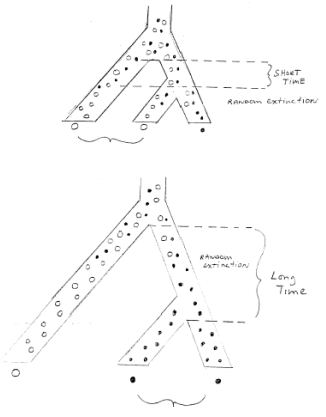
Accommodating Biases in DNA Data

Chris Simon
Ecology & Evolutionary Biology
University of Connecticut

DNA data do not guarantee the correct phylogeny

1. The problem of shared ancestral polymorphisms
2. Multiple substitutions at a single site hide earlier substitutions



Misleading DNA evolution

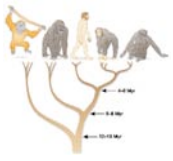
Shared ancestral polymorphisms (lineage sorting, gene-tree vs species-tree disagreement)

Random extinction

Depends on time between speciation events

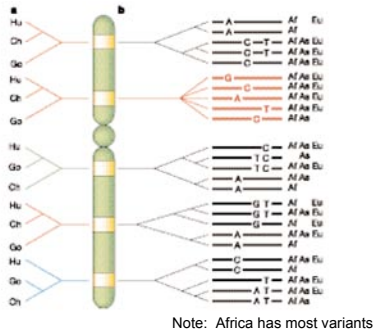
History varies among gene regions

Human, Chimp, Gorilla



Svante Pääbo
Nature. 421.
23 Jan 2003

Among Human Poplns



Note: Africa has most variants

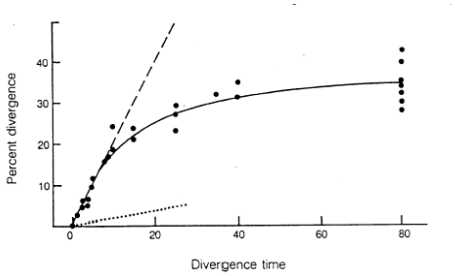
Other reasons for gene tree not matching species tree.

- Hybridization (organelle genomes move across species boundaries and introgress faster than nuclear genes)- recent or ancient
- Horizontal gene transfer

Solution: use methods that create species trees from multiple gene trees (Deep coalescence, gene tree parsimony/consensus tree, Bayesian coalescent gene tree conditioned on species tree, e.g., BEST)

Misleading DNA evolution

2. Multiple substitutions hide previous changes



Brown et al. 1979. PNAS 76:1967

Corrections for multiple substitutions

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)

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

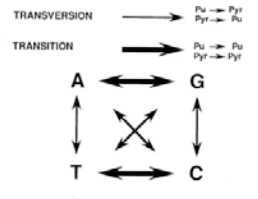
Jukes-Cantor Assumptions Incorrect

1. Nucleotide bias is common

   mtDNA honey bee 84.9% A+T;
   D. yakuba 78.6%;
   Evanoid wasps 77.8% (low for hymenoptera)

   Some bacterial lineages are A+T rich, others G+C rich.

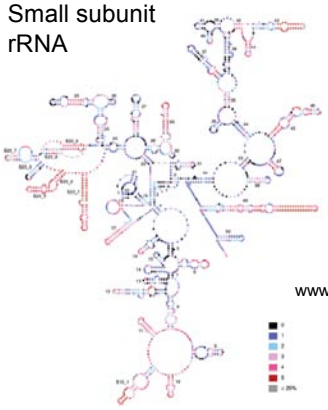
2) Transitions more common than transversions



Pur - Pyr mispairs lead to transitions

In next round of replication

Small subunit rRNA



3) Substitution probabilities vary along a molecule in relation to structural and functional constraints

www.psb.ugent.be/rRNA/varmaps

500 eukaryotes mapped onto yeast structure, Yves Van de Peer

Simon et al. 2006. AREES.

Protein coding genes

Primary structure substitutions influenced by triplet code

```
                    50                                                  60
      T  G  L  F  L  A  M  H  Y  S  P  D  A  S  T  A  F  S  S  I  A
HUMAN ACA GGA CTA TTC CTA GCC ATG CAC TAC TCA CCA GAC GCC TCA ACC GCC TTT TCA TCA ATC GCC
  R1  T.. ..C ... ..T ... ..T ..A ..T ..T A.. ... ... A.. CTC ..G ..A ... ... ... G.T A..
  R2  T.. ..C ... ..T ... ..T ..A ..T ..T A.. ... ... A.. CTC ..G ..A ... ... ... G.T A..
  R3  T.. ..C ... ... ..G ..T ..A ..T ... A.. ... ... A.. ATC ..G ..A ... ... ... G.T A..
  R4  T.. ..C ... ..T ... ... ..A ... ... A.. ... ..T A.. CTC ..A ..A ..C ... ... G.T ...
  R5  T.T ..C ... ... ... ... ..A ... ..T A.. T.. ..T AG. ..T .A. AG. ... ... ... G.A A..

  B1  ... ..C ... C.. T.. ... GCA ..T ..T A.. G.T ..T A.T ..C CTA ... ..C G.T ..C G.A ...
  B2  ... ..C ... C.. ... ... GCA ... ... A.. G.T ..T A.T ..C CTA ... ..C G.T ..C G.A A..
  B3  ... ..C ..G C.. ... ..T GCA ... ... A.. G.T ... A.. ..C CTA ... ... G.C ..C G.A ...
  B4  ... ..C ... ... ... ... GCA ..T ... A.. G.C ... A.. ..C CT. ... ..C AAC ..C G.. ...
  B5  ... ..C ... C.A ... ... ... ... ... A.. G.C ... A.. ..C CTA ... ..C A.C ... G.A ...

  F1  ... ..C ..T ... ..T ..A ..A ... ... A.T T.C ..T AT. G.. ..A ... ... ... ..C G.. ...
  F2  ... ..C ..T ... ..T ..A ..A ... ... A.T T.C ..T AT. G.. ..A ... ... ... ..C G.. ...
  F3  ... ..C ..T ... ..T ..A ..A ... ... A.T T.C ..T AT. G.. ..A ... ... ... ..C G.T ...
  F4  ... ..C A.. ... ..T ..A ..A ..T ... A.T T.T ... AT. G.. ..A ... ... ... ..C G.T ...
  F5  ... ..C ..T ... ... ... ..A ... ..T A.C T.C ... AT. G.C ... ... ... ..C ..C G.. ...
```

Rodent, Bird, Fish **Kocher et al. 1989. PNAS 86:6196**

Rate of AA substitution also varies among genes (most obvious at deeper levels)

Table 1. Percentage of amino acid similarity at various divergence levels

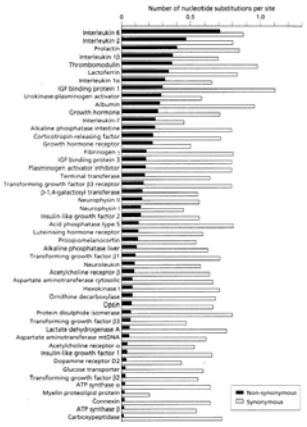
Gene	*D. melanogaster–D. yakuba*[a] 3–19[b, c]	Human–cow[d] >90[b, c]	Apis–*Drosophila*[f] >280[b, g]	*Drosophila*–Locust[h] >300[b, i]	Mouse–*Drosophila*[m] >700[b, k]
COI	99	91	70	82	75
COII	98	73	55	66	57
COIII	98	87	53	73	64
Cyb	96	79	53	76	68
ND1	96	78	47	68	45
ND4	95	74	45	58	42
ND3	95	74	49	62	43
ND5	99	70	42	64	40
ND4L	99	74	37	51	36
A6	96	78	47	73	34
ND2	94	63	27	47	34
A8	96	52	46	41	26
ND6	91	63	31	45	17

Simon et al. 1994. Annals ESA

Rate of AA replacement also varies among nucl. genes; silent sites less variable among genes

Artiodactyls, rodents, & primates

Nei 1987 Text P. 240



Not so simple to identify "silent sites." Third position not completely degenerate

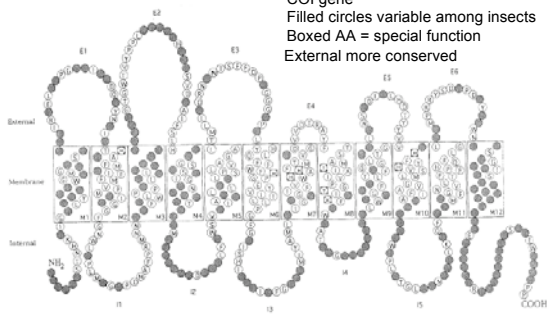
	AGA								UUA	
	AGG								UUG	
GCA	CGA						GGA		CUA	
GCC	CGC						GGC		CUC	
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	CUG	AAA
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	CUU	AAG
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Leu	Lys

			AGC							
			AGU							
		CCA	UCA	ACA			GUA			
		CCC	UCC	ACC			GUC	UAA		
	UUC	CCG	UCG	ACG		UAC	GUG	UAG		
AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA		
Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Stop		

Different probability of change in 2 vs 3 vs 4 vs 6-codon families; "Universal" Genetic Code (nuclear)

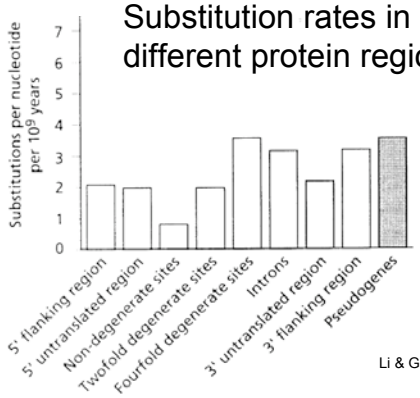
Rate of evolution of AA related to structure and function

COI gene
Filled circles variable among insects
Boxed AA = special function
External more conserved



Lunt, Zhang, Szymura, Hewitt. 1996. MBE 5:153-165

Substitution rates in different protein regions



Li & Graur 1991

Accommodating biases

- Can be done by weighting in parsimony analyses
- Using models of evolution in maximum likelihood and Bayesian analyses (models that improve on Jukes Cantor)

Biased substitutions accommodated in parsimony by weighting

Substitution bias: By weighting (step matrices) but very difficult to do well and sophisticated weighting slows MP analysis significantly.

ASRV: By using successive approximations or other iterative weighting but highly dependent on initial tree. Or by down-weighting variable regions.

Simon et al. 1994. Annals ESA

History of ASRV: Weighting to accommodate ASRV

Less variable characters (less homoplasy) given more weight

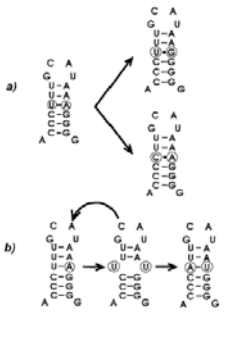
Weight 1st & 2nd position more heavily than third (Irwin et al. 1991) Clearly important

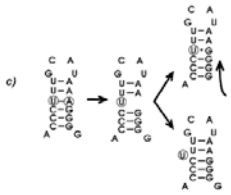
Weight stems more heavily than loops (or vice versa) (DeSalle et al. 1987, Vawter & Brown 1993, Springer & Douzery 1996). DO NOT DO THIS

History of ASRV: Weighting to accommodate ASRV

Weight paired regions less than unpaired regions due to compensatory changes (Wheeler & Honeycut 1988, Hillis & Dixon 1991) How much to down-weight? Fifty percent too much. Hillis & Dixon recommended less.

Down-weighting 50% is too much





Hickson, Simon, Cooper, Spicer, Sullivan, & Penny. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. MBE 13: 150-169.

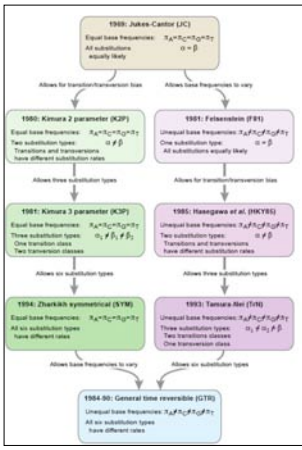
Huelsenbeck & Ronquist include a doublet model in MrBayes

- Accommodates paired, correlated changes
- Kjer 2004 found this model improved results of 18S rRNA phylogeny of insect orders

History of ASRV: Weighting to accommodate ASRV

Divide a rRNA molecule into rate variability classes (Simon 1991, Hickson 1993)
Problem: Where to divide?

Calculate probability of substitution at individual sites….



Models of Evolution

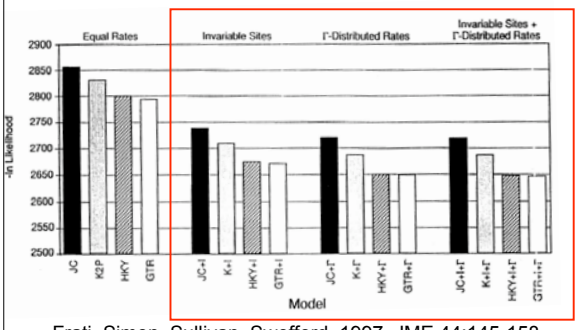
Improving on Jukes Cantor

- GTR 84-90 most general. Implementation delayed due to computational complexity
- HKY 85 & Felsenstein 84 similar
- All are reversible
- Some allow unequal nucleotide frequency but all assume nucleotide bias is same in all taxa
- None include ASRV but it can be added as G, I, or I+G

Note: Tamura-Nei 93 paper did include ASRV.

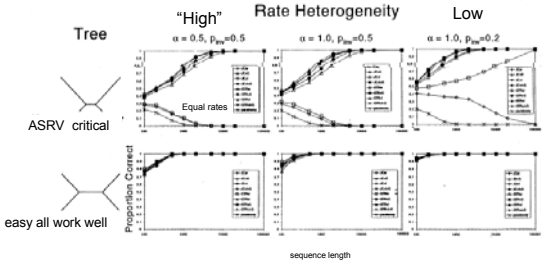
Simon et al. 2006 AREES

ASRV >> fit improvement than models



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

Model effectiveness in phylogeny construction depends on ASRV & tree shape



Sullivan & Swofford 2001. Syst. Biol. 50: 723-729

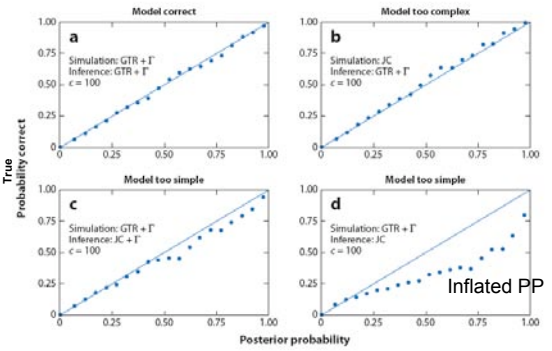
Yang 1996. TREE. Summarized effects of ignoring ASRV

- Under estimates ts/tv rate ratio
- Lower phylogenetic informativeness
- Causes Tajima's D for neutrality to mimic patterns of population expansion.
- Results in under- or over-estimates of divergence times

More effects of ignoring ASRV

- Strongly affects bootstrap support (Buckley et al. 2001 Syst. Biol., Cunningham & Buckley 2002. MBE)
- Strongly affects Bayesian posterior probabilities in simulations (Lemmon & Moriarty 2004. Syst. Biol.)
- Reduces effectiveness of models of AA evolution (Susko et al. 2003. Syst. Biol.)
- Results in incorrect inferences of early, rapid cladogenesis (Ravell et al. 2005. Syst. Biol.)

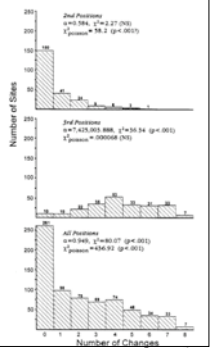
Ignoring ASRV in too-simple model --> inflated PP (Huelsenbeck & Rannala 2004)



Problem: Modeling all genes combined creates a single average model that doesn't fit any one partition very well

Partitioned models avoid "average ASRV" problem

- Partitioned models w/ ASRV improved branch supports & tree likelihood
  - Nylander et al. 2004
  - Castoe et al. 2004
  - Brandley et al. 2004
- In mtDNA gene boundaries can be ignored



Partitioned Models

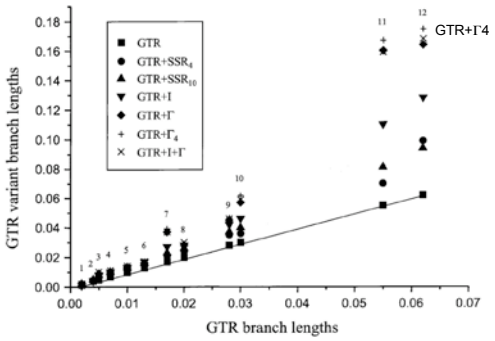
- An attempt to accommodate different rate variation in different partitions.
- Partitioned likelihood available early in PAML but slow
- Partitioned likelihood could only be done in PAUP* by summing across likelihood runs*
- RaxML allows partitions
- Partitioning in MrBayes since 2003

*suggested by Swofford and described in Schwartz et al. 2003.

Partitioned Models: early days

- SSR models in PAUP* accommodate different rate variation in different partitions.
- But have restrictive assumption- the rate of evolution of all sites within each partition is equal
- Newer SSR-gamma or partitioned gamma calculates gamma parameters separately for each partition

How ASRV is accommodated matters



Buckley, Simon & Chambers. 2001. Syst. Biol.

Buckley et al. 2001. Conclusions

- Important to correct for ASRV
- Large difference among corrections in branch lengths & bp
- SSR model branch lengths not much different from uncorrected
- The more distantly related the taxa, the more the ASRV correction matters

Difficult to know how to partition rRNA

Solution: Mixture Models. Pagel & Meade 2004

- Characters viewed as generated by one of a number of distributions
- Each of these distributions contributes to the likelihood during analysis
- Parameters of each model distribution & weights assigned to them estimated from the data
- At end of analysis, each character assigned a probability of belonging to each model

Pagel & Meade 2004

TABLE 6. Numbers of stem and loop sites fitted best by respective rate matrices: mammalian 12S data.

Secondary structure	Q1	Q2	Q3	Q4
Stem	76	21	71	296
Loop	133	112	83	249

Each Q matrix was GTR + G (partitioned & not); parameters varied among the four

- No one model fitted stems vs loops
- Instead, various stem & loop sites were distributed across 4 models with model 4 fitting best the largest number of sites

Pagel & Meade 2004

TABLE 6. Numbers of stem and loop sites fitted best by respective rate matrices: mammalian 12S data.

Secondary structure	Q1	Q2	Q3	Q4
Stem	76	21	71	296
Loop	133	112	83	249

Advantages of Mixture Models

- Patterns are emergent- not pre-specified
- Not constrained
- Not dependent on knowledge of secondary structure

Mixture Models vs Model Averaging*

- Model averaging- several models. Weights and models fixed (e.g., from AIC scores). Models often differ in number of parameters.
- Mixture models- Weights are estimated during analysis; multiple models often the same but differ in the parameter values

* Thanks to Paul Lewis