DNA AND CELL BIOLOGY
Volume 23, Number 10, 2004

Pp. 701–705

The Ambush Hypothesis: Hidden Stop Codons Prevent Off-Frame Gene Reading

HERVÉ SELIGMANN[1,2] and DAVID D. POLLOCK[1]

ABSTRACT

Coding sequences lack stop codons, but many stops appear off-frame. Off-frame stops (stops in −1 and +1 shifted reading frames, termed hidden stops) terminate frame-shifted translation, potentially decreasing energy, and resource waste on nonfunctional proteins. Benefits may include reduced waste elimination costs and avoidance of potentially cytotoxic frame-shifted products. Our "ambush" hypothesis suggests that hidden stops are sometimes selected for. Codons of many amino acids can contribute to hidden stops, depending on the synonymous position state and adjacent codons. In vertebrate mitochondria, 31.75% of all amino acid combinations can form hidden stops. Codons with more potential to form hidden stops have greater usage frequency and bias in their favor among synonymous codons. Among primates, predicted mitochondrial rRNA secondary structure stability correlates negatively with the number of hidden stops in the mitochondrial genome. The taxonomic distribution of genetic codes suggests that +1 frameshifts might be more frequent than −1 frameshifts. This is confirmed by analyses of primate mitochondrial genomes: species with unstable rRNAs have more +1 stops, but the correlation is weak for −1 stops. High hidden stop density seems to be an adaptation in species with slippage prone ribosomes (unstable rRNAs). Hidden stops may thus compensate for reduced efficiency of some parts of the biosynthetic machinery. Some experimental data confirm our hypothesis: gene expression increases with the experimentally manipulated number of stops in the promoter region of a gene, suggesting biotechnological applications.

INTRODUCTION

THIS STUDY DEALS with a putative mechanism that minimizes production of incorrect (off-frame) translation products. Frame shifts are defined as protein translations that start not at the first, but at the second (+1 frameshift) or the third (−1 frameshift) nucleotide of the codon. "Zero" frameshifts produce the "normal," presumably functional, protein. At zero frameshift, coding sequences usually end with stop codons, and no stop codons occur within the coding region.

Frameshifting is one of three classes of recoding of mRNAs (Baranov *et al*., 2001). Surprisingly, little precise data is available on the frequency of frameshift occurence. In some exceptional cases called programmed frameshifts, frequencies are known because +1 or −1 frameshifts yield functional products that differ from the protein produced by zero frameshift reading of the same coding sequence. For example, a ribosomal −1 frameshift occurs in 16% of transcriptions of the *cdd* gene in *Bacillus subtilis*, at a nucleotide upstream to the zero-frameshift stop codon (Mejhelde *et al*., 1999). This shift extends the protein by 13 amino acids, results in an alternate protein with approximately the same activity, and is induced by a Shine-Dalgarno-like sequence located upstream of the shift site (Mejhelde *et al*., 1999). Shine-Dalgarno sequences are part of promoter sequences, and are usually located upstream of the start codon of the coding sequence. They "prepare" the ribosome for the onset of transcription. It is plausible that even regular (nonprogrammed) frameshifts are not rare events. Presumably, most frameshifts would yield nonfunctional proteins; cases where the alternative protein is functional are probably exceptions. Hence, frameshifts lead to waste of energy, resources, and activity of the biosynthetic machinery. Waste elimination is also an energetic cost, and some peptides synthesized after frameshifts are probably cytotoxic.

[1]Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, Louisiana.
[2]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois.

Considering these costs, it seems likely that mechanisms may exist to stop frameshifted translation as soon as possible, depending on the frequency of occurrence. Inclusion of stop codons at close intervals within unused reading frames would reduce the costs of accidental frameshifts, but the costs must be large enough for selection to operate on codon usage, especially the usually synonymous third codon position.

When a frameshift occurs, zero frameshift codons can potentially contribute to two frameshifted codons, depending on adjacent sequence. In +1 frameshifts, what was the first codon position (N1) becomes the third codon position (N3) in the frameshifted sequences, and what were the last two codon positions (N2 and N3) become N1 and N2 in the subsequent codon of the frameshifted sequence. In −1 frameshifts, N1 and N2 become N2 and N3 in the frameshifted sequence, and N3 becomes N1 in the subsequent codon. Due to combinatorics, any codon (including a stop codon, and regardless of the specific genetic code) has 19 codons that can potentially contribute to it in an off-frame context, and 45 that cannot. For example, in a −1 frameshift, the leftmost off-frame codon has $4 \times 4 = 16$ possible contributing codons, while the rightmost off-frame codon has only four possible contributing codons, since two positions are specified, but one of these 20 codons is included in both the possible sets. Contributing codons that code for off-frame stops do not affect the original +0 frame.

ALTERNATIVE GENETIC CODES

Table 1 displays the numbers of codons that can contribute to off-frame stop codons for frameshifts +1 and −1 in the seven combinations of stop codon assignments observed in known genetic code variants (Elzanowski *et al*., 2000; for review, see also Santos *et al*., 2004). This is 19 for codes with only one stop, but for codes with multiple stops the number varies, depending on the overlaps among the stops. For example, among the 64 codons in the standard genetic code, only 20 cannot become part of a stop codon in frameshifted conditions (42 for −1 frameshifts, and 28 for +1 frameshifts). Some codons can contribute to hidden stop codons in several ways, with up to six possibilities for AGU (serine) and AAU (asparagine). For the vertebrate mitochondrial code, adequate choice of synonymous codons can create an off-frame stop codon in 127 (31.75%) of all 400 possible adjacent amino acid combinations.

BIASES IN CODON USAGE

Assuming that hidden stop codons terminate sequence reading after accidental frameshifts, one would expect that there is an advantage in using codons that can be part of hidden stop codons, and that this advantage might allow selection for such codons. This "ambush" hypothesis predicts a positive correlation between the usage of codons and the number of ways codons can be part of hidden stops. Indeed, among 100 organisms from all major taxonomic groups (viruses, archaea, bacteria, and metazoa), a positive, statistically significant ($P < 0.05$, one-tailed test) correlation between the number of ways a codon can contribute to a hidden stop and the mean genome-wide usage frequency of the codons exists in 38 organisms (genomic data available at *http://www.kasuza.or.jp/codon/*; Nakamura *et al*., 2000). Figure 1 presents the results for the bacteria *Borrelia burgdorferi* and the metazoan fungi *Saccharomyces cerevisiae*. In bacteria, correlations were positive in 29 among 46 species, with 17 positive correlations significant at $P < 0.05$, and no significant negative correlations (two to three are expected significant at $P < 0.05$ for random data); in Archaea, correlations were positive in 8 out of 13 species, with four positive correlations significant at $P < 0.05$ and no significant negative correlations (one significant case expected; the exact binomial probability considering multiple comparisons is extremely low). The phenomenon was strongest among viruses (11 positive correlations among 13 viruses, with seven significant at $P < 0.05$ and seems weakest in metazoans (only six positive correlations among 15 species, but four among these were significant at $P < 0.05$ (one significant case expected in both groups)). All correlations for 186 vertebrate mitochondria were positive. Similar analyses with synonymous codon bias indices, rather than codon frequencies, yielded very similar correlation coefficients, qualitatively and quantitatively.

These results suggest three potential adaptive phenomena. (a) At the level of genetic codes, ancient adaptive events may have adjusted codon assignments to increase frequencies of codons that can be part of hidden stops, by assigning codons

Table 1. Alternative Genetic Codes with Different Stop Codon Assignments

Stop codons	*−1*	*+1*	*Tot.*	*Genetic codes*
TAG	19	19	34	Mt flatworm
TGA	19	19	34	Nucl. ciliate
TAA TAG	22	34	42	Nucl. euplotid, Mt ascidian, echinoderm, invertebrate, mold, trematode, yeast
TAA TGA	22	22	37	Mt chlorophycean, nuclear *Blepharisma*
TAA TAG TGA	22	36	44	Standard
TAG TCA TGA	25	25	43	Mt *Scenedesmus obliquus*
AGA AGG TAA TAG	46	42	50	Mt vertebrates

Columns 2–3 show the number of codons that can contribute to hidden stops in −1 and +1 frameshift contexts; column 4 is the number of codons that can contribute to either or both frameshift types. The last column indicates the genetic systems in which the genetic code is found.

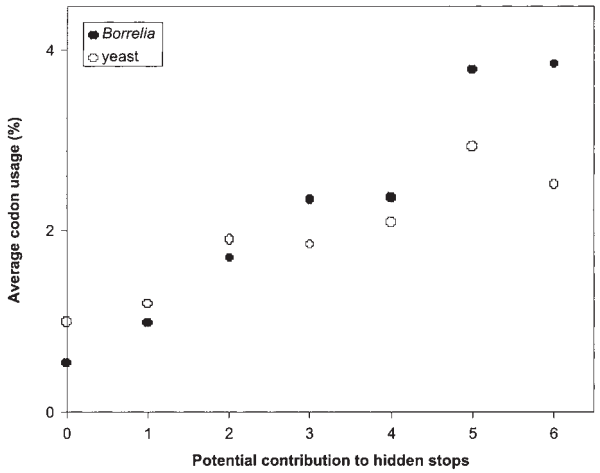

FIG. 1. Codon usage frequencies in genomes of *Borrelia* and *Saccharomyces cerevisiae* as a function of the potential number of ways that these codon can contribute to stop motifs in off-frame contexts.

that can be part of hidden stops to frequently used amino acids. Although these ancient adaptations are nearly impossible to prove one way or another, the large number of complete mitochondrial genomes available shows some patterns that are suggestive of adaptations, because low codon frequencies can be related to codon reassignment (Knight *et al*., 2001). At least in one case, the intermediate stage where a codon is ambiguous, is selectively favored in *Candida* sp. (Santos *et al*., 1999). At the level of sequence microevolution, synonymous substitutions that yield −1 or +1 frameshift stop codons may not be neutral, and organisms may adapt by (b) increasing the usage of synonymous codons proportionally to the number of ways these specific codons can be part of hidden stop codons. They may also select for (c) usage of specific codons that form hidden stops in specific sequences.

OPTIMIZING ACCORDING TO +1 and −1 FRAMESHIFT FREQUENCIES

Note that in Table 1, the taxonomically more common combinations of stop codon assignments have greater stop codon potentiality for +1 than −1 frameshifts. Comparing TAA+TAG with TAA+TGA, and TAA+TAG+TGA with TAA+TCA+TGA, the first stop codon assignment is phylogenetically more widespread than the latter stops, and yields more potential +1 than −1 off-frame stops. This suggests that stop codon assignments might be optimized according to the natural occurrence of +1 and −1 frame shift, and that +1 frameshifts are more frequent than −1 frameshifts.

This hypothesis is corroborated by observations in vertebrate mitochondria. The correlation between codon usage and their potential contribution to hidden stops was greater for potential contribution to +1 than −1 off-frame stop codons in 176 among 186 vertebrate species for which the complete mitochondrial genome is sequenced. The exceptions do not show clear phylogenetic trends. There are five exceptions among lower vertebrates (*Chimaera monstrosa*, *Lepidosiren paradoxa*, *Mustelus manzano*, *Petromyzon marinus*, *Zu cristatus*), and among five amphibian genomes, only *Ranodon sibiricus* showed greater −1 than +1 stop codon optimization. There were no such cases among reptiles and birds (8 and 23 genomes analyzed, respectively), and there were four cases among 72 mammals (*Echinops telfairi*, *Erinaceus europaeus*, *Myoxus glis*, and *Ornythorhynchus anatinus*).

Off-frame stop codon density and ribosomal secondary structure stability

The ambush hypothesis implies that the need for hidden stops increases with the probability of frameshifts. It seems plausible that ribosomes with relatively unstable structures are more likely to frameshift than more stable ribosomes. To determine if available data is consistent with this working hypothesis, we counted the number of hidden stops in all 13 coding sequences from 16 available complete primate mitochondrial genomes and two near-primate outgroup genomes, *Tupaia* and *Cynocephalus*, and plotted this count as a function of the predicted stability, or ΔG (Mathews *et al*., 1999; Zuker *et al*., 1999; Zuker, 2003) of the secondary structure formed by 12s and 16s mitochondrial rRNAs of the same genomes (Table 2). Correlation analyses show that in primates, low rRNA stabilities (high ΔG) associate with high counts of hidden stops, and stable rRNAs with low counts of hidden stops (12s rRNA, $r = 0.83$, $df = 16$, $P < 0.01$; 16s rRNA, $r = 0.33$, $df = 16$, not significant).

Further analyses also corroborate the hypothesis, suggested by the comparisons between alternative genetic codes (Table 1), that +1 frameshifts are more frequent than −1 frameshifts (see previous section). In multiple regression analyses with primate mitochondrial genomes, taking the stability of 12s rRNA as the dependent and the numbers of −1, and +1 frameshifted hidden stop codons as two independents, the regression coefficient of the number of −1 frameshifted hidden stop codons is weaker (0.26, $t = 1.53$, $P = 0.07$) than that for the number of +1 frameshifted hidden stop codons (0.34, $t = 1.81$, $P = 0.046$). Both variables together explain 69% of the variation in 12s rRNA stability.

EXPRESSION EFFICIENCY

Early termination of off-frame transcription should increase the efficiency of expression of a gene, because less time and resources are invested in unproductive off-frame contexts: if the reading frame of the ribosome is not zero, the earlier a stop terminates translation, the earlier mRNA and ribosome are available for interacting correctly. In two experiments where artificial variation has been introduced in the sequence between the promoter and the start codon (Gheysen *et al*., 1982; Ozbudak *et al*., 2002), there is a positive correlation between the number of stop motifs that would be off-frame according to the start codon, and the measured expression levels of the gene. There also is a negative correlation between the number of stop codons in that region that are in frame with the start codon of the gene, and the expression level of the gene in that experi-

Table 2. Numbers of Off-Frame Stop Codons in the 13 Coding Sequences of 13 Primate Mitochondrial Genomes and Two Outgroups, *Cynocephalus* and *Tupaia*

Species	*al|tg*	*at|lg*	*t|laa*	*t|lag*	*a|lga*	*a|lgg*	*ta|la*	*ta|lg*	*ag|la*	*ag|lg*	*12s*	*16s*	*P*
Cebus albifrons	52	38	81	15	83	101	238	146	2	5	−233.2	−345.5	180
Cercopithecus aethiops	51	39	59	17	71	84	210	137	12	10	−237.5	−352.9	73
Colobus guereza	57	44	71	16	70	99	214	155	9	6	−225.5	−379.6	196
Cynocephalus variegatus	57	35	55	14	77	105	201	157	17	12	−217.5	−372.9	60
Gorilla gorilla	43	38	55	10	77	99	186	161	5	3	−246.8	−359.7	265
Homo sapiens	48	27	46	15	75	100	190	158	6	5	−250.4	−362.2	270
Hylobates lar	47	21	46	12	68	90	197	134	10	11	−252.3	−372.2	210
Lemur catta	60	42	80	19	86	121	225	190	9	5	−205.7	−342.9	135
Macaca sylvanus	49	36	57	14	64	88	219	128	5	12	−256.0	−365.1	170
Nycticebus	63	36	72	18	81	95	206	159	13	6	−243.2	−377.0	90
Pan paniscus	55	31	55	13	76	103	200	162	2	5	−239.6	−345.3	244
Pan troglodytes	53	34	60	12	76	107	192	159	5	2	−238.9	−349.1	231
Papio hamadryas	44	26	54	10	65	84	210	131	14	4	−238.8	−359.6	180
Pongo abelli	52	25	39	10	71	102	200	156	11	10	−266.4	−374.8	259
Pongo pygmaeus	52	16	44	15	71	94	198	143	11	8	−250.4	−371.5	249
Tarsius bancanus	64	33	77	16	79	108	242	170	9	4	−204.7	−354.7	180
Trachypithecus obscurus	57	41	94	15	68	90	214	149	8	11	−228.6	−359.3	145
Tupaia belangeri	65	30	63	19	74	106	196	194	13	4	−228.6	−398.6	48

Columns 2–3: +1 and −1 start codons; columns 4–7: +1 frameshift stop codons; columns 8–11: −1 frameshift stop codons; columns 12–13: delta G of optimal secondary structure of 12s and 16s rRNA; column 14: gestation time in days.

ment. These qualitative results were statistically significant in both systems, and our working hypothesis thus provides an explanation for these experimental results.

OFF-FRAME CONSTRAINTS, GENE SIZE, AND EXPRESSION LEVELS, AND CODON USAGE OPTIMIZATION

If the ambush hypothesis is correct, hidden stops should be more frequent for large and frequently expressed genes, since costs of off-frame translation are likely to increase with gene size and expression levels. In addition, the costs of off-frame translation are probably higher when frameshifts occur close to the (5′) start of the coding sequence, rather than close to its 3′ end of the coding sequence. Such associations are indeed evident (e.g., the relative proportion of hidden stops in *Halobacterium* genes is greater for the first sector of the gene, close to its 5′ start). These predictions are, however, not strong tests of the hypothesis because they are similar to those predicted by selection on codon usage, and hidden stops are highly concordant with codon usage biases in most organisms (see section above and Fig. 1). Codon usage biases have already been shown to correlate positively with expression levels of genes (in *Escherichia coli*, Gouy and Gautier, 1982; Ikemura, 1982; in *Saccharomyces cerevisiae*, Bennetzen and Hall, 1982) and the gene's size (in *Drosophila*, Comeron *et al*., 1999; yeast, Kliman *et al*., 2003; *Arabidopsis*, *E. coli*, *Halobacterium*, and *Homo*, Seligmann, 2003). Even the gene sector prediction is confounded by codon usage biases: in *Drosophila*, the region proximal to the 5′ end of the mRNA is mostly optimized (Comeron *et al*., 1999). Hence, although noteworthy, these associations do not necessarily offer further proof of the ambush hypothesis and are therefore not presented in detail. However, predictions for correlations between predicted rRNA stability and hidden stop numbers (from Table 2) are not confounded by codon usage optimization.

Nucleotide combinations at the third codon position and the next nucleotide, at the first codon position in the next codon have been observed as part of the organism-wide DNA signature (Karlin and Mrazek, 1996). Some combinations are universally avoided (Antezana and Kreitman, 1999), decreasing mRNA degradation potential by endoribonuclease cleavage that target sites specifically according to dinucleotide combinations (Duan and Antezana, 2003). Interestingly, the avoided cleavage site, combinations of U and subsequent A, is the most common dinucleotide combination in assigned stop codons. Hence, our results indicate that most organisms follow more the ambush than the cleavage hypotheses, although variation among organisms in the relative effects of each phenomenon on codon choice seems probable.

DEVELOPMENTAL RATES IN PRIMATES

It is also interesting to note that the total number of hidden stops in the 13 mitochondrial proteins correlates negatively with gestation time in primates ($r = -0.53$, same sources for gestation times as in Seligmann, 2003, data in Table 2). There are many intermediate steps linking molecular transcription efficiency and the rate of morphogenesis at the whole organism level, but it is possible to speculate that the ambush mechanism might decrease the biosynthetic costs associated with development, and that more hidden stops somehow allow faster development by more effectively stopping off-frame synthesis, and, at the same time, enabling translation to occur with a less stable, but perhaps faster ribosome.

CONCLUSION

By putatively increasing biosynthetic efficiencies, hidden stop codons could increase metabolic efficiency at the molecular level. We describe correlative evidence that supports the hypothesis that this "ambush" mechanism is selected for in some organisms, and we discuss how this hypothesis may explain evidence from previous experimental manipulations. It may even carry over to the whole organism level, affecting development times. Further exploration of coevolution between rRNA sequences and structures and hidden stops could help elucidate the types of substitutions responsible for changes in ribosomal accuracy. Further analyses may demonstrate concrete links between the specific primate rRNA variant stabilities and the densities of hidden stop codons, and may explain sequence variation in different species.

ACKNOWLEDGMENTS

This work was partly supported by grants from the National Institutes of Health (GM065612 to D.D.P.), the State of Louisiana Board of Regents Millennium Research Program's Biological Computation (to D.D.P.), and Visualization Center and the State of Louisiana's Governor's Biotechnology Initiative (to D.D.P.).

REFERENCES

ANTEZANA, M.A., and KREITMAN, M. (1999). The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences J. Mol. Evolut. **49,** 36–43.

BARANOV, P.V., GURVICH, O.L., FAYET, O., PRÈRE, M.F., MILLER, W.A., GESTELAND, R.F., ATKINS, J.F., and GIDDINGS, M.C. (2001). RECODE: A database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucleic Acids Res. **296,** 264–267.

BENNETZEN, J.L., and HALL, B.D. (1982). Codon selection in *Saccharomyces cerevisiae*. J. Biol. Chem. **257,** 3026–3031.

COMERON, J.M., KREITMAN, M., and AGUADÈ, M. (1999). Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151,** 239–249.

DUAN, J.B., and ANTEZANA, M.A. (2003). Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J. Mol. Evol. **57,** 694–701.

ELZANOWSKI, A., OSTELL, J., LEIPE, D., and SOUSSOV, V. (2000). The Genetic Codes. Bethesda, MD: National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG2)

GHEYSEN, D., ISERENTANT, D., DEROM, C., and FIERS, W. (1982). Systematic alteration of the nucleotide-sequence preceding the translation initiation codon and the effects on bacterial expression of the cloned sv40 small-t antigen gene. Gene **17,** 55–63.

GOUY, M., and GAUTIER, C. (1982). Codon-usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res. **10,** 7055–7074.

IKEMURA, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **146,** 1–21.

KARLIN, S., and MRAZEK, J. (1996). What drives codon choices in human genes? J. Mol. Biol. **262,** 459–472.

KLIMAN, R.M., IRVING, N., and SANTIAGO, M. (2003). Selection conflicts, gene expression, and codon usage trends in yeast. J. Mol. Evol. **57,** 98–109.

KNIGHT, R.D., LANDWEBER, L.F., and YARUS, M. (2001). How mitochondria redefine the code. J. Mol. Evol. **53,** 299–313.

MATHEWS, D.H., SABINA, J., ZUKER, M., and TURNER, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288,** 911–940.

MEJHEKDE, N., ATKINS, J.R., and NEUHARD, J. (1999). Ribosomal-1 frameshifting during decoding of *Bacillus subtilis cdd* occurs at the sequence CGA AAG. J. Bacteriol. **181,** 2930–2937.

NAKAMURA, Y., GOJOBORI, T., and IKEMURA, T. (2000). Codon-usage tabulated from the international DNA sequence databases: Status for the year 2000. Nucleic Acids Res. **28,** 292.

OZBUDAK, E.M., THATTAI, M., KURTSER, I., GROSSMAN, A.D., and VAN OUDENAARDEN, A. (2002). Regulation of noise in the expression of a single gene. Nat. Genet. **31,** 69–73.

SANTOS, M.A., CHEESMAN, C., COSTA, V., MORADAS-FERREIRA, P., and TUITE, M.F. (1999). Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp. Mol. Microbiol. **31,** 937–947.

SANTOS, M.A., MOURA, G., MASSEY, S.E., and TUITE, M.F. (2004). Driving change: The evolution of alternative genetic codes. Trends Genet. **20,** 95–102.

SELIGMANN, H. (2003). Cost minimization of amino acid usage. J. Mol. Evolut. **56,** 151–161.

ZUKER, M., MATHEWS, D.H., and TURNER, D.H. (1999). *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology.* J. Barciszewski and B.F.C. Clark, eds. (Kluwer Academic Publishers) NATO ASI Series.

ZUKER, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31,** 3406–3415.

Address reprint requests to:
Hervé Seligmann, Ph.D.
Department of Biological Sciences
Biological Computation and Visualization Center
Louisiana State University
Baton Rouge, LA 70803

E-mail: hselig@lsu.edu