Estimation of DNA Sequence Context-dependent Mutation Rates Using Primate Genomic Sequences: Application to Estimation of Selection Bias in Protein (Human TP53) Evolution

Wei Zhang

Dept. of Microbiology & Molecular Genetics

The University of Vermont, Burlington, VT 05405

Sequence context-dependency of DNA Mutations

- *Biochemical Evidence*

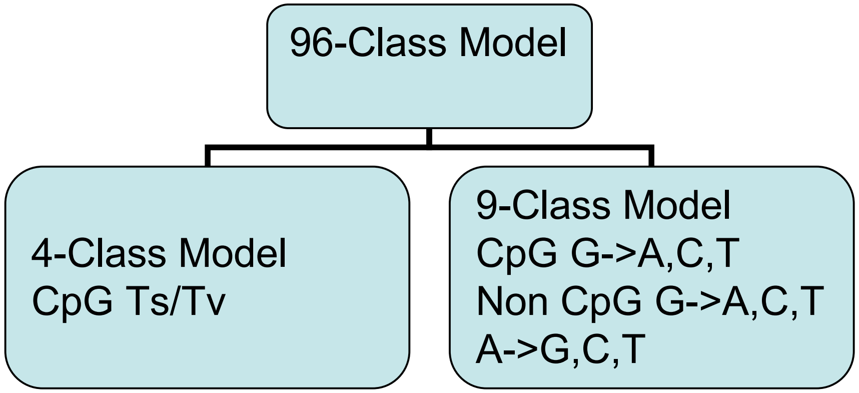
- Formation of pyrimidine dimers
- Misincorporation of nucleotides during translesion synthesis
- DNA polymerase-lesion interactions
- Methylation C and deamination of methylcytosine in vertebrates

- *Sequence Analysis Evidence*

- Plant chloroplast DNA
- Mammalian gene-pseudogene pairs

Purpose of the Study

- Quantify a DNA mutation model that accounts for sequence context-dependency of mutations, i.e effects of immediate neighbors on a mutation
- Can we find a less complicated model (with less parameters) for the 96 classes of mutations in the form of abc->adc, where a,b,c,d are nucleotides and b≠d?
- Can we separate selection and mutation biases in a protein eg. human p53?



96-Class Model

4-Class Model
CpG Ts/Tv

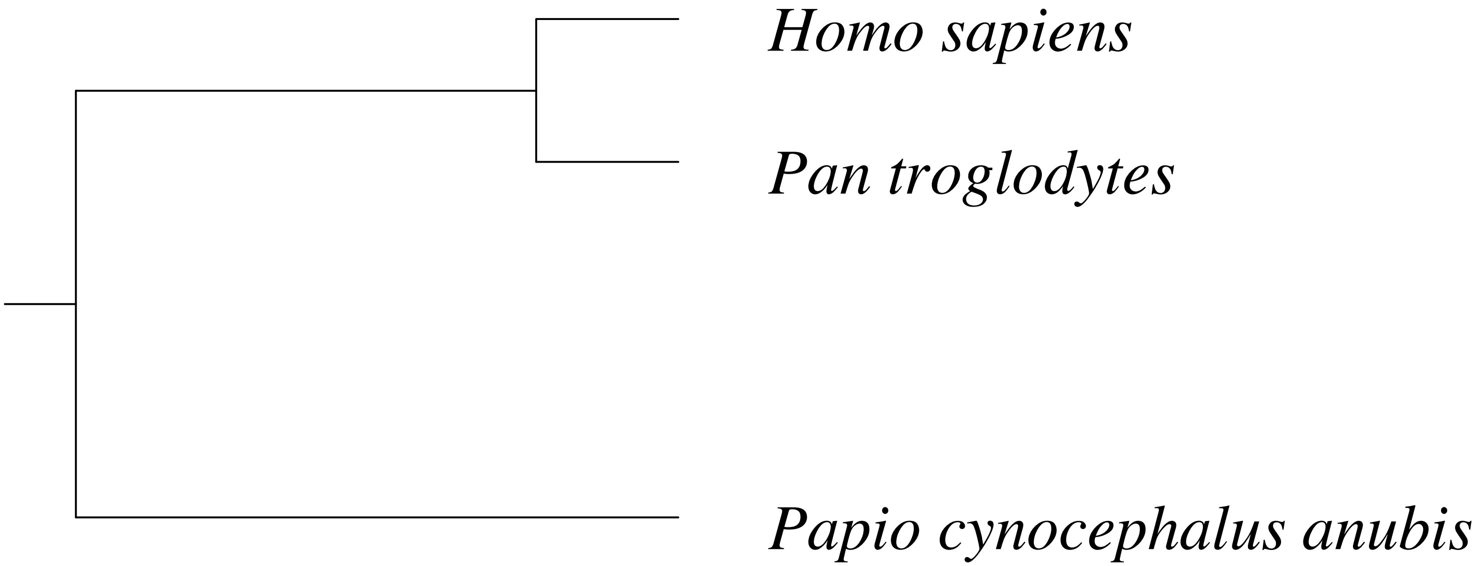
9-Class Model
CpG G->A,C,T
Non CpG G->A,C,T
A->G,C,T

Primate Genomic Sequences
NISC Taget-1 & Target-2

Target (Version)	Location (Human)	Size (Mb)	Number of Known Genes	No. of Additional Predicted Genes	Selected Genes	Gap-free Alignment (Nt)
1 (Jan 2003)	7q31	~1.5	10	24	CFTR, WNT2, MET, ST7	1162626
2 (May 2003)	7q11, 7q22	~3	19	N/A*	ELN, LIMK, p47-PHOX, ZPA3	197213

*: Target-2 is not annotated yet

The Primates



Maximum Likelihood Estimation

Under the mutation model the probability of observing homologous nucleotides *c*, *d*, and *e* in the phylogeny

$$s = \left(C,\left(D,E\right):t_2\right):t_1$$

conditional upon invariance of *u* and *v* in the extant sequences, is

(1)

$$P^{uv}_{cde}\left(Q,s\right) = \sum_a \rho^{uv}_a \, p^{uv}_{ca}\left(t_1\right)\left(\sum_b p^{uv}_{ab}\left(t_1 - t_2\right) p^{uv}_{db}\left(t_2\right) p^{uv}_{eb}\left(t_2\right)\right) + \sum_{\omega \in \Omega} P_\omega$$

where *Q* is a substitution rate matrix with elements defined by

(2)

$$q_{yx} \equiv \left.\frac{dP\left(y \mid x,t\right)}{dt}\right|_{t=0}$$

Our approximation, then is to neglect the second term

(3)

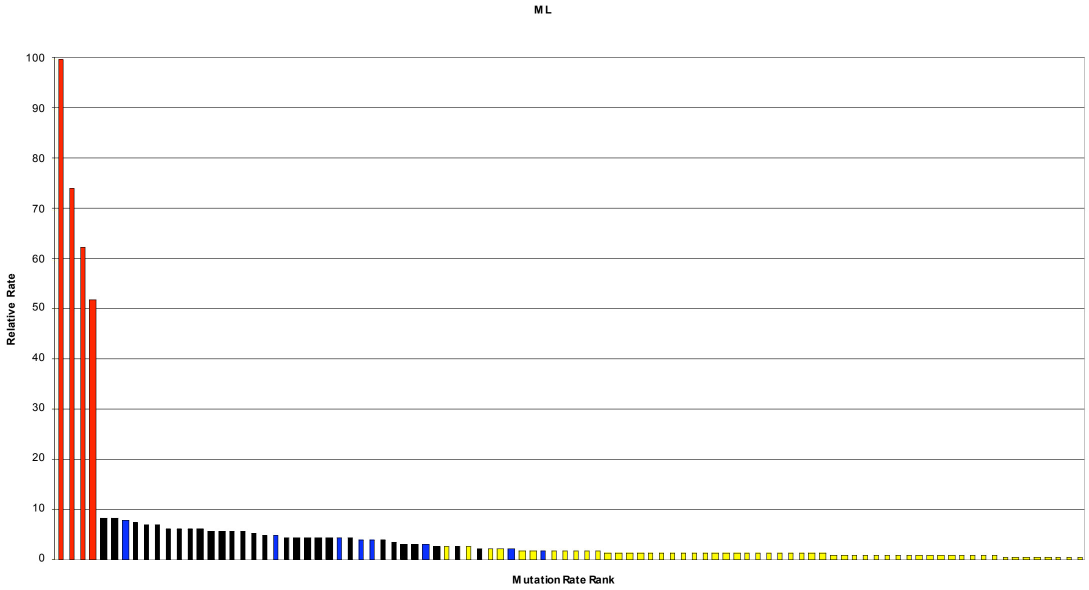
$$P^{uv}_{cde}\left(Q,s\right) \approx \sum_a \rho^{uv}_a \, p^{uv}_{ca}\left(t_1\right)\left(\sum_b p^{uv}_{ab}\left(t_1 - t_2\right) p^{uv}_{db}\left(t_2\right) p^{uv}_{eb}\left(t_2\right)\right)$$

With this approximation the log probability of the data given the model is multinomial,

(4)

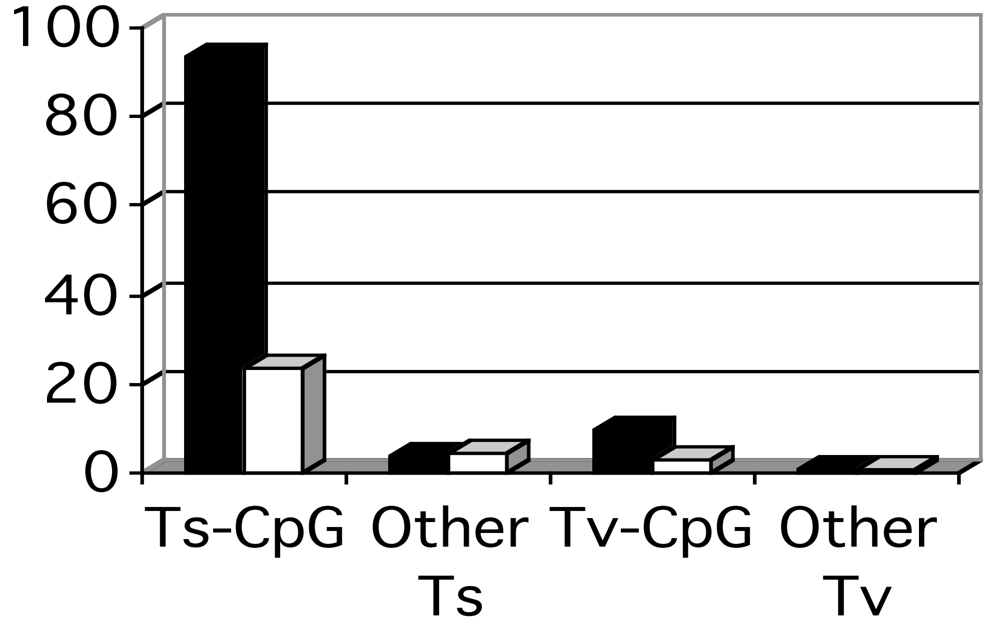
$$\log P\left(\left\{n_{cdeuv}\right\} \mid u,v,s\right) = \sum_{c,d,e} n_{cdeuv} \log P^{uv}_{cde}\left(Q,s\right) + C\left(\tilde{n}\right)$$

96-Class Mutation Spectrum
NISC Target-1



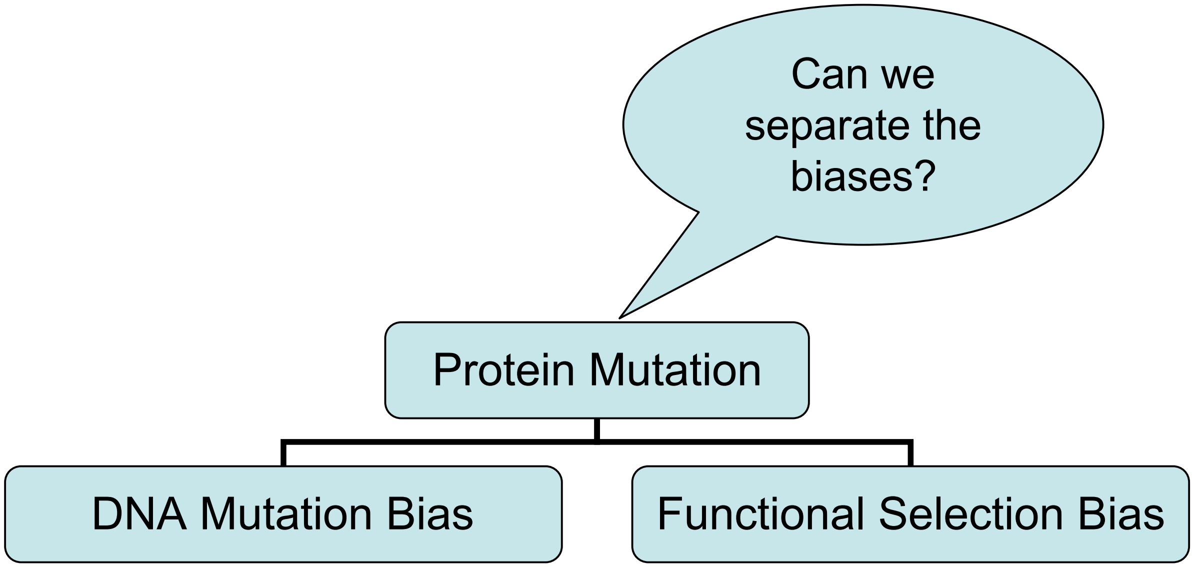
Mutation classes: 1) Ts-CpG; 2) Other Ts; 3) Tv-CpG; 4) Other Tv

4-Class Mutation Model
NISC Target-1& Target-2



Ts-CpG: Transitions at CpG sites
Other Ts: Transitions at non-CpG sites
Tv-CpG: Transversions at CpG sites
Other Tv: Transversions at non-CpG sites
Black columns: Target-1
White columns: Target-2
Other Tv = 1.0

Protein Mutation Problem



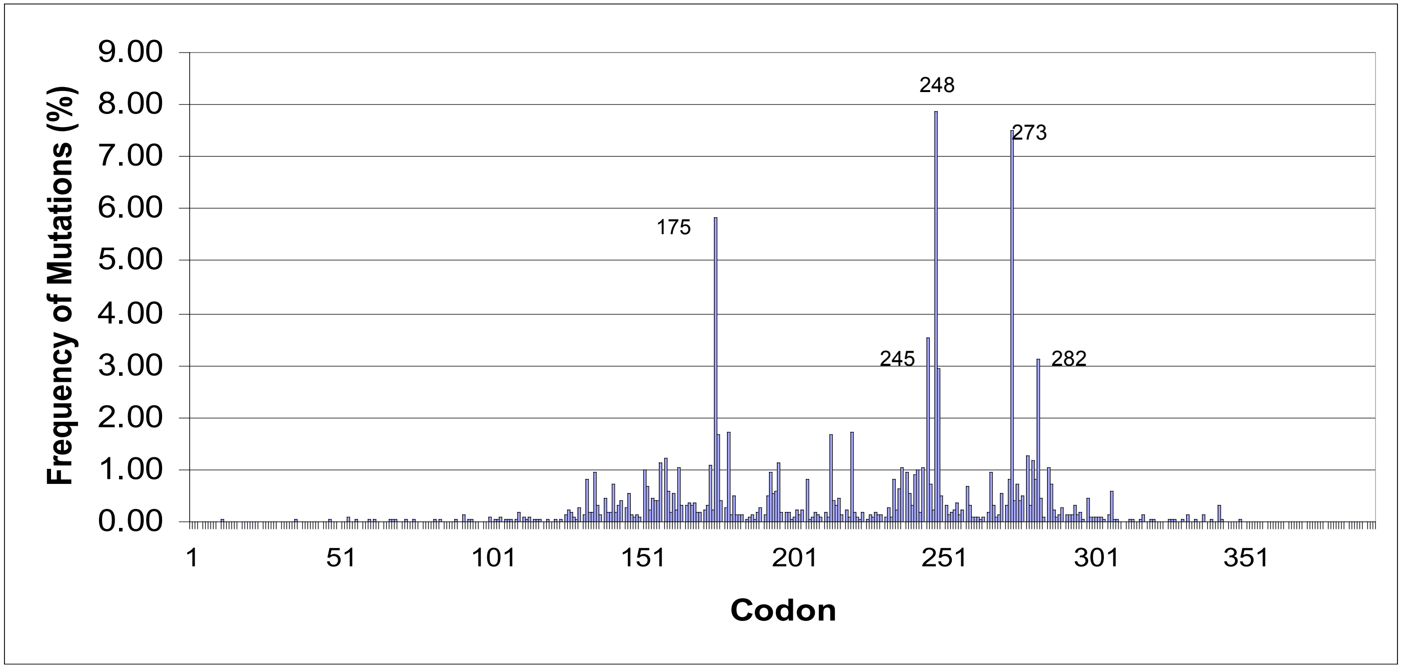
Can we separate the biases?

Protein Mutation

DNA Mutation Bias

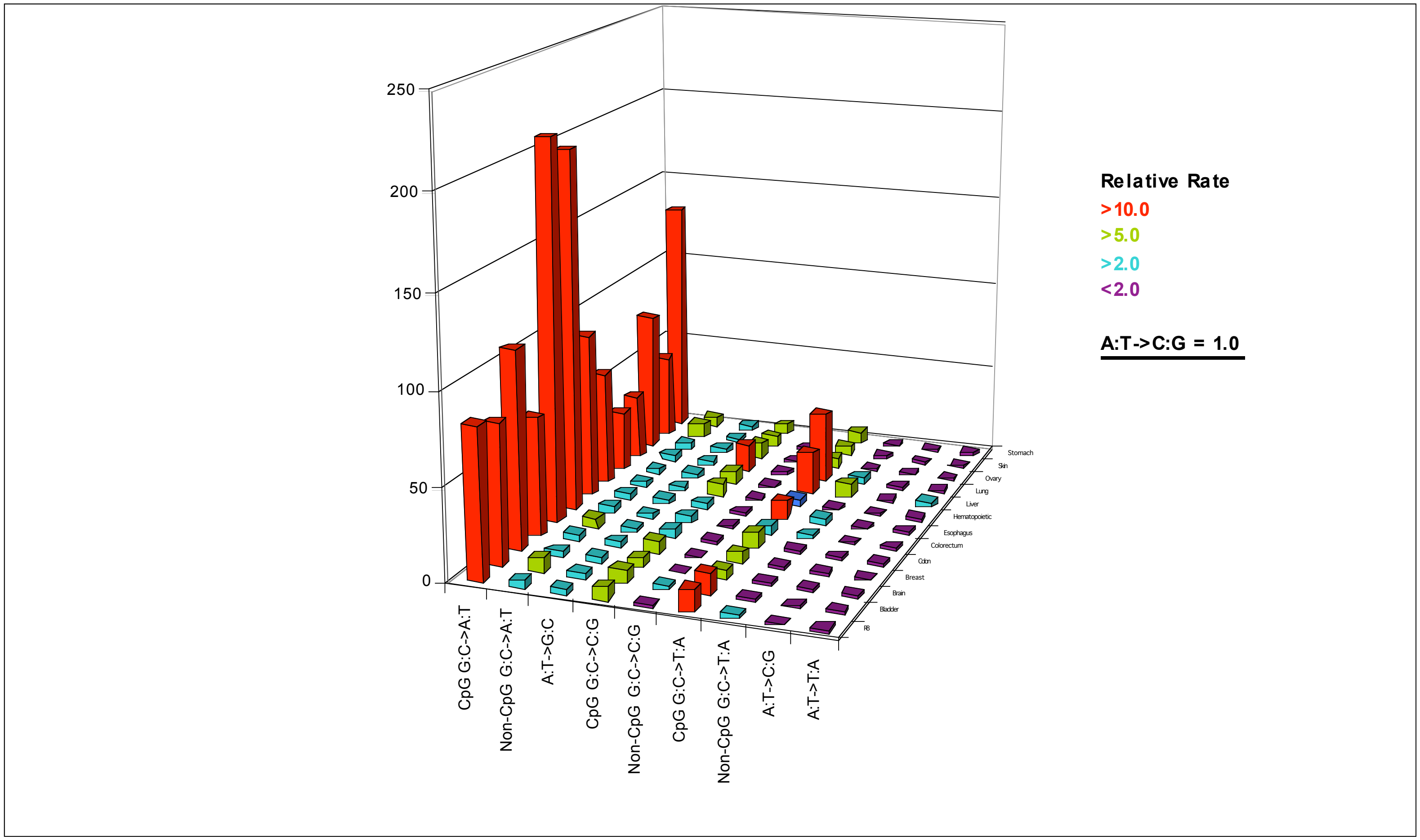
Functional Selection Bias

IARC/WHO Human TP53 Somatic Mutation Database



Release 8, June 2003

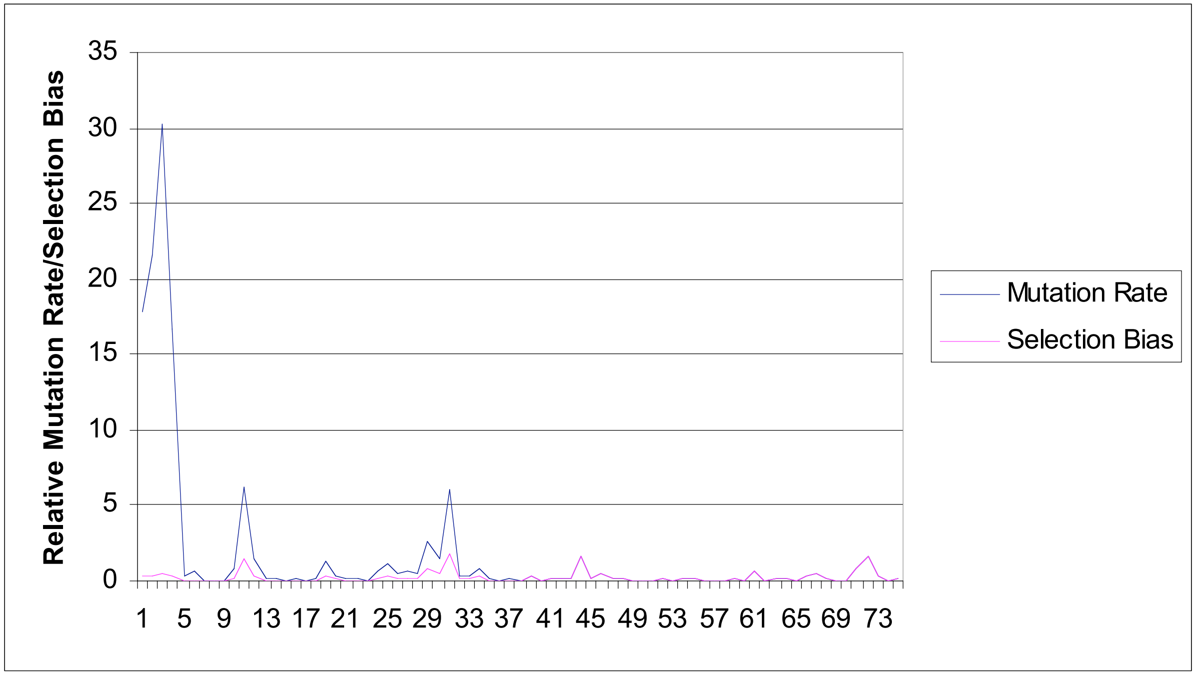
Mutation Rates of Single Nucleotide Point Mutation – Partitioned by Primary Tumor Type



Removal of DNA mutation bias in different groups of residues that are subject to strong selection

Residues	No. of Codons	No. of Distinct Mutations	Before removal of mutation bias		After removal of mutation bias		Ratio of Standard Dev.
			Mean	Standard Deviation	Mean	Standard Deviation	
Zinc Binding	4	29	0.62	0.80	0.41	0.64	1.21
		25 (excluding G:C->T:A)	0.49	0.62	0.25	0.17	0.54
DNA Binding	14	75	1.65	5.03	0.25	0.38	0.50
		62 (excluding G:C->T:A)	1.80	5.48	0.21	0.34	0.53
Glycine	6	29	0.72	1.98	0.18	0.26	0.53
		19 (excluding G:C->TA)	0.82	2.40	0.09	0.12	0.46
Conserved	63	347	0.58	2.24	0.21	0.46	0.57
		280 (excluding G:C->T:A)	0.63	2.47	0.18	0.39	0.55
R8	393	1196	0.37	1.81	0.125	0.31	0.51

An example: 14 DNA Binding Codons Subject to Strong Selection



Acknowledgements

Jeffrey P. Bond, Ph.D., University of Vermont

Gerard Bouffard, Ph.D. , NISC/NIH

Eric Green, Ph.D., NISC/NIH

Susan S. Wallace, Ph.D., University of Vermont

US Dept of Energy EPSCoR Vermont Graduate Research Fellowship