nGICE Phylogenomics Course

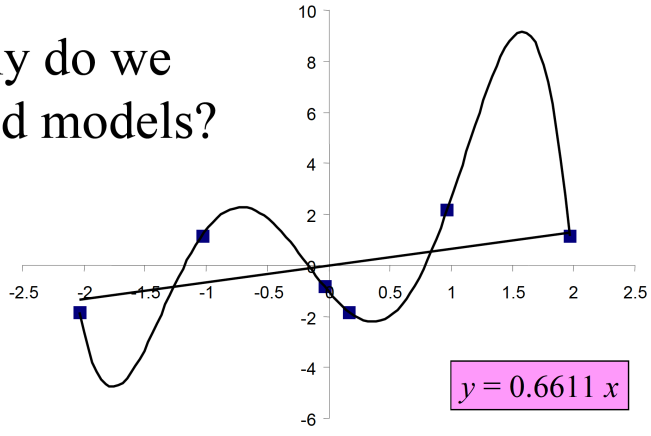
Introduction to model-based methods

Niklas Wahlberg

Slides by Jadranka Rota, Paul Lewis and Chris Simon (University of Connecticut, USA)

$y = -1.5972\,x^5 + 23.167\,x^4 - 126.18\,x^3 + 319.17\,x^2 - 369.22\,x + 155.67$

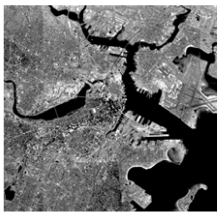
Why do we need models?



A very *practical* MBTA subway map

17

A very *realistic* MBTA subway map



18

- **Which is more useful?**

Models

- Models help us intelligently **interpolate between our observations** for purposes of predicting future observations
- **Adding parameters** to a model generally increases its fit to the data
- **Underparameterized** models lead to poor fit to observed data points
- **Overparameterized** models lead to poor prediction of future observations
- Criteria for choosing models include likelihood ratio tests, AIC, BIC, Bayes Factors, etc.
  - all provide a way to choose a model that is neither underparameterized nor overparameterized

Modelling nucleotide substitution

- **With thousands of genomes sequenced**
  - **Good understanding of how DNA sequences evolve**
  - **Different regions of the genome have their own substitution dynamics**
  - **Different lineages may have their own substitution dynamics**

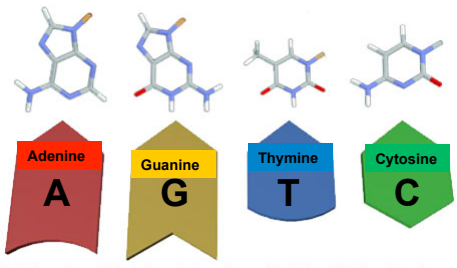
Main Challenge

- **DNA has only four characters**

Purines

Pyrimidines

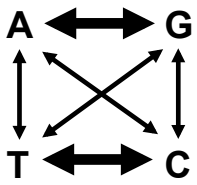
Figure B-3: The Four Nitrogenous Bases



Each base has a distinct shape that can be used to distinguish it form the others. 3D representations of the four bases are shown, with the corresponding chemical structures drawn above.

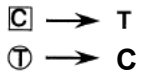
Substitution types

- Purines: A, G
- Pyrimidines: C, T
- Transversions
  - Pu → Pyr
  - Pyr → Pu
- Transitions – more common
  - Pu → Pu
  - Pyr → Pyr



Pur - Pyr mispairs lead to transitions

In next round of replication



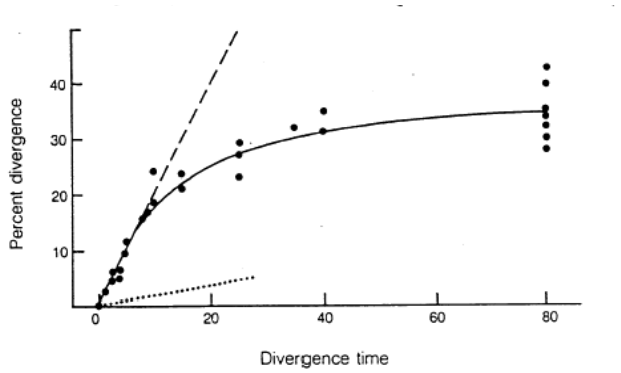
Slide by Chris Simon 2005

Saturation in sequence data:

- Saturation is due to multiple substitutions at the same site subsequent to lineage splitting
- Models of evolution attempt to infer the missing information through correcting for "multiple hits"
- Most data will contain some fast evolving sites which are potentially saturated
  - e.g. in protein-coding genes codon position 3
- In severe cases the data become essentially random and all information about relationships can be lost
- Probabilistic models of sequence evolution are used to calculate expected changes

Misleading DNA evolution

Multiple substitutions hide previous changes



Brown et al. 1979. PNAS 76:1967

Slide by Chris Simon 2005

Difference between mutation and substitution

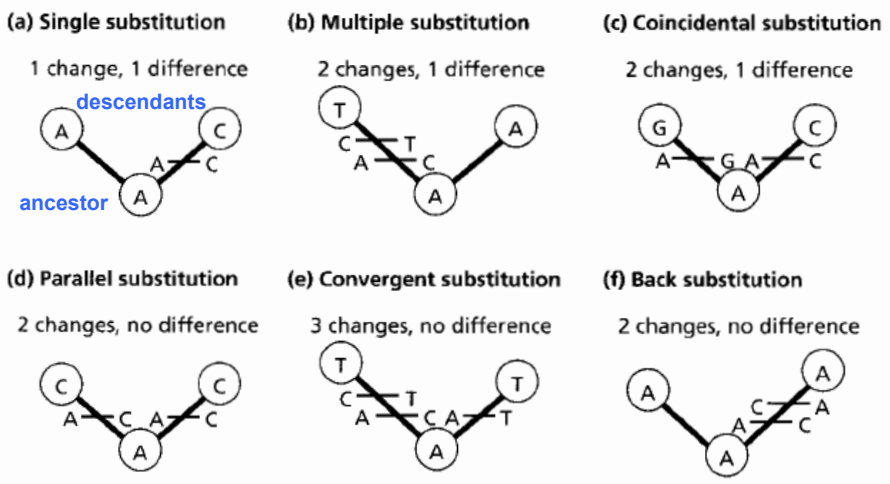
- **Substitutions** = mutational changes observed in populations
- **Mutations** = not all observed in populations, randomly distributed

  1) removed by proof reading enzymes

  2) cause death of cell, gamete, embryo

Slide by Chris Simon 2005

Types of Substitutions



Page, R. and E. Holmes. 1998. Molecular Evolution: A phylogenetic Approach. Blackwell.

Modelling nucleotide substitutions

- These dynamics can be modelled over a tree and they are incoporated into distance methods, maximum likelihood, and Bayesian inference
- Models incorporate information about the rates at which each nucleotide is replaced by each alternative nucleotide
  - For DNA this can be expressed as a 4 x 4 rate matrix (known as the Q matrix)
- Other model parameters may include:
  - Site by site rate variation (aka among-site rate variation - ASRV) - often modelled as a statistical distribution - for example a gamma distribution

Corrections for multiple substitutions:
First DNA substitution model

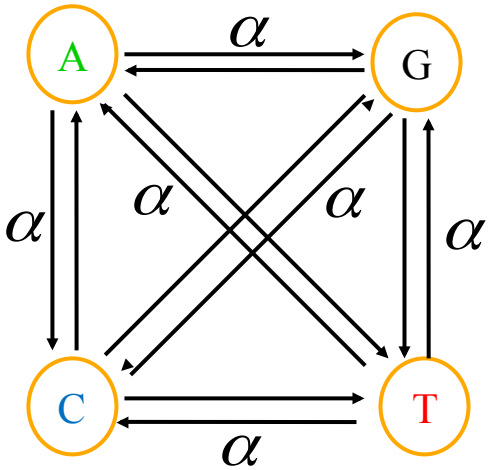
Jukes & Cantor (1969) assumptions:

1. A = T = G = C No nucleotide bias
2. Every base changes to every other base with equal probability (no TS/TV bias)
3. All sites change with the same probability (no ASRV - among-site rate variation)

Also: probability of substitution & base composition remains constant over time/across lineages

Slide by Chris Simon 2005

Jukes-Cantor model



t = time

- α = the rate of substitution (α changes from A to G every t)
- The rate of substitution for each nucleotide is 3α
- In t steps there will be 3αt changes

The Q matrix

From \ To	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

The Jukes-Cantor model: the simplest model

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

JC model: one parameter model

1) It assumes that all bases are equally frequent (p=0.25)
2) It assumes that all sites can change and they do so at the same rate – α

The Jukes-Cantor model: the simplest model

	A	C	G	T
A	–	α	α	α
C	α	–	α	α
G	α	α	–	α
T	α	α	α	–

JC model: one parameter model

1) It assumes that all bases are equally frequent (p=0.25)
2) It assumes that all sites can change and they do so at the same rate – α

Improvements on Jukes-Cantor

- Allow base frequencies to be unequal
- Allow transitions to be more common than transversions, in fact, allow separate estimates of the probability of change of all six possible nucleotide substitutions
- Allow the probability of substitution to change along the molecule - ASRV

Slide by Chris Simon 2005

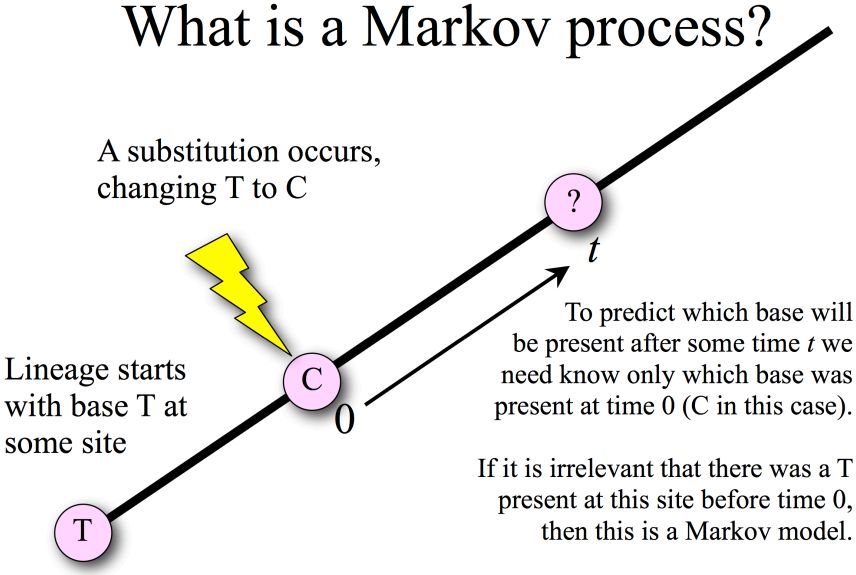
Parameters we are interested in

- The mean instantaneous substitution rate
  =the general mutation rate + rate of fixation in population
- The relative rates of substitution between each nucleotide
- The average frequencies of each base in the dataset
- Topology (part of the model) and branch lengths

Time-homogenous time-continuous stationary Markov models: Assumptions

- Rate of change from base *i* to base *j* is independent of the base that occupied a site prior to *i* (Markov property)

What is a Markov process?



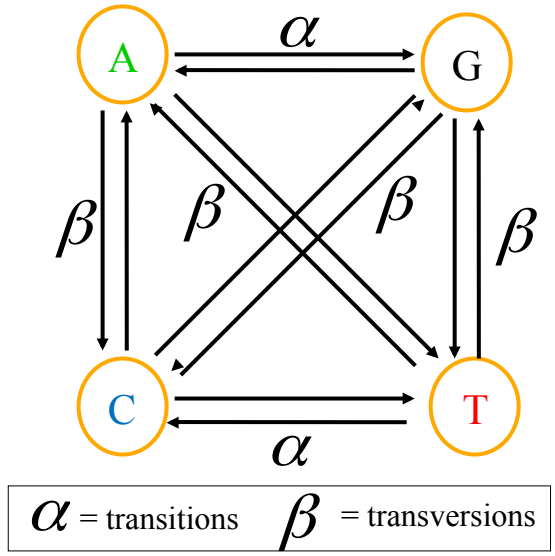
To predict which base will be present after some time *t* we need know only which base was present at time 0 (C in this case).

If it is irrelevant that there was a T present at this site before time 0, then this is a Markov model.

Time-homogenous time-continuous stationary Markov models

- Rate of change from base *i* to base *j* is independent of the base that occupied a site prior to *i* (Markov property)
- **Substitution rate does not change over time (homogeneity)**
- **Relative frequencies of A, G, C, and T are at equilibrium (stationarity)**
- **Rate of change from base *i* to base *j* is identical to the rate of change from base *j* to base *i* (time reversibility)**

Kimura (1980) model: K2P



The Kimura model has 2 parameters

	A	C	G	T
A	–	β	α	β
C	β	–	β	α
G	α	β	–	β
T	β	α	β	–

K2P model is more realistic, but still

1) **It assumes that all bases are equally frequent (p=0.25)**
2) **There are two substitution types (transitions – α and transversions - β**

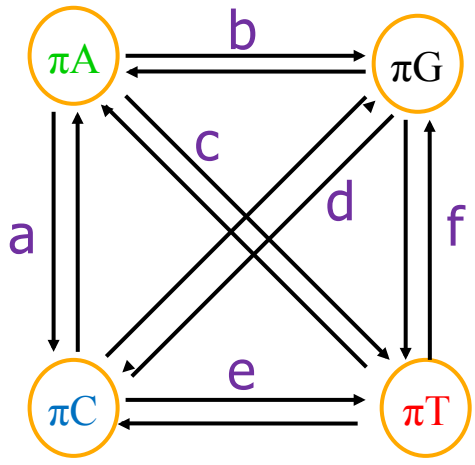
The Hasegawa-Kishino-Yano model

	A	C	G	T
A	–	$\pi_C\beta$	$\pi_G\alpha$	$\pi_T\beta$
C	$\pi_A\beta$	–	$\pi_G\beta$	$\pi_T\alpha$
G	$\pi_A\alpha$	$\pi_C\beta$	–	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\alpha$	$\pi_G\beta$	–

HKY model:

1) **Base frequencies are allowed to vary: πA, πC, πG, πT**
2) **There are two substitution types (transitions – α and transversions - β**

The General Time-Reversible model



The General Time-Reversible model (GTR)

	A	C	G	T
A	–	$\pi_C a$	$\pi_G b$	$\pi_T c$
C	$\pi_A a$	–	$\pi_G d$	$\pi_T e$
G	$\pi_A b$	$\pi_C d$	–	$\pi_T f$
T	$\pi_A c$	$\pi_C e$	$\pi_G f$	–

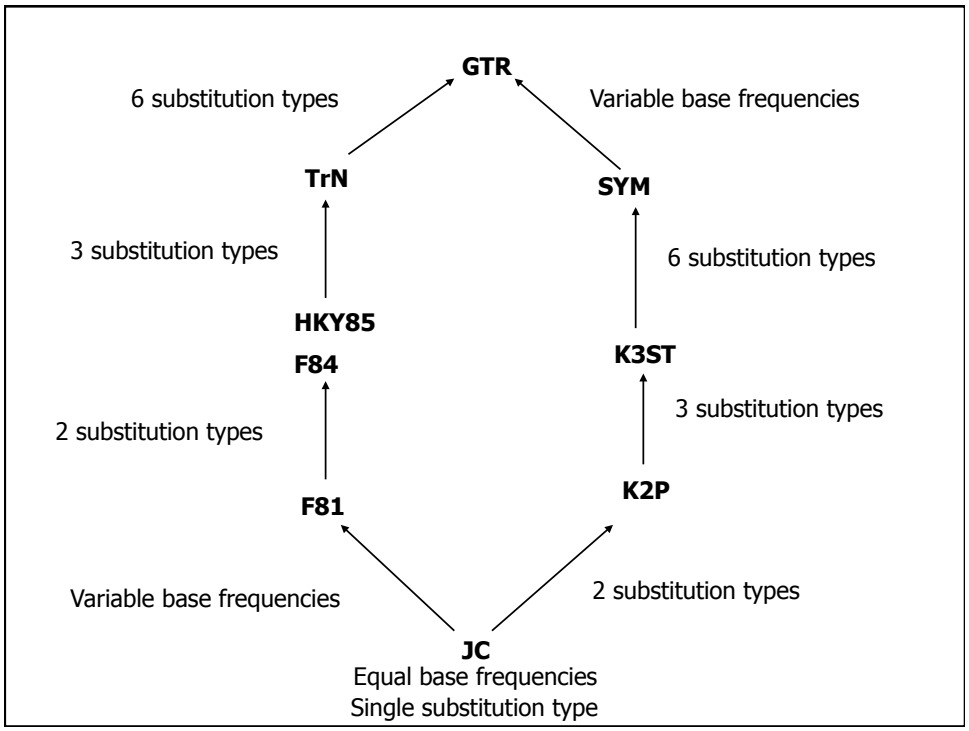
GTRmodel:

1) **Base frequencies are allowed to vary: πA, πC, πG, πT**
2) **There are six substitution types: a, b, c, d, e, f**

The most commonly used models

- **Almost all models used are special cases of one model:**
  - **The general time reversible model - GTR**

ACAGGTGAGGCTCAGCCAATTTGAGCTTTGTCGATAGGT



Models

- **Model parameters can be:**
  - **estimated from the data (using a likelihood function)**
  - **can be pre-set based upon assumptions about the data (for example that for all sequences all sites change at the same rate and all substitutions are equally likely - e.g. the Jukes-Cantor model)**
  - ***wherever possible avoid assumptions* which are violated by the data because they can lead to incorrect trees**

Modelling among-site rate variation (ASRV)

- All of the models so far assume that the rate of change is the same for every position in the alignment
- Biggest difference in substitution rate between variable and "invariable" sites
- Two classes of "invariable sites"
  - Highly restricted "not free to vary"
  - not observed to vary but in fact variable
    - due to convergence or reversal
    - % invariable sites can't be calculated by simple sequence comparison.

Slide by Chris Simon 2005

ASRV, Yang 1996, TREE 11(9):367-372

Why is modelling ASRV important?

- Protein-coding genes – 1st, 2nd, 3rd codon positions evolve differently from each other
- RNA molecules – stems and loops
- Introns vs. exons

RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Amino Acids

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

Typical pattern of variation among codon positions

E.g. in mtDNA in Collembola

56.7% of all variable sites are located in third positions

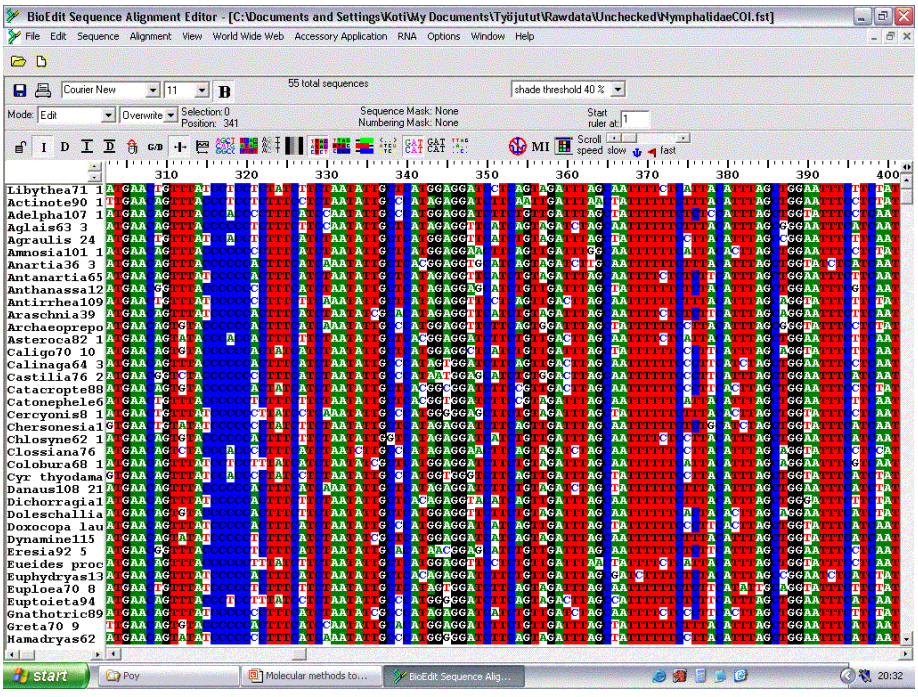
1st 27.9% 2nd 15.4% 3rd 56.7%

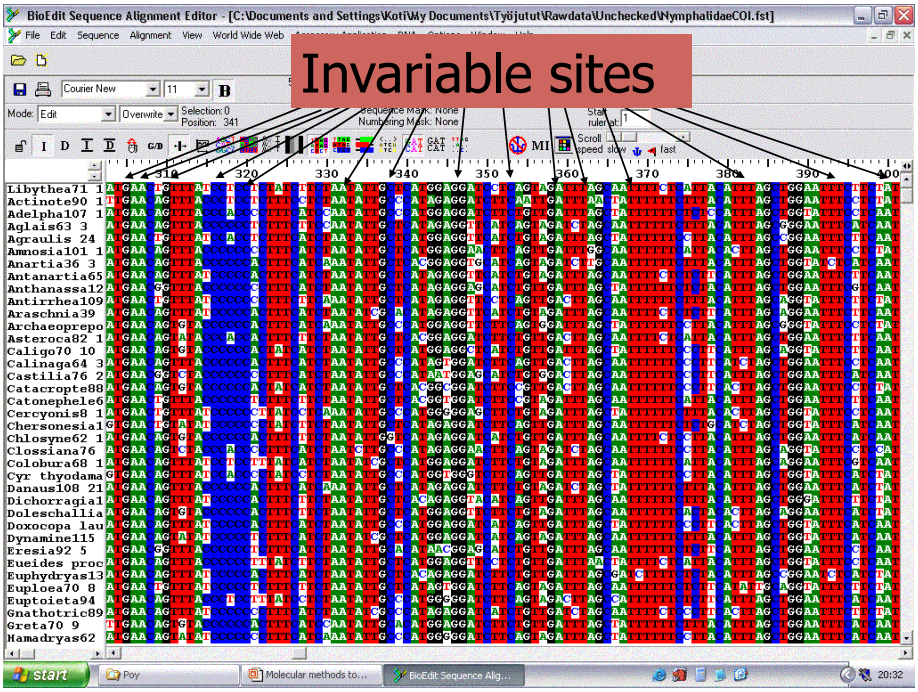
96.9% of all third positions are variable

1st 47.8% 2nd 26.3% 3rd 96.9%

Frati et al. 1997. J. Mol. Evol.

Slide by Chris Simon 2005





Modelling among-site rate variation (ASRV)

- The most common additional parameters are:
  - A correction for the proportion of sites which are invariable (parameter *I*)
  - A correction for variable site rates at those sites which can change (parameter gamma, *G*)
- All models can be supplemented with these parameters (e.g. GTR+*I*+*G*, HKY+*I*+*G*)

Modelling ASRV in variable sites

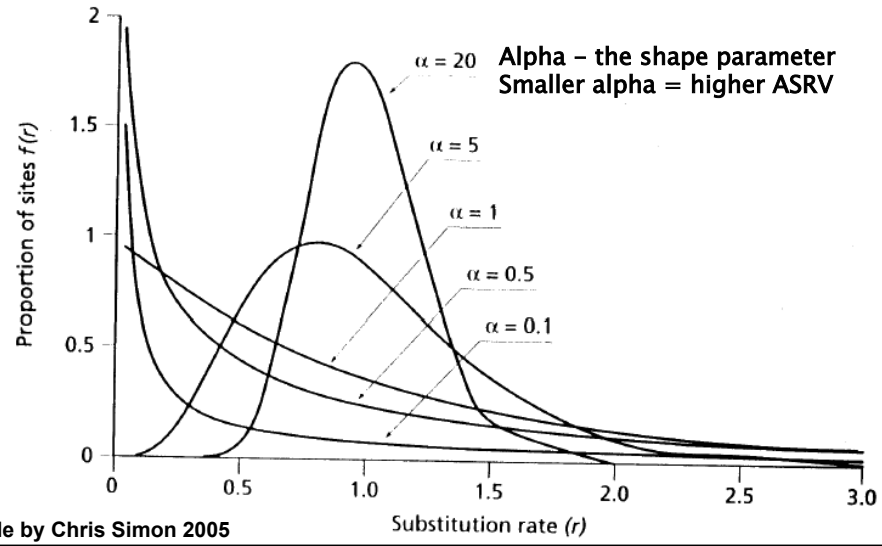
- ASRV in variable sites commonly modelled with a gamma distribution
- Alpha – the shape parameter of this distribution

Slide by Chris Simon 2005

ASRV, Yang 1996, TREE 11(9):367-372

Gamma distribution:

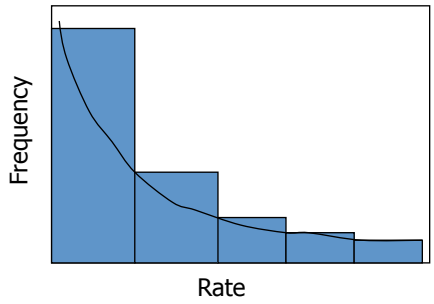
Relative substitution rates for different α values



Slide by Chris Simon 2005

Gamma distribution computationally costly

- Computational difficulties in using continuous distribution
- Most programs use discrete categories

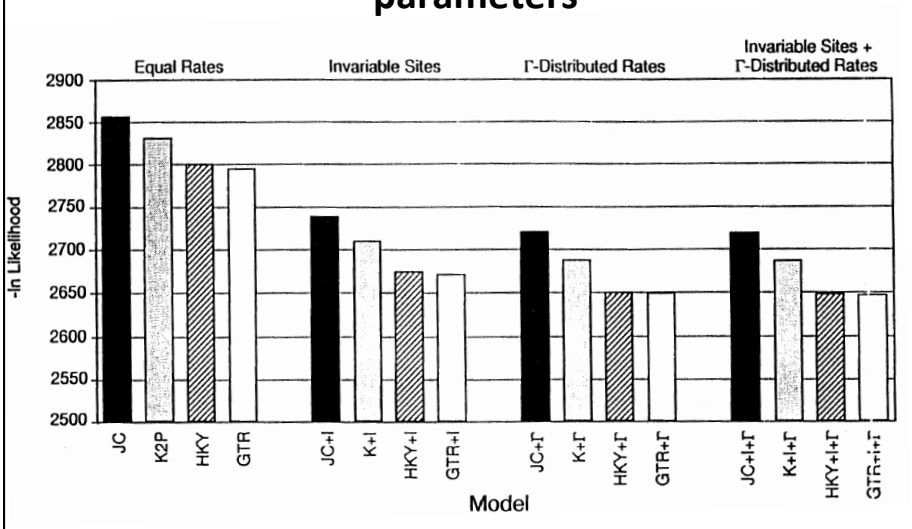


ASRV: Yang discrete model

- Continuous data divided into "n" discrete rate classes (generally 4)
- If $\alpha < 0.2$ Yang recommends more rate classes
- Less computer intensive than obtaining likelihoods by integrating over the continuous gamma distribution

Slide by Chris Simon 2005

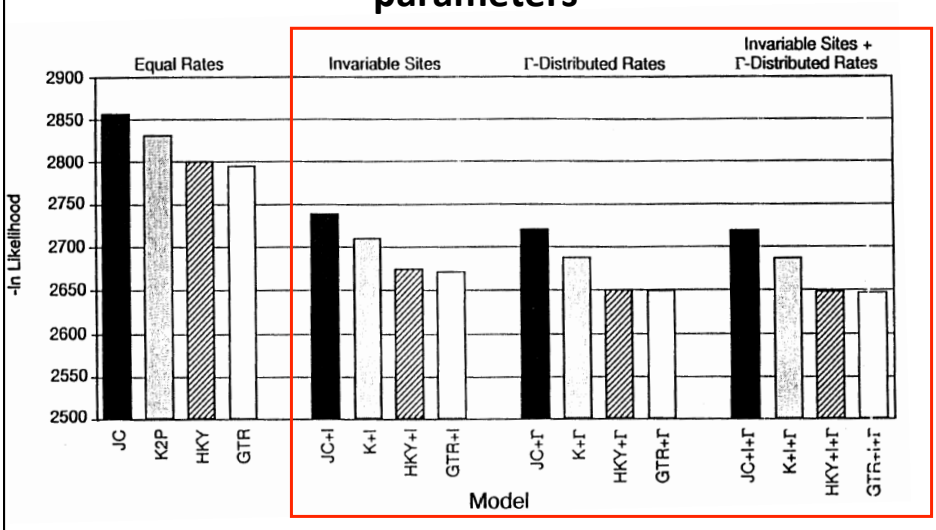
ASRV >> fit improvement than by other parameters



Frati, Simon, Sullivan, Swofford. 1997. JME 44:145-158

Slide by Chris Simon 2005

ASRV >> fit improvement than by other parameters



Frati, Simon, Sullivan, Swofford. 1997. JME 44:145-158

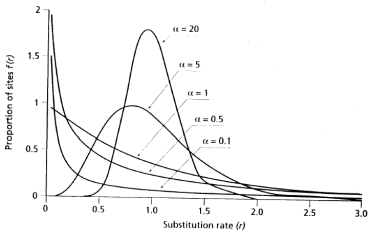
Slide by Chris Simon 2005

Difficulties in estimating ASRV

- The parameters *I* and *G* covary!
- (*I* + *G*) can be estimated, but the values of *I* and *G* are not easily teased apart
- Parameter *G* takes *I* into account, *I* not needed (in many/most? datasets)

Another method for modelling ASRV

- Gamma distribution is always unimodal
  - Not necessarily the case in our dataset!



- Flexible rate heterogeneity across sites model
  - Probability distribution free model so that you can find the distribution that fits your data (FreeRate Model)
  - Implemented in IQ-TREE

Kalyaanamoorthy et al. 2017 (Nature Methods) doi:10.1038/nmeth.4285

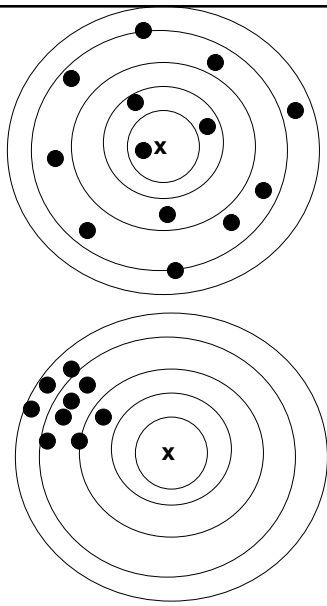
Parameters in models of DNA evolution

- **Numbers of parameters estimated:**
  - **Substitutions (up to 5; 1 fixed, 5 estimated)**
  - **Base composition (1 fixed, 3 estimated)**
  - **Among-site-rate variation**
    - **Gamma shape parameter = 1 parameter**
    - **Invariant sites = 1 parameter**
    - **Gamma + I = 2 parameters**
  - **Partitioned models – add up parameters of each partition**

Slide by Chris Simon 2005

Models can be made more parameter rich to increase their realism

- **But the more parameters estimated, the more time needed, and the more sampling error accumulates**
  - **One might have a realistic model but large sampling errors**
  - **Realism comes at a cost in time and precision!**
  - **Fewer parameters may give an inaccurate estimate, but more parameters decrease the precision of the estimate**
  - **In general use the simplest model which fits the data**



Dave Swofford's Target Analogy

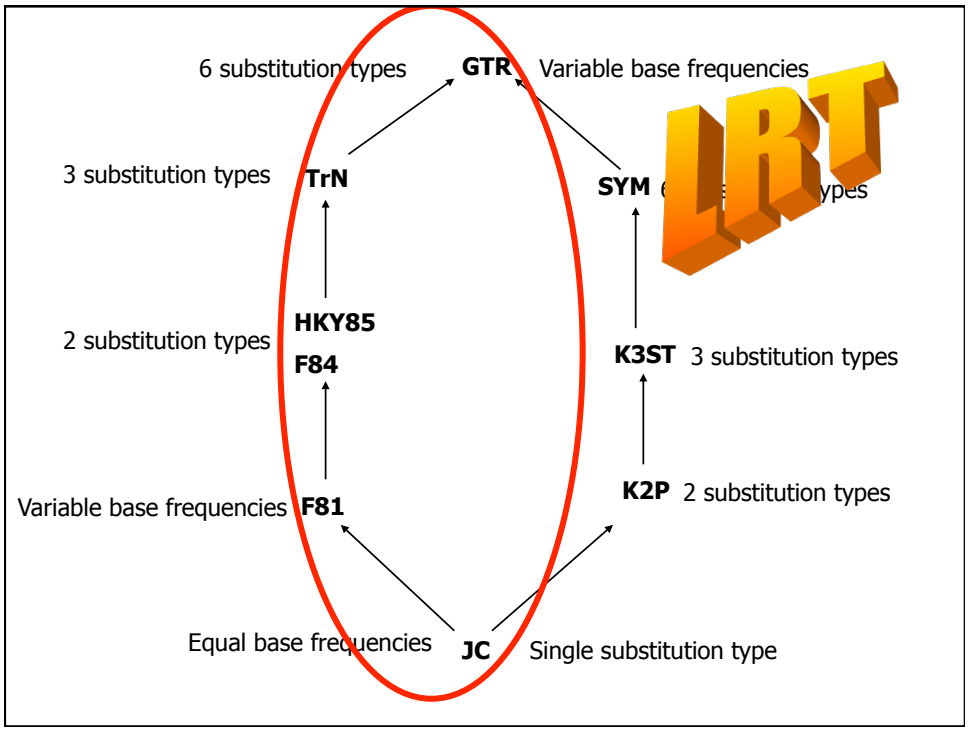
Trade-off between highly parameterized models & model error variance

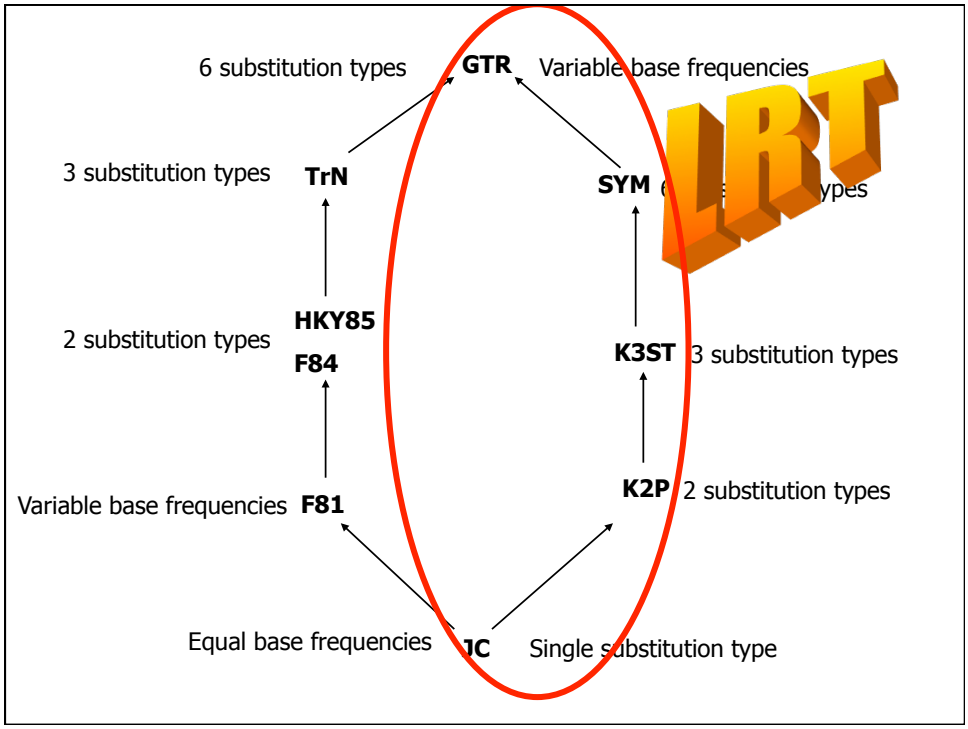
- Many parameters, higher error variance but clustered around the true value (higher accuracy, lower precision)
- Few parameters, lower error variance but may not be centered around the mean (lower accuracy, higher precision)

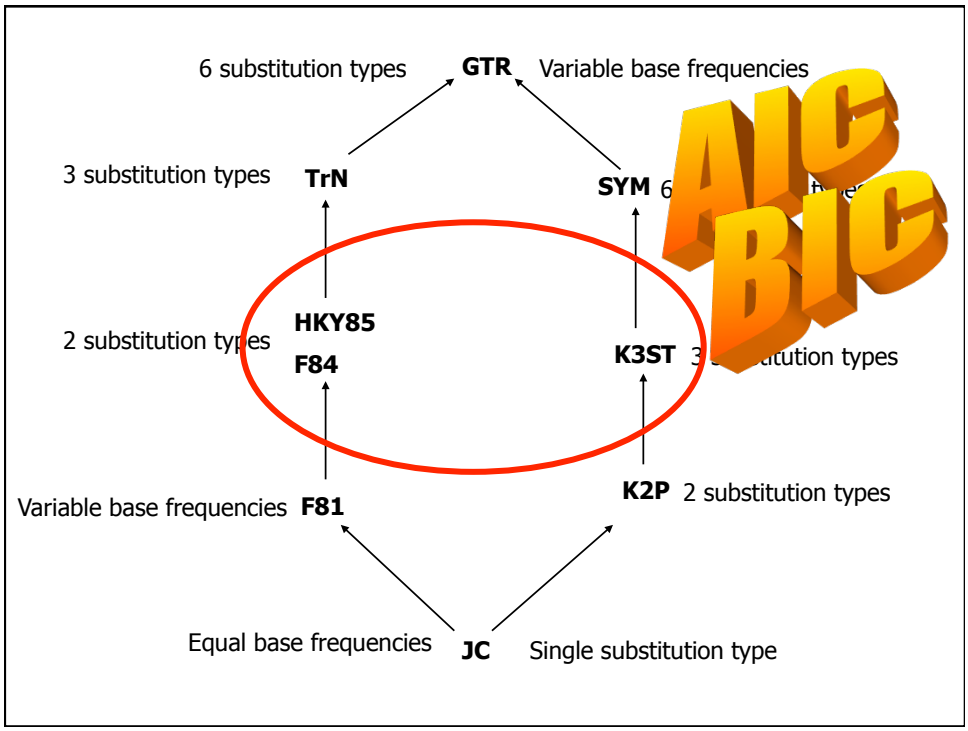
Slide by Chris Simon 2005

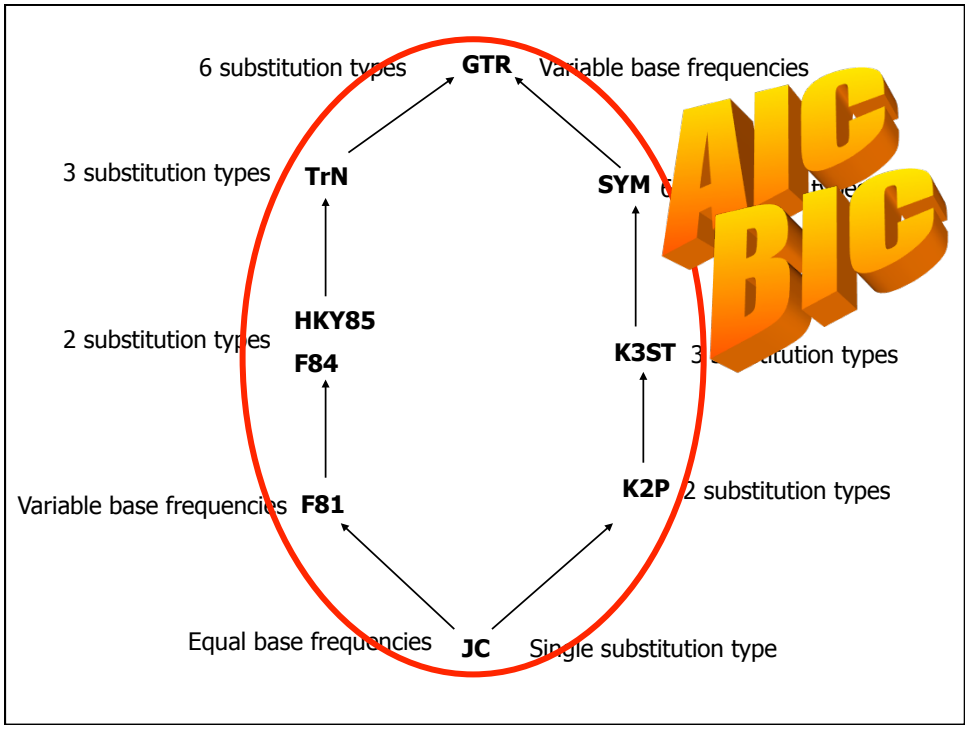
Choosing between models

- Tools to determine whether the model can estimate parameters from the data
- When models are nested
  - Likelihood ratio test (LRT)
- When models are not nested
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)


6 substitution types
GTR
Variable base frequencies
3 substitution types
TrN
SYM
LRT
2 substitution types
HKY85
F84
K3ST
3 substitution types
K2P
2 substitution types
Variable base frequencies
F81
Equal base frequencies
JC
Single substitution type


6 substitution types
GTR
Variable base frequencies
3 substitution types
TrN
SYM
LRT
2 substitution types
HKY85
F84
K3ST
3 substitution types
K2P
2 substitution types
Variable base frequencies
F81
Equal base frequencies
JC
Single substitution type


6 substitution types
GTR
Variable base frequencies
3 substitution types
TrN
SYM
HKY85
2 substitution types
F84
K3ST
K2P
2 substitution types
Variable base frequencies
F81
Equal base frequencies
JC
Single substitution type


6 substitution types
GTR
Variable base frequencies
3 substitution types
TrN
SYM
HKY85
2 substitution types
F84
K3ST
K2P
2 substitution types
Variable base frequencies
F81
Equal base frequencies
JC
Single substitution type

Estimation of substitution model parameters

- Yang (1995) has shown that parameter estimates are reasonably stable across tree topologies provided trees are not "too wrong"
- Thus one can obtain a tree using a quick method and then estimate parameters on that tree
- These parameters can then be used to calculate the likelihood of a model for model comparison

Model-testing programs

- **Modeltest**
  - Posada & Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9): 817-818.
- **jModeltest**
  - Darriba et al. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8), 772.
- **PartitionFinder**
  - Lanfear et al. 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. MBE 34(3), 772 − 773.
- **ModelFinder built into IQ-Tree**
  - S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, and L.S. Jermiin (2017) ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates, Nature Methods, 14:587–589. https://doi.org/10.1038/nmeth.4285

Model testing easier nowadays

- Bayesian statistical framework
  - MrBayes has a model jumping feature
  - It samples over all possible models based on their probabilities
  - No longer necessary to test for which model is optimal
- Maximum Likelihood framework
  - IQ-Tree - ModelFinder implemented

Partitioned models (1/2)

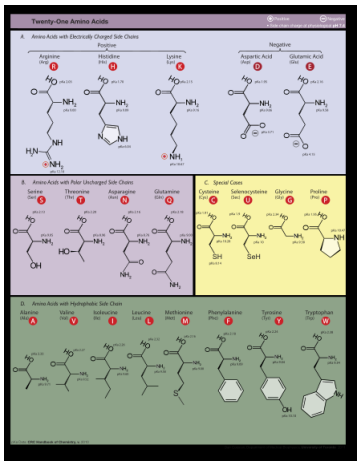
- Today's datasets tend to be large, including hundreds or thousands of genes
- Unrealistic to have the same model for the whole dataset (underparameterization)
- Modelling DNA substitution for separate sections of the data (partitions)
  - E.g. different genes, codon positions, introns/exons, etc.
- To avoid overparameterization, partitions with similar properties can be merged

Partitioned models (2/2)

- This approach allows us to accommodate heterogeneity across data subsets in overall rate and in substitution model parameters
- In some programs also possible to unlink topology and branch lengths so that each data subset evolves differently from each other
- Built into IQ-Tree

Models of amino acid substitution

- Empirical and mechanistic models
- Empirical models: based on empirical AA replacement with matrices from different taxa
  - 20 amino acids – 20x20 matrix too big for estimation
  - Examples: JTT, WAG, LG, MtREV (for mitochondria), Blosum62
- Mechanistic models:
  - e.g. codon models (61x61 matrix
  - Tend to outperform empirical models BUT
  - Computationally very intensive



Recommended reading

- Christoph Bleidorn (2017) Phylogenomics: An Introduction (DOI: 10.1007/978-3-319-54064-1)
- Hoff et al. 2016. Does the choice of nucleotide substitution models matter topologically? BMC Bioinformatics 17: 143. doi.org/10.1186/s12859-016-0985-x
- Kainer & Lanfear. 2015. The Effects of Partitioning on Phylogenetic Inference. Molecular Biology and Evolution, 32(6), 1611–1627. doi.org/10.1093/molbev/msv026