Introduction

RNA VIRUSES

- Largest virus group; contains most dreaded human pathogens (HIV, Ebola, SARS, Dengue, Hanta, Influenza)
- High mutation rate confers high adaptability to changing conditions and environments
- Easy escape from human intervention (drugs)
- Few effective treatment options

Introduction

Live attenuated virus vaccines - the traditional approach

Attenuation: weakening of a pathogenic virus to a degree that is safe for human administration while retaining ability to elicit protective immune response

- Many rounds of passaging of pathogenic virus through diverse non-human cell cultures and animal hosts
- Virus acquires mutations to adapt to new host conditions
- Virus loses (or not) its pathogenic potential in humans

Introduction

Disadvantages of the traditional approach

- Few attenuating mutations each having a large effect- easily revert to virulence
- Function of attenuating mutations poorly defined or not understood at all
- Costly and time consuming

- Example: The poliovirus vaccine strain Sabin1 was derived by 52 rounds of monkey infections and 16 rounds of monkey kidney cell culture passages, requiring several years of work at prohibitive cost
  (A total of over 100,000 monkeys @ $10,000 = $ 1 Billion)

Sequence Design – Novel sequence design

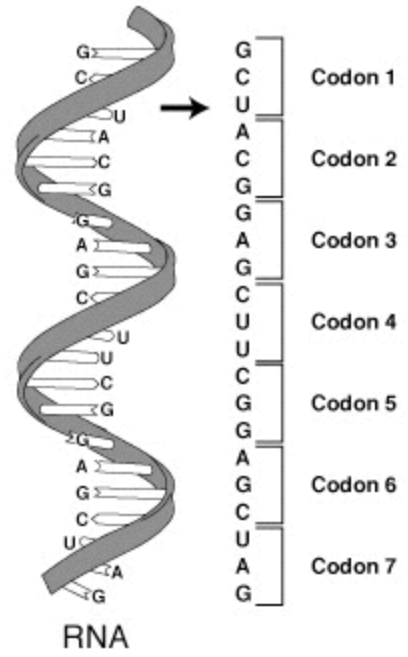
Introduction

How to engineer an attenuated genetically more stable vaccine virus?

"Beat the virus at it's own game" by introducing large number of mutations that the virus can not revert

Sequence Design – Novel sequence design

Introduction



Second base of codon

First base of codon	U	C	A	G	Third base of codon
U	UUU Phenylalanine phe UUC UUA Leucine leu UUG	UCU Serine ser UCC UCA UCG	UAU Tyrosine tyr UAC UAA STOP codon UAG	UGU Cysteine cys UGC UGA STOP codon UGG Tryptonphan trp	U C A G
C	CUU Leucine leu CUC CUA CUG	CCU Proline pro CCC CCA CCG	CAU Histidine his CAC CAA Glutamine gin CAG	CGU Arginine arg CGC CGA CGG	U C A G
A	AUU Isoleucine ile AUC AUA AUG Methionine met (start codon)	ACU Threonine thr ACC ACA ACG	AAU Asparagine asn AAC AAA Lysine lys AAG	AGU Serine ser AGC AGA Arginine arg AGG	U C A G
G	GUU Valine val GUC GUA GUG	GCU Alanine ala GCC GCA GCG	GAU Aspartic acid asp GAC GAA Glutamic acid glu GAG	GGU Glycine gly GGC GGA GGG	U C A G

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Sequence Design – Novel sequence design

Introduction

Species specific synonymous codon bias

- Synonymous codons are used at unequal frequencies
- Rarely used codons =
  rare tRNAs =
  inhibition of protein translation
- Replacing unfavorable (rare) codons with favorable synonymous codons leads to improved translation = Codon Optimization

Codon	Human	Drosophila	E. coli
Arginine:			
AGA	22 %	10 %	1 %
AGG	23 %	6 %	1 %
CGA	10 %	8 %	4 %
CGC	22 %	49 %	39 %
CGG	14 %	9 %	4 %
CGU	9 %	18 %	49 %
Total number of arginine codons	2403	506	149
Total number of genes	195	46	149

Sequence Design – Novel sequence design

Introduction

Does large scale codon de-optimization of poliovirus genome result in attenuation?

- Debilitate virus genome translation/ replication by increasing the number of rare (unfavorable) synonymous codons in the virus genome
- Cumulative phenotype of many mutations each with a small effect - difficult to revert; genetically stable
- 7.9×10^{442} possible synonymous encodings for a poliovirus capsid gene of 881 amino acids (compare this to an estimated 1.3×10^{79} atoms in the known universe)

Sequence Design – Novel sequence design

Experiments

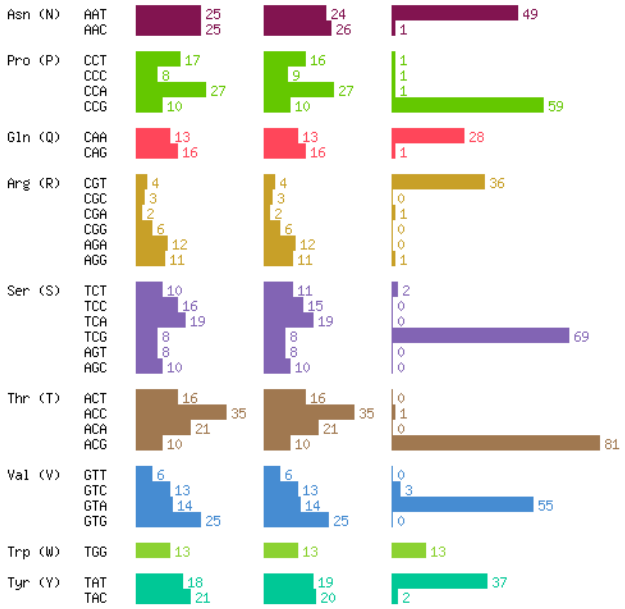
- We tested the hypothesis that underrepresented codons reduce translation efficiency by creating a novel polio capsid design (PV-AB) which:
  - Encoded the same amino-acid sequence
  - Used only the least frequent codon for each amino-acid in human brain specific genes (and in human tissues in general).

  Total number of silent mutations: 680

- We also created another design (PV-SD) which maximized the hamming distance of the capsid encoding, while keeping the same codon frequency distribution.
  - Total number of silent mutations: 934

- Why alter only the capsid coding region?

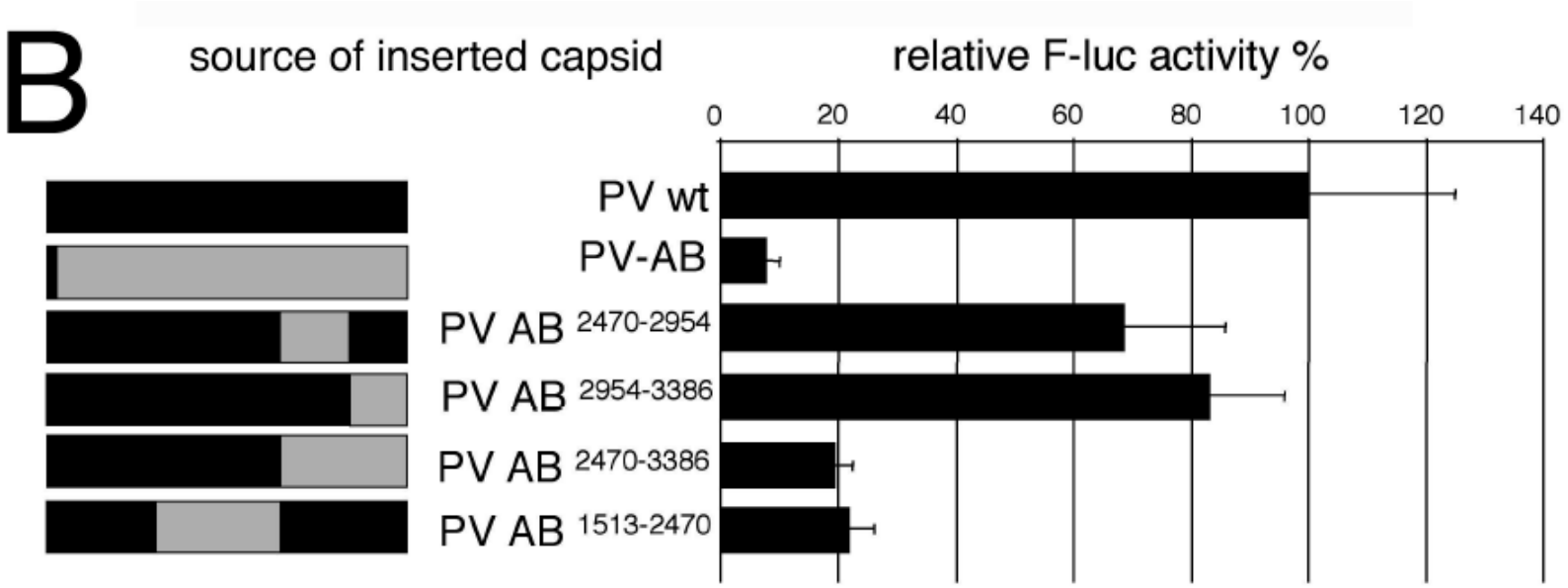
  No cis-acting structural RNA elements



Sequence Design – Novel sequence design

Experiments

- The “shuffled” polio design translates relatively well and is as potent in killing mice as the wildtype.
- The brain-hostile design translates minimally, but use of smaller segments leads to attenuated strains.

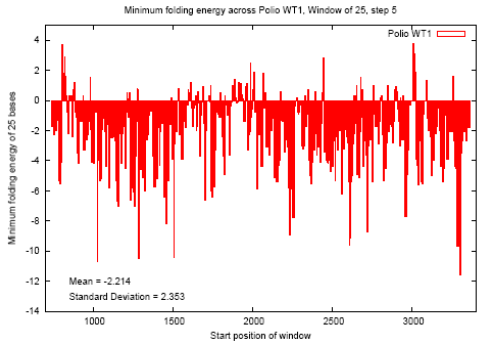


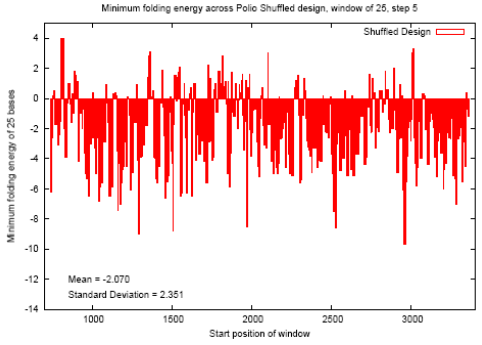
Sequence Design – Novel sequence design

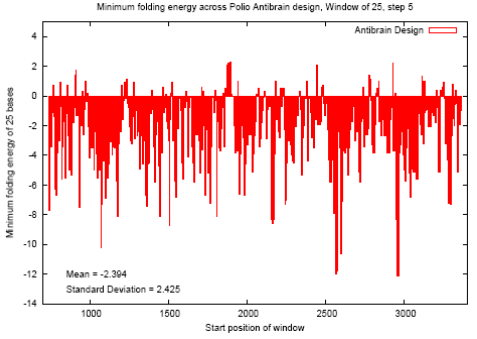
Methods

- To achieve maximum hamming distance without altering codon bias, we used maximum weight bipartite matching between codon positions and codons, using as weight the number of bases changed.
- Restriction sites were inserted uniquely (inserted in specific areas and then eliminated everywhere else).
- Certain regions were locked to preserve secondary structure.

Evaluation of secondary structure:







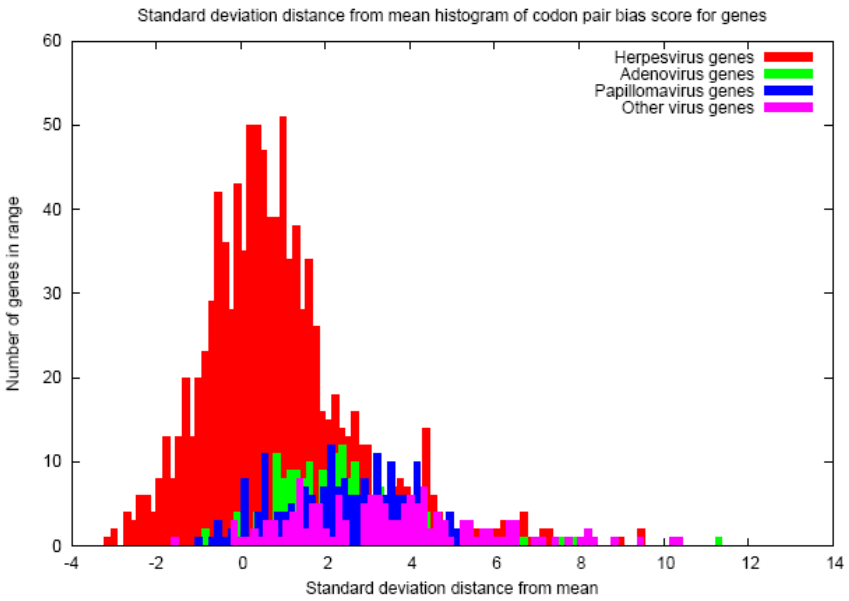
Sequence Design – Novel sequence design

Codon pair bias

- According to Hatfield et al., another source of translation (in)efficiency is the codon pair bias.
- We can measure the bias with the following score:

$$codon\ pair\ score = \log \frac{observed\ codon\ pair\ occurences}{expected\ codon\ pair\ occurences}$$

- Many viruses actually are using overrepresented codon pairs (in human) to encode their genes.



Sequence Design – Novel sequence design

Codon pair bias

- There also seems to be significant correlation between related eucaryotes and codon pair bias.

Eucaryotes codon pair bias correlation

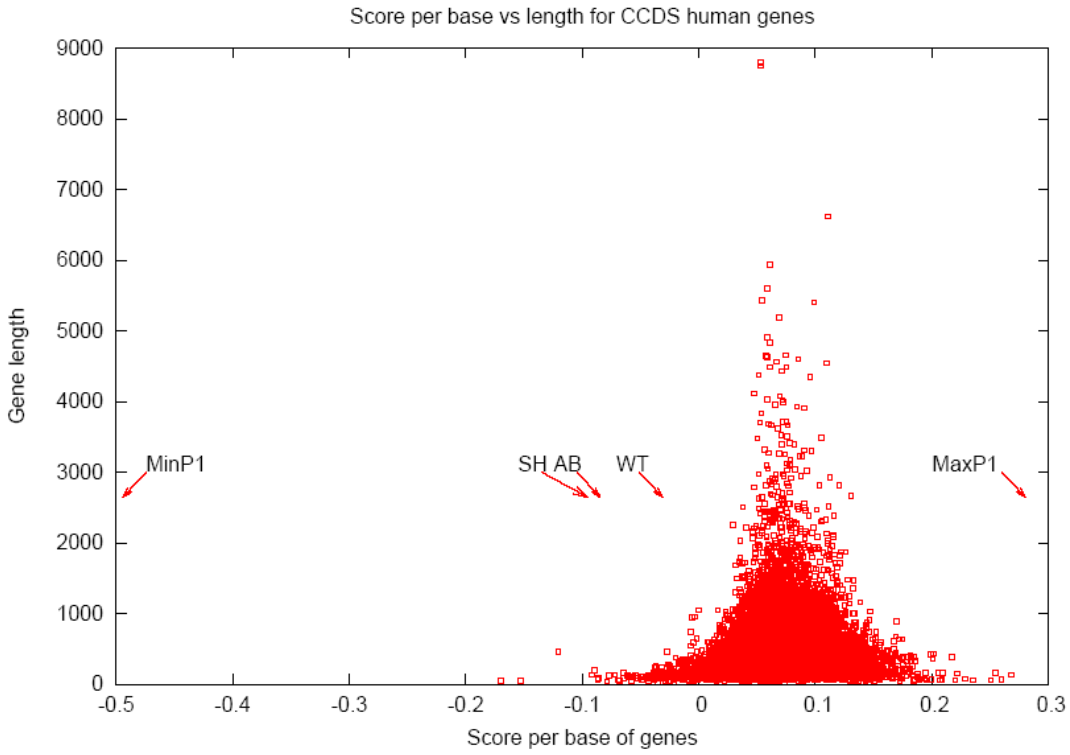
	Saccharomyces cerevisiae	Anopheles gambiae	Apis mellifera	Drosophila melanogaster	Caenorhabditis briggsae	Caenorhabditis elegans	Ciona intestinalis	Fugu rubripes	Tetraodon nigroviridis	Danio rerio	Xenopus tropicalis	Gallus gallus	Mus musculus	Rattus norvegicus	Bos taurus	Canis familiaris	Pan troglodytes	Homo sapiens
Saccharomyces cerevisiae (baker's yeast)	1	0.338	0.248	0.396	0.39	0.422	0.381	0.358	0.357	0.369	0.301	0.423	0.360	0.370	0.330	0.376	0.363	0.363
Anopheles gambiae (malaria mosquito)	0.338	1	0.461	0.573	0.396	0.456	0.375	0.278	0.358	0.263	0.164	0.216	0.124	0.131	0.188	0.212	0.137	0.144
Apis mellifera (honey bee)	0.248	0.461	1	0.431	0.341	0.476	0.365	0.291	0.346	0.234	0.229	0.301	0.204	0.181	0.316	0.32	0.287	0.293
Drosophila melanogaster (fruitfly)	0.396	0.573	0.431	1	0.521	0.641	0.476	0.452	0.542	0.489	0.377	0.421	0.436	0.407	0.397	0.456	0.462	0.463
Caenorhabditis briggsae (soil nematode)	0.39	0.396	0.341	0.521	1	0.867	0.47	0.373	0.395	0.396	0.29	0.325	0.312	0.317	0.294	0.32	0.297	0.295
Caenorhabditis elegans (soil nematode)	0.422	0.456	0.476	0.641	0.867	1	0.574	0.424	0.476	0.449	0.374	0.418	0.401	0.385	0.37	0.425	0.417	0.413
Ciona intestinalis (sea squirt)	0.381	0.375	0.365	0.476	0.47	0.574	1	0.537	0.54	0.552	0.551	0.539	0.498	0.487	0.465	0.507	0.501	0.498
Fugu rubripes (pufferfish)	0.358	0.278	0.291	0.452	0.373	0.424	0.537	1	0.942	0.901	0.776	0.813	0.765	0.792	0.806	0.799	0.793	0.738
Tetraodon nigroviridis (spotted green pufferfish)	0.357	0.358	0.349	0.542	0.395	0.476	0.54	0.942	1	0.867	0.719	0.783	0.759	0.763	0.787	0.795	0.755	0.761
Danio rerio (zebrafish)	0.369	0.263	0.234	0.489	0.396	0.449	0.552	0.901	0.867	1	0.809	0.837	0.813	0.825	0.792	0.801	0.776	0.777
Xenopus tropicalis (pipid frog)	0.301	0.164	0.229	0.377	0.29	0.374	0.551	0.776	0.719	0.809	1	0.914	0.881	0.869	0.868	0.881	0.853	0.854
Gallus gallus (chicken)	0.423	0.216	0.301	0.421	0.325	0.418	0.539	0.813	0.783	0.837	0.914	1	0.92	0.93	0.928	0.94	0.904	0.906
Mus musculus (mouse)	0.360	0.124	0.204	0.436	0.312	0.401	0.498	0.765	0.759	0.813	0.881	0.92	1	0.977	0.912	0.949	0.97	0.960
Rattus norvegicus (rat)	0.370	0.131	0.181	0.407	0.317	0.385	0.487	0.792	0.763	0.825	0.869	0.93	0.977	1	0.937	0.956	0.929	0.93
Bos taurus (cow)	0.330	0.188	0.316	0.397	0.294	0.37	0.465	0.806	0.787	0.792	0.868	0.928	0.912	0.937	1	0.973	0.905	0.914
Canis familiaris (dog)	0.376	0.212	0.32	0.456	0.32	0.425	0.507	0.793	0.795	0.801	0.881	0.94	0.949	0.956	0.973	1	0.949	0.954
Pan troglodytes (chimp)	0.363	0.137	0.287	0.462	0.297	0.417	0.501	0.733	0.755	0.776	0.853	0.904	0.97	0.929	0.905	0.949	1	0.998
Homo sapiens (human)	0.363	0.144	0.293	0.463	0.295	0.413	0.498	0.738	0.761	0.777	0.854	0.906	0.959	0.93	0.914	0.954	0.998	1

0.8 - 1.0 | 0.6-0.8 | 0.4 - 0.6 | 0.2 - 0.4 | -0.2 - 0.2

Sequence Design – Novel sequence design

Our designs

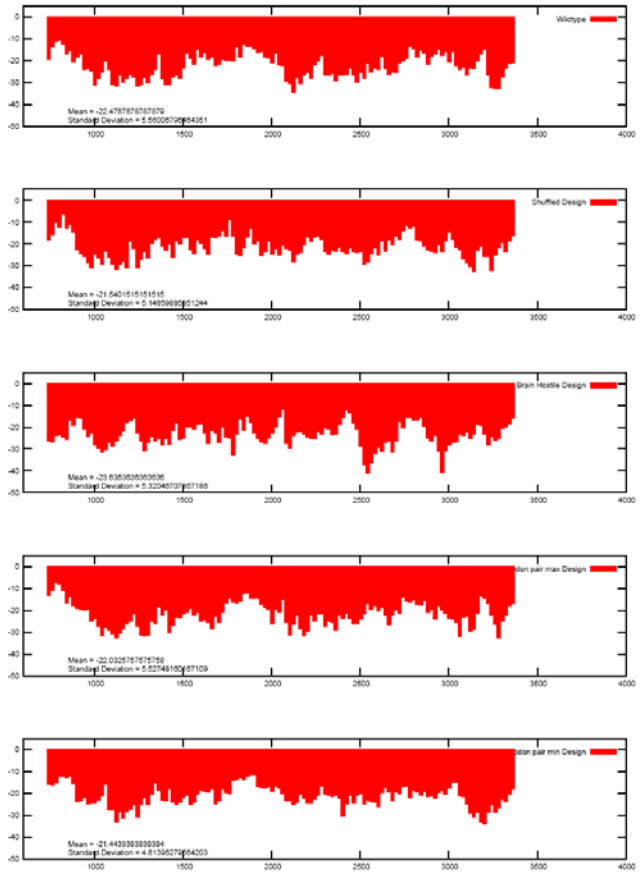
- We designed two polio capsid sequences that optimize the usage of over-represented and under-represented codon pairs in human.



Sequence Design – Novel sequence design

Our designs

- The design consisted of the following steps:
  - Same codon frequency distribution
  - Optimized codon pair score
  - Restriction site uniqueness and elimination
  - Secondary structure folding energy minimization
  - Splice site elimination
- Goals achieved with simulated annealing, optimization passes and manual intervention.



Sequence Design – Novel sequence design

Our designs

How do the new codon pair biased designs behave?

- maxP1 (using overrepresented codon pairs) translates as well as the wildtype.
- minP1 (using underrepresented codon pairs) translates poorly.

Investigation on the influence of other signals, such as CG dinucleotide content, which are inherent in such biased constructs, is also important, in order to determine the origin of the translation rate differences of different designs.