

Bioinformatics 1

**Distances and models.
Synonymous and non-synonymous
substitutions.
Basics of the neutral theory.**

Claudia Acquisti

Evolutionary Functional Genomics
Institute for Evolution and Biodiversity, WWU Münster
claudia.acquisti@uni-muenster.de

Contacts

Claudia Acquisti: claudia.acquisti@uni-muenster.de

Wojciech Makałowski: wojmak@uni-muenster.de

Robert Fuerst: rfuerst@uni-muenster.de (lab coordinator)

office hours - see the web site

<http://www.bioinformatics.uni-muenster.de/teaching/courses-2011/bioinf1/index.hbi>

DNA damage

Copying errors

MUTATIONS: heritable changes to the genome,
essential for evolution.

LARGE
SCALE

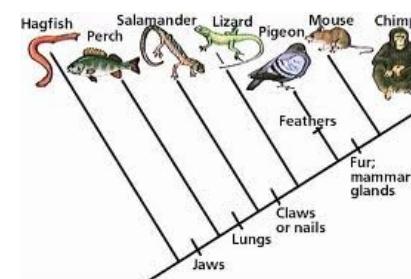
Chromosomal
rearrangements

SMALL
SCALE

Point
mutations



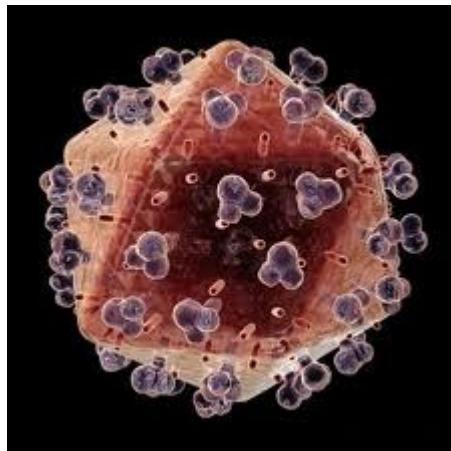
POLYMORPHISMS



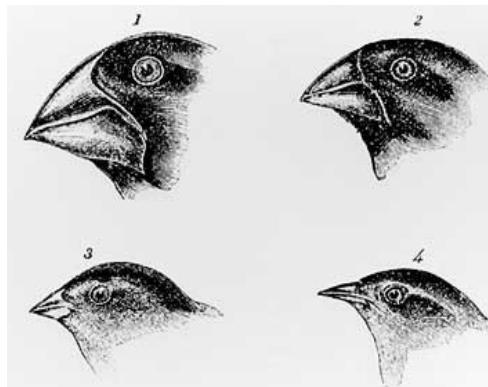
SUBSTITUTIONS

Selection on

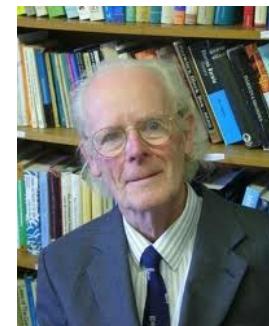
De novo mutations vs. Standard genetic variation



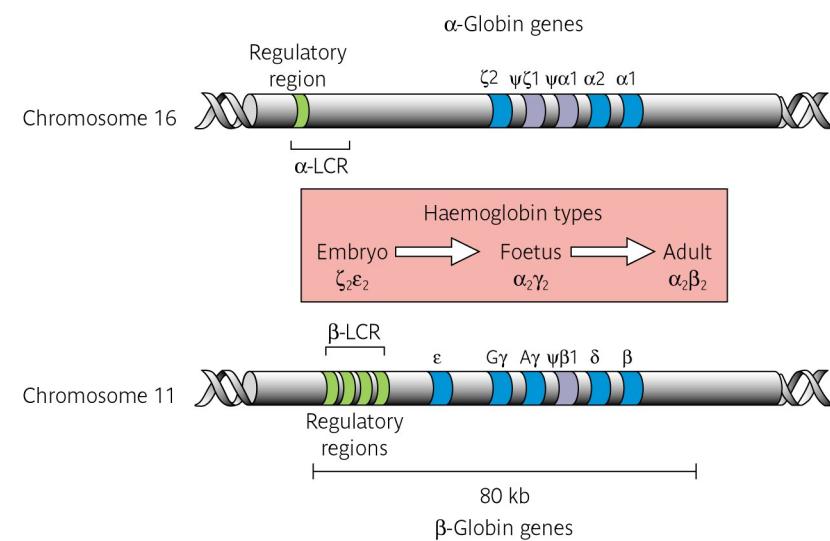
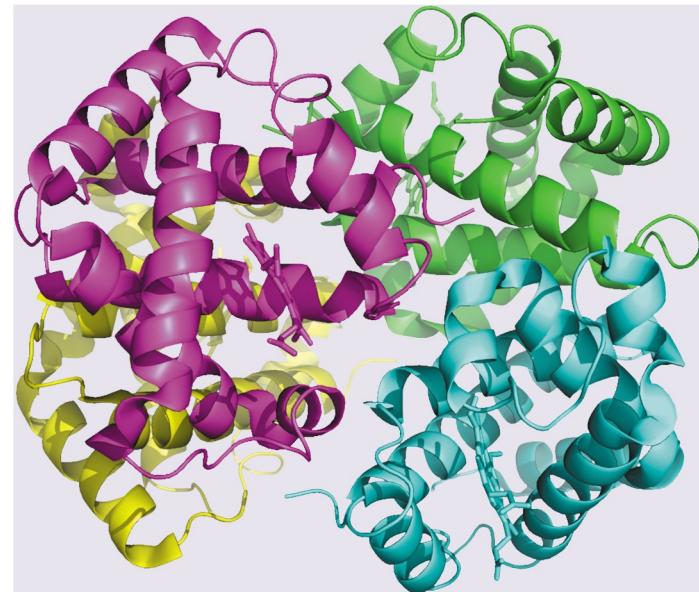
HIV virus evolution



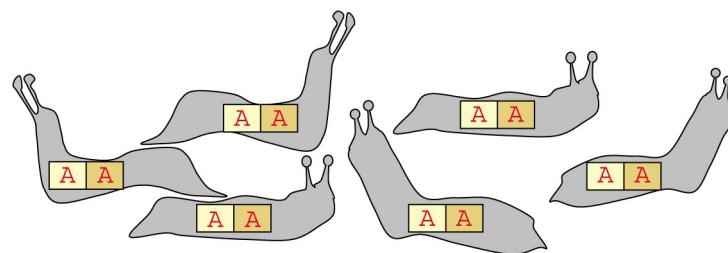
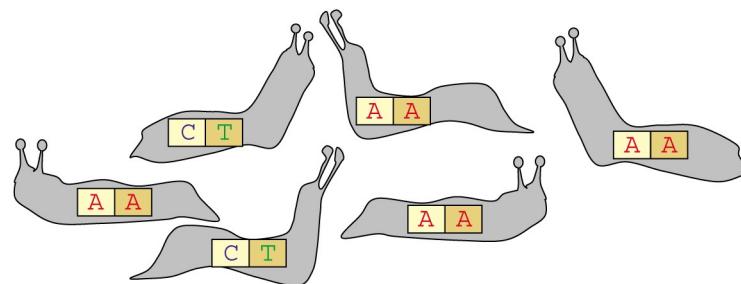
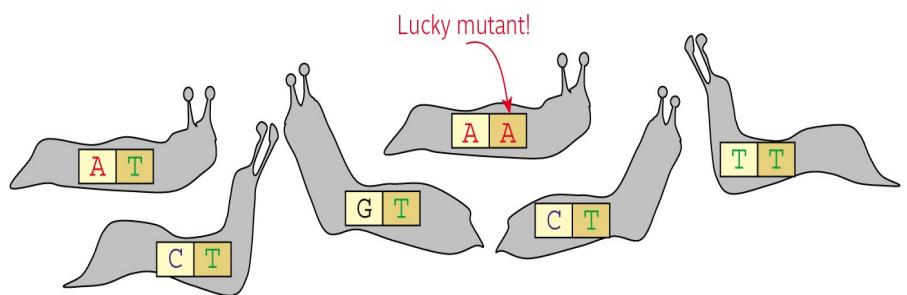
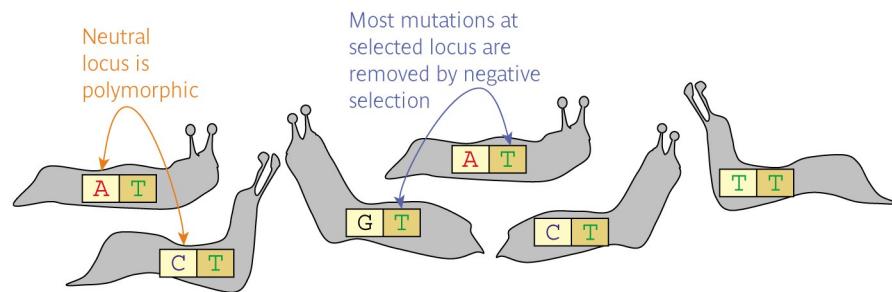
Darwin's finches
The effect of the environment



Genetic background of a mutation



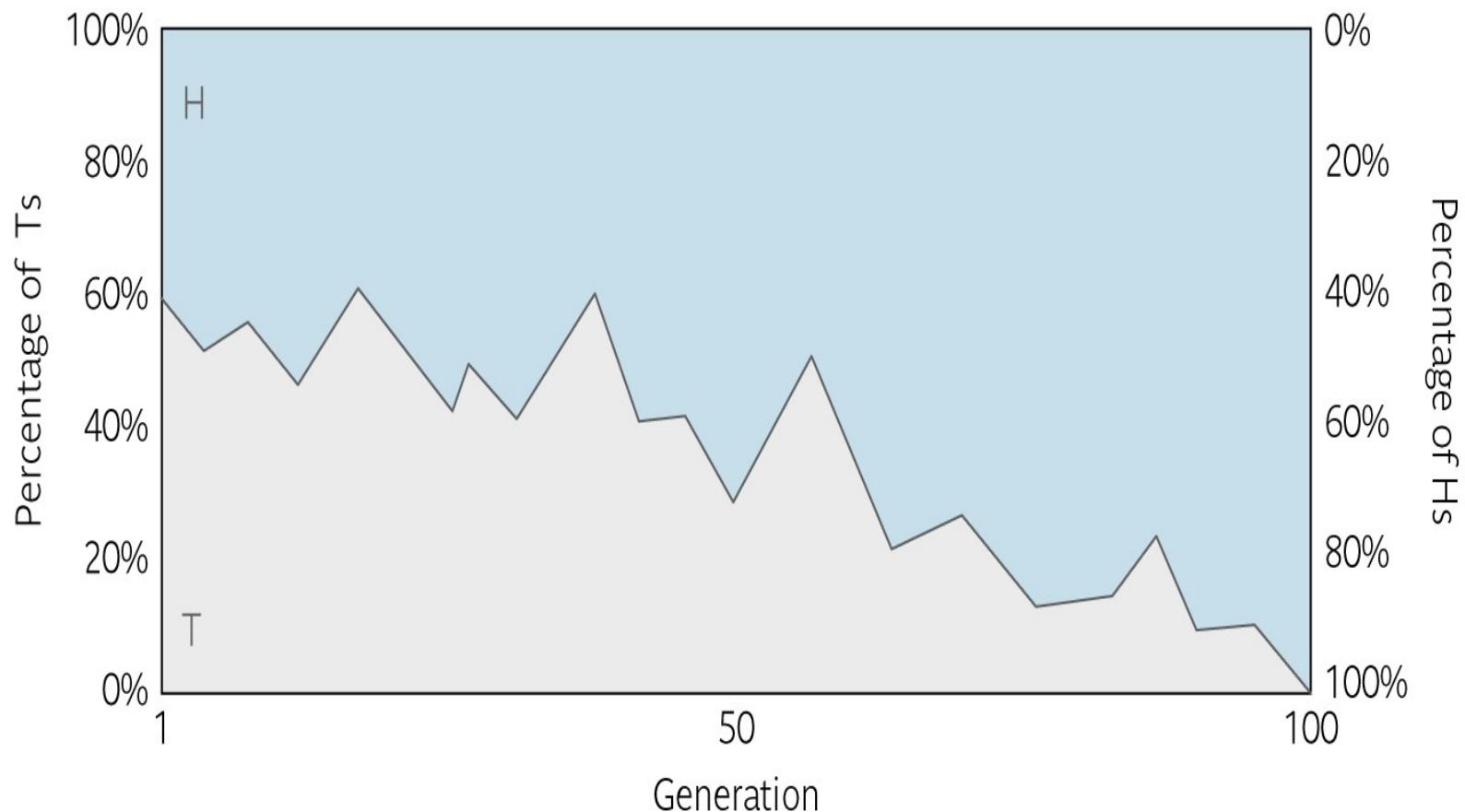
Linkage



Hitchhiking



Fixation without selection: Drift



Neutral theory

Genetic drift is the main force changing allele frequencies.



M. Kimura

Distinguishing between different types of selection

NEGATIVE SELECTION: Conservation of sequences

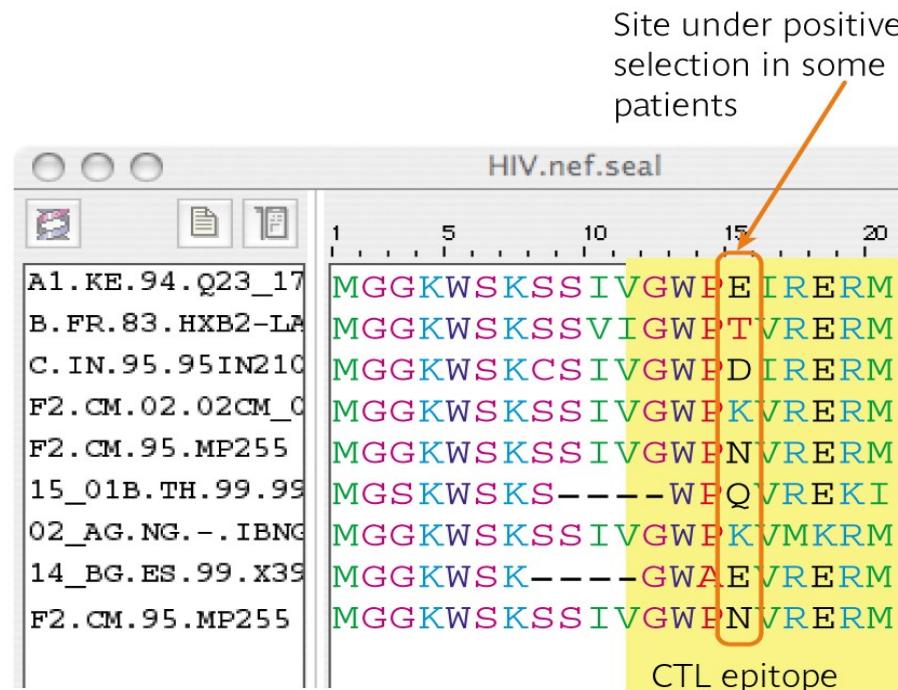
Homo sapiens	GRSRDGGLRF GEME
Rattus norvegicus	GRSRDGGLRF GEME
Drosophila melanogaster	GRARDGGLRF GEME
Neurospora crassa	GRARDGGLRF GEME
Oryza sativa	GRKYGGGIRF GEME
Escherichia coli	GKAQF GGQRF GEME

Homo sapiens	GGTAGATCTCGTGATGGTGGCCTGCGTTTGGAGAAATGGAA
Rattus norvegicus	GGCAGATCGCGTGATGGTGGCCTGCGCTTGGAGAAATGGAG
Drosophila melanogaster	GGTCGTGCTCGTGATGGTGGCTTGCCTTCGGTGAGATGGAG
Neurospora crassa	GGTCGTGCCAGAGACGGTGGCTCCGTTTCGGTGAAATGGAA
Oryza sativa	GGAAGGAAAATACGGTGGAGGGATT CGGTTCCGGTGAGATGGAG
Escherichia coli	GGTAAGGCACAGTTGGTGGTCAGCGTTTCGGGGAGATGGAA

Distinguishing between different types of selection

POSITIVE SELECTION: Rapid substitution

E.g., HIV CTL Epitopes



How can we quantify?

#Human	V-LSPADKTN	VKAAGKVGA	HAGEYGAEAL	ERMFLSFPTT	KTYFPFH-DL	SHGSAQVKGH
#HorseA.....S...GG....A.
#CowA....G.G	..A.....
#KangarooA....GH	...I.....GA..G.	..T.H.....IQA.
#Newt	MK..AE..H.	..TT.DHIKG	.EEAL.....	F...T.L.A.	R....AK...	.E..SFLHS.
#Carp	S...DK..AA	..I..A.ISP	K.DDI.....	G..LTVY.Q.A.WA..	.P..GP..-.
#Human	GKKVA-DALT	NAVAHVDDMP	NALSALSDLH	AHKLRVDPVN	FKLLSHCLLV	TLAAHLPAEF
#HorseG..	L..G.L..L.	G...D..N..S	...V...ND.
#Cow	.A....A....	K..E.L..L.	G...E.....S...	...S...SD.
#Kangaroo	...I.....G	Q..E.I..L.	GT..K.....F....GDA.
#NewtM.G..SI..ID	A..CK...K.	.QD.M...A.	.PK.A.NI..	VMGI..K.HL
#CarpIMG.VG	D..SKI..LV	GG.AS..E..	.S.....A.	..I.ANHIV.	GIMFY..GD.
#Human	TPAVHASLDK	FLASVSTVLT	SKYR			
#HorseS.....			
#CowN.....			
#Kangaroo	..E.....	...A.....			
#Newt	.YP..C.V..	..DV.GH...			
#Carp	P.E..M.V..	.FQNLALA.S	E...			

Statistical measures of evolutionary distance between amino acid sequences

```
#Human      V-LSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHF-DL SHGSAQVKGH
#Horse      ....A..... ....S...G .....G.....G....A.
#Cow       ....A...G. ....G.....A..... .
#Kangaroo   ....A...GH ...I.....G .....A..G. ..T.H..... .IQA.
#Newt        MK..AE..H. ..TT.DHIKG .EEAL..... F...T.L.A. R....AK... .E..SFLHS.
#Carp        S...DK..AA ..I..A.ISP K.DDI..... G..LTVY.Q. ....A.WA.. .P..GP...-
```

Elimination of all sites with indels from the computations

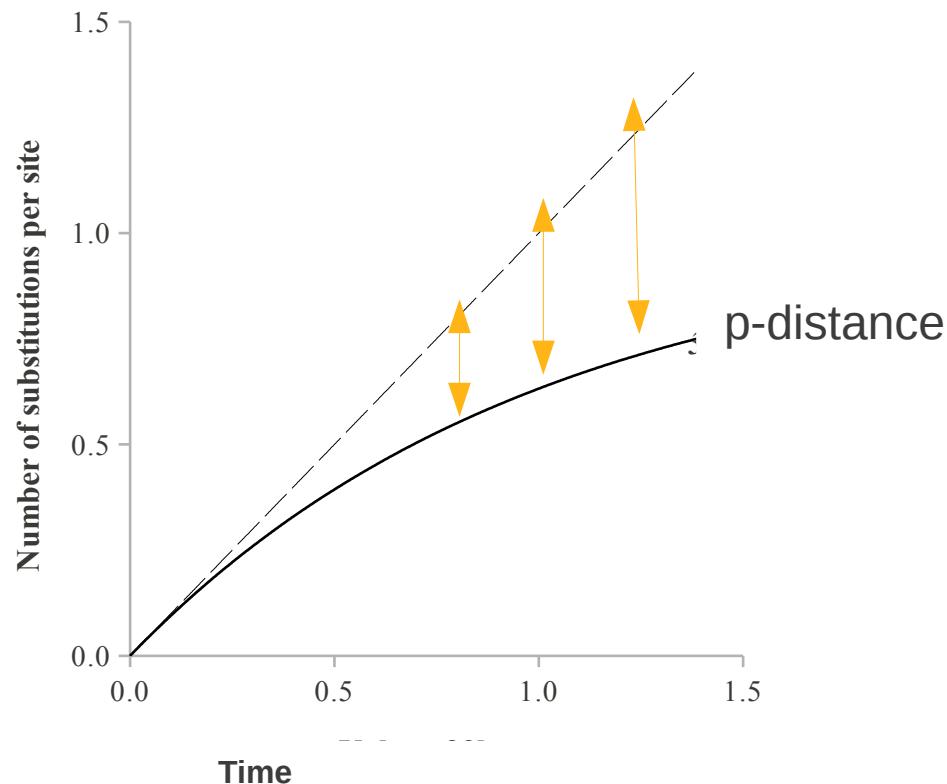
$$P = n_d / n$$

$$P = n_d / 140$$

	Human	Horse	Cow	Kangaro o	Newt	Carp
Human		17	17	26	61	68
Horse	0.121		17	29	66	67
Cow	0.121	0.121		25	63	65
Kangaroo	0.186	0.207	0.179		66	71
Newt	0.436	0.471	0.450	0.471		74
Carp	0.486	0.479	0.464	0.507	0.529	

Easy, and clean!

Is the p-distance a well defined distance?

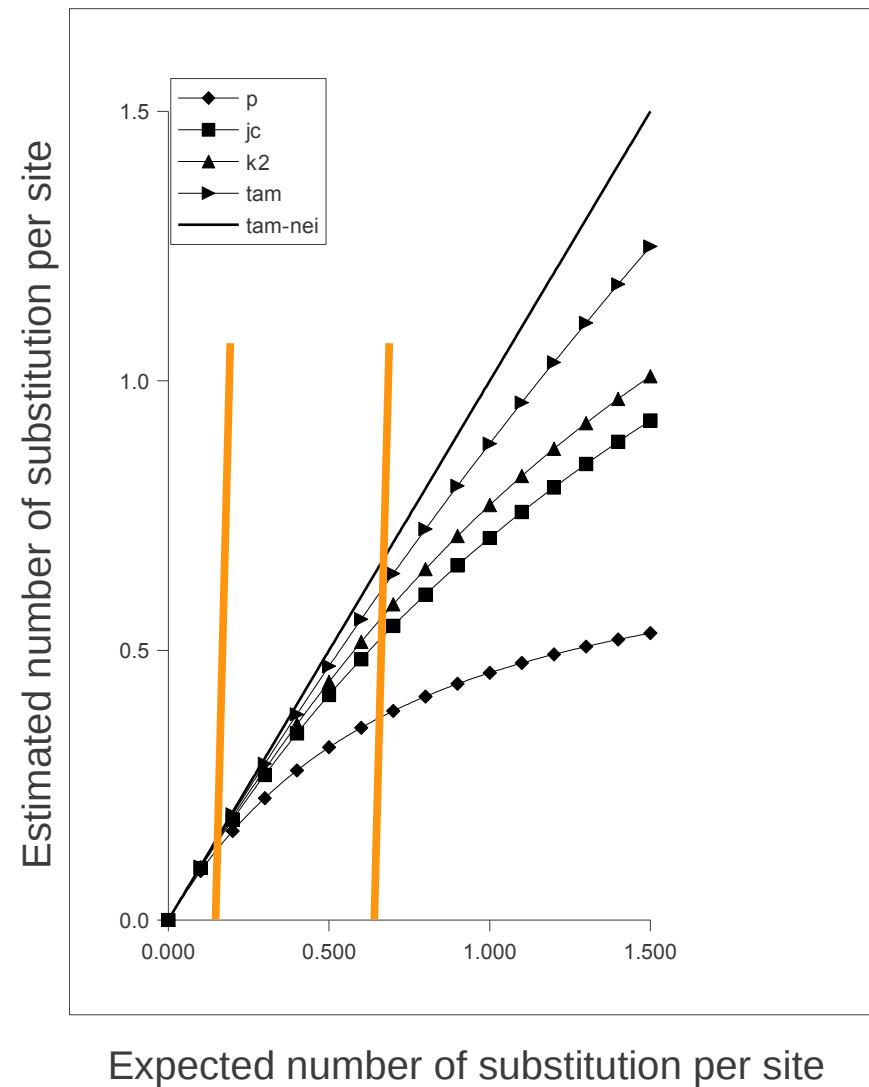


Gradual
underestimation
of the real
distance



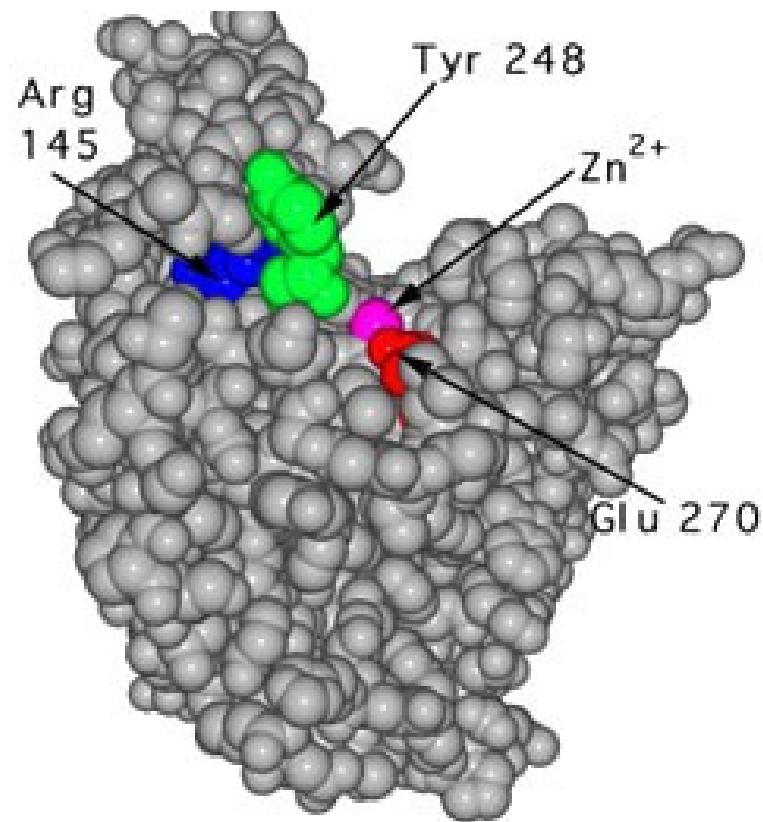
WHY?

Models



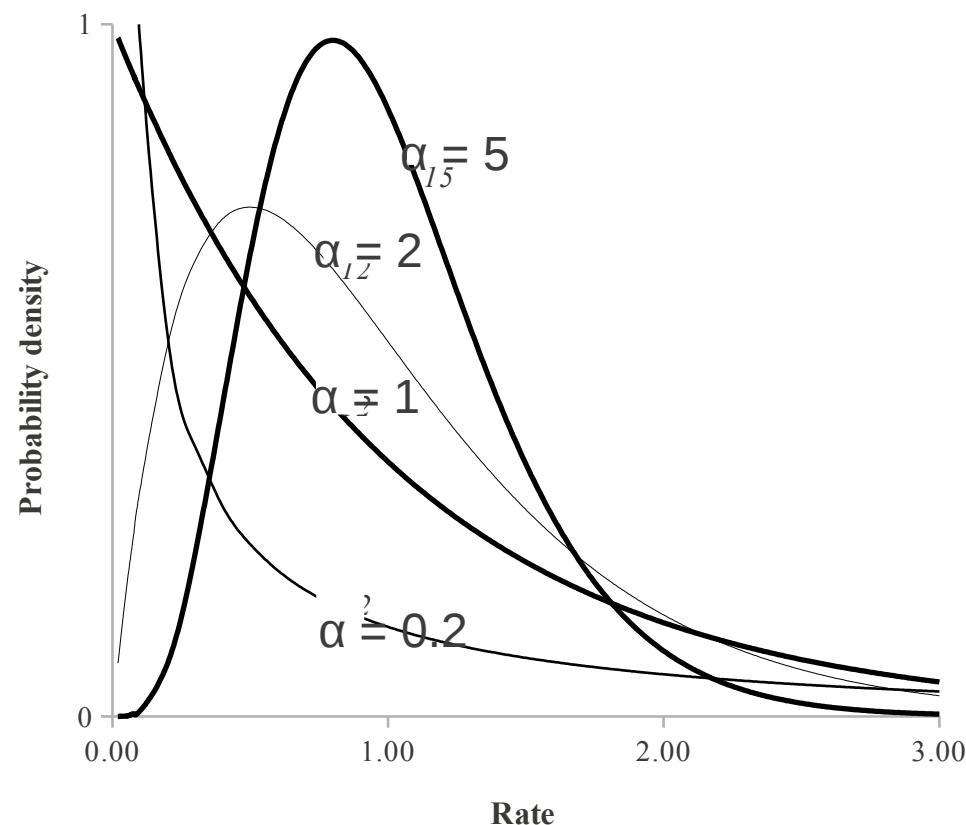
How do we deal with functional heterogeneity between sites?

Functional-structural constraints



**What is the rate of
amino acid substitution
in different regions?**

Gamma distribution (Uzzell and Corbin 1971)



Heterogeneity between sites: a well understood case

Synonymous and non-synonymous sites

		2 nd					
		U	C	A	G		
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	

Neutral substitution rate is determined by mutation rate

Statistical measures of rate of Syn NonSyn substitutions

Approximation:

1st + 2nd vs. 3rd codon positions

Count number of Syn and NonSyn sites!

Exact calculation based on the genetic code

$$D_s = \# \text{ Syn Subst} / \# \text{ Syn sites}$$

$$D_n = \# \text{ NonSyn Subst} / \# \text{ NonSyn sites}$$

What do we expect?

Synonymous substitution to be more frequent than Nonsynonymous substitutions

Rate of Synonymous substitutions to be more similar between genes than the rate of Nonsynonymous substitutions

Dn/Ds measure of positive selection

Sliding window approach to detect local signal of positive selection