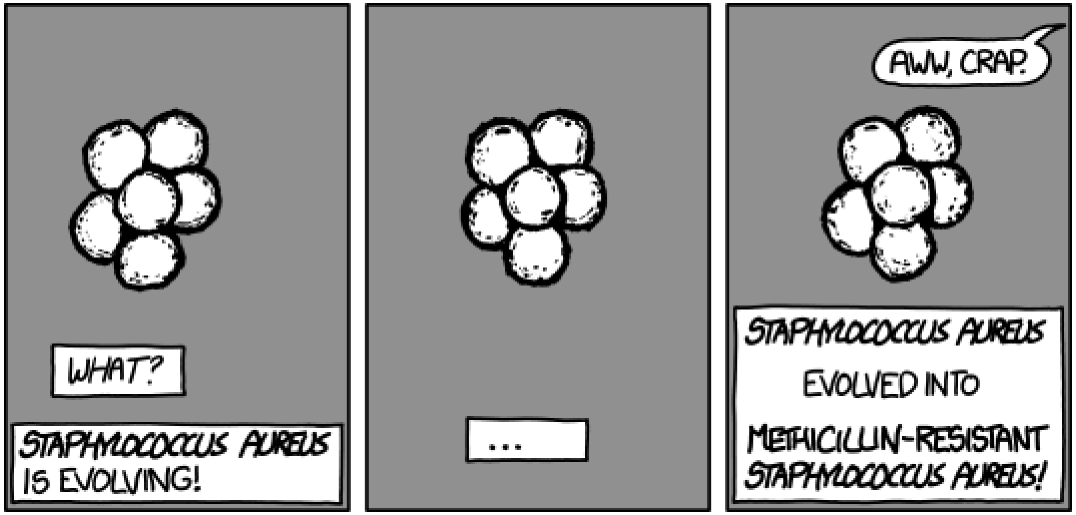
MIMM4750G

Diversity and rates of evolution


WHAT?
STAPHYLOCOCCUS AUREUS IS EVOLVING!
...
AWW, CRAP.
STAPHYLOCOCCUS AUREUS EVOLVED INTO METHICILLIN-RESISTANT STAPHYLOCOCCUS AUREUS!

Measuring diversity

- Genetic distances are useful
- What if there are many sequences (lots of pairwise comparisons!)
- Next-generation sequencing yields thousands to millions of reads.
- What if the sequences are very long?
- Bacterial genomes can be over 14Mbp long.
- Shortest animal genome is about 19.6Mbp (*Pratylenchus coffeae*, parasitic nematode of plants).

Simple diversity measures

- Proportion of polymorphic sites in an alignment.
- Nucleotide diversity - the expected p-distance (p) between a random pair of sequences:

  $$\pi = \sum_i \sum_j f_i f_j p_{ij}$$

  where f_i is the frequency of the i-th sequence variant.
- Shannon entropy - very common in bioinformatics (see next).

Shannon Entropy

- Based on information theory, Shannon entropy is calculated from the frequencies of variants indexed by i:

$$S = -\sum_i p_i \log(p_i)$$

- If most frequencies are near zero, S approaches 0.
- S is greatest when frequencies are equal.
- Often averaged across nucleotide or amino acid sites of an alignment.

Application of Shannon entropy to characterize 16S rRNA gene diversity using PacBio (long read) and MiSeq (short read) NGS platforms.

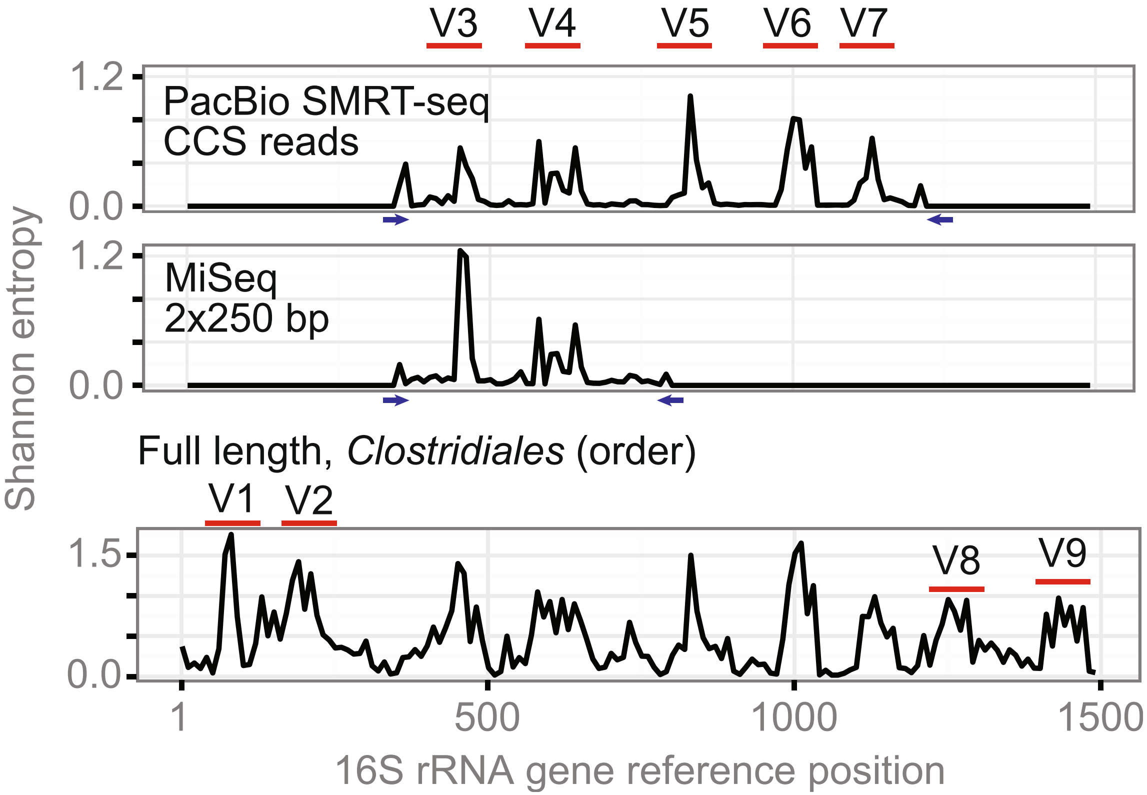


Figure from O Franzén *et al.* (2015) Microbiome 3:43.

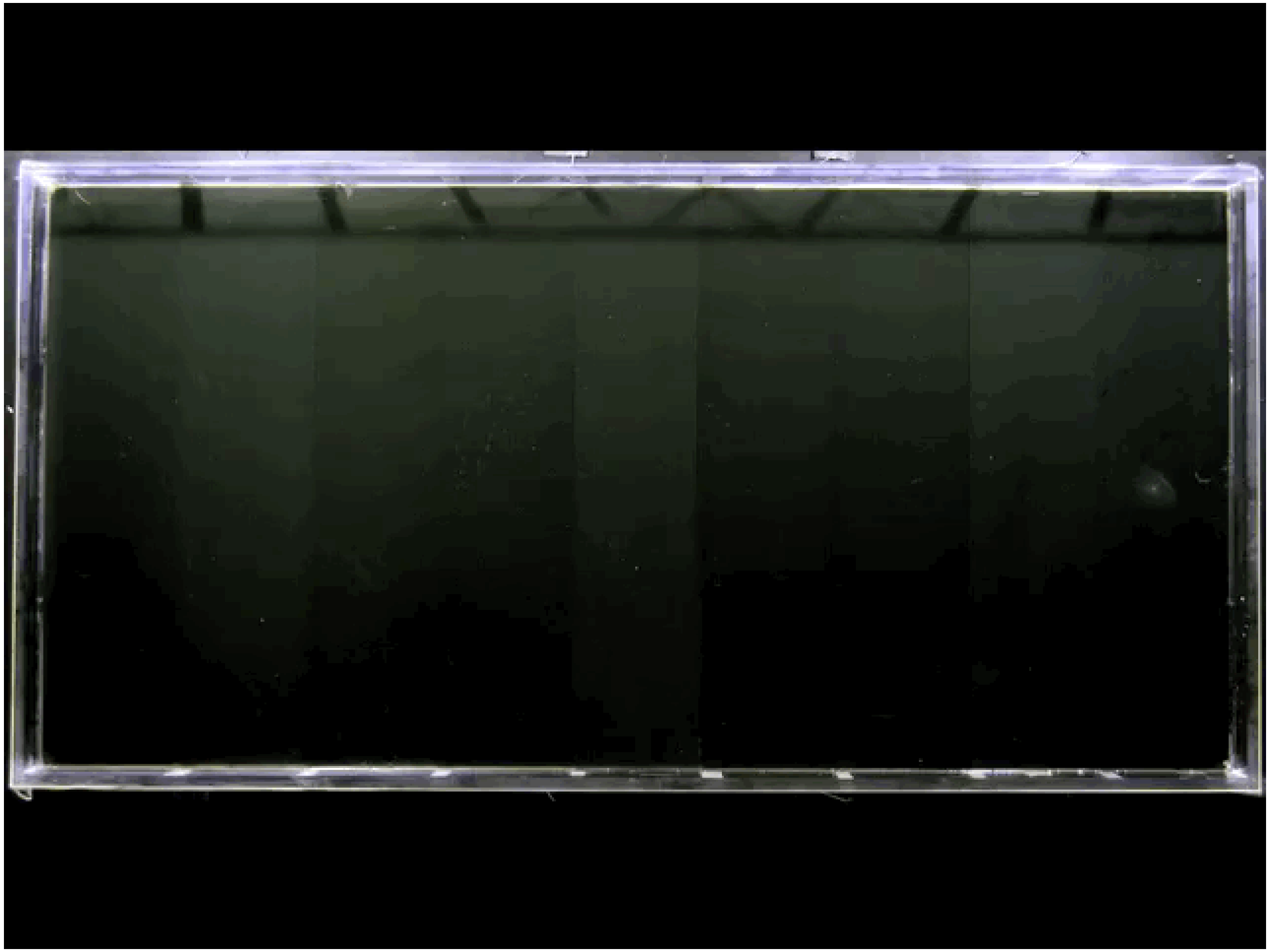
Limitations of diversity measures

- Many diversity measures assume every sequence is an independent observation.
- We can badly overestimate diversity if most sequences inherited their variation from the same common ancestors.
- Consider two infections that differ by one nucleotide.
  - Infection A is ancestral to 10 other infections.
  - Infection B is ancestral to another 10.
  - If we compare this site across these 20 descendants, it will look very diverse!

Rates of evolution

- Counting mutations is the key to measuring the rate of evolution
- **Why do rates matter?**
- Sites that evolve faster than others can reveal targets of selection.
- Rate of evolution may determine which variant survives.
- We can use the rate of evolution to extrapolate back in time.

Spread of *Escherichia coli* on a "megaplate" with gradients of antibiotic trimethoprim.



Media from M Baym *et al* (2016) Science 353: 1147.

HIV-1 protease homodimer, colored by rate of evolution (blue fastest).
Protease inhibitor (yellow) bound at active site.

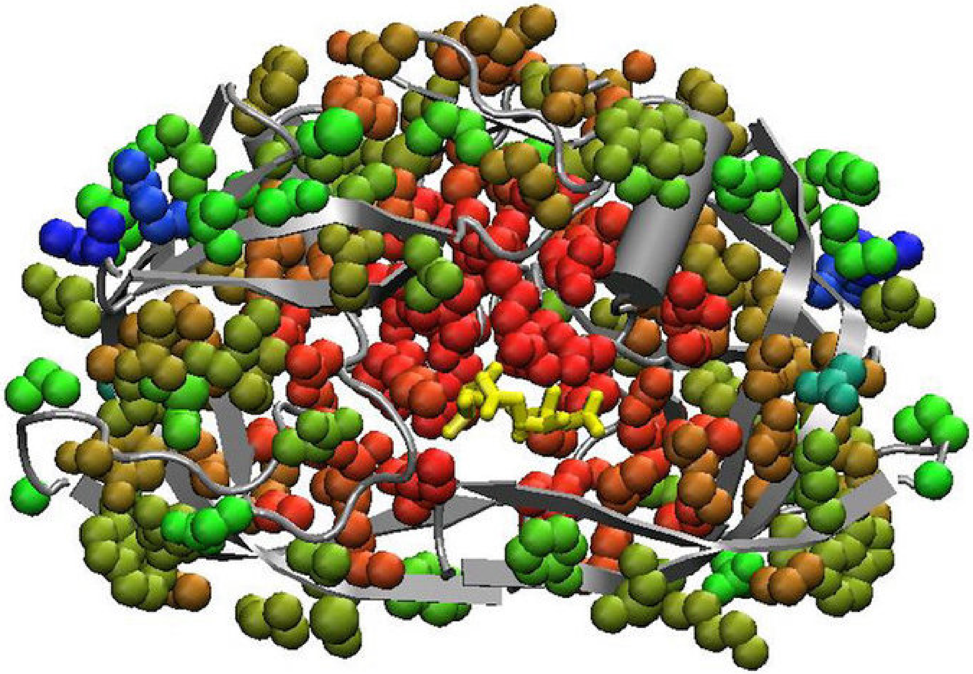
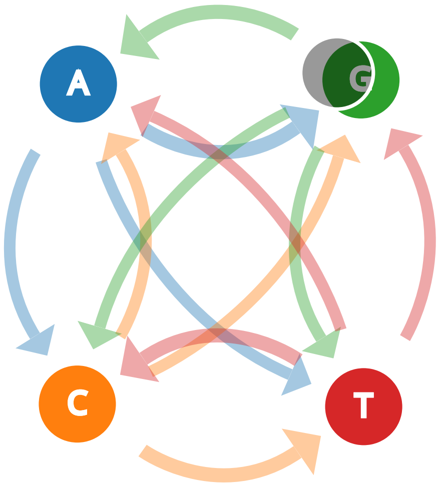


Image from Kuiken *et al.*, Los Alamos National Laboratory HIV Sequence Compendium 2003.

Modeling evolution

- Recall that sequence evolution is often modeled as a *continuous-time Markov chain*
- Constant rate of evolution - exponential waiting times.



Based on JS by Victor Powell

Substitution models

- The Jukes-Cantor model can be expressed by the following rate matrix:

$$\begin{pmatrix} * & \mu & \mu & \mu \\ \mu & * & \mu & \mu \\ \mu & \mu & * & \mu \\ \mu & \mu & \mu & * \end{pmatrix}$$

- The diagonal entries $*$ are set to -3μ so that each row sums to 0.

Other models

- The Hasegawa-Kishino-Yano (HKY85) model allows for unequal base frequencies (π_i) and a transition/transversion rate bias (κ).

$$\begin{array}{cc} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{pmatrix} * & \kappa\pi_C & \pi_G & \kappa\pi_T \\ \kappa\pi_A & * & \kappa\pi_G & \pi_T \\ \pi_A & \kappa\pi_C & * & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \kappa\pi_G & * \end{pmatrix} \end{array}$$

Generalized models

- In general, there are six rates for a time-reversible (symmetric rates) model:

  $$\begin{pmatrix} * & a & b & c \\ a & * & d & e \\ b & d & * & f \\ c & e & f & * \end{pmatrix}$$

  where these rates are assigned in alphabetical order — a is the rate from $A \leftrightarrow C$, b is $A \leftrightarrow G$, etc.

Model specification

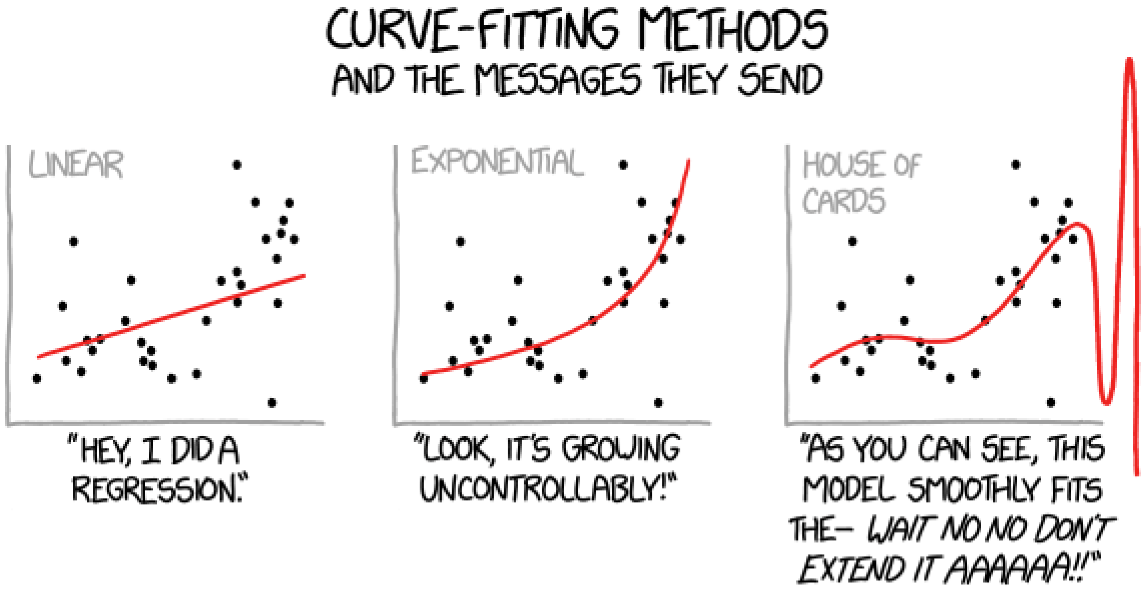
- PAUP* was a popular commercial software package for reconstructing phylogenies.
- It used a six-digit number ($abcdef$) to represent any kind of time-reversible nucleotide substitution model:
- *e.g.*, HKY85 becomes 010010.
- This scheme is still used by other software, such as HyPhy and PhyML.

Why does the model matter?

- There are an enormous number of possible time-reversible models of nucleotide substitution.
- Using the wrong model (*model misspecification*) can bias estimates of other model parameters, *e.g.*, reconstructing the correct tree.
- The process of figuring out which model is best supported by the data is called *model selection*.

Model selection

- We want to choose the model that has the best fit to the data.
- Adding parameters to the model improves the fit.
- We need to justify additional parameters!


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Likelihood ratio test

- The *likelihood ratio test* (LRT) is a method of model selection that applies when one model is a special case of another.
- *e.g.*, the JC69 model is a special case of HKY85 where $\kappa = 1$.
- If the likelihood of model 1 is L_1 and model 2 is L_2, then this test statistic:

  $$-2\log\left(\frac{L_1}{L_2}\right) = -2(\log L_1 - \log L_2)$$

  follows a χ^2_k distribution.
- k is the difference in the number of parameters.

Akaike information criterion

- What if the models are not nested?
- The Akaike information criterion (AIC) penalizes the model's likelihood by the number of parameters
- *There is no statistical distribution*! The best model minimizes the AIC.

$$\mathrm{AIC} = 2k - 2\log(L)$$