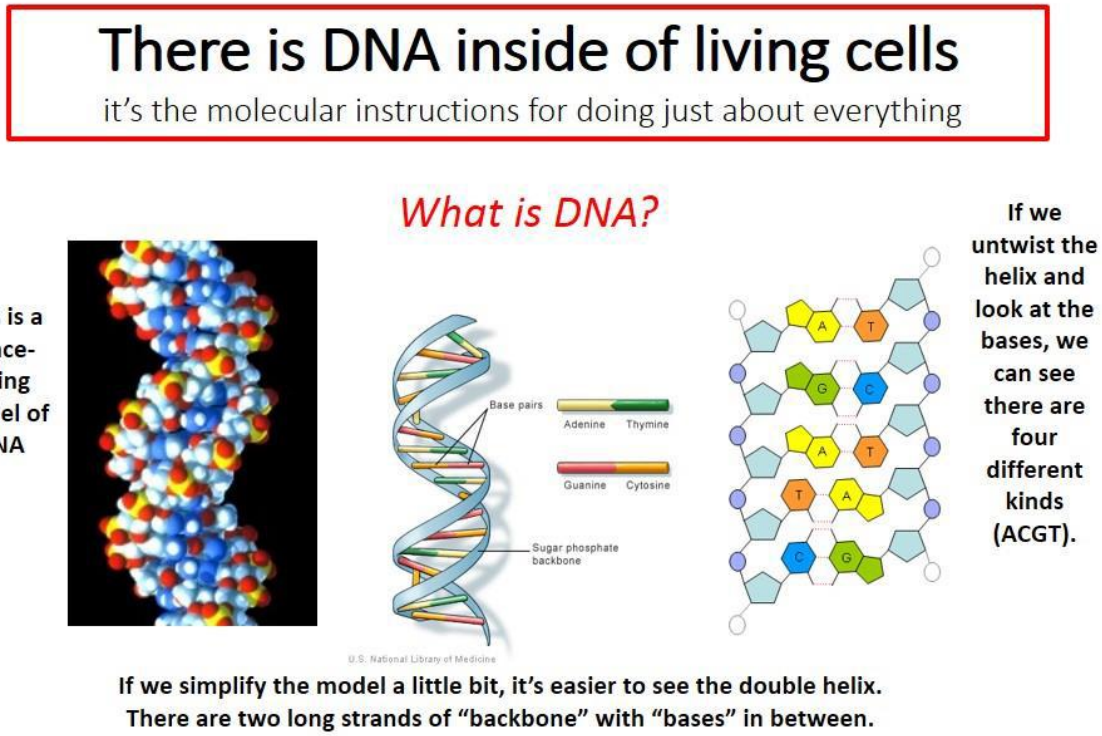
DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

Part A: DNA Sequence Evolution Simulation

Background:

DNA is a polymer. That is, a strand of DNA is a chain of many linked simple building blocks. There are only four different types of building blocks, or nucleotides, in a molecule of DNA. All four include the same backbone made of alternating sugars (deoxyribose) and phosphates. The only difference between the four building blocks is the part called the "base". The four different bases are called adenine, cytosine, guanine and thymine, or A, C, G, and T, for short.

A complete molecule of DNA includes two strands of these repeating blocks twisted into a double helix. The bases on the two strands are held together by hydrogen bonds. Each base can only bind to one other type of base. A always pairs with T, and G always pairs with C. This is what allows living cells to replicate their DNA before cell division. This base pairing also means that we can describe a DNA sequence in short hand, just by specifying the order of building blocks on one of the two strands. For example, if I tell you that one strand of a DNA sequence is GGAATTCC, you automatically know the other strand is CCTTAAGG. During replication, each strand serves as the template to make a copy of the other strand. By copying the two strands of a single molecule, a cell can produce two identical copies of the original DNA molecule…unless a mutation occurs, in which case the two new copies will be slightly different from each other. A common type of mutation replaces a single building block with one that contains a different base. For example, an A can change to a G.



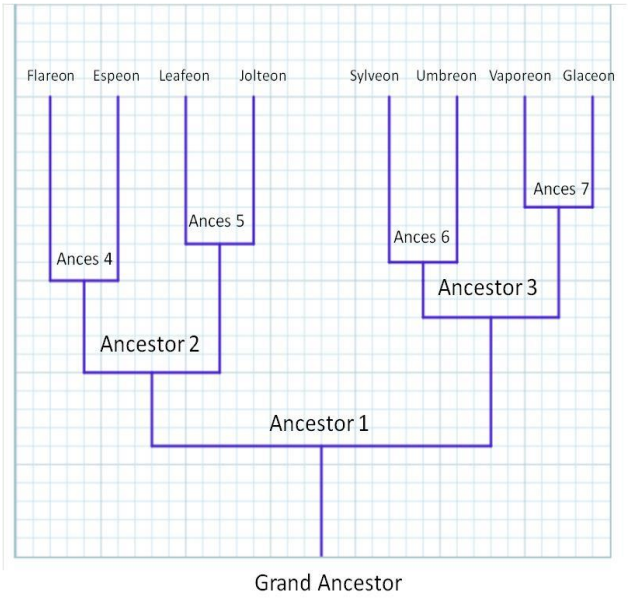
Overview: In this exercise, you are going to simulate evolution of DNA sequences. Simulations are an effort to imitate a process, in this case, mutation of DNA sequences over evolutionary time. Simulations play an important role in science. For example, many new methods in bioinformatics (a field of inquiry dedicated to developing new computational methods to analyze biological data) are tested on simulated data sets. An advantage of working with simulated data is that many aspects of the correct answer are known, and often that is not the case with real data. Testing a new analysis method on simulated data allows scientists to make statements like "If real data arises through a natural process that is accurately imitated by this simulation, my new method of analyzing it gets the correct answer."

Instructions:

1. Start with group of 8 students.
2. Simulate evolution starting with a 100 base ancestral DNA sequence. This is already provided to you in page 1 of the Excel file and is labelled Grand Ancestor. You will work on the "Grand Ancestor" sequence.
   Note: Page 2 of the Excel file has a sequence called "Pidgey". **DO NOT** use that sequence. You will use it later in PART B.
3. You will now draw three random numbers (you can use the "=randbetween(..,..) function in excel).
4. Draw the first random number between 2-7 (both numbers inclusive). This will determine the number of mutations that will occur in the DNA sequence before the group of 8 will split into two groups of four each.
5. Draw the second random number between 1-100 (both numbers inclusive). This will determine the position in the DNA sequence that will undergo mutation.
6. Draw the third random number between 1-4 (both numbers inclusive). This will determine the new nucleotide that will replace the existing nucleotide at the position specified in point (5) above. Use the following code 1-A, 2-T, 3-G, 4-C. If the random number that you have picked does not cause a change at the specified position (say, that already there is an A at the specified position and you pulled number 1. Replaceing A with A does not lead to a mutation.), pick a random number between 1 and 4 again till it causes a change in the base.
7. Repeat points (5) and (6) till the number of mutations specified in point (4) are completed.
8. Note down (or save in an Excel sheet) the evolved sequence. Give the sequence a name.
9. Now split your group into two groups of four students each. Each group will take the evolved sequence and work on it independently.
10. Each group will independently repeat points 5 to 7.
11. The groups will keep splitting into two till there are 8 groups of one person each.
12. Note down (or save) the sequences at each point.
13. At the end, there should be eight evolved sequences.
14. **EXTREMELY IMPORTANT: The total number of mutations along any line should be exactly 25.** This means that when you are working alone, the number of

mutations that you have to make in the DNA sequence is already fixed by the previous three choices (*ie* the random numbers that you pulled when the group size was 8, 4 and 2).

15. **RANDOMLY** assigns one of the following names to each of the eight sequences- (Vaporeon, Jolteon, Flareon, Espeon, Umbreon, Leafeon, Glaceon, Sylveon). Please note that the assignment of names **HAS TO BE RANDOM**.
16. Since you have all the information, draw a Phylogenetic tree representing the evolutionary relationships between different sequences. An example is given below. Please note that the tree is to scale with each unit on the graph paper being equal to one mutation. Additionally, total number of mutations along any path of the tree from bottom to tip is 25.



Modified from:

The UPGMA portion was inspired by a lesson (http://csunplugged.org/wpcontent/uploads/2014/12/PhylogeneticsUnplugged.pdf) developed by Tru Women in Computer Science (TWiCS) http://twics.truman.edu Truman State University, Kirksville, Missouri, USA (Mariya Davidkova Amy McNabb Molly Smith Julia Stefani Michelle VanKleeck Allie Wehrman and Jon Beck) and contributed to csunplugged.org by Katrin and Jim Becker, Mount Royal College; Calgary, Alberta, Canada. The dice-based simulation, instructor's guide and student handouts were created by Nicole T. Perna and Jeremy D. Glasner from the J.F. Crow Institute for the Study of Evolution at the University of Wisconsin – Madison ***for non-commercial purposes only.*** . CSUnplugged and this exercise are distributed under a Creative Commons BY-NC-SA License, which makes it easy to copy, adapt and share https://creativecommons.org/licenses/by-nc-sa/3.0/.

Part B: Constructing a phylogenetic tree using UPGMA method.

Instructions:

1. You will attempt to reconstruct the phylogenetic tree of sequences evolved by another group.
2. Save the eight sequences that you evolved in Part A with random names as suggested before.
3. Give the sequences that your group evolved to another group and take the eight sequences that they evolved.
   Important: Make sure that you only take (or give) the eight sequences. Do not take the ancestral sequences or the tree developed by the other group.
4. Copy the "Pidgey" sequence provided to you on page 2 of the excel file. Therefore you now have a collection of nine sequences.
5. Do a pair-wise comparison of the nine sequences and note down the number of nucleotide differences between each sequence pair.
6. Fill in the table provided below to construct a 'Distance Matrix".
7. Detailed step-by-step instructions for constructing a phylogenetic tree using UPGMA method con be found here (http://www.slimsuite.unsw.edu.au/teaching/upgma/).

Distance Matrix: Fill in the number of nucleotide differences between the sequences.

	Pidgey	Vaporeon	Jolteon	Flareon	Espeon	Umbreon	Leafeon	Glaceon	Sylveon
Pidgey									
Vaporeon									
Jolteon									
Flareon									
Espeon									
Umbreon									
Leafeon									
Glaceon									
Sylveon									

Part C: Constructing Phylogenetic Tree using real DNA sequences.

Introduction: A phylogenetic tree is a visual representation of the relationship between the organisms being considered for making the tree. It depends mainly on two different parameters. First, the data that is used to make the tree and second the method by which it is made. The data can be from morphology, protein sequence, DNA sequence or a combination of them. Today, we are going to use DNA sequences. The other factor is the method in which a tree is made. There are many different statistical models and hundreds of programmes to build a tree and today we are going to use a programme called Mr. Bayes which uses Bayesian statistics to make a tree.

What will you need?
Download the following programmes: Bioedit, Mr. Bayes and Treeview. Bioedit is a DNA alignment software and you would need to align your DNA sequences before making a tree. Mr. Bayes is a phylogenetic programme by which you are going to make the tree. Treeview is another programme by which you will visualize the tree and save it for printing. The DNA sequences to be used will be given to you.

What do you need to do for this practical?

1. Align the DNA sequences and show it to any of the instructors.
2. Run the alignment in Mr. Bayes to make a tree and show it to any of the instructors.
3. Hand draw the tree and label the nodes with the posterior probabilities (we will explain it in the class) and show it to any of the instructors.
4. Visualize the tree in Treeview and show it to any of the instructors.
5. Play around with the many different kinds of trees that can be made through Treeview using the same data.
6. In the next class your laboratory note book must come with the write up of this lesson along with a print out of the tree with hand written posterior probabilities for each node.

Doing an alignment:

1. You have been given a file (Haem.fas) which has partial DNA sequences from the haemoglobin alpha chain gene from a bunch of different animals.
2. Open Bioedit. (File> New Alignment>Open>locate where the file Haem.fas is). You might have to indicate the type of file your DNA sequences are in. FYI, it is a .fas file. Indicate that in the drop down menu.
3. Visualize the sequences. How many animals do you see?
4. Now you will have to make the alignment.
5. Go to Edit>select all sequences. All the sequences in your file should now be selected.
6. Go to accessory application>ClustalW Multiple Alignment>Run ClustalW. A window will appear. Click OK.
7. ClustalW is a programme that aligns DNA and protein sequences.
8. After aligning a new window will appear with the alignment. Save this file as Haem_aligned.fas.
9. View the alignment as “View conservation by plotting identities to a standard as a dot”.
10. Now you will have to edit your alignments. Switch the mode button now to Edit.

11. Now you can edit your sequences using the delete and backspace buttons of your laptop.
12. If you would like to insert anything then just go the position and right click.
13. Delete all the regions that have missing data and save the file.
14. Show it to any of the instructors.

Running a phylogenetic tree:

1. Locate the Haem.nex file and put it in the same folder as Mr. Bayes.
2. This is the same file that you have aligned but is in a different format which Mr. Bayes likes!
3. Click on the mrbayes_x86 icon. Mr. Bayes window will appear.
4. Type exe Haem.nex. Mr. Bayes will now process the file.
5. At the Mr. Bayes> prompt, type lset nst=6 rates=invgamma
6. At the Mr. Bayes> prompt, type mcmc ngen=1000000 samplefreq=1000 printfreq=1000 diagnfreq=10000.
7. The analysis will start and Mr. Bayes will print a bunch of files in the same folder. Do not delete these files!
8. When it is done it will ask whether to continue with the analysis. Type no.
9. Type sump.
10. Type sumt.
11. Mr. Bayes will now show you the phylogenetic file.
12. Draw a hand diagram of the tree with the posterior probabilities. These are numbers in each node. They tell you how statistically robust your analysis has been.
13. Show it to the instructors.

Visualizing a Phylogenetic tree:

1. Open Treeview. File>Open and select Haem.nex.con.tre file from the Mr. Bayes folder.
2. In Treeview go to Tree>Define outgroup. A window will appear. Select pig from the ingroup list and put it in the outgroup list. Click OK.
3. Go to Tree>Root with Outgroup. Do you see any changes?
4. Now play around with the many kinds of trees.
5. Keep changing the outgroup and see what changes you see in the tree.
6. Show it to your instructors.

Links:

1. Mr Bayes: (http://mrbayes.sourceforge.net/index.php)
2. BioEdit: (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)
3. a. Fig Tree (http://tree.bio.ed.ac.uk/software/figtree/)
   b. Tree View (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html)

P.S. If you ever need any Evolution related software, look here:
(http://evolution.genetics.washington.edu/phylip/software.html)