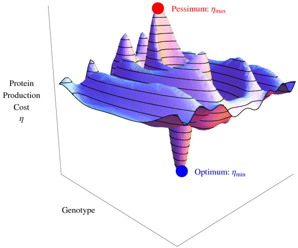
Evolutionary Bioinformatics
Fitting Biological Models to Genomes
The Need for HPC

Michael A. Gilchrist

Dept. of Ecology & Evolutionary Biology
University of Tennessee
Knoxville, TN 37996



Introduction

Evolution & Biology

“Nothing in Biology Makes Sense
Except in the Light of Evolution”

Dobzhansky (1973)

“Nothing in Evolution Makes Sense
Except in the Light of Population Genetics”

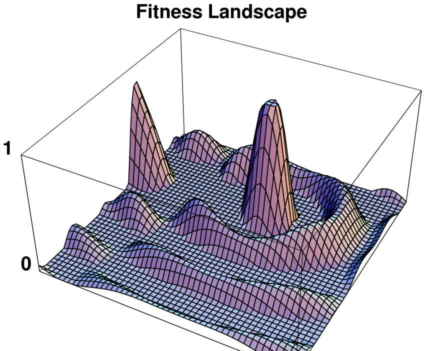
Lynch (2007)

Introduction

Fitness Landscape

A metaphor for describing evolution

- Selection = Directed force (hill climbing)
- Drift = Undirected, diffusive effect
- Mutation = Introduces new genotypes to population on landscape

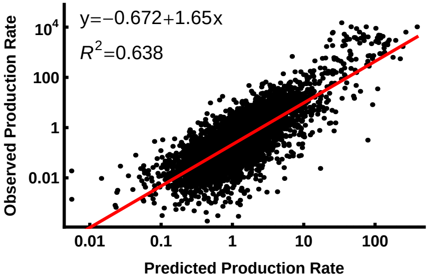


General Goal

Use fundamental concepts in ecology, evolution, and cellular biology to improve interpretation of biological datasets

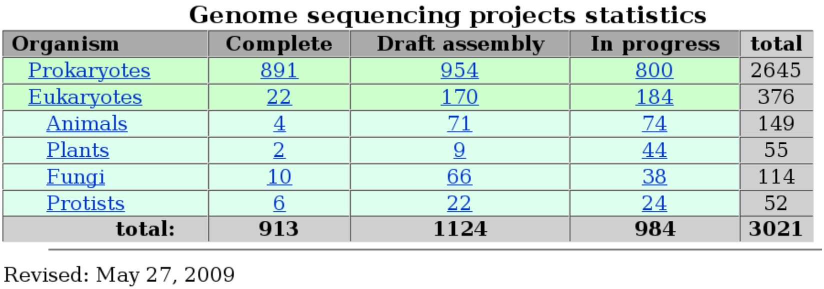
Specific Goal

Use mechanistic models of protein translation & evolution to extract biologically important information from genomic datasets



Current Situation

Explosion of sequenced genomes = lots of data



Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	891	954	800	2645
Eukaryotes	22	170	184	376
Animals	4	71	74	149
Plants	2	9	44	55
Fungi	10	66	38	114
Protists	6	22	24	52
total:	**913**	**1124**	**984**	**3021**

Revised: May 27, 2009

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

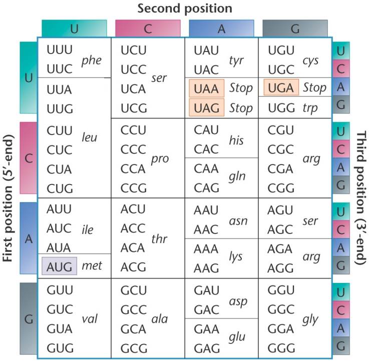
Background

The Origin of Codon Bias

Codon Redundancy & the Genetic Code

- DNA uses 4 types of nucleotides (A, T, G, & C)
- Proteins use 20 types of amino acids: (Phe, Leu, Ile, . . . Gly)

Because $4^2 = 16 < 20 < 64 = 4^3$ the genetic code is redundant.



Second position

First position (5'-end)	U	C	A	G	Third position (3'-end)
U	UUU *phe*	UCU *ser*	UAU *tyr*	UGU *cys*	U
	UUC	UCC	UAC	UGC	C
	UUA *leu*	UCA	UAA *Stop*	UGA *Stop*	A
	UUG	UCG	UAG *Stop*	UGG *trp*	G
C	CUU	CCU *pro*	CAU *his*	CGU *arg*	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA *gln*	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU *ile*	ACU *thr*	AAU *asn*	AGU *ser*	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA *lys*	AGA *arg*	A
	AUG *met*	ACG	AAG	AGG	G
G	GUU *val*	GCU *ala*	GAU *asp*	GGU *gly*	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA *glu*	GGA	A
	GUG	GCG	GAG	GGG	G

Distribution of Codon Redundancy

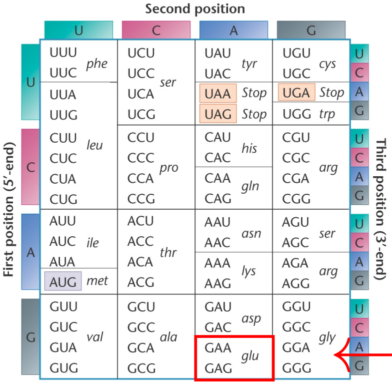
# Codons	# of AA
1	2
2	9
3	1
4	5
5	0
6	3*

Background

Codon Bias & Information Theory

Redundancy Means Information

- Codon redundancy → codon usage can encode information
- Non-uniform codon usage implies meaningful information
- Goal: Extract this information



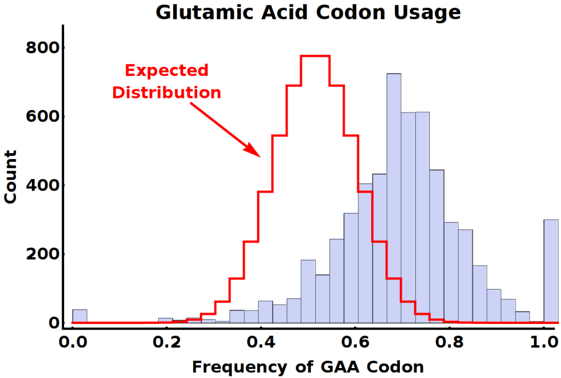
First position (5′-end)	Second position								Third position (3′-end)
	U		C		A		G		
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	trp	G
C	CUU		CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Glu: GAA or GAG

General Definition of Codon Bias

Non-uniform synonymous codon usage within a gene

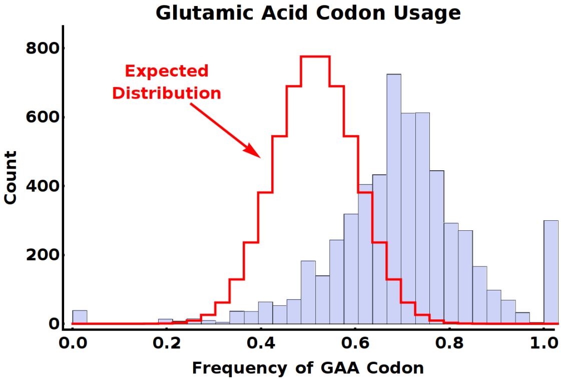
Example: Preference for GAA over GAG in *S. cerevisiae*



Information Encoded in Codon Bias

If CUB is caused by some systematic processes,

then the CUB of a gene will contain information on this process.



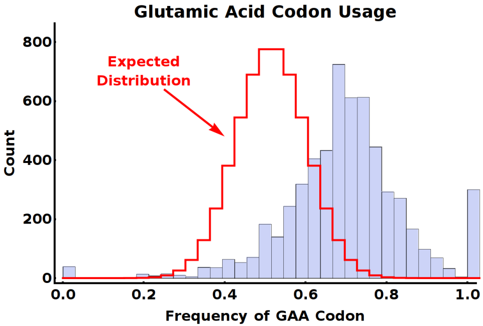
Explanations of Codon Usage Bias (CUB)

Non-Adaptive

- Biased mutation
- Genetic drift

Adaptive

- mRNA stability
- DNA packaging
- Translational efficiency



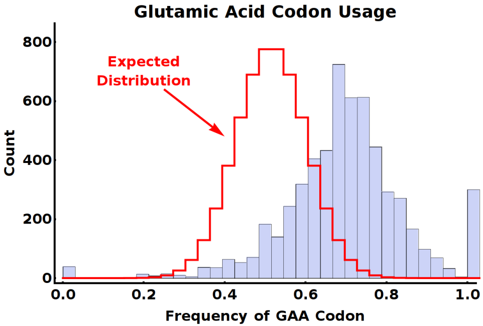
Explanations of Codon Usage Bias (CUB)

Non-Adaptive

- Biased mutation
- Genetic drift

Adaptive

- mRNA stability
- DNA packaging
- Translational efficiency

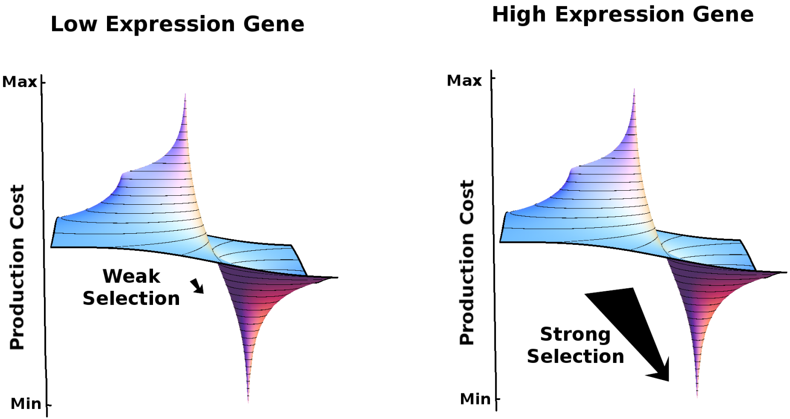


Background

Explaining Codon Bias: Translational Efficiency

Selection for Translational Efficiency

- Natural selection favors genes with CUB that reduce protein production costs.
- Strength of selection related to gene expression level.



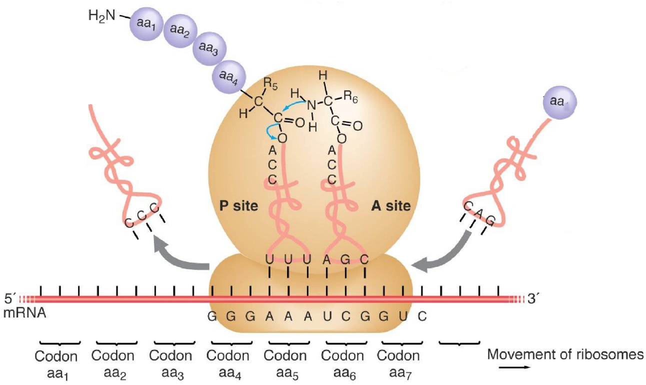
Low Expression Gene

High Expression Gene

Calculating the Cost of Protein Production

Direct costs

- Ribosome assembly on mRNA = 4 ATPs
- Elongation step = 4 ATPs



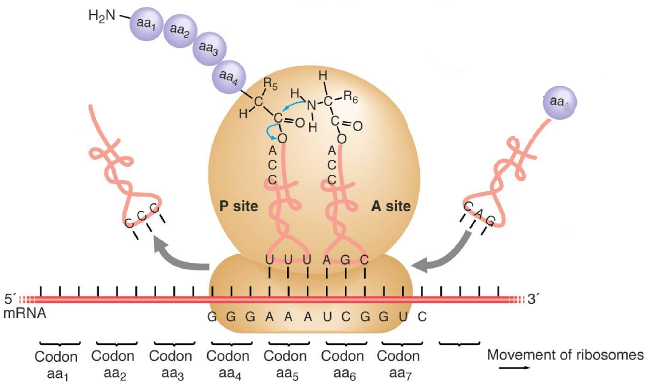
Background

Translational Efficiency

Calculating the Cost of Protein Production

Indirect costs

- Ribosomes are expensive to create
- Have limited lifespans

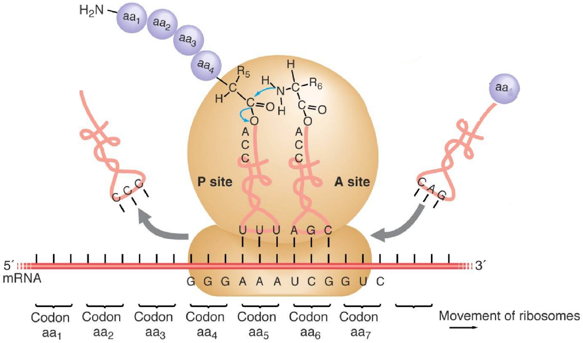


Background

Explaining Codon Bias: Translational Efficiency

Ways of Increasing Translational Efficiency

- Minimize ribosome wait time

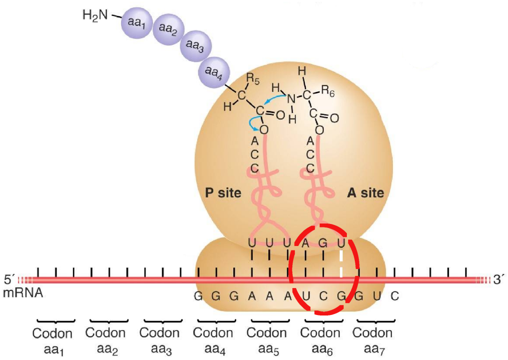


Background

Explaining Codon Bias: Translational Efficiency

Ways of Increasing Translational Efficiency

- Minimize ribosome wait time
- Minimize Pr. wrong amino acid inserted: “Missense Errors”

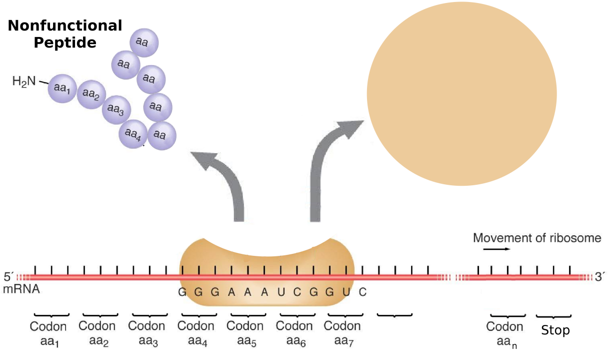


Background

Explaining Codon Bias: Translational Efficiency

Ways of Increasing Translational Efficiency

- Minimize ribosome wait time
- Minimize Pr. wrong amino acid inserted: “Missense Errors”
- Minimize Pr. premature termination: “Nonsense Errors”

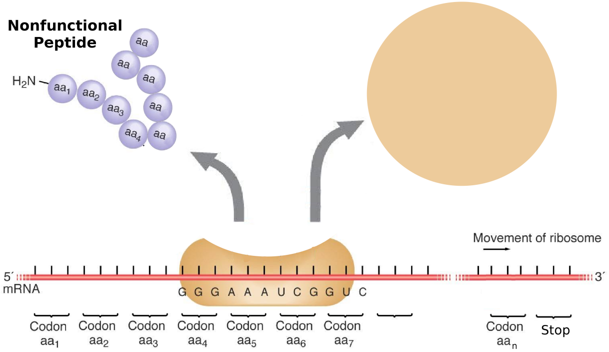


Background

Explaining Codon Bias: Translational Efficiency

Ways of Increasing Translational Efficiency

- Minimize ribosome wait time
- Minimize Pr. wrong amino acid inserted: “Missense Errors”
- Minimize Pr. premature termination: “Nonsense Errors”



Linking Genotype to Phenotype

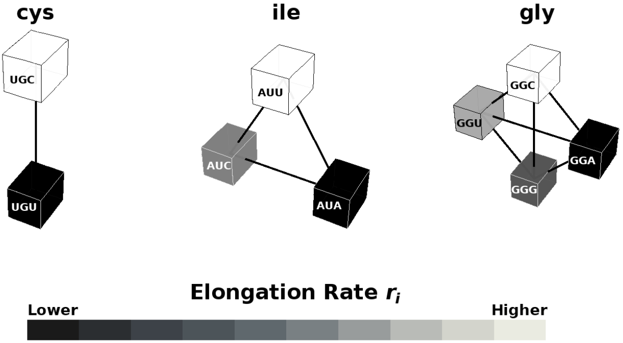
Model Assumptions

Variation in Elongation Rates: r

Let,

$r_i =$ Elongation rate of codon i

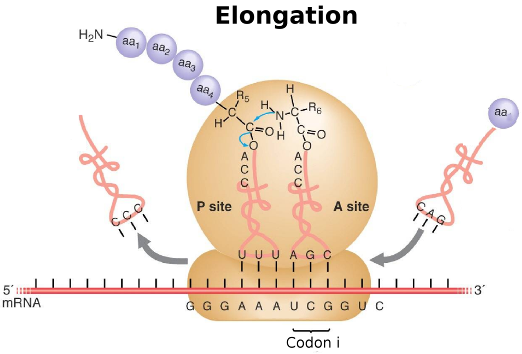
$\Rightarrow$ E(Elongation Time of codon i) $= 1/r_i$

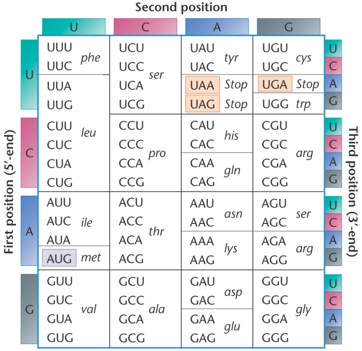


Linking Genotype to Phenotype
Model Assumptions

Problem: Elongation Rates

- Complete codon specific estimates of r do not exist for any organism.
- Goal: Estimate r_i for each codon based on genome's CUB pattern.





	Second position				
First position (5′-end)	U	C	A	G	Third position (3′-end)
U	UUU phe	UCU ser	UAU tyr	UGU cys	U
	UUC	UCC	UAC	UGC	C
	UUA leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG trp	G
C	CUU	CCU pro	CAU his	CGU arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU ile	ACU thr	AAU asn	AGU ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA lys	AGA arg	A
	AUG met	ACG	AAG	AGG	G
G	GUU val	GCU ala	GAU asp	GGU gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA glu	GGA	A
	GUG	GCG	GAG	GGG	G

Linking Genotype to Phenotype

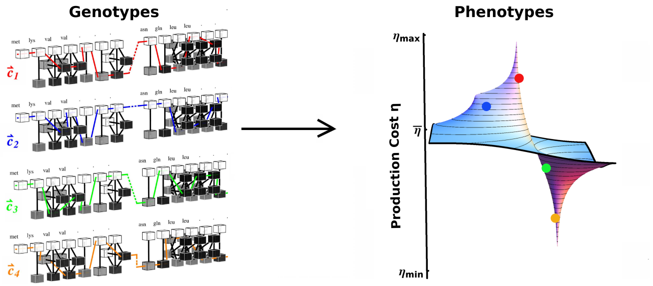
Model Result

Ribosome Overhead Production Cost

Using our elongation rates we can map genotype to phenotype

$$\eta(\vec{r}) = \mathrm{E}(\text{Cost of Protein Production}) = q\sum_{i=1}^{n}\frac{1}{r_i}$$

$q =$ Ribosome overhead cost $n =$ Protein length

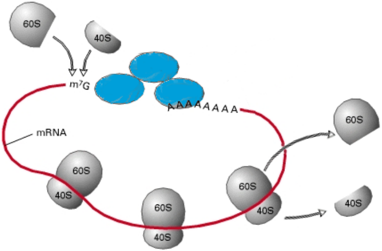


Linking Phenotype to Fitness

Protein Production Rates & Cost

- Selection favors alleles with lower production costs $\eta(\vec{c})$.
- Strength of selection increases with gene's production rate ϕ.

$$\text{Energy Usage}|\ \vec{r} \quad = \quad \underbrace{\eta(\vec{r})}_{\text{Production cost}} \quad \times \quad \overbrace{\phi}^{\text{Production rate}}$$

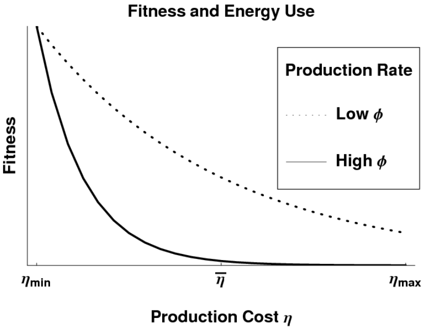


Linking Phenotype to Fitness

Fitness & Energy Usage

Assume fitness decreases exponentially with energy expended to meet target production rate ϕ

$$\begin{aligned} w &= \text{Fitness} \\ &\propto \exp\left[-q \times \eta(\vec{r}) \times \phi\right] \end{aligned}$$



Definitions

q = Scaling term

η = Production cost

$\vec{r}$ = Elongation rates

ϕ = Production Rate

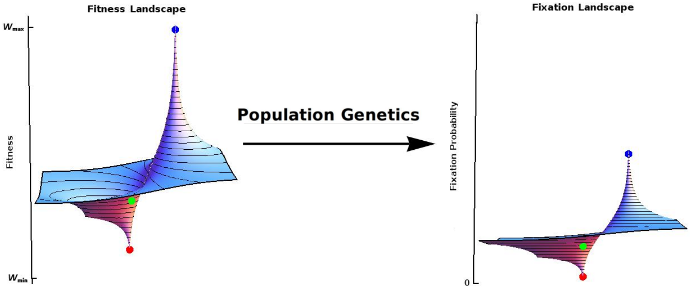
Linking Fitness to Fixation Probability

Fitness and Gene Fixation

Fixation probability of a genotype $\vec{r}$ function of

- Population size N_e
- Fitness Landscape w
- Mutation bias

Wright 1968, Sella & Hirsh (2005)



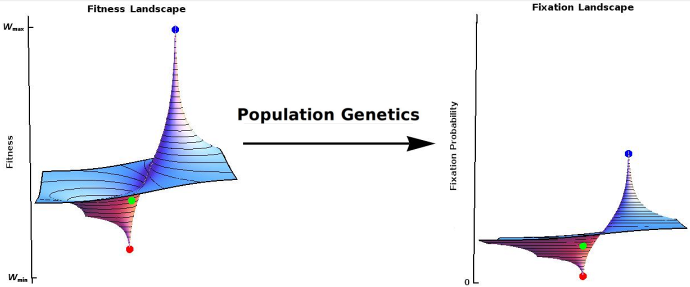
Linking Fitness to Fixation Probability

Gene Fixation and Fitness

$$
\begin{aligned}
P\left(\eta|\phi\right) &= \mathrm{Pr}(\text{Genotype } \vec{r} \text{ fixed in population}) \\
&= \frac{\exp\left[-q\,\phi\,\eta(\vec{r})N_e\right]}{\sum_{i\in\mathbb{C}} \exp\left[-q\,\phi\,\eta\left(\vec{r}_i\right)\ N_e\right]}
\end{aligned}
$$

ϕ = Protein production rate q = Scaling term

N_e = Population size $\mathbb{C}$ = Set of synonymous genotypes

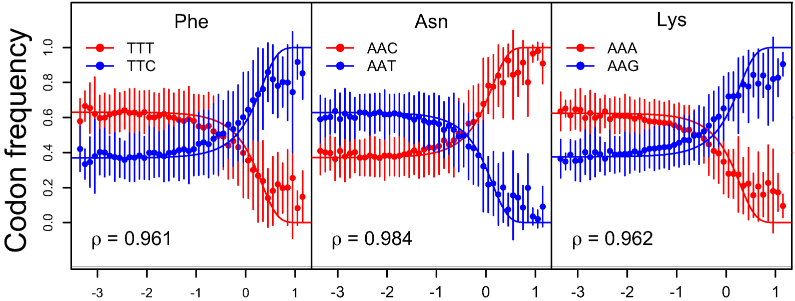


Model Fitting

Likelihood Function: Known ϕs

If we have data on ϕ for multiple genes, we can estimate r_i using a simple likelihood function

$$\text{Lik}(\vec{r}|\vec{\phi}) = \prod_j \frac{\exp\left[-q\,\phi_j\,\eta(\vec{r}_j)N_e\right]}{\sum_{i\in\mathbb{C}} \exp\left[-q\,\phi\,\eta\left(\vec{r}_i\right)\ N_e\right]}$$

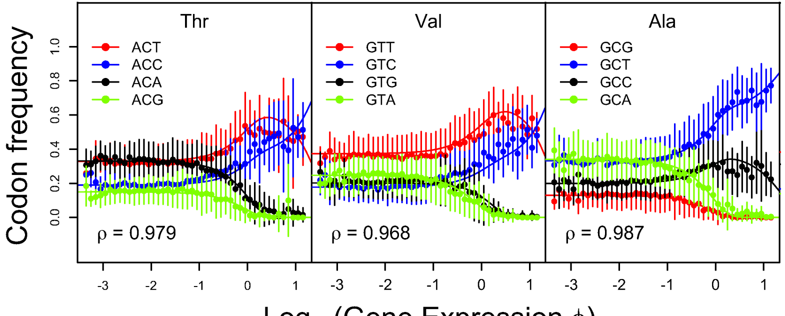


Model Fitting

Likelihood Function: Known ϕs

If we have data on ϕ for multiple genes, we can estimate r_i using a simple likelihood function

$$\text{Lik}(\vec{r}|\vec{\phi}) = \prod_j \frac{\exp\left[-q\,\phi_j\,\eta(\vec{r}_j)N_e\right]}{\sum_{i\in\mathbb{C}} \exp\left[-q\,\phi\,\eta\left(\vec{r}_i\right)\ N_e\right]}$$



Model Fitting

Likelihood Function: Unknown ϕ

Problem: For many organisms, estimates of ϕ don't exist.

Solution: Assume prior for ϕ integrate it out.

Complex Likelihood Function

$$\text{Lik}(\vec{r}|\vec{\alpha}) = \prod_j \int_0^{\phi_{\max}} \frac{\exp\left[-q\,\phi_j\,\eta(\vec{r}_j) N_e\right]}{\sum_{i\in\mathbb{C}} \exp\left[-q\,\phi\,\eta\left(\vec{r}_i\right)\ N_e\right]}\,\pi\left(\phi|\vec{\alpha}\right) d\phi$$

where $\pi(\phi|\vec{\alpha})$ =Prior on ϕ

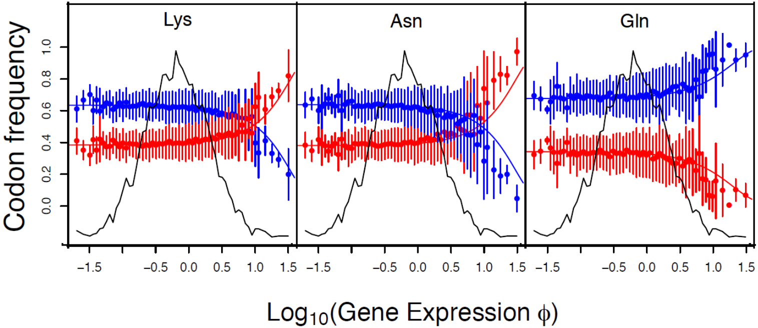
Gilchrist & Shah, In Prep

Model Fitting

Likelihood Function: Unknown ϕ

Problem: For many organisms, estimates of ϕ don't exist.

Solution: Assume prior for ϕ integrate it out.



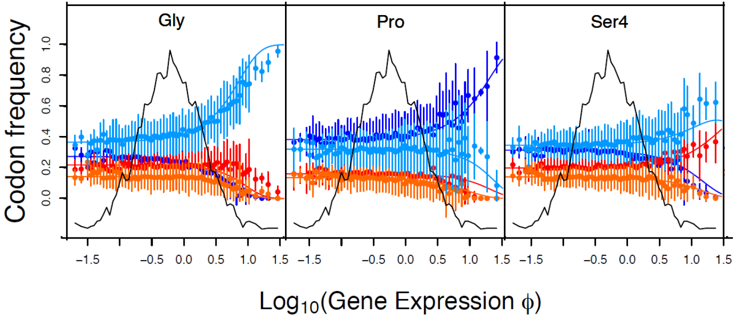
Gilchrist & Shah, In Prep

Model Fitting

Likelihood Function: Unknown ϕ

Problem: For many organisms, estimates of ϕ don't exist.

Solution: Assume prior for ϕ integrate it out.



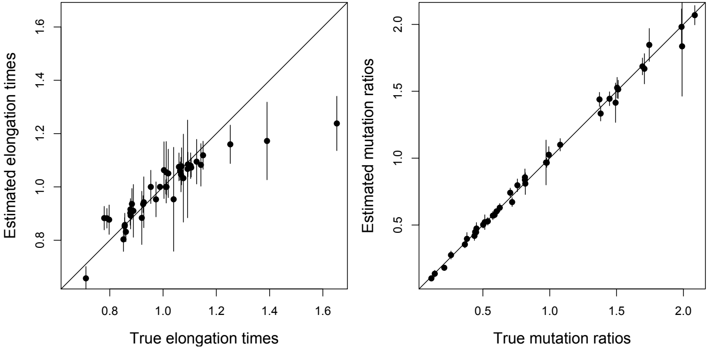
Gilchrist & Shah, In Prep

Model Fitting

Likelihood Function: Unknown ϕ

Problem: For many organisms, estimates of ϕ don't exist.

Solution: Assume prior for ϕ integrate it out.



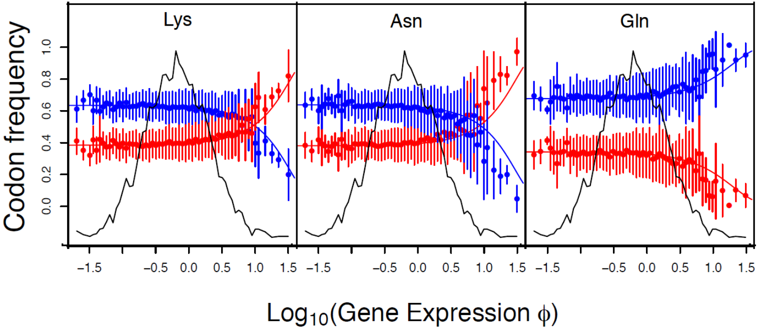
Gilchrist & Shah, In Prep

Model Fitting

Likelihood Function: Unknown ϕ

New Problem: Very computationally intensive!
Run times are $>$ 7 days, making it hard to modify assumptions.

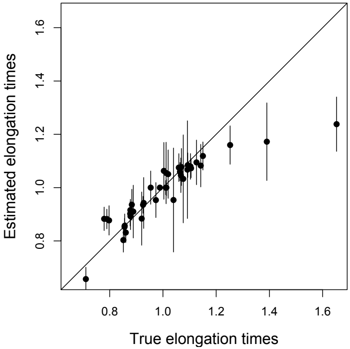
New Solution: HPC!
Expect to get results in hours!

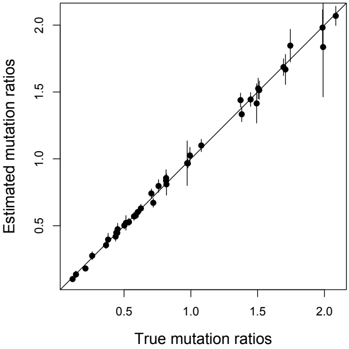


Model Fitting

Computation

- We have analytic solutions for very restricted assumptions.
- Problem is embarrassingly parallel.
- Most genes have little data so need to use many of them.
- Integration routines take most of the time.





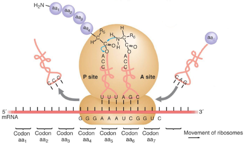
Future Work

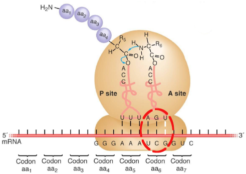
Model Expansion

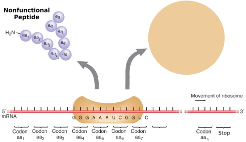
Generalize ribosome model to include multiple costs

- Minimize ribosome wait time
- Minimize Pr. wrong amino acid inserted: “Missense Errors”
- Minimize Pr. premature termination: “Nonsense Errors”

$$\eta(\vec{c}) = \frac{\sum_{i=1}^{n+1} \left(\beta_{i-1} + \gamma_{i-1}(\vec{c})\right) \sigma_{i-1}(\vec{c})\, p(c_i)}{\frac{1}{n}\left(\sum_{i=1}^{n+1} \left(\sum_{j=1}^{i-1} F(c_j)\right) u_{i-1}\, \sigma_{i-1}(\vec{c})\, p(c_i)\right)}$$





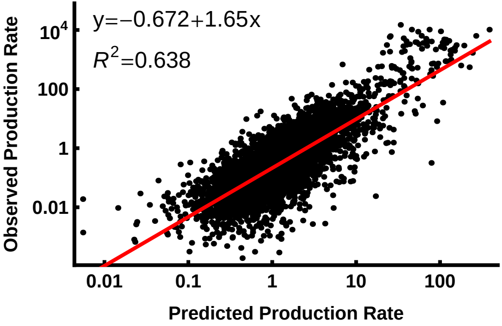


Applications

Long Term Goals

Fit models to genome sequence data to get species specific

- Genome wide estimates of protein production rates.
- Estimates of codon elongation and error rates.
- Quantify the role different forces play in driving CUB.



Acknowledgements

- Premal Shah
- Russell Zaretzki
- Mike Saum
- NIMBioS
- Tennessee Science Alliance JDRD Fund