Two Proteins for the Price of One: The Design of Maximally Compressed Coding Sequences*

Bei Wang[1], Dimitris Papamichail[2], Steffen Mueller[3], and Steven Skiena[2]

[1] Dept. of Computer Science, Duke University, Durham, NC 27708
beiwang@cs.duke.edu
[2] Dept. of Computer Science, State University of New York, Stony Brook, NY 11794
{dimitris, skiena}@cs.sunysb.edu
[3] Dept. of Microbiology, State University of New York, Stony Brook, NY 11794
smueller@ms.cc.sunysb.edu

Abstract. The emerging field of *synthetic biology* moves beyond conventional genetic manipulation to construct novel life forms which do not originate in nature. We explore the problem of designing the provably shortest genomic sequence to encode a given set of genes by exploiting alternate reading frames. We present an algorithm for designing the shortest DNA sequence simultaneously encoding two given amino acid sequences. We show that the coding sequence of naturally occurring pairs of overlapping genes approach maximum compression. We also investigate the impact of alternate coding matrices on overlapping sequence design. Finally, we discuss an interesting application for overlapping gene design, namely the interleaving of an antibiotic resistance gene into a target gene inserted into a virus or plasmid for amplification.

1 Introduction

The emerging field of *synthetic biology* moves beyond conventional genetic manipulation to construct novel life forms which do not originate in nature. The synthesis of poliovirus from off-to-shelf components [1] attracted worldwide attention when announced in July 2002. Subsequently, the bacteriophage PhiX174 was synthesized using different techniques in only three weeks [2], and Kodumal, et al. [3] recently set a new record for the longest synthesized sequence, at 31.7 kilobases. The ethics and risks associated with synthetic biology continue to be debated [4], but the pace of developments is quickening. Indeed, Tian, et al. [5] have just proposed a method for DNA synthesis based on microarrays and multiplex PCR that promises a substantial reduction in cost.

Once you can synthesize an existing genome from scratch, you can do the same for new and better designs as well. In this paper, we explore an interesting problem in genome design, namely designing the provably shortest genomic sequence to encode a given set of genes, by exploiting alternate reading frames and the redundancy of the genetic code. Theoretically, up to six proteins can be

* This research was partially supported by NSF grants EIA-0325123 and DBI-0444815.

A. Carbone and N.A. Pierce (Eds.): DNA11, LNCS 3892, pp. 387–398, 2006.

encoded on the same genomic sequence using three alternate reading frames on both strands. Indeed, long gene overlaps occur frequently in nature.

Our contributions in this paper are:

- *Finding Shortest Encodings for Given Protein Pairs* – We present an algorithm for designing the shortest DNA sequence simultaneously encoding two given amino acid sequences. Our algorithm runs in worst-case quadratic time, but we provide an expected-case analysis explaining its observed linear running time when employing the standard DNA triplet code.
- *Comparing Natural and Synthetic Coding-Pair Sequences* – We compare the overlapping gene designs constructed by our algorithm to those occurring in natural viral sequences. We show that the coding sequence of naturally occurring pairs of overlapping genes in general approach maximum compression, meaning that it is impossible to design overlapping shorter coding sequences for them which save more than 1-2% over independent genes. This counterintuitive result has natural explanation in terms of the evolutionary mechanics of overlapping gene sequences.

  Further, we show interesting differences between the preferred phase (reading frame), strand, and orientation of natural and optimized overlapping sequences.
- *Impact of Alternate Coding Matrices on Overlapping Sequence Design* – Protein designs are not immutable; indeed, certain pairs of amino acids share such similar physical/chemical properties they can be fairly freely substituted without altering protein function. This freedom can be exploited to design substantially shorter encodings for a given pair of proteins.

  We investigate the impact of increasingly permissive amino acid substitution matrices (derived from the well-known PAM250 matrix) on the potential for constructing tight encodings. Extremely tight encodings are often possible while largely preserving the hydrophobicity of the associated residues. Further, the encodings designed under each of these matrices shows interesting differences between the preferred phase (reading frame), strand, and orientation.
- *Biotechnology Applications of Nested Encodings* – We propose an interesting application for overlapping gene design, namely the interleaving of an antibiotic resistance gene into a target gene inserted into a virus or plasmid for amplification. Selective pressures tend to quickly remove such target genes as disadvantageous to the host. However, coupling such a target with a resistance gene provides a means to select for individuals *containing* the arbitrarily selected target gene.

  To demonstrate the feasibility of this technique, we apply our algorithm to encode each of five important antibiotic resistance genes within the body of the Hepatitis C virus. In fact, we demonstrate there are many possible places to encode each resistance gene within the virus, assuming a sufficiently (but not excessively) permissive codon replacement matrix.

These sequence design problems naturally arise in our project, currently underway, to design and synthesize weakened viral strains to serve as candidate vaccines [6]. This work also follows our previous efforts to design encoding sequences for proteins which minimize or maximize RNA secondary structure [7] and avoid restriction sites [8].

This paper is organized as follows. In Section 2, we survey the literature concerning why gene overlaps occur in nature and how they evolve. We present our algorithm for constructing optimal encodings in Section 3, with associated analysis, and compare our synthesized designs with wildtype viral encodings. In Section 4, we study the impact of alternate codon substitution matrices on the size and parity of minimal pairwise gene encodings. Finally, in Section 5, we present our results on encoding antibiotic resistance genes within viral coding sequences.

2 Overlapping Genes in Nature

Overlapping genes are adjacent genes whose coding regions are at least partly overlapping. They occur most frequently in prokaryotes, bacteriophages, animal viruses and mitochondria, but are seen in higher organisms as well. Gene overlapping presumably results from evolutionary pressure to minimize genome size and maximize encoding capacity [9]. For viruses, this is manifested in two ways; first when genome size substantially affects the speed of replication, and second when an upper bound on the genome size is imposed by packaging.

Overlapping genes are common for viruses with prokaryotic hosts because they must be able to replicate sufficiently fast to keep up with their host cells [10]. As an example, many bacteriophages have compact genomes which maximize coding information into the minimum genome size [10].

In term of evolutionary pressure to minimize genome size, packaging size pressure (the packaging size of the virus particle as the amount of nucleic acid which can be incorporated into the virion) sets the genome size upper bound for viruses with eukaryotic hosts [10].

Overlap between genes is very common in genomes mutating at high rates, such as bacteria and mitochondria, but especially viruses. Although a mutation in an overlapping region can impair more than one protein and would be naturally selected against, there are several reasons overlapping genes can benefit an organism:

- By reducing the size of the genome, without affecting the number of genes encoded.
- By generating new (or sometimes more complex) proteins without increasing the size of the genome.
- By coordinating the expression levels of functionally related genes.
- By coordinating the expression levels of genes, where the expression of one gene requires the deactivation of the other.

The first two functions are supported by the theory of “overprinting”, which attempts to describe the origin of new genes from an existing genome with minimal mutational change [11]. Size reduction is considered important under the assumption that replication rate is inversely related to genome length, since it has an obvious effect in increased rates of replication and minimization of mutation load [12].

Overlapping reading frames can serve to expedite efficient translation. Overlaps can bring translation machinery close to both overlapping genes, which can co-ordinate or co-regulate their expression [12]. In other cases an overlap can bring the termination site of one gene into the same region as the translation initiation site for the next gene [13].

The rate of evolution can be expected to be slower in overlapping genes [14]. Since point mutations in overlapping regions can affect two genes simultaneously, a mutant variant produced with a mutation in an overlapping region will have a lower growth rate and in most cases cannot compete with the wild type variant [15].

Although high mutation rates and selection towards a compacted genome would indicate that overlapping genes should occur mostly in viral and cellular prokaryotic genomes and mitochondria, recent studies show that mammalian genomes have relatively frequent occurrences of overlapping genes too. The observed 774 overlapping genes in the human and 542 overlapping genes in the mouse genome [16] do not compare favorably with the 806 overlapping gene pairs in the genome of E.Coli [9], since the latter genomes is three orders of magnitude smaller. Nevertheless, the same mechanisms of evolution, like rearrangements or loss of parts and utilization of neighboring gene signals, provide explanation for the origin of these overlaps.

Overlapping genes offer an efficient way to study how coding and control sequences have evolved. With direct comparison of the overlapping genes for related species, one can determine how the overlaps evolved and under which conditions, like neighboring gene distance (for example, in closely related bacterial species it has been observed that most of the overlapping genes were generated or degraded in gene pairs that have a short intergenic region [17]). By comparing gene overlaps that are not conserved between related species, the mutational changes that caused the diversion can often be identified. In other cases further species sequencing are necessary to decipher the evolutionary mechanisms and tendencies (see [9] [16] and [17]).

In bacterial species it has been observed that the total number of overlapping genes depends on the genome size or the total number of genes, which could imply that the rates for the accumulation and degradation of overlapping genes are universal among bacterial species [17].

Overlapping gene regions can also provide information for evolution patterns among classes of organisms and seem to converge with ribosomal RNA phylogenetic methods' results [18]. For certain bacterial species, the extent of conservation of unidirectional overlaps correlates with the evolutionary distances between pairs of species [9]. Gene overlaps have even been correlated with certain human disease genes; further genomic rearrangements are likely to occur within overlapping regions, possibly as a consequence of anomalous sequence features prevalent in these regions [19].

3 Finding Maximally Compressed Gene-Pair Encodings

Our algorithm for constructing the maximally compressed encoding for a given pair of amino acid sequences P_1 and P_2 can be most succinctly described via a dynamic program. We consider the canonical case where the encoding of P_1 starts to the left (5' end) of P_2 as shown in Figure 1; the reverse case follows by simply relabeling the proteins. We present only the algorithm for the case of same-strand encodings; the case of alternate strand encodings follows analogously.

Let P_1 contain n residues and P_2 m residues, respectively. Let o_1, o_2, and o_3 denote possible DNA sequences of 0 to 3 bases in length. There are two general cases:

- We say that $C[i, j, o_1, top]$ is *realizable* iff there exists a pair of sequences o_2, o_3 such that o_1o_2 codes for residue $P_1[i]$, o_2o_3 codes for residue $P_2[j]$, and $C[i+1, j, o_3, bottom]$ is realizable or $i = n$.
- We say that $C[i, j, o_1, bottom]$ is *realizable* iff there exists a pair of sequences o_2, o_3 such that o_1o_2 codes for residue $P_2[j]$, o_2o_3 codes for residue $P_1[i]$, and $C[i, j+1, o_3, top]$ is realizable or $j = m$.

An exception occurs only when the residues are aligned, where only one case is needed, in which we advance both indices i and j and we check for reaching both ends of the proteins.

The basis cases for the canonical labeling assert that an overlap is attainable ($C[n, j, o_1, top]$ or $C[i, m, o_1, bottom]$ is realizable) iff $C[j, 1, o_1, top]$ is realizable for some $1 < j < n$.

Since there are only a constant number of possible short strings o_1, o_2, and o_3, it takes constant time to evaluate a given value of $C[i, j, o, b]$ given the solution of all smaller cases. With $\Theta(mn)$ values to evaluate, the algorithm runs in worst case $\Theta(mn)$ time.

By ceasing evaluation once no realizable values remain, the longest overlap can be computed in $O((n+m)l)$, where l is the length of the longest overlap between the protein sequences. Below, we argue that l should in general be of constant length on non-degenerate substitution matrices; this states that on average this algorithm should run in linear time on such matrices.

We say that two overlapping proteins are *in-phase* if the overlap length is congruent to 0 mod 3, i.e. they align along codon boundaries. Non-trivial

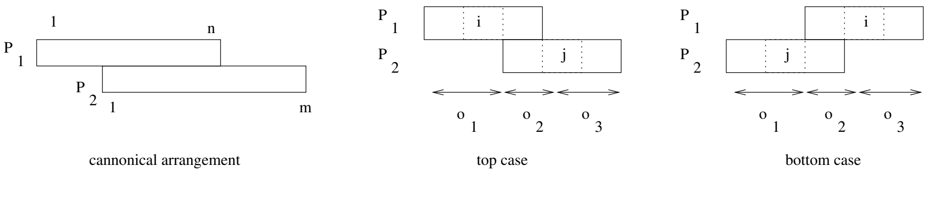

Fig. 1. Notation for the gene encoding algorithm: the canonical encoding (left), with the top (center) and bottom (right) overhang cases

in-phase, same-strand overlapping designs are in principle forbidden by the fact that proteins must end with stop codons. However, we consider an abstraction of this case to simplify the analysis.

Here we consider the expected length of the maximal overlap as a function of the *residue equivalence probability*, defined as the probability that two randomly selected amino acids have an equivalent codon between them. This residue equivalence probability p is a function both of the codon substitution matrix and the distribution of amino acids in the proteins.

Assuming independence of the protein sequences, the expected length of the longest left-right overlap $E(O)$ of two random sequences P_1 and P_2 is given by

$$E_1(O) = \sum_{l=0}^{\infty} l p^l \prod_{i=l+1}^{\infty} (1 - p^i) \tag{1}$$

For the case of two-sided overlaps (i.e. either P_1 or P_2 may occur on the left side of the alignment),

$$E_2(O) = \sum_{l=0}^{\infty} l(2p^l - p^{2l}) \prod_{i=l+1}^{\infty} (1 - (2p^i - p^{2i})) \tag{2}$$

The above analysis demonstrates that the expected maximum overlap length remains quite small until the residue equivalence probability approaches 1. This suggests that two arbitrary proteins are unlikely to permit substantially compressed in-phase encodings except under a forgiving (degenerate) coding matrix.

Still, all is not lost. Our analysis of both wildtype and synthetic overlaps demonstrates that out-of-phase encodings are likely to be substantially longer than in-phase encodings. This phenomenon appears to be difficult to analyze in general because it strongly depends upon the properties of the codon equivalence matrix.

Each amino acid is encoded by a minimum of one and a maximum of six different codons. In total, 61 of the 64 codons encode 20 amino acids while the other three are stop codons, a termination point for protein-synthesizing machinery. Thus there is an approximate 1-to-3 correspondence between amino acids and their codon encodings. It is this redundancy that offers the flexibility in amino acid sequence encoding.

To study the extent of gene overlapping in viruses, we analyzed all 1058 completely sequenced viral genomes available in Genbank as of February 22, 2004. After excluding 273 genomes containing a single annotated gene (and hence not a candidate for gene overlapping) and 108 genomes with sequence ambiguity or obvious annotation errors, we were left with 677 viruses of interest.

In total, these viruses contained 3,232 pairs of overlapping genes, 2,407 of which had overlaps of length greater than four bases.[1]

Figure 2 presents the frequency distribution of gene overlaps by length, the tail of which demonstrates that long overlaps occur with surprisingly high frequency.

[1] Overlaps of less than four bases are not particularly interesting, since the possible overlaps are restricted to the start and stop codons possessed by every gene.

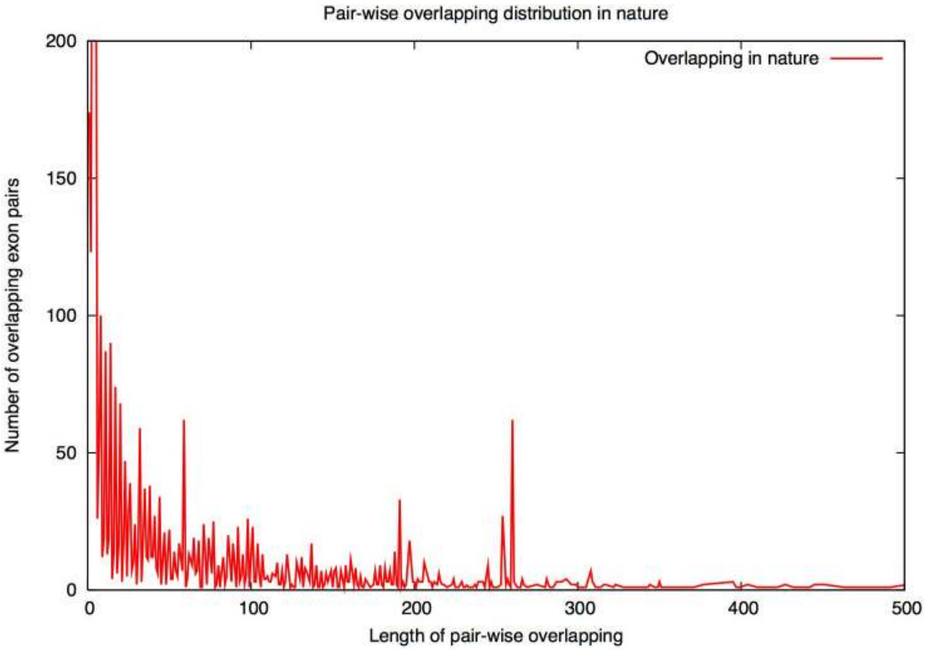

Fig. 2. Length distribution of pairwise-overlapping genes in viral genomes

Table 1 partitions these overlaps into disjoint cases, distinguished by whether the genes occur on the same strand, or are head-to-head or tail-to-tail on opposing strands. Same strand overlaps dominate in the sample. Table 1 also partitions these overlaps by the length mod 3. In-phase overlaps are understandably rare (any stop codon breaks both same strand sequences), but there is also a clear preference for 2 mod 3 parity over 1 mod 3.

Table 1. Parities of natural gene overlaps, ties discarded. All 3232 gene pairs (left). The 2407 gene pairs with overlap > 4.

	All overlaps, parity mod 3					Length > 4, parity mod 3			
Pattern	0	1	2	All		0	1	2	All
SAME	0.0%	23.1%	39.9%	63.0%		0.0%	12.9%	53.3%	66.2%
HH	4.4%	1.9%	4.8%	21.1%		5.9%	4.9%	6.5%	17.3%
TT	3.0%	1.9%	11.0%	15.9%		4.0%	2.6%	9.9%	16.5%
Total	7.4%	36.9%	55.7%	100%		9.9%	20.4%	69.7%	100%

Using our gene pair encoder, we attempted to find more compressed representations of the wildtype gene pairs. In general we failed badly, with the vast majority of cases having zero or insignificant improvement (recall that approximately one third of all natural overlaps were of length 4 or less). In no case were we able to increase the overlap length of such an overlapping gene pair by more than 20 bases.

The lesson here is that gene overlaps occur because the proteins evolved together – significant potential overlaps are extremely unlikely to arise in unrelated sequence pairs because the genetic code does not provide sufficient flexibility. Figure 3 presents the results of optimally encoding 135,869 pairs of unrelated proteins. In no case were we able to reduce the length of an overlapping gene pair by more than 30 bases.

More interesting is the breakdown of our optimized encodings by strand and parity, reported in Table 2. The optimized encodings show sharply different preferences than the wildtype encodings. Functional demands likely constrict the choice of same strand encodings, although it is less obvious why there is such dramatic difference in head-to-head and tail-to-tail preferences. The difference in preferred parity is largely explained by the change in strand encoding distribution.

Table 2. Longest optimized overlap using codon matrix, ties discarded. All 135,869 overlapping gene pairs (left), the 14,925 overlapping gene pairs of length > 4 (right).

	All overlaps, parity mod 3					Length > 4, parity mod 3			
Pattern	0	1	2	All		0	1	2	All
SAME	0.01%	31.92%	1.47%	33.40%		0.05%	3.04%	13.16%	16.25%
HH	5.09%	35.51%	2.31%	42.91%		45.55%	7.77%	21.07%	74.39%
TT	0.08%	0.91%	22.70%	23.69%		0.62%	8.28%	0.46%	9.36%
Total	5.18%	68.34%	26.48%	100.00%		46.22%	19.09%	34.69%	100.00%

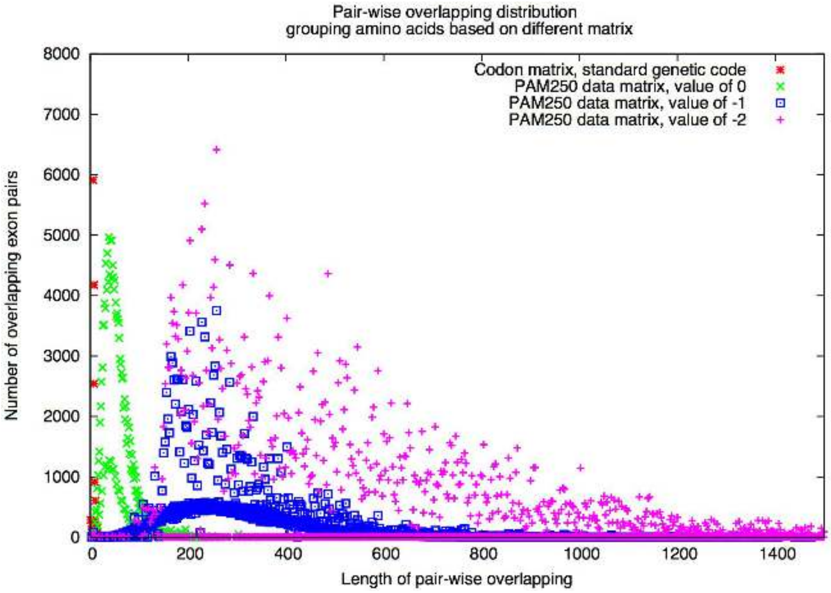

Fig. 3. Distribution of maximum overlaps under four different codon substitution matrices

4 Experiments in Synthetic Gene Encoding

Recent studies [20] have demonstrated that the genetic code maximizes the likelihood that a gene mutation will not harm and may even improve the protein. In general, the code is resilient to random mutations leading to significant changes of the affected amino-acid properties, so that a misread codon often codes for the same amino acid or one with similar biochemical properties. Furthermore, simulations by Gilis et al. [21] have shown that taking the amino-acid frequency into account further increases the resilience of the code compared to random codings. It is also known that proteins with a limited number of point mutations which lead to non-synonymous substitutions fold in similar ways, in a degree that homology database search can detect function similarity in proteins differing in up to 50% of their amino-acid compositions [22].

Based on these results, we decided to further investigate the pairwise gene overlapping possibilities using non-synonymous amino-acid substitution matrices, which increase the combinatorial possibilities of compressed overlapping representations at the cost of minor changes in the residues in the underlying proteins.

Our substitution matrices are derived from the well-known PAM 250 amino acid substitution scoring matrix. The value of each entry describes the reward or penalty in replacing an instance of the first amino acid with the second in aligned sequences. Positive values contribute favorably to an alignment, and negative values unfavorably. We may derive a permissive codon equivalence matrix from PAM 250 as a function of a threshold t by permitting replacement of amino acid x with y if the score is $\geq t$. By decreasing t, we can define a sequence of increasingly permissive substitution matrices for our experiments.

Clearly other substitution matrices are possible (e.g. Levitt's hydrophobicity scoring matrix [23]), and perhaps even preferable. Our primary interest is establishing the flexibility for compressed sequences as a function of more tolerant substitution matrices.

The results of our overlapping experiments with the use of the alternate substitution matrices are shown in Figure 3. One can observe the significant increase in both the number and frequency of long overlaps with increasing length as the matrices become more permissive. In particular, almost arbitrarily long overlaps appear possible under $t \geq -3$ substitution.

5 Hiding Short Genes in Long Genes

Here we report on proof-of-concept simulations of two related biotechnology applications for carefully designed overlapping of synthetic gene sequences:

- *Plasmid incorporation into mammalian cells* – A common technique for incorporating target gene expression into mammalian cell involved plasmid incorporation and mammalian cell transfection. Initially, the plasmid containing the target gene is propagated in bacteria. The naked plasmid DNA is extracted and then introduced into the mammalian cell by transfection.

Typically the target gene is paired with an antibiotic resistance gene, so as to create a marker for selection in the eukaryotic cell. All cells not expressing this marker can be eliminated with the corresponding antibiotic drug (ex. geneticin or G418), to isolate cells expressing the target protein. Sometimes, however only one gene is expressed, such as when the cell fails to incorporate the entire plasmid. Because the plasmid is linearized to be incorporated in a chromosome, the cut may also occur in the target gene location.

By overlapping the target and marker genes, we reduce the probability that either the cut will eliminate the target gene but the not the marker, as well as the probability that the two genes will be separated.

- *Foreign gene incorporation into viruses* – RNA viruses are very prone to recombination, so an added sequence has a high probability to be deleted. Since RNA viruses are streamlined to perform a limited number of specific tasks, the addition of a gene slows down the virus processes, merely by extending slightly its length. Since the foreign gene is undesirable, its deletion will result in a faster produced replicon that will eventually outgrow the engineered virus we implanted.

  Interleaving the target gene into a gene that the virus needs can prohibit its deletion through reversion.

Positive indications in the direction of gene overlap engineering are the recent results of [20], which show that the amino acid code minimizes the effects of mutations and maximizes the likelihood that a gene mutation will improve the resulting protein. Additionally, methods of local sampling (see [24] and [25]) can help us simulate the behavior of slightly altered proteins in respect to folding and docking, so that we can test the codon substitution effects without lab experimentation.

To evaluate the potential for such synthetic overlap encodings, we attempted to find maximal encodings of five important antibiotic genes (whose length ranges from 375 to 1026 nucleotides) within the coding region of the Hepatitis C virus (HCV). Consistent with the results of the previous section, only trivial overlaps can be obtained using synonymous substitutions.

However, multiple complete encodings are possible under $t \geq -2$ and $t \geq -3$ substitution for each of the five antibiotic resistance genes, as reported in Table 3. There is a strong bias for alternate strand encodings, although all five

Table 3. Number of fully-enclosed $t \geq -2$ and $t \geq -3$ encodings of antibiotic resistance genes within the Hepatitis C virus, same strand (SS) and alternate strand (AS)

			$t \geq -2$ encodings		$t \geq -3$ encodings	
Gene	Accession	Length	SS	AA	SS	AS
Hygromycin	X03615	1026	0	1	4	1
Neomycin	M55520	795	0	1	2	3
Puromycin	X92429	600	0	11	16	25
Blasticidin	AYI96214	423	56	250	217	442
Zeocin	A31902	375	35	132	163	175

antibiotic resistance genes offer same strand encodings for $t \geq -3$. In fact, the preferable target for the inserted gene encoding (and promoter region) in the virus application is the minus strand, so this bias appears fortunate.

Based on these results, we are pursing more rigorous designs for intended synthesis and implementation.

Acknowledgments

We thank Eckard Wimmer for his interest and support. We also thank Chen Zhao, Huei-Chi Chen, and Rahul Sinha for discussions and contributions to this research.

References

1. J. Cello, A. Paul, and E. Wimmer. Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template. *Science*, 297:1016–1018, 2002.
2. H. Smith, C. Hutchison, C. Pfannkoch, and J. C. Venter. Generating a synthetic genome by whole genome assembly: phix174 bacteriophage from synthetic oligonucleotides. *Proc. Nat. Acad. Sci.*, 100:15440–15445, 2003.
3. S. Kodumal, K. Pael, R. Reid, H. Menzella, M. Welch, and D. Santi. Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Nat. Acad. Sci.*, 44:15573–15578, 2004.
4. P. Ball. Starting from scratch. *Nature*, 431:624–626, 2004.
5. J. Tian, H. gong, N. Sheng, Z. Zhou, E. Gulari, X. Gao, and G. Church. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, 432: 1050–1054, 2004.
6. S. Skiena and E. Wimmer. Gene design for vaccines and theraputic phages. NSF ITR Award 0325123, 2003.
7. B. Cohen and S. Skiena. Natural selection and algorithmic design of mrna. *J. Computational Biology*, 10:419–432, 2003.
8. S. Skiena. Designing better phages. *Bioinformatics*, 17:253–261, 2001.
9. Y. Fukuda, T. Washio, and M. Tomita. Evolution of overlapping genes: Comparative genomics of mycoplasma genitalium and mycoplasma pneumoniae. The Ninth Workshop on Genome Informatics, 1998.
10. Cann A.J. *Principles of Molecular Virology.* Academic Press, 1993.
11. P. Keese and A. Gibbs. Origins of genes: “big bang” or continuous creation? *Proc. Natl. Acad. Sci.*, 89:9489–9493, 1992.
12. D. C. Krakauer. Evolutionary principles of genomic compression. *Comments on Theor. Biol.*, 2002.
13. D. Oppenheim and C. Yahofsky. Translational coupling during expression of the tryptophan operon of e. coli. *Genetics*, 95:785–795, 1980.
14. T. Miyata and T. Yasunaga. Evolution of overlapping genes. *Nature*, 272:532–535, 1978.
15. D. C. Krakauer. Stability and evolution of overlapping genes. *Evolution*, 54(3): 731–739, 2000.

16. V. Veeramachaneni, W. Makalowski, M. Galdzicki, R. Sood, and I. Makalowska. Mammalian overlapping genes: The comparative method. *Genome Research*, 14:280–286, 2004.
17. Y. Fukuda, Y. Nakayama Y, and M. Tomita. On dynamics of overlapping genes in bacterial genomes. *Gene*, 323:181–7, 2003.
18. I. Rogozin, A. Spiridonov, A. Sorokin, Y. Wolf, J. King, R. Tatusov, and E. Koonin. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, 18(5):228–232, 2002.
19. S. Karlin, C. Chen, A. Gentles, and M. Cleary. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc. Natl. Acad. Sci.*, 99(26):17008–13, 2002.
20. S. Freeland and L. Hurst. Evolution encoded. *Sci Am.*, 290(4):84–91, 2004.
21. D. Gilis, S. Massar, N.J. Cerf, and M. Rooman. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol.*, 2(11), 2001.
22. M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct.*, 29:291–325, 2000.
23. M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104:59–107, 1976.
24. R. Elber and M. Karplus. Enhanced sampling in molecular dynamics: Use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.*, 112:9161–9175, 1990.
25. V. Hornak and C. Simmerling. Generation of accurate protein loop conformations through low-barrier molecular dynamics. *Proteins*, 51:577–590, 2003.