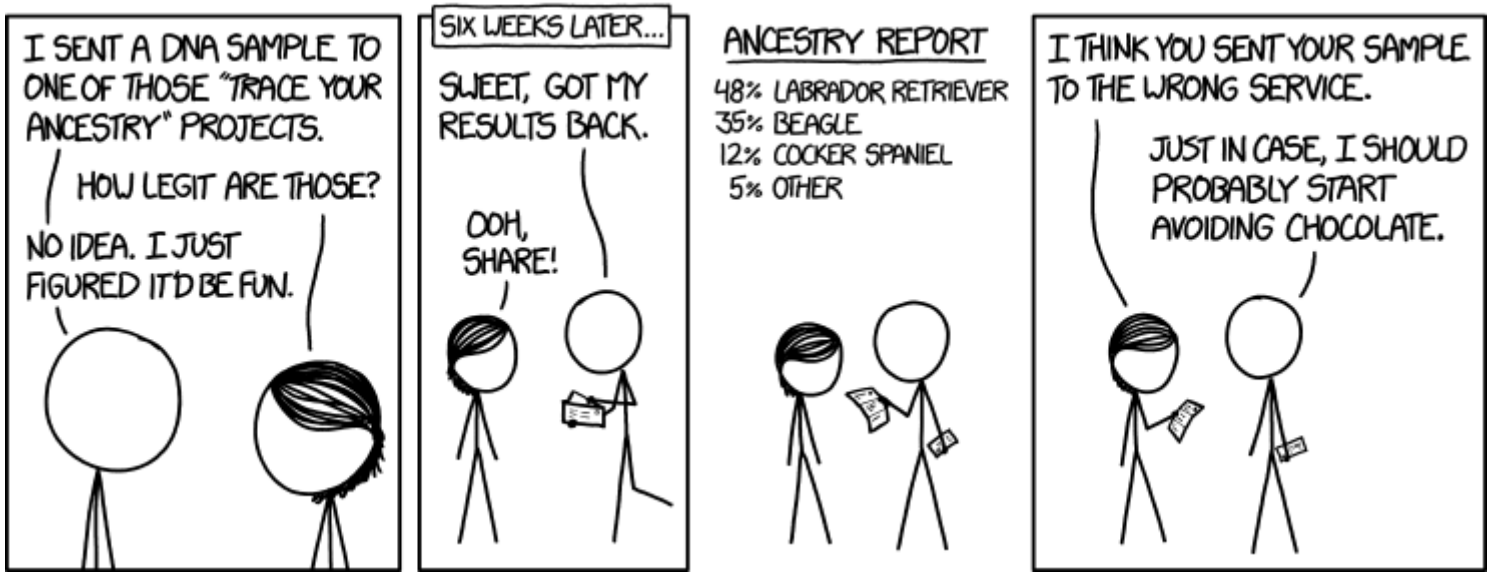
**MIMM4750g**

# Genetic distances



## Aligned sequences

- Now that we can align sequences, we can make biologically meaningful comparisons.
- Which sequences are more closely related than others?
- It is far easier to measure similarity when the sequences are aligned.

```
GGGTTGCGCTCGTTG    GGGTTGCGCTCGTTG
||  |         |     ||| ||| |||| ||
GGTTGCGCTCGTTGA    GGGATGCACTCGCTG
```

## p-distances

- The simplest distance is to count the number of different residues.

```
GGGTTGCGCTCGTTG
|||  |||  ||||  ||   = 3 differences
GGGATGCACTCGCTG
```

- This is called the Hamming distance.
- Hamming distance (HD) increases with sequence length.
- We can divide the HD by sequence length. This gives us the p-distance (*p* is for *proportional*).

> *What is the p-distance for the above example?*

## Multiple hits

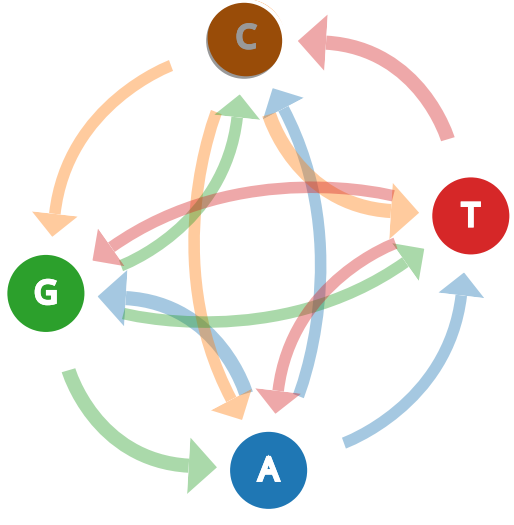
- A big problem with the Hamming and *p*-distances is that they tend to underestimate the amount of evolution.
- Suppose we are tracking the evolution of a sequence `AAAA`
- A single mutation occurs resulting `AGAA` ($p = 0.25$)
- As we continue to accumulate mutations, the chance that we mutate a site *that has already undergone a mutation* increases.
- Multiple hits mask evidence of previous evolution ($\texttt{A} \longrightarrow \texttt{G} \longrightarrow \texttt{A}$).

## Modeling evolution

- Let's make a few lousy assumptions:
    1. Each residue in a sequence evolves independently of the others.
    2. A residue mutates to another at a rate that is constant over time.
    3. A residue is equally likely to mutate to any of the other residues.
    4. The frequency of every residue is the same.
- These define the *Jukes-Cantor* model.

## Jukes-Cantor model

- Based on a computer program written by grad student Charles Cantor in 1966.
- It is an example of a *continuous time Markov model*.




**Based on JS by Victor Powell**

## Markov processes

- The Jukes-Cantor model describes a Markov process.
- A process has the *Markov property* if the probability of state at time *t* depends only on the state at a previous time *and no further*.
- *i.e.*, the system has no memory.
- For example, Snakes and Ladders is a Markov process.

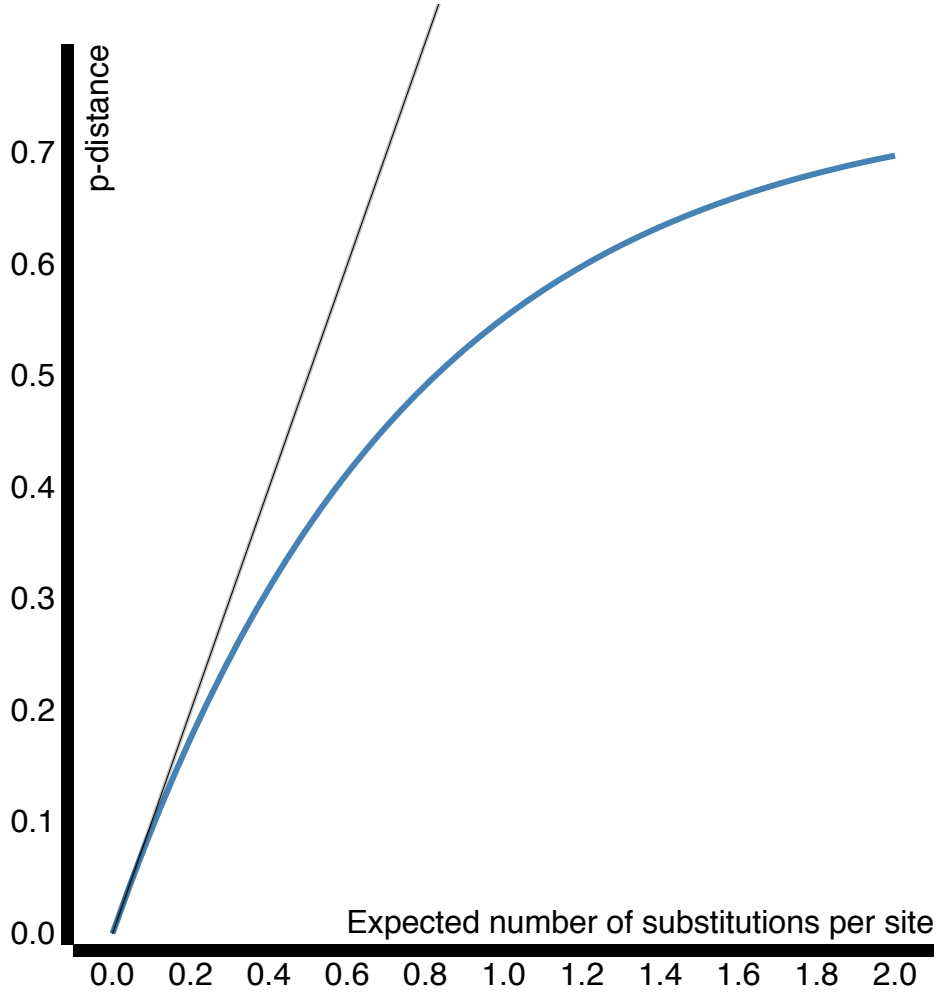
> *What is another example of a process with no memory? A process* ***with*** *memory?*

## Jukes-Cantor formula

- Because of multiple hits, the actual number of mutations tends to be *greater* than the number of visible differences.
- Given a p-distance ($p$) between two sequences, the JC estimated number of mutations ($\hat{d}$) is:

$$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

# Jukes-Cantor simulation




AAAAAAAAAA

Evolve Reset

Number of mutations = 0

p-distance = 0

Predicted num. mutations = 0

Jukes-Cantor adjustment = 0

## INCA Question #3

> *Why does this calculator occasionally report a "Jukes-Cantor adjustment" of `NaN`? (Not a Number)*

Hint: think about some of the assumptions of the Jukes-Cantor model.

Another hint: The log of $x \leq 0$ is not a number (undefined).

## Why does this matter?

- The Jukes-Cantor model enables us to estimate the divergence time of two populations (species or infections) more accurately.
- Two distantly related species might otherwise look about the same as more closely related species.

## Improvements to Jukes-Cantor

- Kimura's two-parameter distance (1980, K2P) has different rates for transitions and transverions.
- Tajima-Nei's (1984) distance allows unequal nucleotide frequencies.
- Tamura 3-parameter distance (1992) extends K2P to allow for GC-content bias.
- Tamura-Nei (1993, TN93) has two rates for transitions and a transverion rate, and unequal nucleotide frequencies.

## Which distance should I use?

- It is fairly likely that the assumption of equal nucleotide frequencies is broken.
- The HIV-1 genome is roughly 40% A's.
- The Actinobacteria (including *Streptomyces*) are also known as "high G+C Gram-positive bacteria".
- Transition/transversion bias is ubiquitous.
- Nowadays we seldom see distances other than TN93 in use.

## Software for calculating distances

- MEGA - user-friendly software for sequence analysis.
- `dist.dna` function in the *R* package `ape`
- tn93 - a very fast TN93 calculator in C++

# BioPython demonstration