How Molecules Evolve

Guest Lecture: Principles and Methods of Systematic Biology

11 November 2013

Chris Simon

Approaching phylogenetics from the point of view of the data …

Understanding how sequences evolve informs how we model DNA data.

Advantages of Molecular Data for Tree Building

- In all living organisms
- most basic level of organization
- Genes relatively easy to homologize
- Knowledge of molecular evolution
  - a) transmission
  - b) mutation/substitution

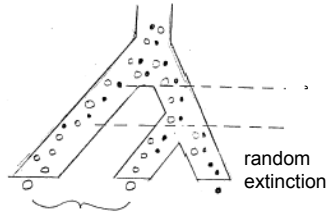
Advantages of Molecular Data for Tree Building

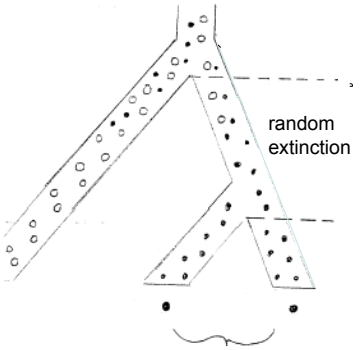
- Can study extinct biota (younger better)
- Large number of characters
- mtDNA and nuclear DNA independently inherited. Some nuclear genes independent of others.

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

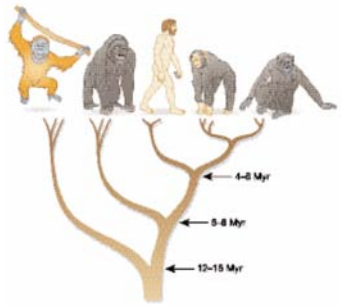
NATURE | VOL 421 | 23 JANUARY 2003 | www.nature.com/nature

The mosaic that is our genome

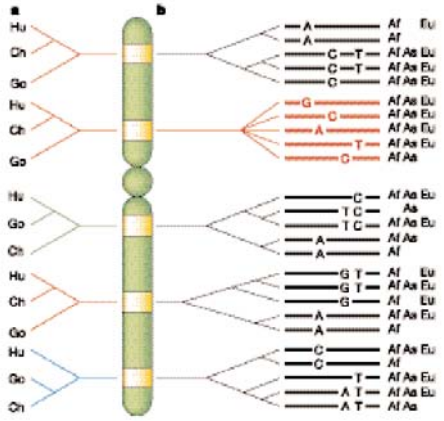
Svante Pääbo

Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany (e-mail: paabo@eva.mpg.de)

The discovery of the basis of genetic variation has opened inroads to understanding our history as a species. It has revealed the remarkable genetic similarity we share with other individuals as well as with our closest primate relatives. To understand what make us unique, both as individuals and as a species, we need to consider the genome as a mosaic of discrete segments, each with its own unique history and relatedness to different contemporary and ancestral individuals.

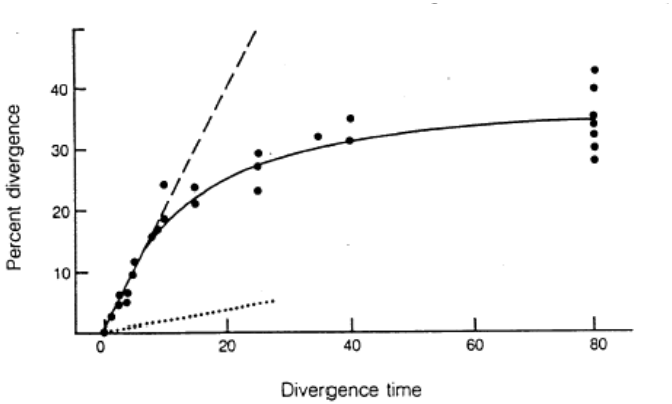


Human, Chimp, Gorilla



Misleading DNA evolution

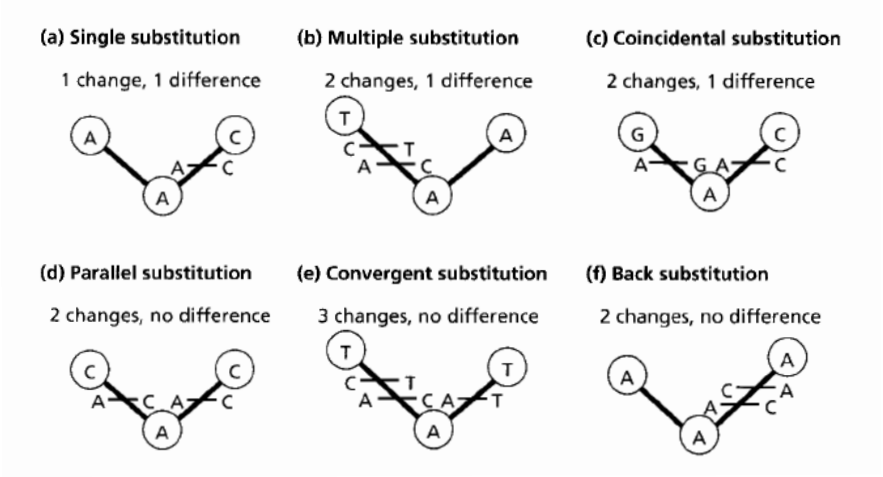
Multiple substitutions hide previous changes



Brown et al. 1979. PNAS 76:1967

Types of Substitutions

Assuming we could know the ancestral states …



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

Difference between mutation and substitution

- Substitution = mutational changes observed in populations (not random)
- Mutations = not all observed in populations, randomly distributed

  1) removed by proof reading enzymes

  2) cause death of cell, gamete, embryo

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)

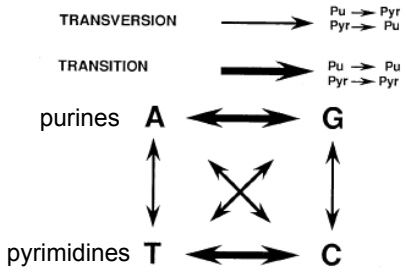
Also: probability substitution & base composition remains constant over time/across lineages

Jukes-Cantor Assumptions Incorrect

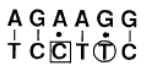
1. Nucleotide bias is common
   mtDNA honey bee 84.9% A+T; D. yakuba 78.6%

2. TS/TV bias is common

3. Substitution probabilities vary along a molecule in relation to structural and functional constraints (ASRV, RAS models)

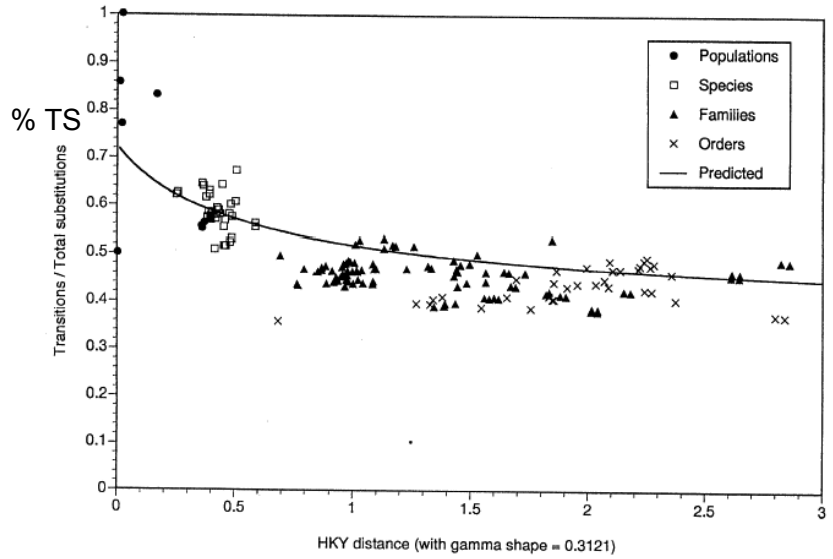


Pur - Pyr mispairs lead to transitions

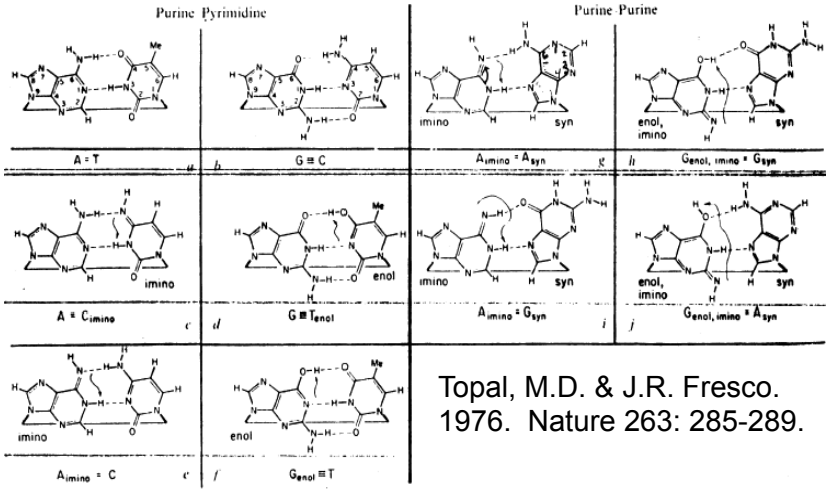


In next round of replication due to Watson-Crick pairing

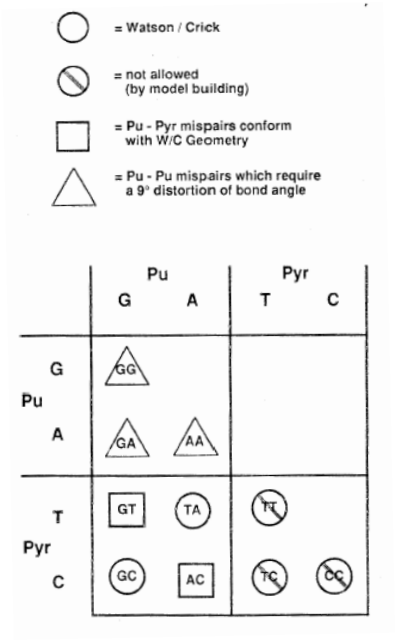
Transitions vs Total Substitutions



Frati, F., C. Simon, J. Sullivan, D. Swofford. 1997. JME 44: 145-158.



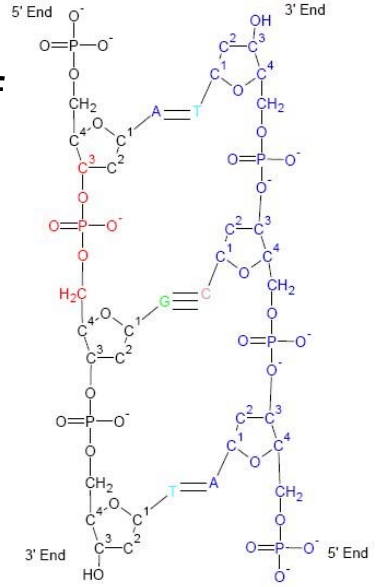
Topal, M.D. & J.R. Fresco. 1976. Nature 263: 285-289.



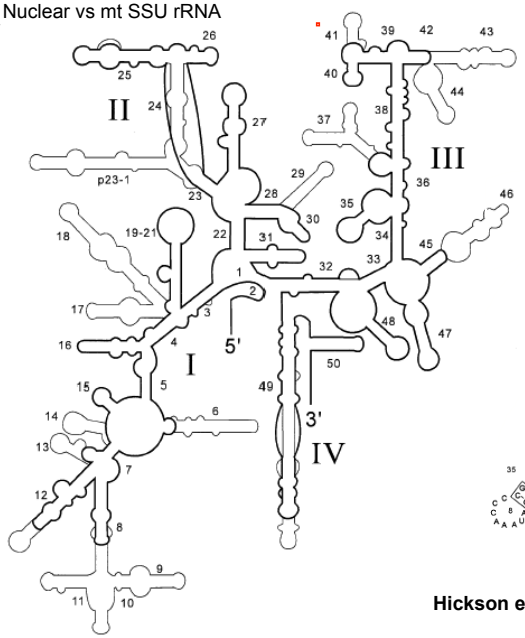
Definition: 3’ vs. 5’ end of DNA

Phosphate groups attached to the carbons number 3 & 5 in the deoxyribose sugar.

The two sides of the helix are antiparallel (run in opposite directions).



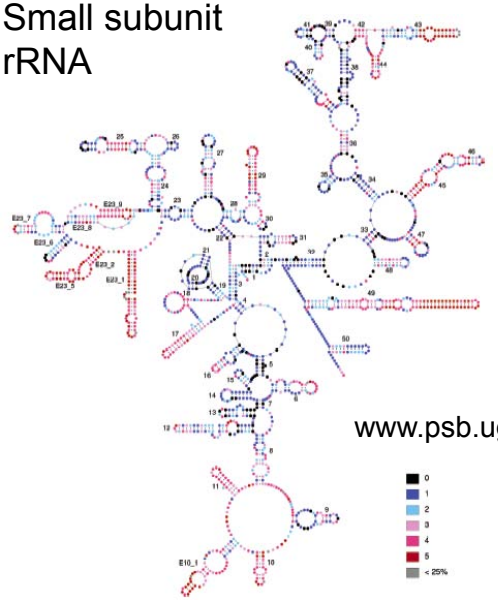
Nuclear vs mt SSU rRNA



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

Hickson et al. 1996. MBE 13: 150-169.

Small subunit rRNA



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

Red = most variable
Blue = least
Invariable = black

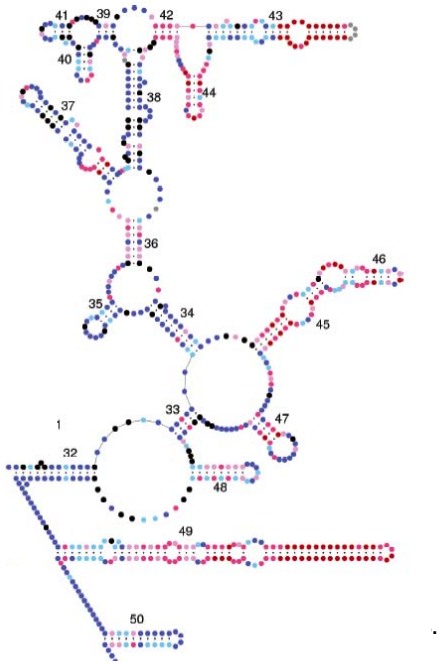
www.psb.ugent.be/rRNA/varmaps

500 eukaryotes mapped onto yeast structure, Yves Van de Peer

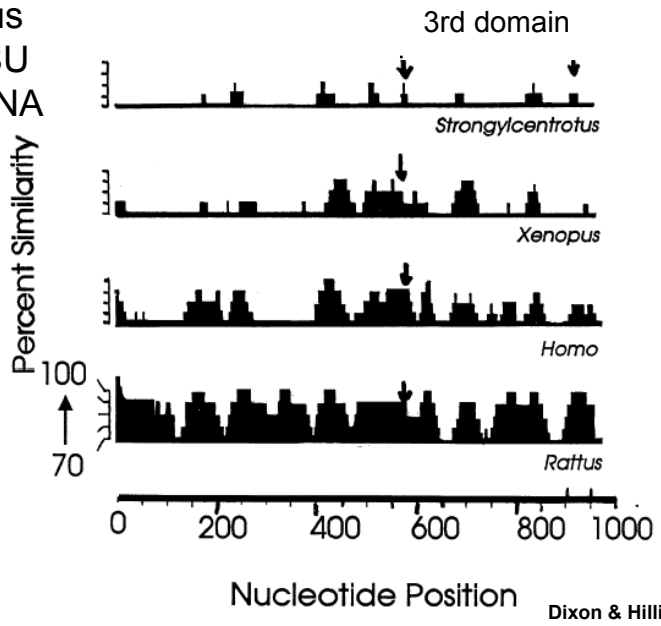
Simon et al. 2006. AREES.

Third domain small subunit rRNA

Red = most variable
Blue = least
Invariable = black

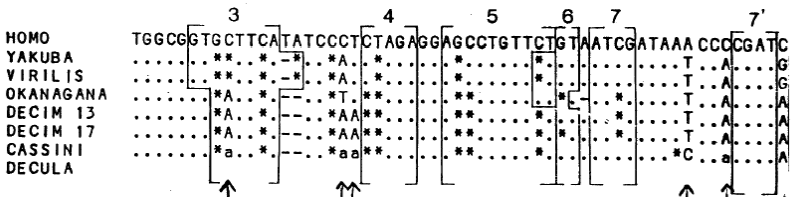


Mus SSU rRNA



Dixon & Hillis 1993 MBE

Transitions in conserved regions



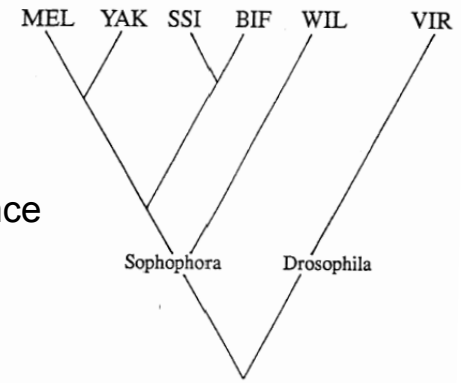
* = transitions
• = identity

Fast sites in slow molecules

Position No.						
14318	A	G	A	G	G	G
14330	T	A	T	T	A	A
14322	T	T	A	A	A	T
14392	A	A	T	T	T	T
14460	A	A	G	A	G	G

12S SSU 3rd domain

Few sites variable (5/300 bp)
At least four experience homoplasy



Simon et al. 1996. MBE

12S SSU

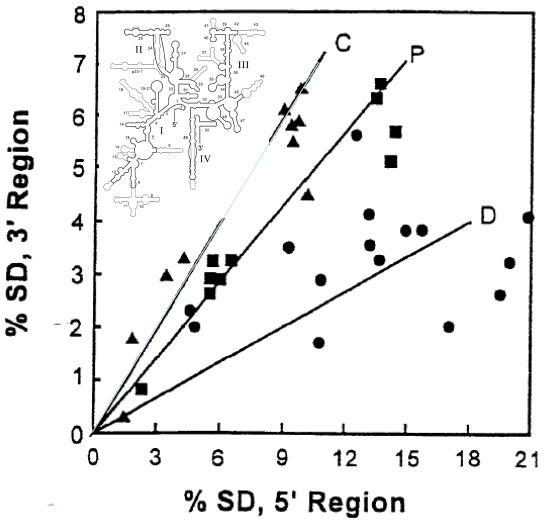
2nd vs 3rd domain Mosaic Evolution

Cicada

Primates

Drosophila

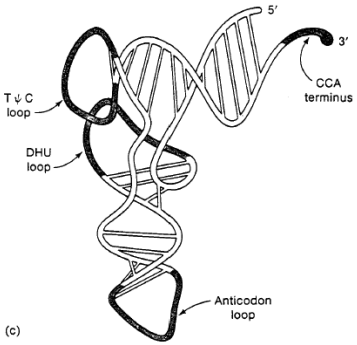
% pairwise sequence divergence

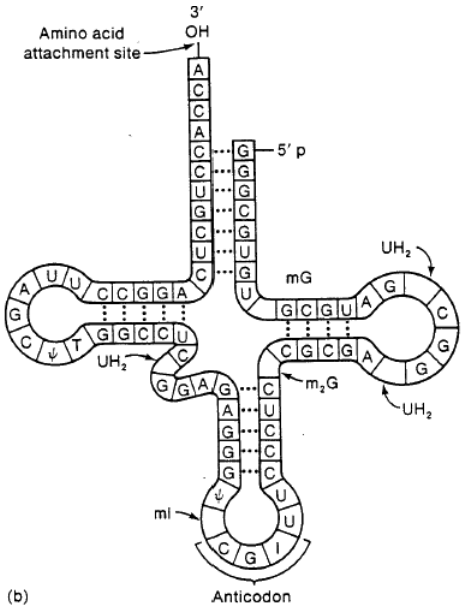


Note different scales.

Simon et al. 1996. MBE

tRNA





Protein coding genes

```
                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.. ...
```

Primary structure substitutions influenced by triplet code

Rodent, Bird, Fish

Kocher et al. 1989. PNAS 86:6196

Third position not completely degenerate

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

Deviation from “Universal” Code; mitochondrial code is different and varies among taxa

In Evanoid wasps: start codon of mitochondrial Nad1 = TTG; might be common across Hymenoptera.

In collembola, ATC, ATT, and ATA are used as start codons in the COII gene.

In Vertebrates both AUA and AUG are Methionine start codons

Vertebrate mitochondrial code.
Red = differences from the Universal code

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Trp	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Met	ACA Thr	AAA Lys	AGA Stop	A
	AUG Met	ACG Thr	AAG Lys	AGG Stop	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Codon bias (nucleotide bias)

TABLE 3 (part)

Codon Usage in COII Genes of 13 Species of Insects

Ser (S)	TCT	62	Tyr (Y)	TAT	92	Cys (C)	TGT	21
	TCC	17		TAC	21		TGC	8
	TCA	96	TER	TAA	7	Trp (W)	TGA	77
	TCG	1		TAG	0		TGG	1
Pro (P)	CCT	56	His	CAT	52	Arg (R)	CGT	22
	CCC	17	(H)	CAC	25		CGC	4
	CCA	52	Gln	CAA	86		CGA	59
	CCG	3	(Q)	CAG	4		CGG	4
Thr (T)	ACT	62	Asn	AAT	178	Ser (S)	AGT	20
	ACC	19	(N)	AAC	40		AGC	4
	ACA	95	Lys	AAA	71		AGA	46
	ACG	3	(K)	AAG	9		AGG	4
Ala (A)	GCT	31	Asp	GAT	95	Gly (G)	GGT	27
	GCC	18	(D)	GAC	29		GGC	8
	GCA	59	Glu	GAA	115		GGA	58
	GCG	1	(E)	GAG	7		GGG	10

Liu & Beckenbach 1992. MPE- codons ending in T & A most used

D. yakuba mtDNA- no CAG, TAG, CGC, AGG in any proteins

Triplet code- primary structure

Collembola display a typical pattern of variation among codon positions

56.7 % of all variable sites are located in third positions

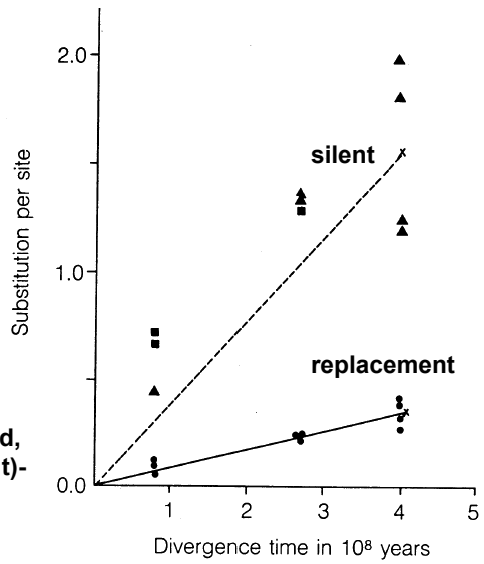
1st 27.9% 2nd 15.4% 3rd 56.7%

96.9% of all third positions are variable

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

Frati et al. 1997. JME

Substitution Rates Insulin Gene

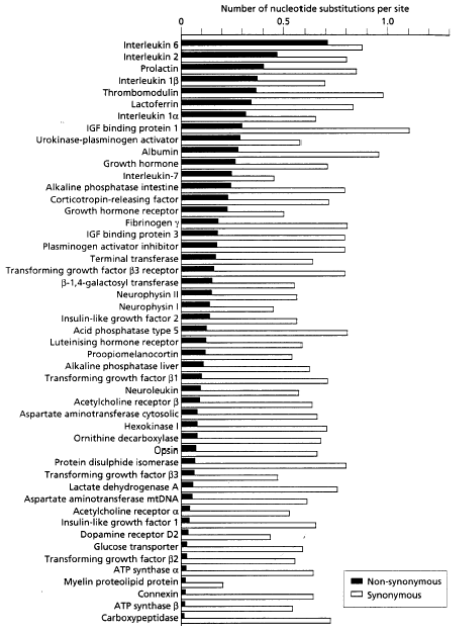


Synonymous (silent)- dashed, Non-synomous (replacement)- Solid line

Nei 1987 Text P. 81

Rates of Evolution of Nuclear Protein Genes

Artiodactyls, rodents, & primates



Nei 1987 Text P. 240

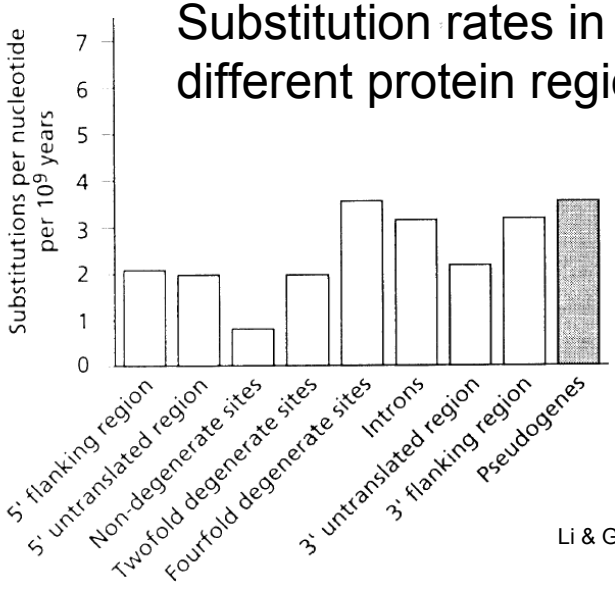
Table 1. Percentage of amino acid similarity at various divergence levels

Gene	*D. melanogaster*–*D. yakuba*[a] 3–19[b, c]	Human–cow[d] >90[b, e]	Apis–*Drosophilia*[f] >280[b, g]	*Drosophila*–Locust[h] >300[b, i]	Human–sea urchin[j] >700[b, k]	Mouse–*Drosophila*[m] >700[b, k]
COI	99	91	70	82	75	75
COII	98	73	55	66	61	57
COIII	98	87	53	73	62	64
Cytb	96	79	53	76	62	68
ND1	96	78	47	68	55	45
ND4	95	74	45	58	43	42
ND3	95	74	49	62	48	43
ND5	89	70	42	64	44	40
ND4L	99	74	37	51	34	36
A6	96	78	47	73	37	34
ND2	94	63	27	47	39	34
A8	96	52	46	41	21	26
ND6	91	63	31	45	30	17

Some genes are more conserved at the AA level; most obvious at deeper levels

Simon et al. 1994. Annals ESA

Substitution rates in different protein regions



Li & Graur 1991

Substitution rates vary within genes…

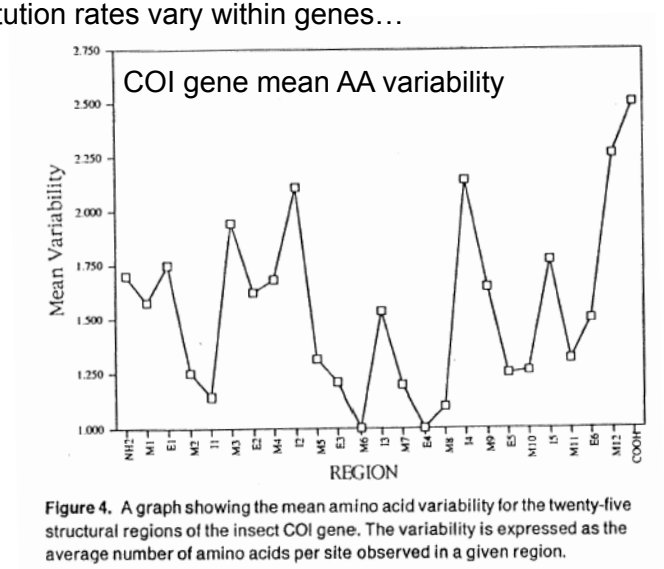
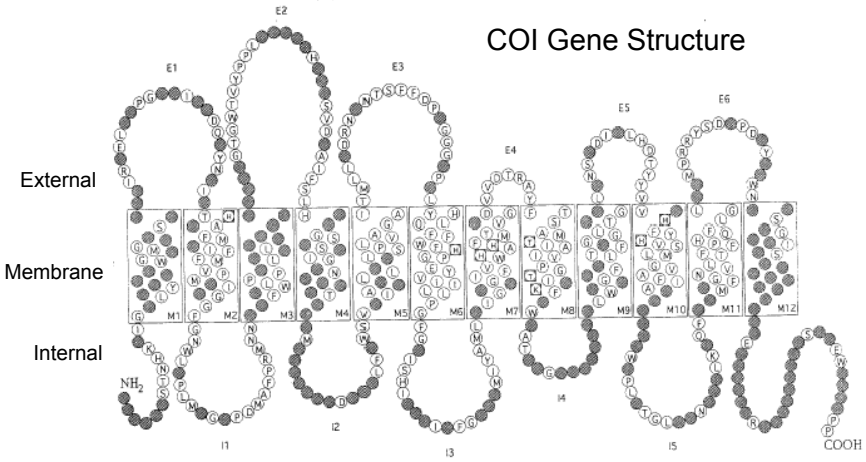


Figure 4. A graph showing the mean amino acid variability for the twenty-five structural regions of the insect COI gene. The variability is expressed as the average number of amino acids per site observed in a given region.

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



Filled circles = AA variable among insect species
Boxed AA = special functional significance

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

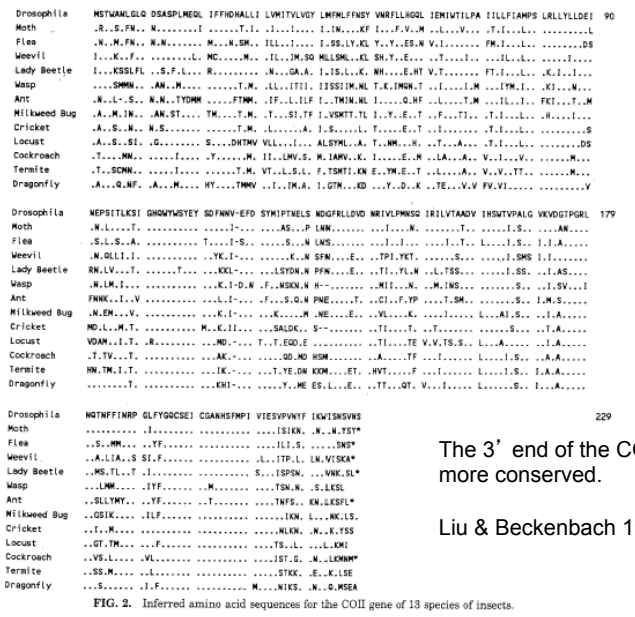


FIG. 2. Inferred amino acid sequences for the COII gene of 13 species of insects.

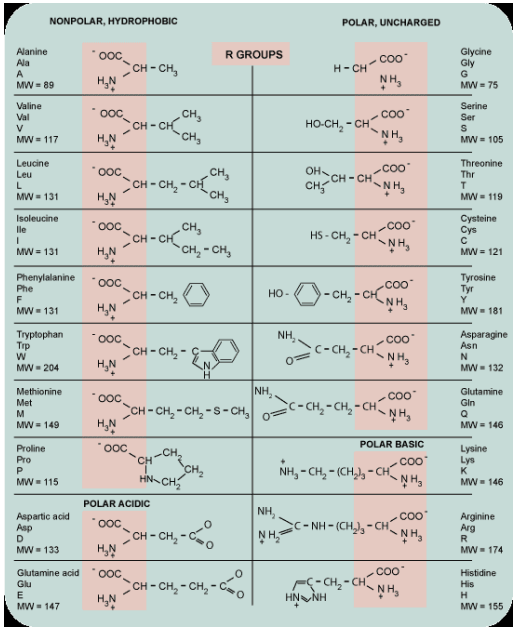
The 3’ end of the COII molecule is more conserved.

Liu & Beckenbach 1992. MPE

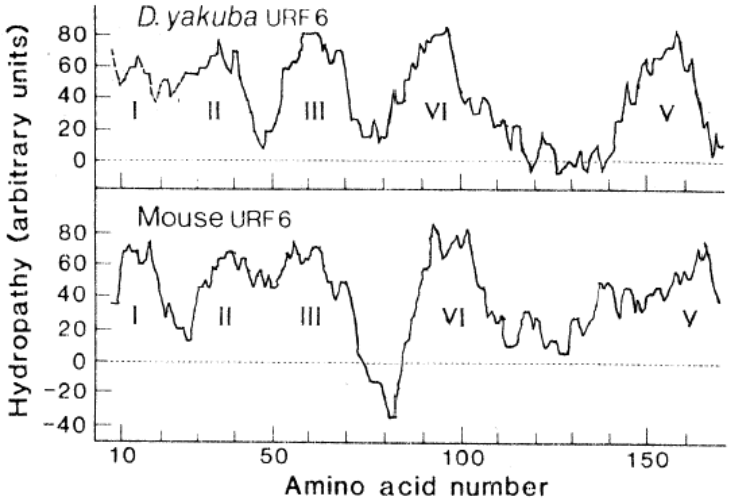
Some AA substitutions easier than others

Four Families of Amino Acids

http://matcmadison.edu/biotech/resources/proteins/labManual/images/amino_000.gif



Hydropathy profiles allow alignment of genes with low AA similarity



Clary & Wolstenholme 1985. JME 22:252-271

Among Site Rate Variation

Definition: Model parameters said to be "nonidentifiable" when data do not contain enough information to estimate the parameter with precision (too many parameters for the data).

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

- If all sites in sequence change at same rate = Poisson Distribution; otherwise, Negative Binomial Distribution
- Biggest difference in rate is between variable and "invariable" sites
- Two classes of "invariable sites"
  - Highly restricted "not free to vary"
  - Variable but not observed to vary due to chance convergence
  - % variable sites can't be calculated by simple sequence comparison.

When calculating genetic distances, must take into account ASRV…

- Walter Fitch 1970's "number of sites free to vary"

```
GATTTCACGAAATATGCATT
..CC.TT..GG.C..A..CC
```

10/20 sites observed to vary = 50% seq diverg

With five sites not free to vary…

10/15 sites vary = 75% sequence divergence

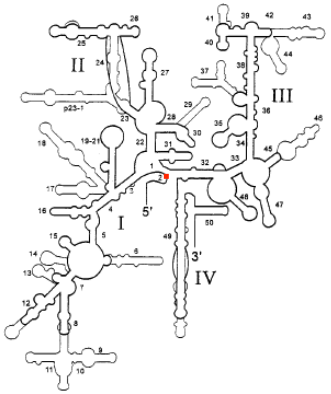
ASRV: “sites free to vary” not accurate

By definition:

- Sites with high probabilities of change will accumulate substitutions first
- Followed by sites with intermediate probabilities of substitutions,
- Followed by sites with low probabilities of substitutions

Homo^{mt}	A	C	G	G	G	C	G	G	T	G	T	G	T
Dros^{mt}	.	.	.	.	.	.	.	A	.	.	.	.	.
E.coli	.	.	.	.	.	.	.	.	.	.	.	.	.
Sacch	.	T	.	A	.	T	.	A	.	T	A	.	.
Trypan	G	T	.	.	.	.	A	A	C	A	A	-	-

Given enough time even sites with low probabilities of change will accumulate substitutions

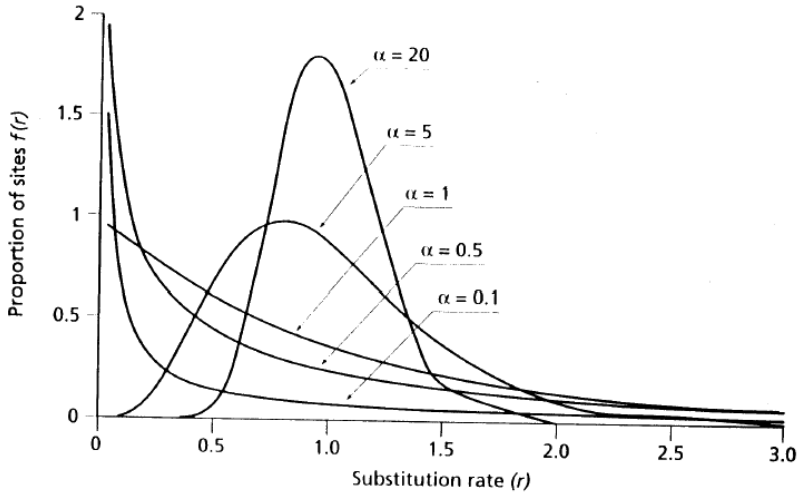


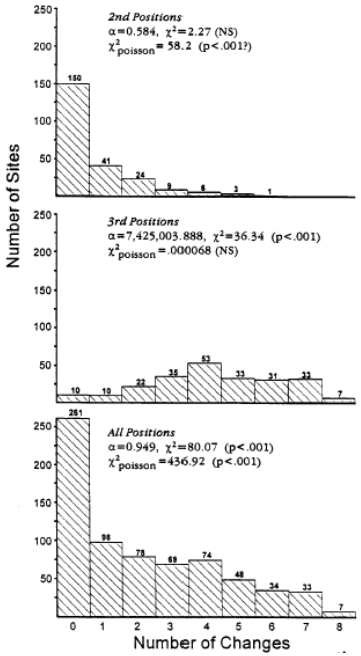
= tRNA attachment site

Better to model ASRV as continuous variation

- Negative binomial distribution with gamma distributed rates
- Described by α shape parameter
- As α increases to infinity --> equal rates
- Smaller alpha = higher ASRV

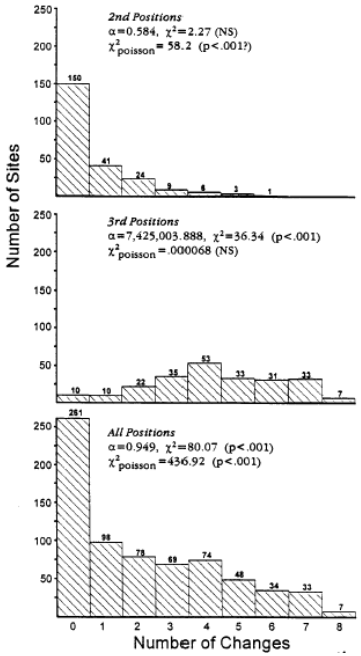
Relative substitution rates for different α values





ASRV varies among codon positions

- 2nd- fit neg binom signf diff Poisson
- 3rd- fit Poisson, not neg binom
- All- Signf dif neg binom & Poisson
- Average value gives poor picture of pattern of ASRV

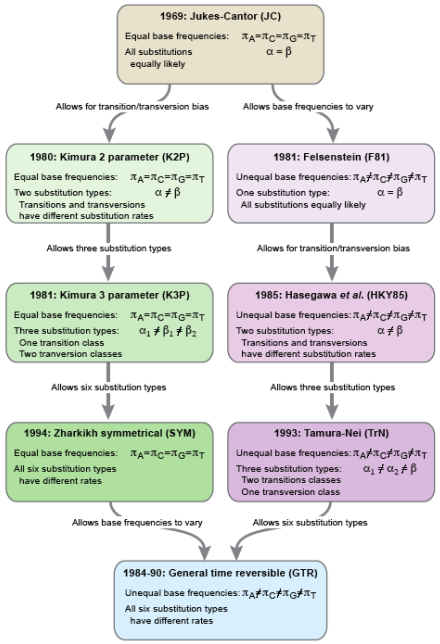


ASRV varies among codon positions

- 2nd- fit neg binom signf diff Poisson
- 3rd- fit Poisson, not neg binom
- All- Signf dif neg binom & Poisson
- Average value gives poor picture of pattern of ASRV

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



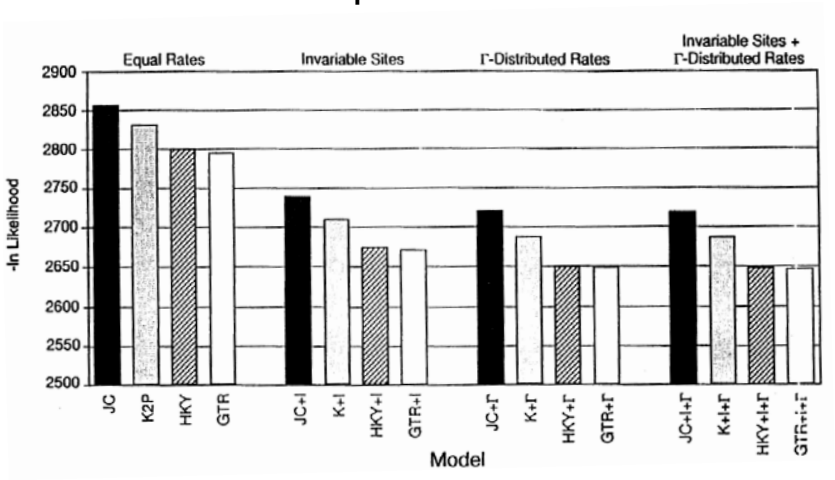
Models of Evolution

- 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

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