Science Fair Speech

This is my project, “A Phylogenetic Analysis of the HIV-1 and HIV-2 Strains.” This particular project focused on strains prevalent in the region of West Africa, which has long been considered the origin of the virus. HIV, human immunodeficiency virus, (POINT TO DIAGRAM) (made of two main subtypes HIV-1 and HIV-2 with the former being much more widespread and pathogenic) is becoming a larger and larger problem in our world, and after years of no definitive treatment, millions have perished after being infected with it and getting AIDS, Acquired Immunodeficiency Syndrome. HIV is known to have evolved from SIV, simian immunodeficiency virus, which is a related disease in monkeys that spread to humans from contact.

The reason I chose to create phylogenetic tree (which, by the way represents relationships between species using various techniques and is represented via discrete branches) on the HIV virus in the West African region was taken is because of two main reasons. The first reason is that after looking through over 30 trees in 50+ research papers, the overwhelming majority are focused on the Western world, as that is where most of the research centers are located. The second reason is that every single one of these trees came with low bootstrap or posterior probability (which will be **EXPLAIN LATER**) or simply did not publish it, meaning very few trees exist with the support necessary to draw accurate conclusions about HIV’s relationships with other similar viruses.

The hypothesis was that I could produce very strong trees using phylogenetic software in the West African region with the expectation that old relationships from past trees would be found as well as new relationships being discovered.

I set out, using a variety of different materials. First and foremost was a modern computer, and I used my very high power laptop. I gathered nucleotide sequence data from the Los Alamos HIV database, and processed it through a variety of programs to produce these trees, most important of which are SeaView, MAFFT, PAUP and the BEAST suite. You may already notice that the two trees displayed come from PAUP and BEAST, which will come up later.

The methods used came in three main steps: 1) Securing and Aligning DNA, 2) Producing the Trees and 3) Revising the Gene Pool.

Securing the DNA meant going to the databases and gathering the HIV-1, HIV-2 and SIV sequences of about 900 nucleotides (A, G, C, or T) to produce a tree. The sequences would then be aligned using SeaView and MAFFT. The reason sequences must be aligned is because the tree is created using differences between sequences; if the difference between two sequences is great, then it is likely they are not closely related. If they are similar, it is likely they are closer (show on the board). Hence, alignment is necessary as if one deletion is made, it can mess up two otherwise very similar sequences (demonstrate with hands) Hence, it provides a more accurate representation of the differences.

Second was the actual production of the trees. PAUP, the first program used, used a technique known as parsimony to make the trees, which is the idea that the tree with the least number of base change events required is most likely the best tree. The second program used was BEAST, which used a combination of Bayesian inference, prior probabilities and Monte Carlo Markov Chains to produce a strong tree. The MCMC chain was set at 20,000,000 and took almost 1 and half days to complete!

Finally, the gene pool was revised about 50 times during the entire procedure. This was done by choosing a set of sequences and running it in PAUP (which is an incredibly fast program) and seeing how it fared. A final gene pool was created and run in BEAST only a few times, which was the best option given the time it took to run.

AND AFTER ALL THAT, THE FINAL TREES WERE PRODUCED! Many hours were put into them, and they are incredibly well supported. The first tree produced by PAUP was subjected to bootstrap testing, which essentially checks the number of times a certain relationship is the same after removing one branch and then reforming the tree again. This is measured by percentage, and if you look on your tree, nearly every branch is above 90, meaning very good support! The BEAST tree on the other hand measures strength of branches with probability, meaning a value closer to 1 is good. A close look at the BEAST tree reveals that it too has incredible support, with all but one branch having a probability of 0.99 or above!

READ OFF CONCLUSIONS

READ OFF APPLICATIOSN

SUMMARIZE FUTURE WORK

INTRODUCE BIB YOU DONE!