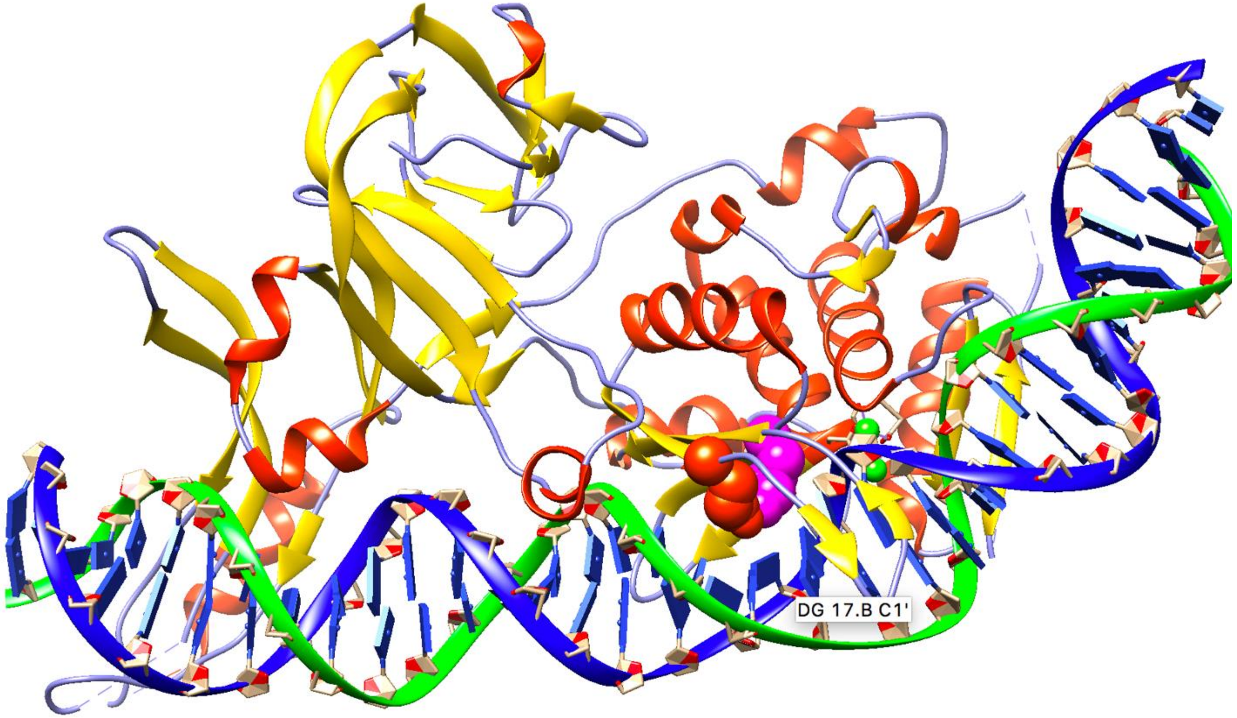
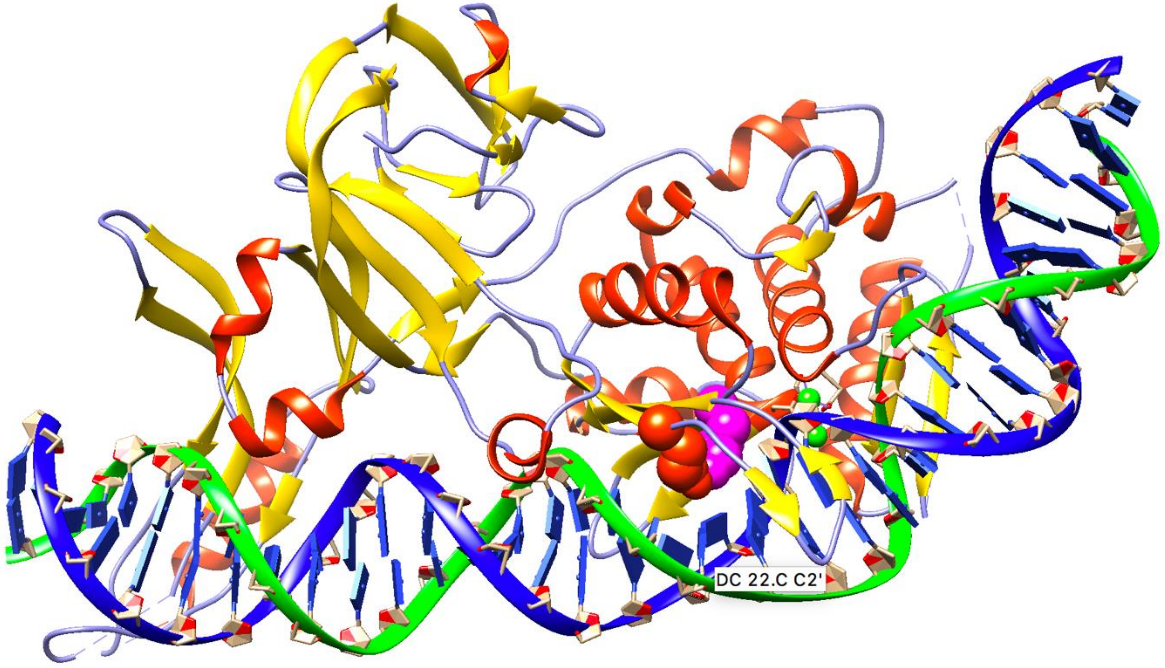
# Class 6
# MCB 3421 - 2020

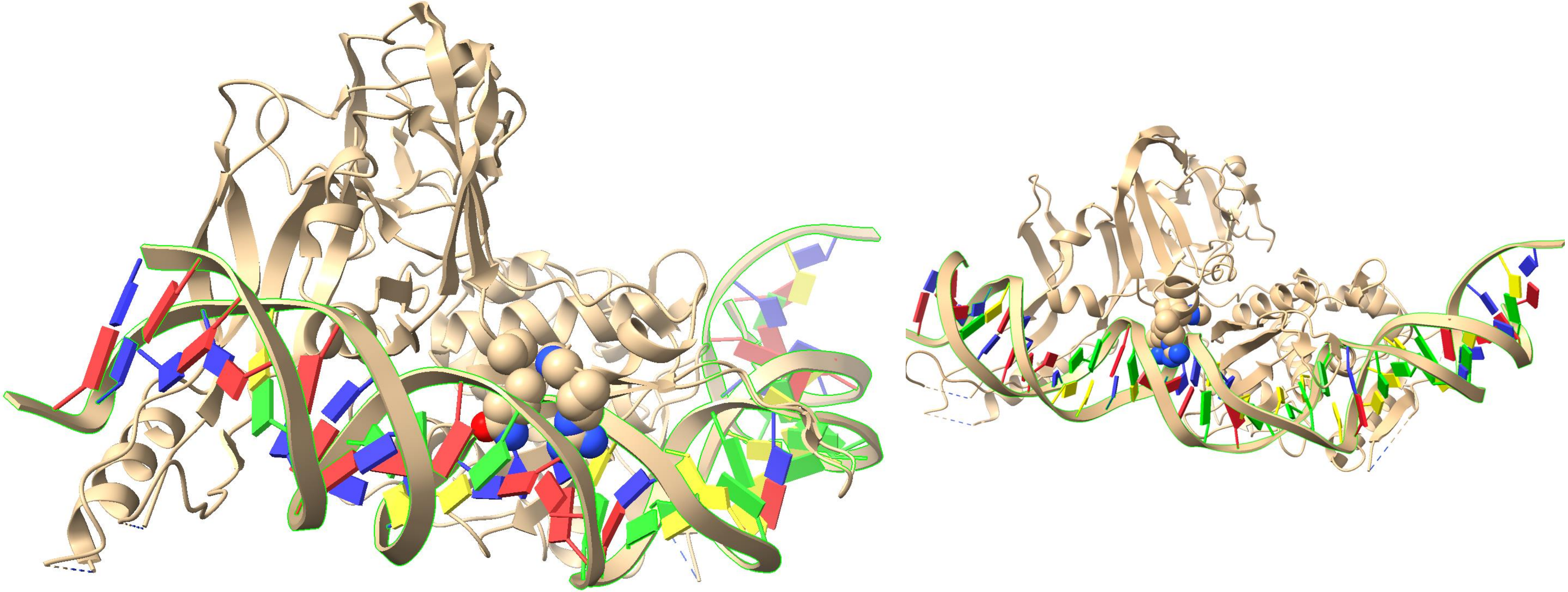
Assignment and Jukes Cantor

# Computer lab

Interactions between aa and DNA: The question in that lab asked for base pairs, i.e. part of the DNA that interacts with aa, Lys 340: magenta, Glu366: red







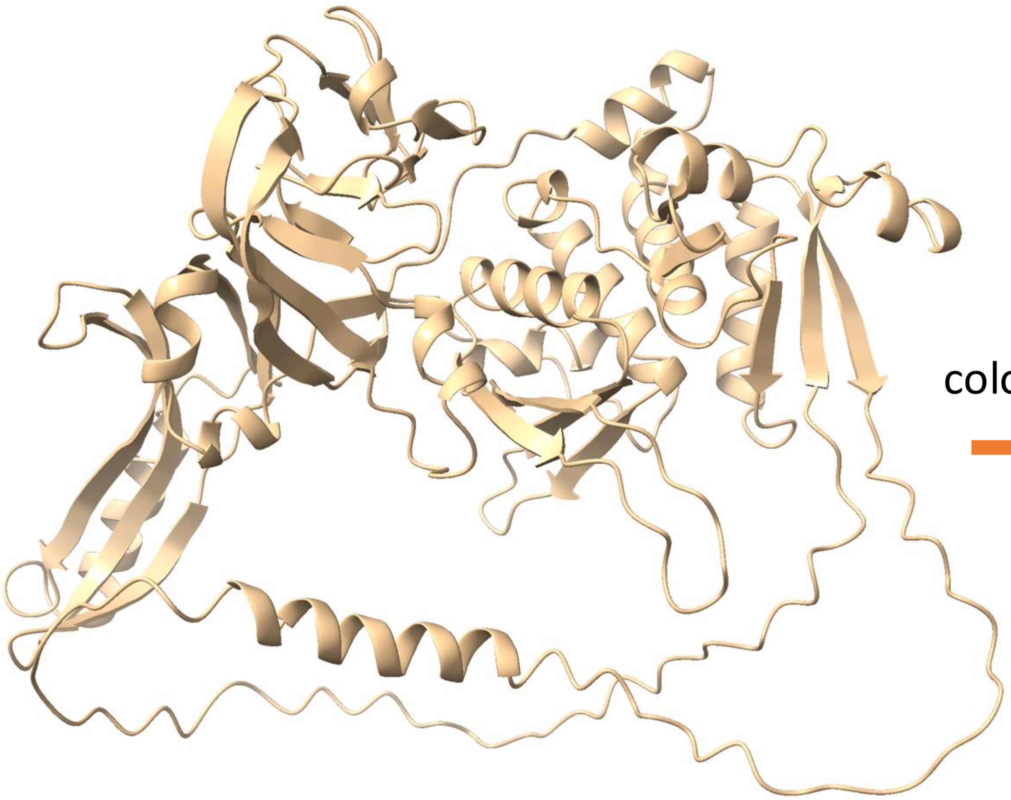
Most inetractions between the HE and the DNA are with the major groove; however, Arg 57 and Gln 55 probe the DNA’s minor groove,

# Coloring structure by confidence of prediction

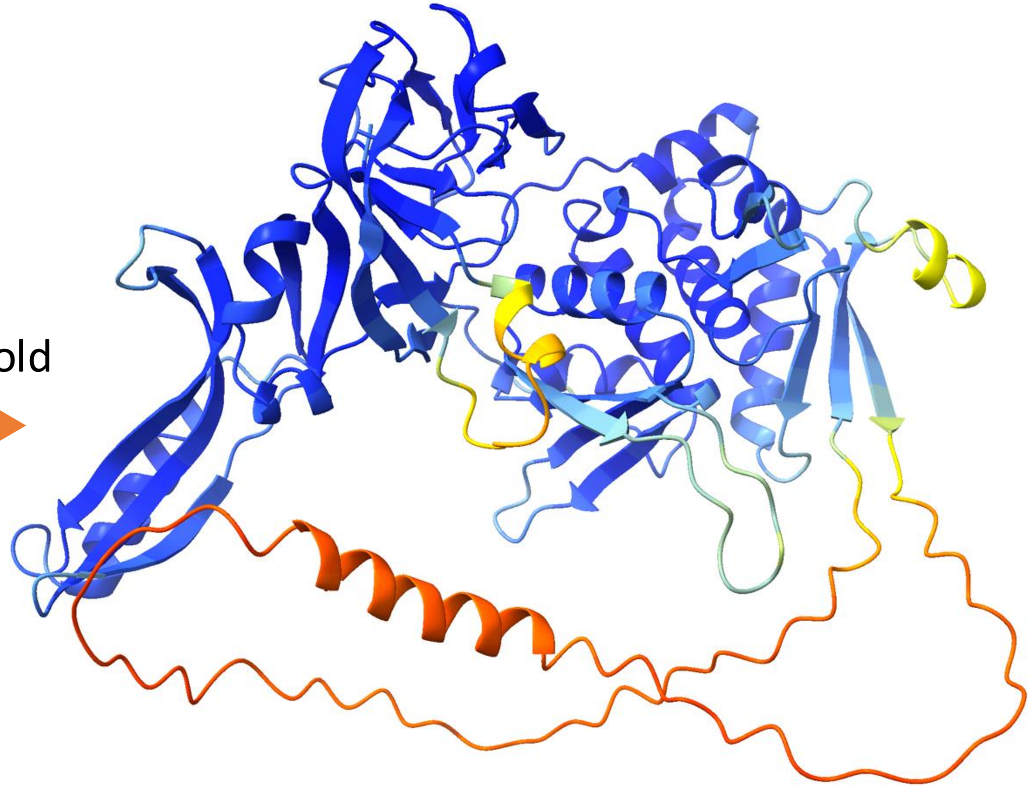
- By default, when alphafold2 is done, it colors the predicted structure by the confidence of the prediction (pLDDT = predicted local distance difference test)
- If you saved the predicted structure as a pdb, and subsequently load the file again, the coloring is gone; however, it is nevertheless saved in the pdb file
- To get the structure colored by pLDDT, load the pdb file, and then use the commandline to execute

```
color bfactor palette alphafold
```

Thanks to https://kpwulab.com/2023/03/09/color-alphafold2s-plddt/



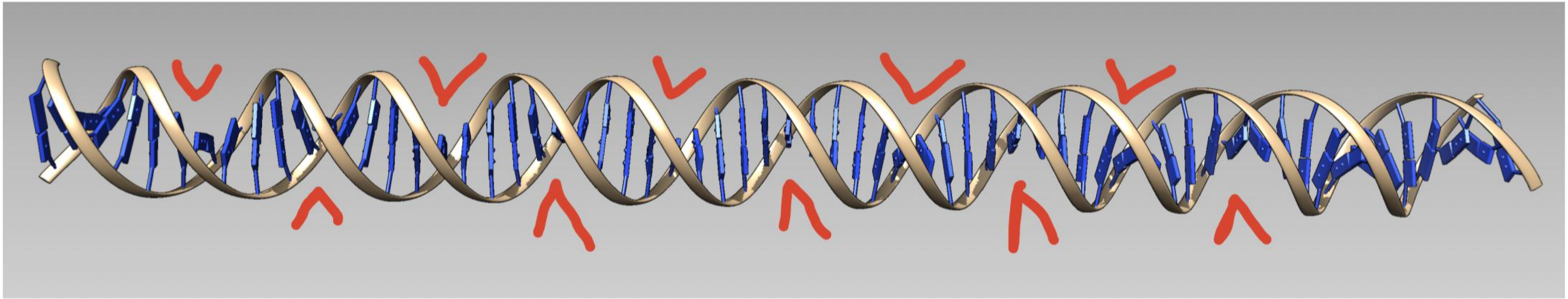
```
color bfactor palette alphafold
```



The best model file saved as pdf
(also saved in Downloads/chimeraX/AlphaFold/Prediction_n)

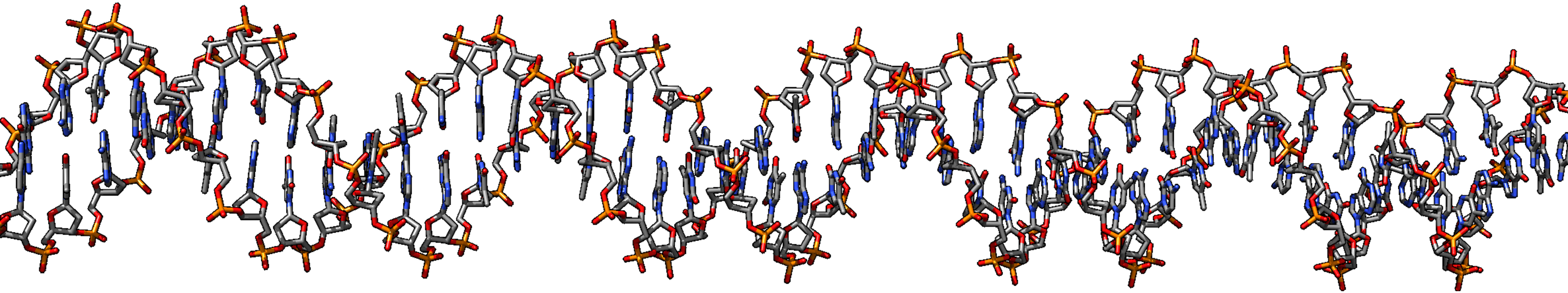
# The Major and Minor Groove in DNA

Maybe the DNA bound to a protein (and bent towards the breaking point (this is what the HE is doing) may not have been the best illustration.  Below is a DNA molecule built in chimera:

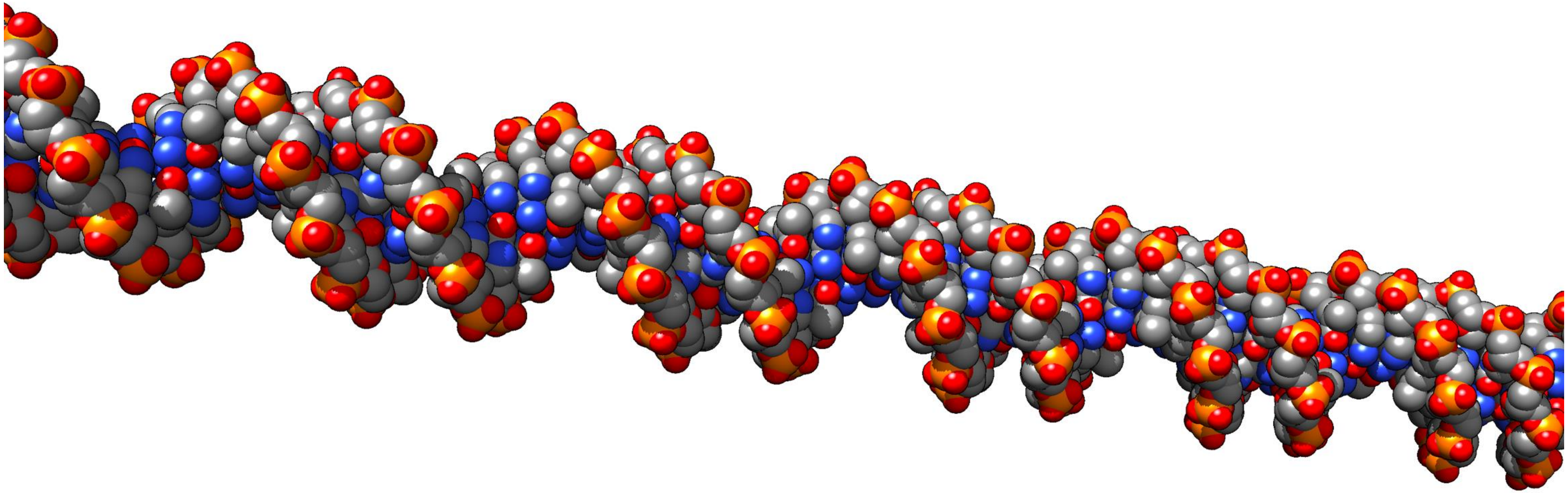


The red arrow tips indicate the major groove, and opposite to it is the minor groove.

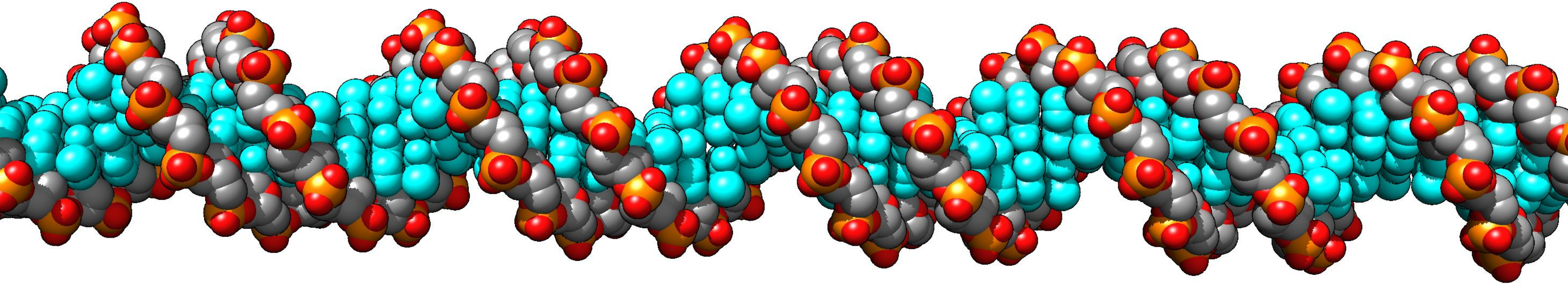
The same as sticks:



And even more impressive as space filling spheres (turned a little so that one look down the grooves) colored by elements:



Below is the same, but the base pairs are colored in cyan, clearly revealing that the base pairs are more accessible through the major groove:



A link to the chimera session is here

# Assignment

**1. Draw a sketch for the relation between the number substitutions that occurred in evolution and the percent identity of the two sequences. (I.e. how does the observed similarity change, as more and more substitutions occur?)**

2.What is the endpoint for 4 letter alphabet - for 20 letter alphabet.

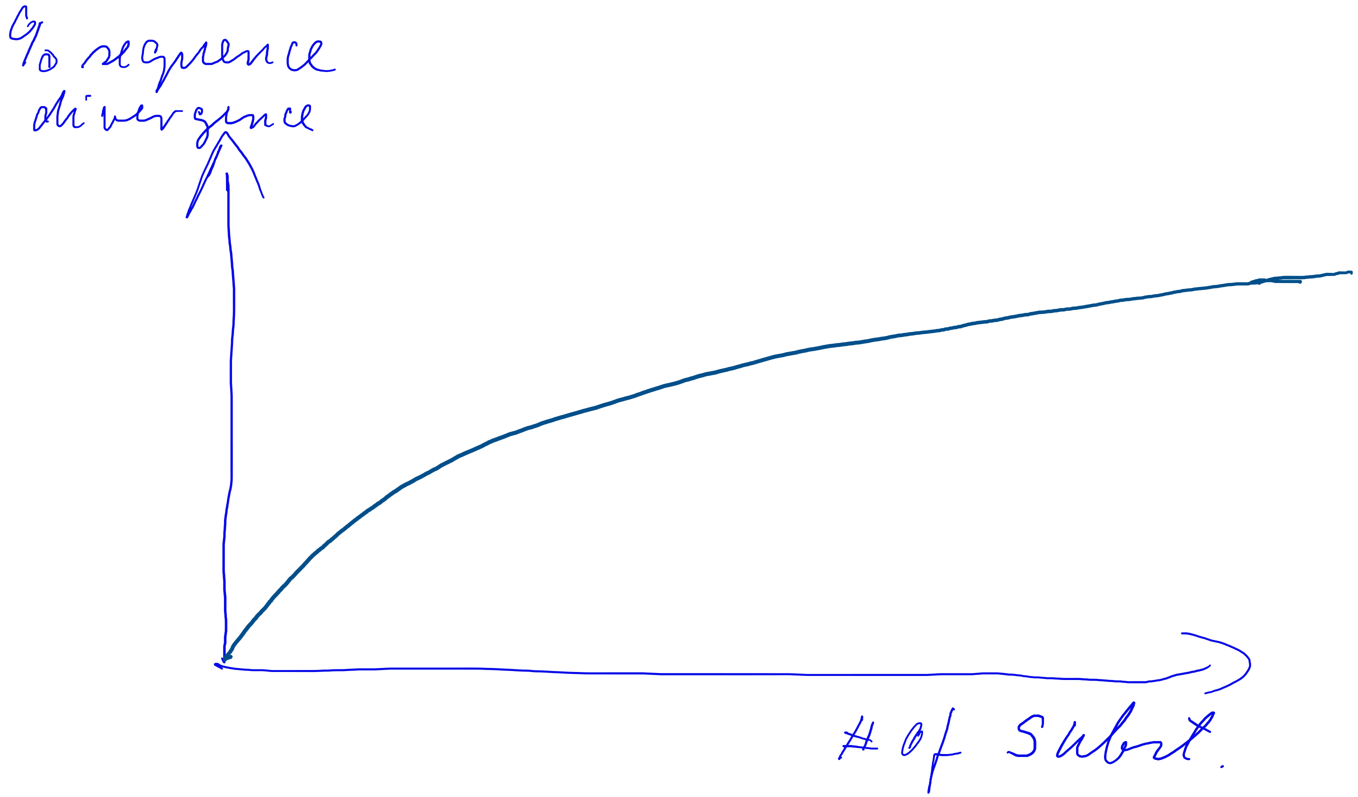
3. How does this relationship change, if some parts of the sequence are so important that the protein becomes non-functional, if a mutation occurs in these positions (i.e., these parts of the sequence are never observed to undergo any change?

4.If you **were** to do a realistic calculation and you were to consider a nucleotide sequence, how long would it take to arrive at 20% identity? (tip: how similar are two random sequences that have not been aligned?)

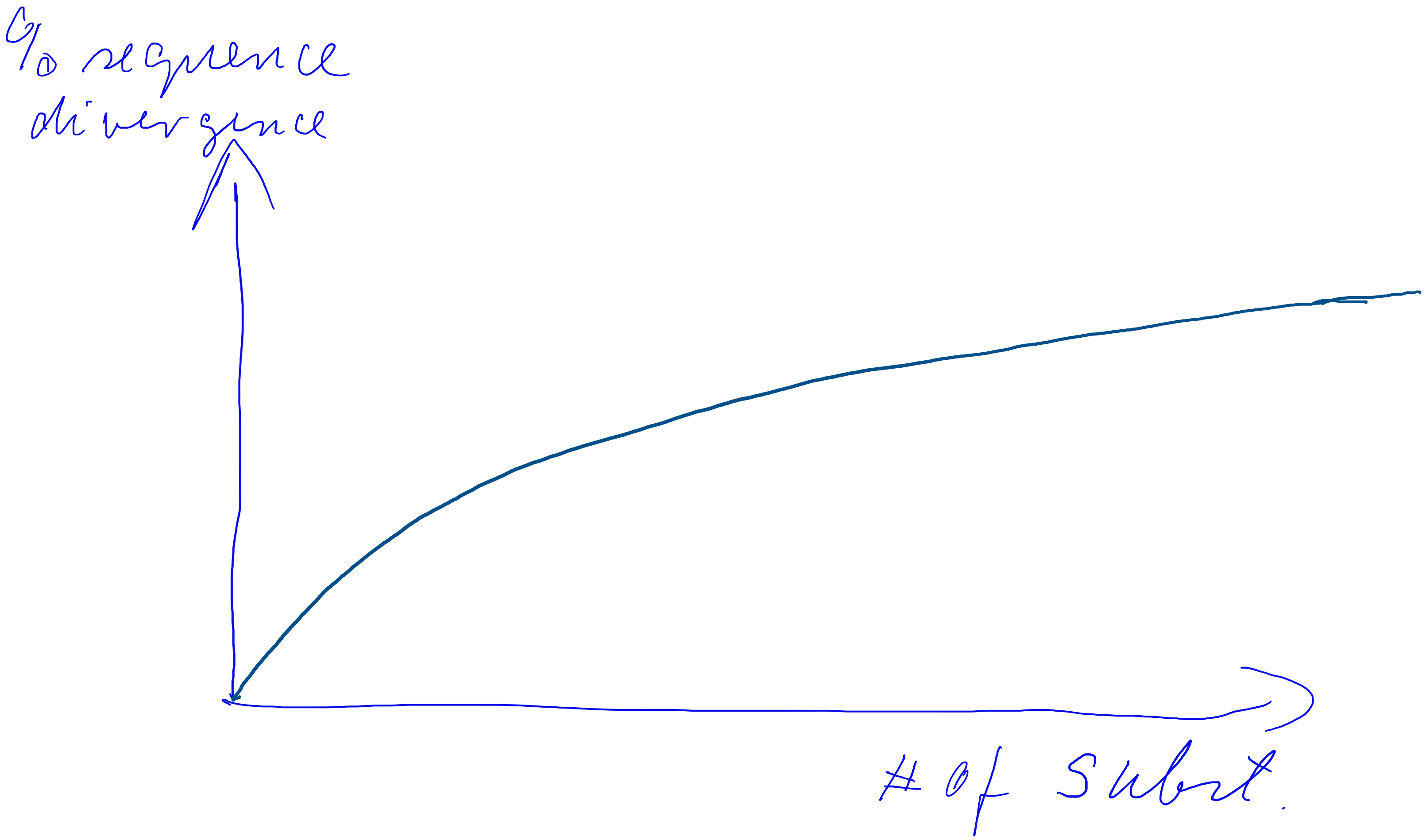
(Note: answering this should not require the use of a calculator or a formula, just common sense.)

**Draw a sketch for the relation between the number substitutions that occurred in evolution and the percent identity of the two sequences. (I.e. how does the observed similarity change, as more and more substitutions occur?)**

**Draw a sketch for the relation between the number substitutions that occurred in evolution and the percent identity of the two sequences. (I.e. how does the observed similarity change, as more and more substitutions occur?)**



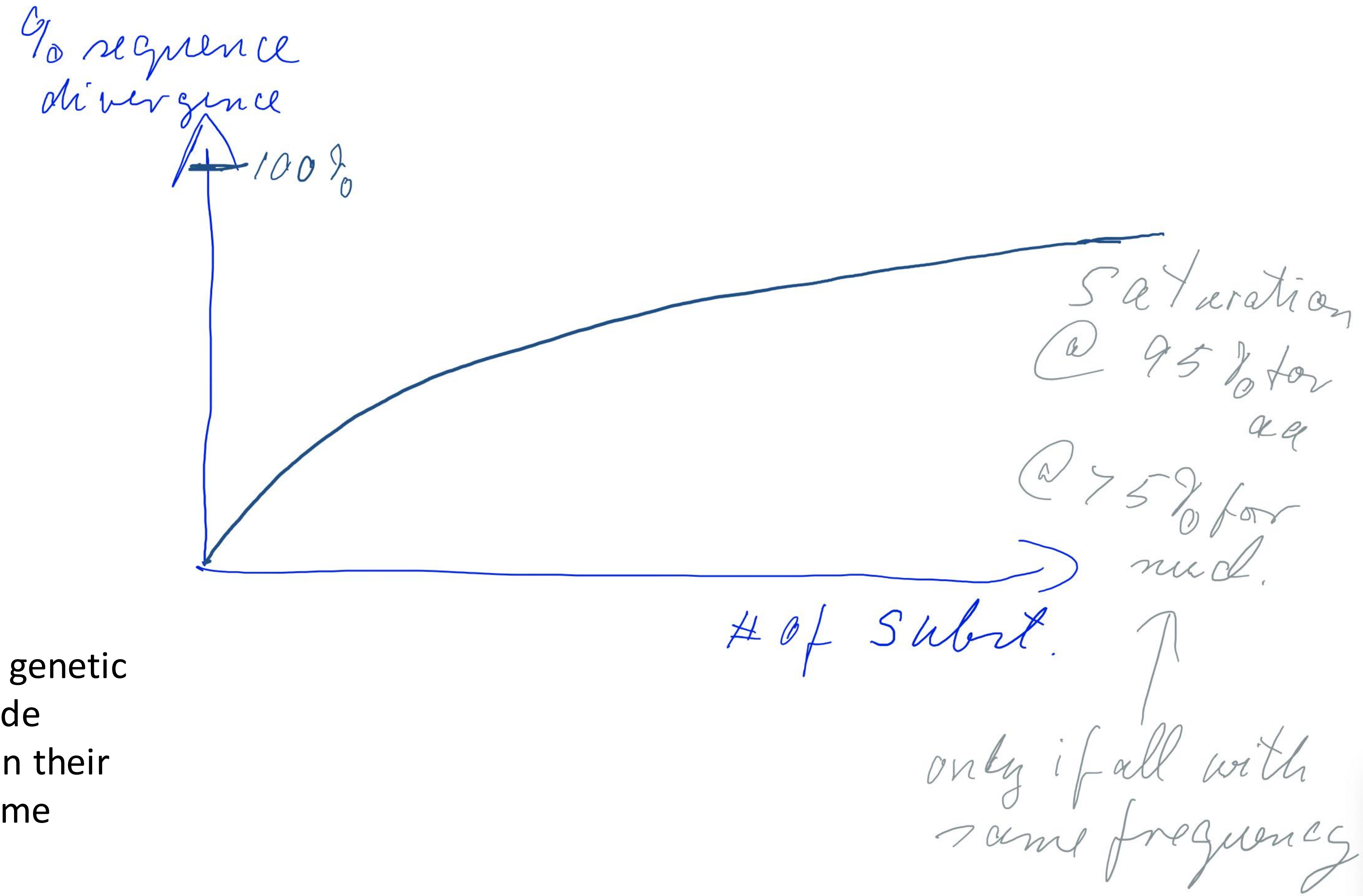
# What is the endpoint for 4 letter alphabet - for 20 letter alphabet.



# What is the endpoint for 4 letter alphabet - for 20 letter alphabet.

If all “letters” occur with equal frequency, and the starting sequence had an A,  then the probability that the sequence that is completely saturated with substitutions also has an A is qual to the frequencies of A (1/4 in case of nucleotides, 1/20 in case of amino acids

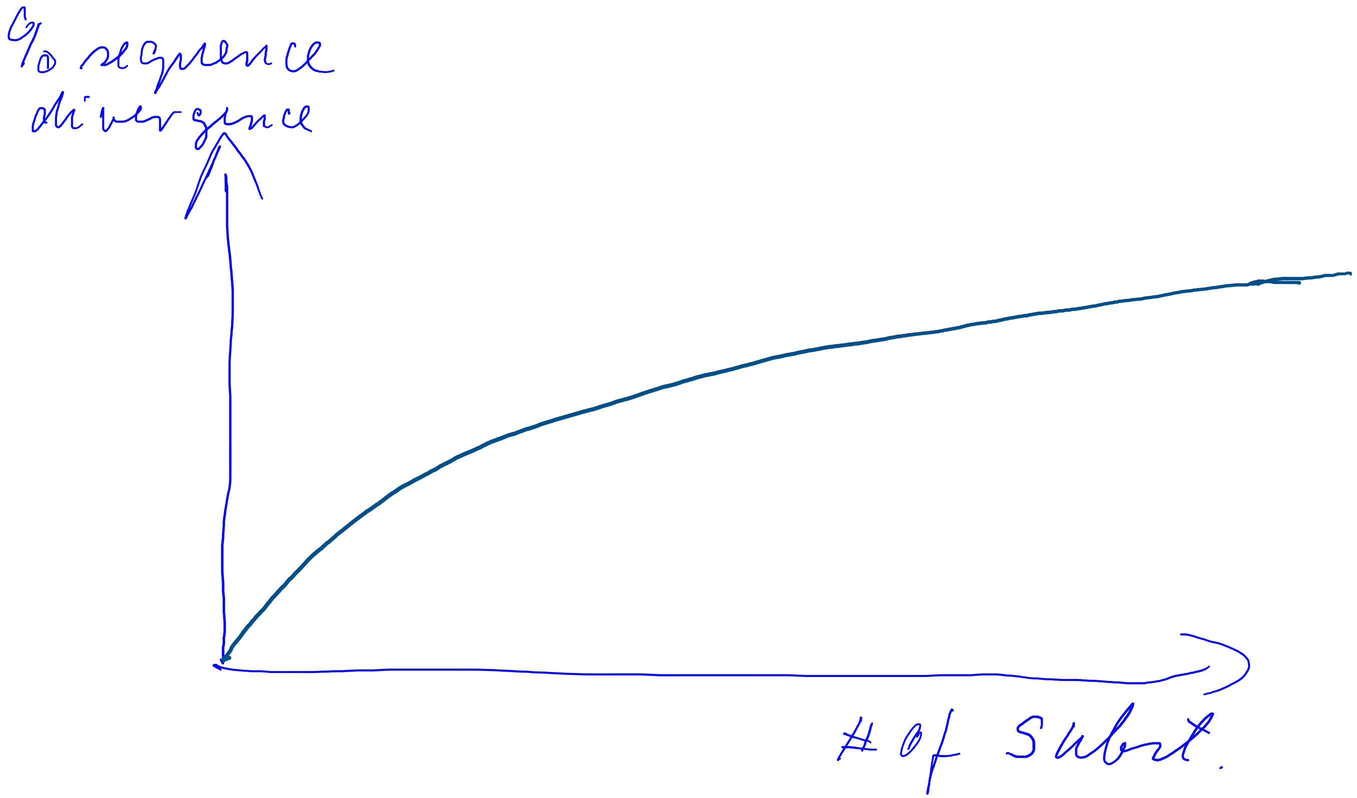
1/4 identity = 75% divergence
1/20 identity = 95% divergence



Note: Because of the redundancy of the genetic code  (61 codon encode 20 aa), nucleotide sequences can be about 30% divergent in their sequence and still encode exactly the same amino acid sequence.

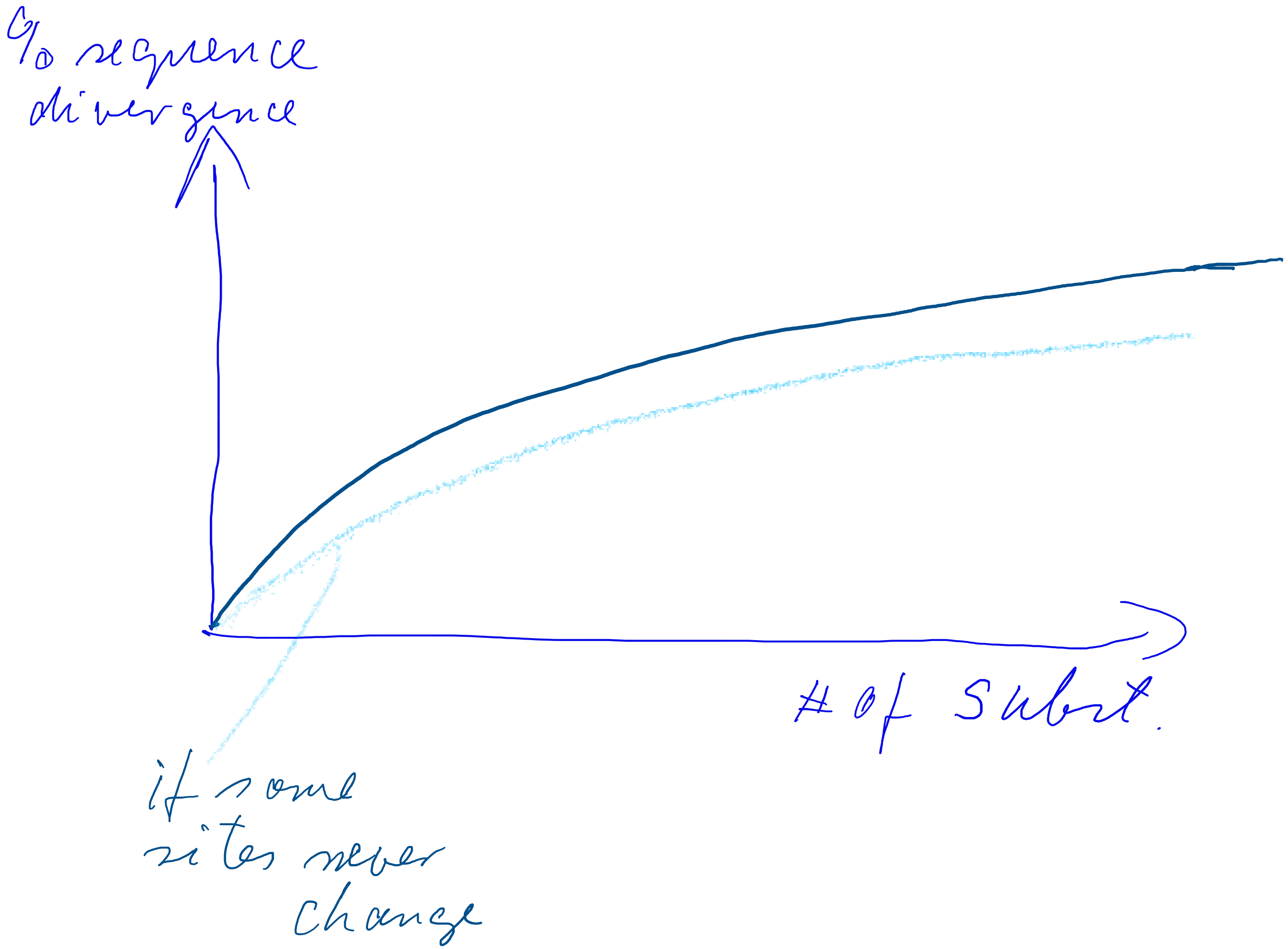
How does this relationship change, if some parts of the sequence are so important that the protein becomes non-functional, if a mutation occurs in these positions (i.e., these parts of the sequence are never observed to undergo any change?

The saturation level is lower by the % of positions that do not change



How does this relationship change, if some parts of the sequence are so important that the protein becomes non-functional, if a mutation occurs in these positions (i.e., these parts of the sequence are never observed to undergo any change?

The saturation level is lower by the % of positions that do not change



We could estimate the number of substitutions per site that occurred between two nucleotide sequence that **diverged 3.5 billion years** ago:

Without selection, two sequences that evolved from a common ancestor 3,500 million years ago (in total separated by 7 billion years), experienced rate x time = 10^-8 (substitutions per site and per year) x 7 x 10^9 (years) = 70 substitutions per site.

I.e., without selection removing substitutions that lead to less functional proteins, sequences that diverged more than a few 100 million years ago would be saturated with substitutions.

# If you **were** to do a realistic calculation and you were to consider a nucleotide sequence, how long would it take to arrive at 20% identity? (tip: how similar are two random sequences that have not been aligned?)

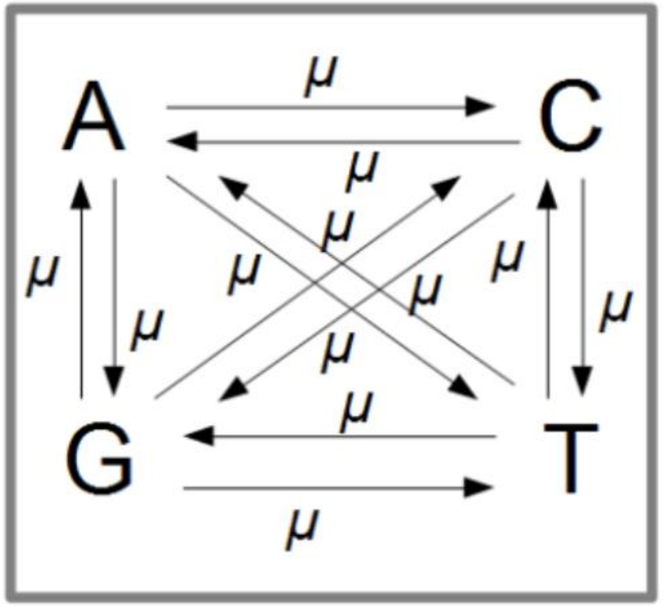
Using the approach ignoring multiple substitutions, we could calculate:

rate * unknown time =0.8
or with the time in years being X:
10^-8 * X=0.8
X=0.5*10^8 = 80 million years.
A common ancestral sequence would have diverged to two extant sequences with that difference in 40 million years.

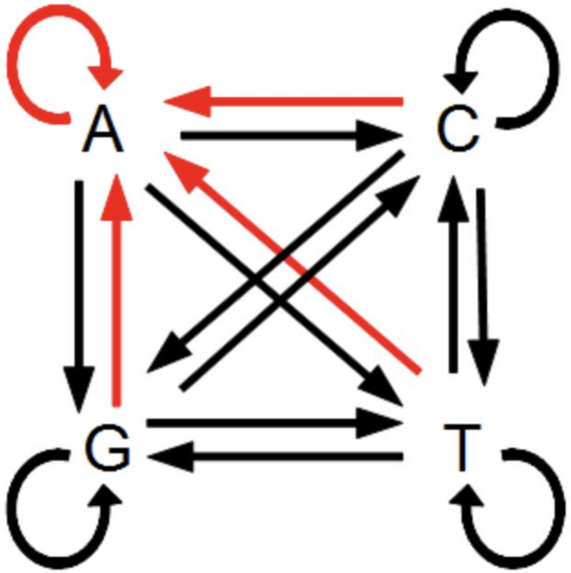
However, **under realistic conditions the answer is “forever”, because to random sequences are already 25% identical.**

# Jukes Cantor

A simple substitution model that takes back mutations and multiple substitutions into account is the Jukes Cantor model.  It assumes that all substitutions occur with equal frequency:



In addition, a position could not experience a change:



# Jukes Cantor continued

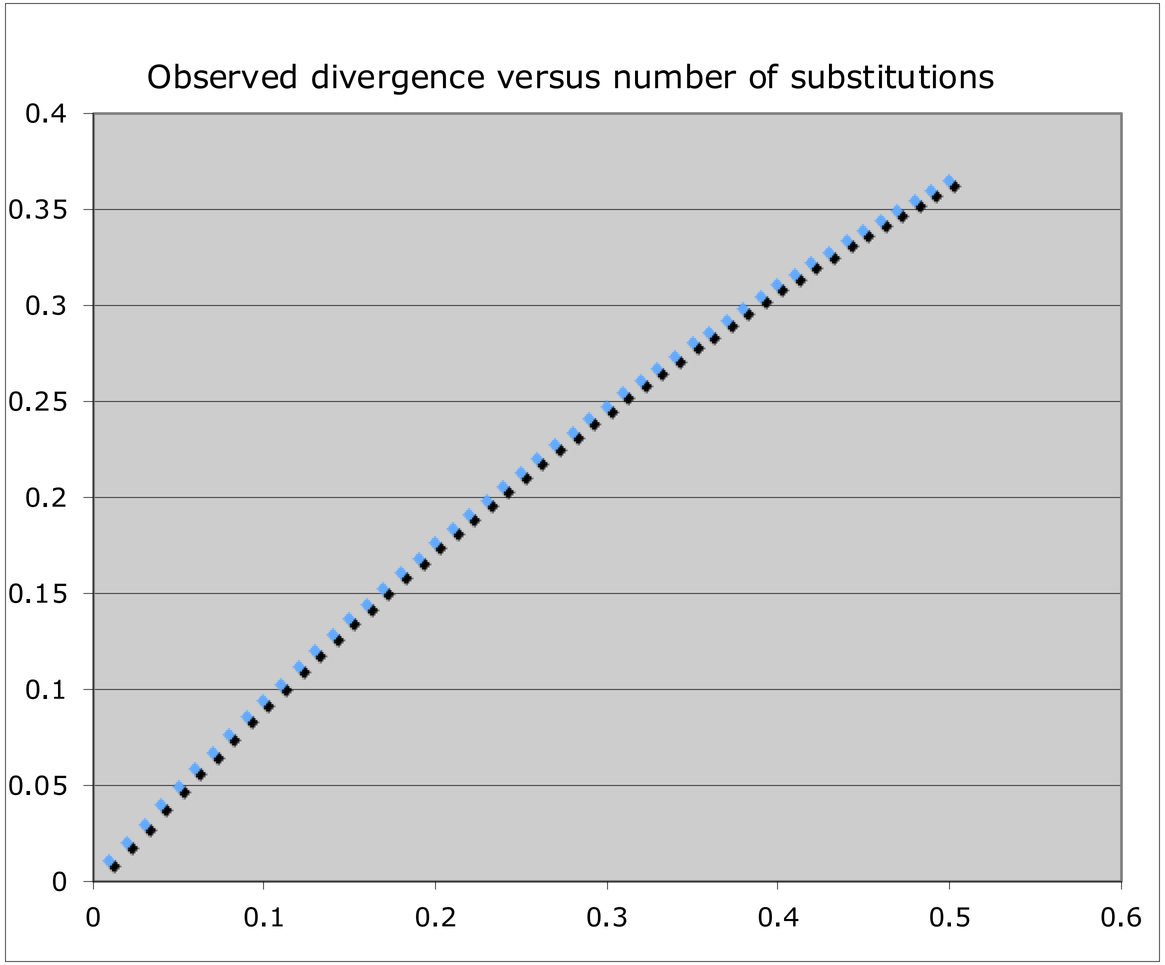
Solving this model for the relationship between **observed differences** p between two sequences and the **number of substitution** events d results in

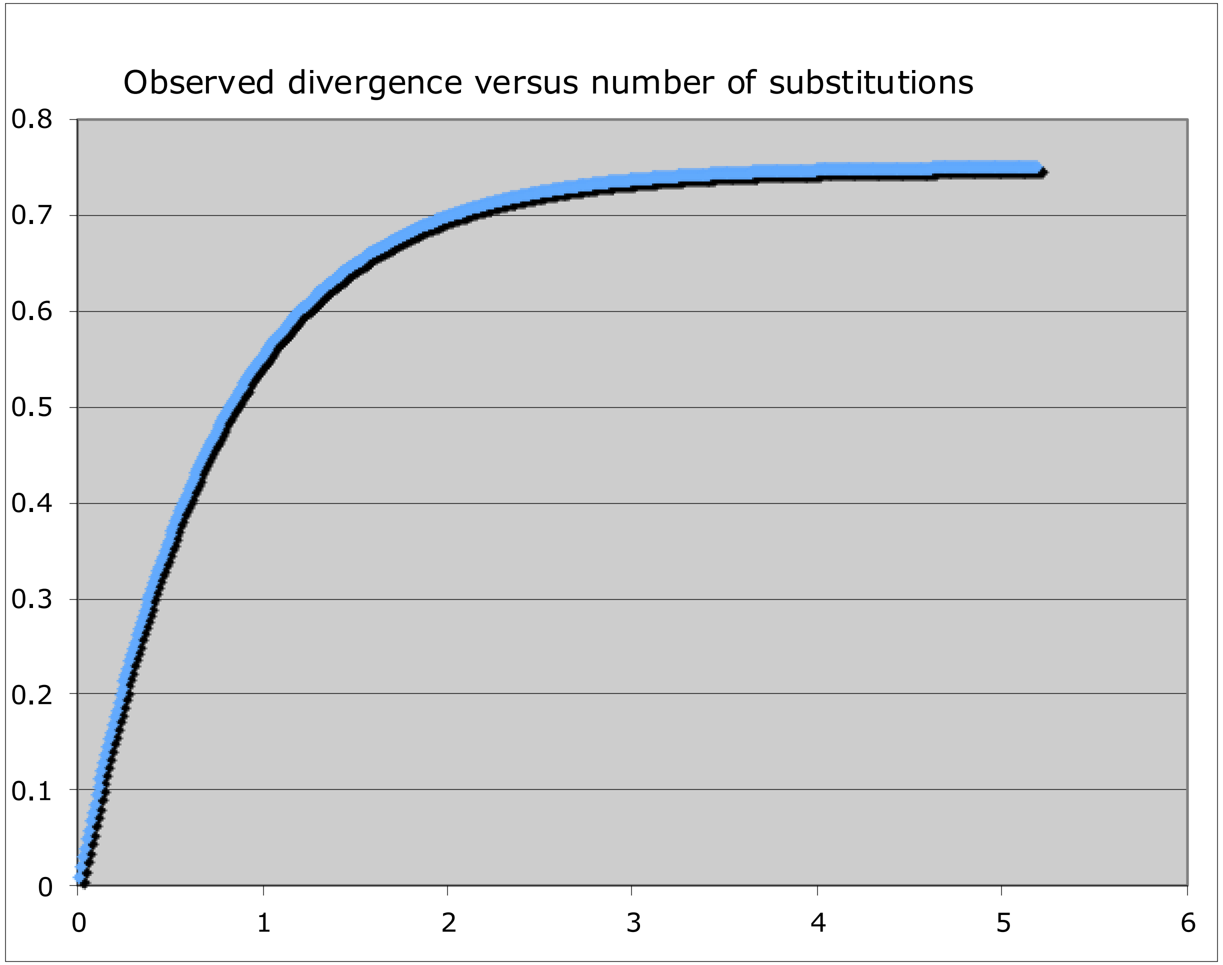
$$d=-(3/4)*\ln(1-4/3p)$$

or

$$p = 3/4-3/4*EXP((-(4/3)*d))$$

# Jukes Cantor continued (2)



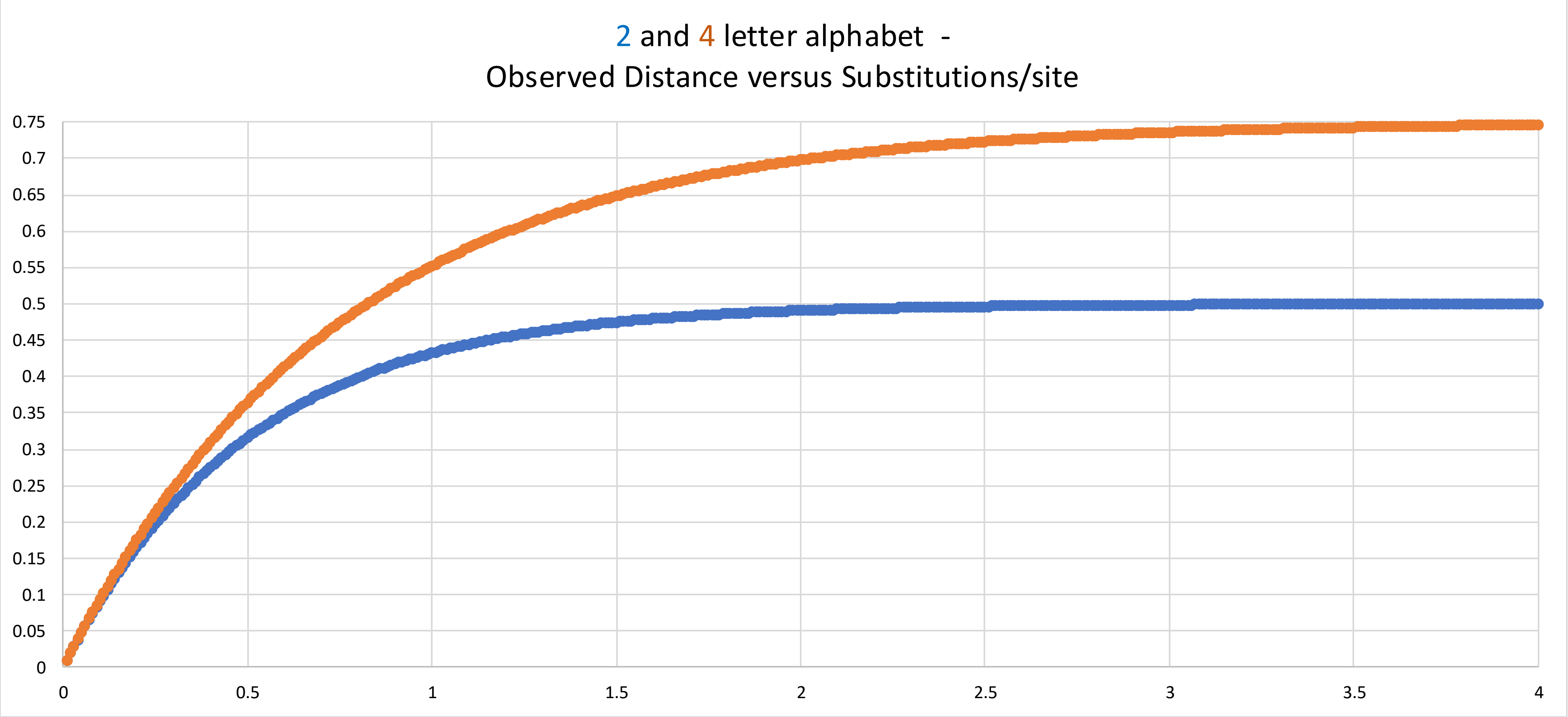


The numbers give substitutions (observed or occurring) per site

Initially, every substitution contributes to observed differences, later on, most substitutions change a residue that had already been changed.  (A spreadsheet is  here – note the different pages)

# How does the relationship change, if the different letters do not occur with the same frequency?

In the extreme we could consider only two nucleotides occurring:



Initially, every substitution leads to a change, but then saturation occurs at a lower level.