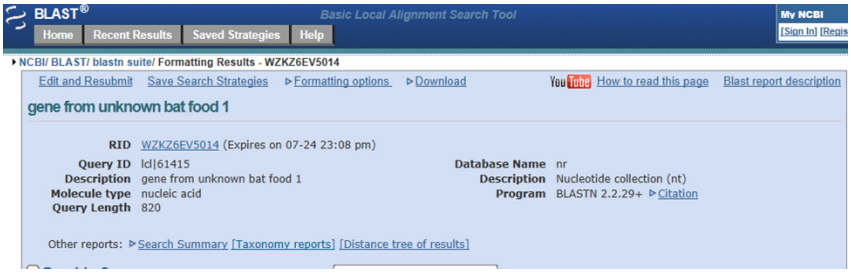


Scroll down the page and click 'Somewhat similar sequences'. Then click BLAST.



You'll need to wait for a few minutes while the programme is running and it is important not to refresh the page or hit the back arrow during this time.

The results will look something like this:



And if you scroll to the bottom of the page you will see the following table:

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Musa acuminata subsp. zebrina ribosomal protein S16 (rps16) gene, intron; chloroplast	1480	1480	100%	0.0	100%	FJ428139.1
☐	Musa rubra ribosomal protein S16 (rps16) gene, intron; chloroplast	1460	1460	100%	0.0	99%	FJ428132.1
☐	Musa acuminata subsp. malaccensis chloroplast complete genome, biomaterial CIRAD:930	1449	1449	100%	0.0	99%	HF677508.1
☐	Musa rosea ribosomal protein S16 (rps16) gene, intron; chloroplast	1449	1449	100%	0.0	99%	FJ428131.1
☐	Musa textilis plastid, complete genome	1433	1433	100%	0.0	99%	KF601567.1
☐	Musa yunnanensis ribosomal protein S16 (rps16) gene, intron; chloroplast	1425	1425	100%	0.0	99%	FJ428143.1
☐	Musa laterita ribosomal protein S16 (rps16) gene, intron; chloroplast	1424	1424	97%	0.0	99%	FJ428136.1
☐	Musa banksii ribosomal protein S16 (rps16) gene, intron; chloroplast	1411	1411	97%	0.0	99%	FJ428138.1
☐	Musa acuminata subsp. burmannica ribosomal protein S16 (rps16) gene, intron; chloroplast	1409	1409	97%	0.0	99%	FJ428135.1
☐	Musa siamensis ribosomal protein S16 (rps16) gene, intron; chloroplast	1409	1409	96%	0.0	99%	FJ428134.1

STUDENT TASK (cont...):

Now use the internet to find out the common name for *Musa acuminata*. This will give you some information about the diet of this bat.

Question: Is this an Irish bat?

Repeat the exercise using **gene for unknown bat food 2** and complete the table.

Question: Is this an Irish bat?

	gene for unknown bat food 1	gene for unknown bat food 2
Scientific name for food found in bat faeces	*Musa acuminata*	
Common name for food found in bat faeces		
This bat has a diet that consists of:		

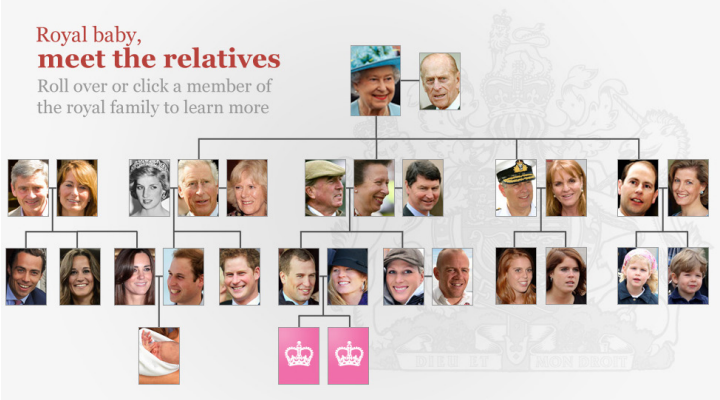
FUN FACT: The smallest bat is the Kitti's Hog-nosed bat which is also known as the Bumblebee Bat. It can be found in parts of Thailand and Burma. This species is considered as the smallest mammal in the world.

3.2 Phylogenetics

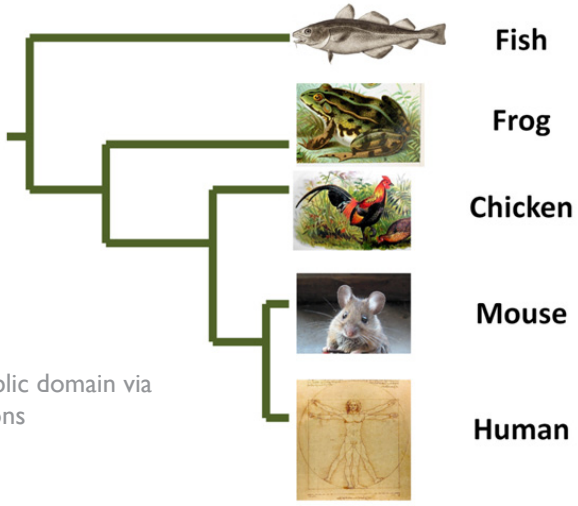
Phylogenetics is the study of how different species are related. We can say that phylogenetics is looking back in the evolutionary time to trace the history of species or a group of species. This evolutionary history of a group of organisms can be represented in a diagram called a phylogenetic tree.

Let's plant a phylogenetic tree

What is a phylogenetic tree? Well, have you ever seen a family tree? Or perhaps you had to prepare a genealogical tree of your family for some school project before? Actually, a phylogenetic tree is a quite similar. Family tree shows how family members are related to each other – it includes both really close relatives (like your mum and dad) and more distant ones (like this far cousin living in Australia, who you've never met). A phylogenetic tree, on the other hand, presents how different species are related to each other (who is our closer relative: monkey or a horse?).



This family tree presents the history of British royal family.
Screenshot from http://edition.cnn.com



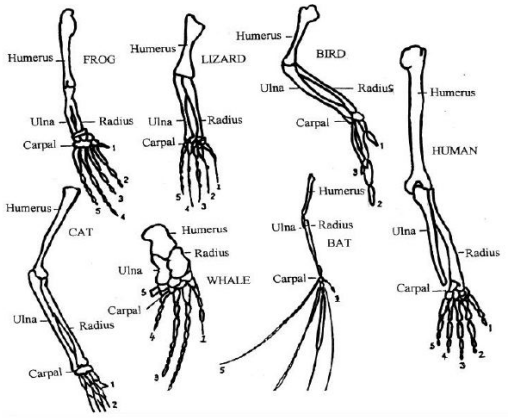
All images used: public domain via Wikimedia Commons

Phylogenetic tree! It shows how different organisms are related.
Are humans more closely related to fish or frog?

When you prepare a family tree you can use different historical documents (like marriage and birth certificates). You can also use genetic tools (if you have, for example, a blood sample) to determine if two people are closely related. When scientists construct phylogenetic trees they can also use different sources of information. They can look at how similar animals are or use a more modern approach – look how similar are certain genes (or other DNA segments) between the species. So both MORPHOLOGY and GENES are useful for constructing a phylogenetic tree!

When Constructing a Phylogenetic Tree - you can use both:

MORPHOLOGY



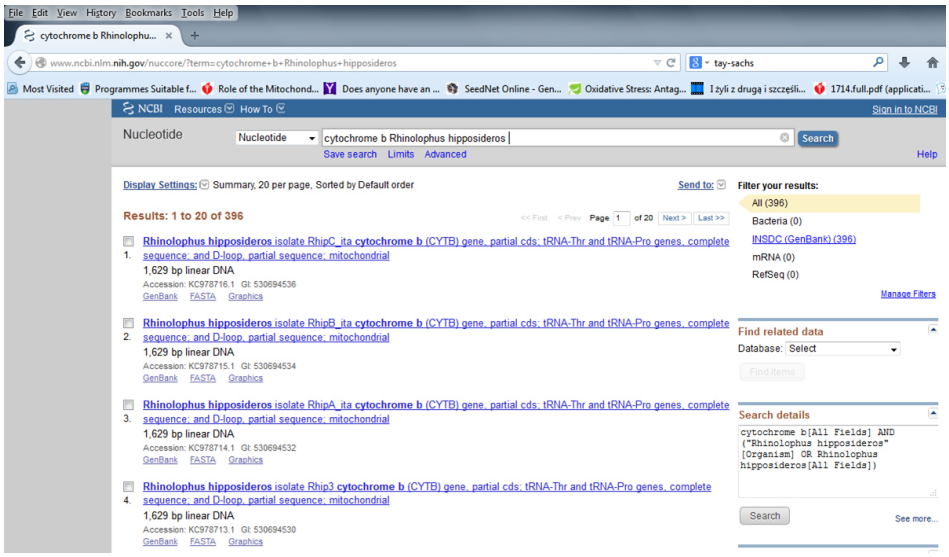
"Anatomy and physiology of animals Various vertebrate limbs:"
Licensed under CC BY 3.0 via Wikimedia Commons

GENES

BAT CASE STUDY:
Let's take a look at a gene called cytochrome b, which is found in the mitochondria. We are going to use this gene to construct a phylogenetic tree to study the evolutionary history between several different species of bat (Lesser Horse-shoe Bat, Black Flying Fox, Bumblebee Bat, Little Brown Bat, Flying Fox, Common Pipistrelle, Brown Long-Eared Bat and Leisler's Bat).

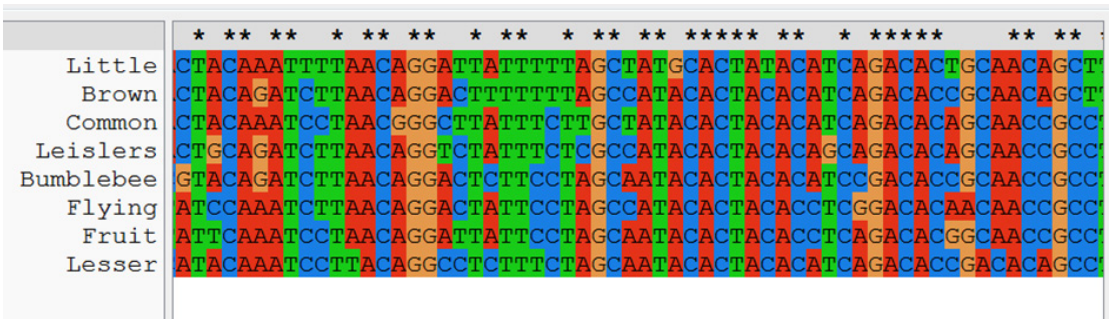
1. First we need to find the cytochrome b gene sequences for all these bat species.
The gene sequences can be obtained using NCBI database. NCBI is a great source of biological information. You can search the available nucleotide sequences at:
http://www.ncbi.nlm.nih.gov/nuccore

Note that usually the scientific names of species are used in such databases – always good to know both common and scientific name of the species you want to study! For example, the Latin name of Lesser Horseshoe Bat is *Rhinolophus hipposideros*. That's how I found the cytochrome b gene sequence of the Lesser Horseshoe Bat. The search query I used was 'cytochrome b *Rhinolophus hipposideros*'. ***Why don't you try to find seqeunces from the other species mentioned above!*** **Remember you should find out what the latin scientfic name is first before searching for the cytochrome b sequence.**
The correct cytochrome b seqeunces will be provided by your teacher (they are in the teacher's pack).



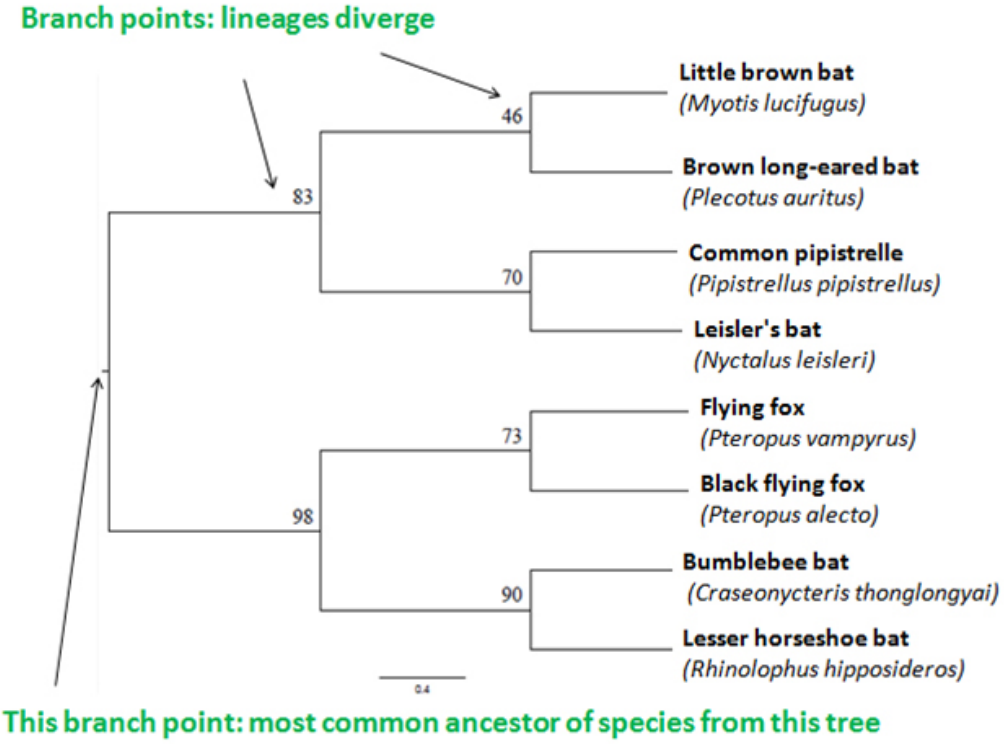
2. Now let's compare these sequences

We can use an online software (such as Clustal: http://www.ebi.ac.uk/Tools/msa/clustalo/) to carry out a process known as alignment. Cut and paste your seqeunces into Clustal, then hit submit (see below for more detail). The cytochrome b genes from the 8 bat species will be arranged in a way that allows for similar regions to be identified. Similarity between the cytochrome b regions may be due to evolutionary relatedness. The more similar the sequences are – probably the more closely related are two species.



3. Construct a phylogenetic tree based on the cyt b sequences

From the above data it is possible to create a phylogenetic tree using a programme called MEGA 6 (http://www.megasoftware.net/). MEGA offers many different methods for construction of a phylogenetic tree from the aligned sequences. See an example tree overleaf.



How can we read this tree?

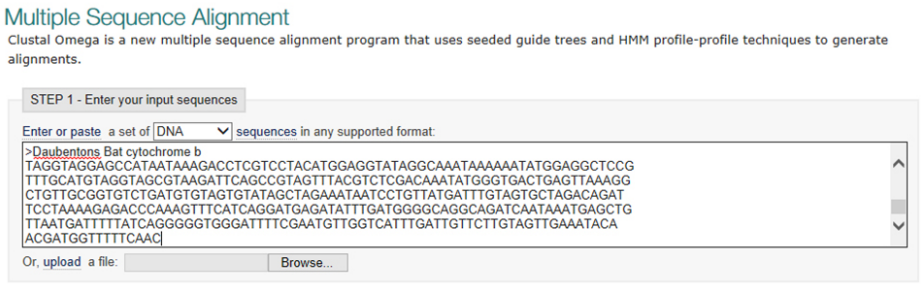
At the right hand side, you can see all the names of the species that have been analysed. These are the 'leaves' of the tree. Each leaf has its own 'branch'. Branch points (tree nodes) tell us where two lineages diverged, explaining the evolutionary history of species presented on this tree. From this tree, you can for example, read that the Bumblebee bat and Lesser Horseshoe bat shared an immediate common ancestor. We don't know who this ancestor was, but at some time in the past it branched off into two separate evolutionary paths resulting in the Bumblebee bat and the Lesser Horseshoe bat.

Do I trust this tree? Let's BOOTSTRAP!

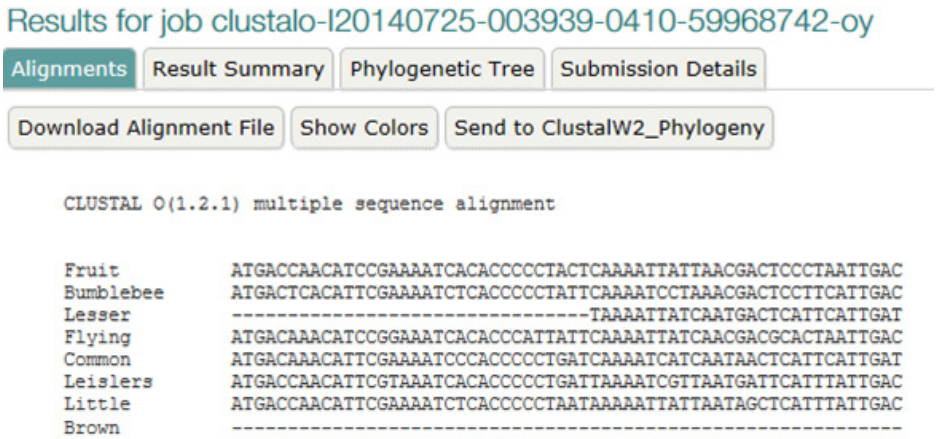
A phylogenetic tree is a model – it presents the relatedness of species based on available data. Different trees can be prepared for the same group of species. There are different statistical methods to assess how good our tree is. Did you notice the numbers displayed at each branch point of this tree? They are the results of a test called Bootstrap. These values tell us how certain we are of each connection, the higher number - the better. For example, there is a number '90' displayed at the branch point at which Bumblebee Bat and Lesser Horseshoe bat diverged. It means that 90 out of 100 times we prepare this tree this node will be the same. Take a look at the tree yourself and use the Bootstrap values to assess different nodes of this tree. Which one you would trust the most? Which one is the least reliable?

STUDENT TASK – CREATE YOUR OWN PHYLOGENETIC TREE

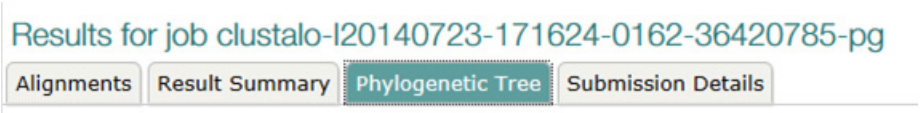
1. Log on to: http://www.ebi.ac.uk/Tools/msa/clustalo/
2. Copy and paste all genes from the 8 species of bat into the space provided.



3. Scroll down the page and click submit. Our results should look like this:



4. Now click on Phylogenetic Tree



5. Scroll down the page to see your tree that you have created. Does it look the same as the tree created above? Why do you think that there might be differences?

You have probably noticed that the phylogenetic tree produced by Clustal online tool is quite different than the one we generated using programme called MEGA. The phylogenetic tree generator tool offered by Clustal will never produce an accurate phylogenetic tree. The major function of Clustal is to align the sequences and the tree is used to guide this alignment. More specialized programmes, like MEGA, should be used to build (and evaluate) more accurate phylogenetic trees.

3.3. Comparitive Biology and Its Uses

Due to all the advancements in modern molecular technologies and the software available, it is now possible for us to sequence large DNA fragments or even your whole genome (all of your DNA in your cell). The process is getting faster and cheaper, which means we can build a huge databases containing DNA sequences from different species.

PGenome 10k Project – Unveiling animal diversity

A project aimed at sequencing 10 000 vertebrate genomes to understand how the complex animal life evolved through changes in DNA. Watch a video presenting this genetic Noah's Ark:

https://genome10k.soe.ucsc.edu/

Mind that DNA sequences alone are only beginning of the story – the big challenge is to figure out how genomes work. Comparative genomics is a field of biological research in which the genomic features of different organisms are compared. Comparative genomics allows the construction of the evolutionary tree of all living organisms (Tree of Life) – with more detail and accuracy than ever before!

By making comparisons between similar DNA sequences from human and from different species, we can also learn a lot about how our own genome works.

Comparative genomics and medicine

Comparative genomics is an amazing tool to predict if a single nucleotide variation, SNV (a change in a single nucleotide), in certain site in the genome is harmful, not-harmful or perhaps beneficial. Each of us have thousands of SNVs in our genomes, could this make us different from each other?.

Think about your colleagues from school. What if I tell you that at certain site in the genome all of your school friends have a nucleotide Guanine (G), and you are the only person with a Thymine (T) instead? Is it good or is it bad? It's hard to say if we don't know what this region in the genome does!

T Instead of G: Good or Bad?

G G G
G T G
G G G
G G

How would you feel if I tell you that this site lies within the gene needed for vision? Let's sequence this gene in other mammals and investigate if they have a G or a T at this site? What can that tell us?.

SNP in vision related region

Humans: G **Bats: T**
Giraffes: G
Gorillas G
Elephants: G
Dogs: G

Bats can do plenty of amazing things – but some don't see very well. Knowing that of all mammals, the ones having T in this particular site do not have the greatest vision may suggest that that this change in humans (from G – like in most of human population to T – like bat) can indicate increased risk of a vision related disease. We can make predictions like that thanks to the power of comparative genomics!

Perhaps in the future, comparative genomics will be used on regular basis in medicine.