Cite as: SM Carr, HD Marshall, T Wareham, D Craig. 2015. pp 265-274 *in* Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology (QN Tran and H Arabnia, eds.) Elsevier. ISBN: 978-0-12-802508-6

CHAPTER

13

The Big ORF Theory: Algorithmic, Computational, and Approximation Approaches to Open Reading Frames in Short- and Medium-Length dsDNA Sequences

Steven M. Carr[1,2,4], H. Dawn Marshall[1], Todd Wareham[2], and Donald Craig[3]
Department of Biology, Memorial University of Newfoundland, St. John's NL, Canada[1]
Department of Computer Science, Memorial University of Newfoundland, St. John's NL, Canada[2]
eHealth Research Unit (Faculty of Medicine), Memorial University of Newfoundland, St. John's NL, Canada[3]
Terra Nova Genomics, Inc., St. John's NL, Canada[4]

1 INTRODUCTION

The cracking of the genetic code by means of a rapid series of experiments and logical inferences is arguably the first instance of a "big science" approach in the history of molecular genetics (Judson, 1996). Theoretical considerations had already indicated that any nucleic acid code words must comprise a minimum of three letters (Crick, 1966). After it was demonstrated in 1961 that an artificial poly-U RNA template directs incorporation of the amino acid proline into a polypeptide, and thus that UUU was the code for PHE, Marshall Nirenberg's lab had by 1963 deduced an incomplete "dictionary" of 50 three-letter code words (Nirenberg *et al*., 1965), and a substantially complete genetic code table was created by 1965 (Nirenberg *et al*., 1966; also see Figure 13.1). The iconic 4 × 4 × 4 table is now a standard

1st Base	2nd Base U	2nd Base C	2nd Base A	2nd Base G	3rd Base
U	PHE*	SER*	TYR*	CYS*	U
	PHE*	SER*	TYR*	CYS	C
	leu*?	SER	TERM?	cys?	A
	leu*, f-met	SER*	TERM?	TRP*	G
C	leu*	pro*	HIS*	ARG*	U
	leu*	pro*	HIS*	ARG*	C
	leu	PRO*	GLN*	ARG*	A
	LEU	PRO	gln*	arg	G
A	ILE*	THR*	ASN*	SER	U
	ILE*	THR*	ASN*	SER*	C
	ile*	THR*	LYS*	arg*	A
	MET*, F-MET	THR	lys	arg	G
G	VAL*	ALA*	ASP*	GLY*	U
	VAL	ALA*	ASP*	GLY*	C
	VAL*	ALA*	GLU*	GLY*	A
	VAL	ALA	glu	GLY	G

FIGURE 13.1

The genetic code, 1965. Note that uncertainties still existed as to the coding properties of UGA (a TERM or stop codon) and UGG (a Leu codon).

feature of biology textbooks and has been incorporated into bioinformatic computational schemes as a fundamental feature.

In this chapter, we consider properties of short segments of the genetic code that are of interest both theoretically, as unexplored computational challenges, and practically, bearing on the evolution and function of the code and coding molecules. Taken together, the solution of these challenges at the intersection of computational and biological science provides reciprocal illumination to each.

2 MOLECULAR GENETIC AND BIOINFORMATIC CONSIDERATIONS

2.1 MOLECULAR GENETICS OF DNA → RNA → PROTEIN

DNA is famously a double-stranded molecule (dsDNA) that comprises two polymeric sequences of four bases (A, C, G, and T) in an aperiodic order that conveys bioinformation. The two strands are arranged in antiparallel 5'→3' directions that are implicit in the deoxyribose component. The strands are held together by noncovalent hydrogen bonds between paired A + T or C + G base pairs. The antiparallel arrangement and base pairing rules ensure that the alternative strands are complementary to each other. This relationship is the basis of DNA as a self-replicating molecule.

One DNA strand, designated the *template strand*, serves as a template for 5'➔3' synthesis (transcription) of a complementary messenger RNA (mRNA) molecule, where RNA differs from DNA in being single-stranded and substituting base U for T. The mRNA molecule is translated in the 5'➔3' direction into a polymer comprising a sequence of amino acids (a *polypeptide*), according to a genetic code (Figure 13.1). In the code, each of the 64 possible three-letter base sequences (*codons*) reads 5'➔3' and specifies a particular amino acid, except that three codons (UAA, UAG, and UGA) do not specify any amino acid and therefore serve as terminators (known as *stops*) to polypeptide synthesis. A common genetic code is universal for the nuclear genomes of all organisms.

2.2 BIOINFORMATIC DATA-MINING

Because the mRNA sequence is complementary to that of the DNA template strand, it necessarily has the same base sequence in the same 5'➔3' direction as the DNA strand complementary to the template strand, except for the substitution of U for T. This DNA strand, designated the *sense strand*, may therefore be read directly from the genetic code table, substituting T for U. As a bioinformatic process, it is straightforward to read the polypeptide sequence directly from the DNA sense strand, without the intermediate molecular steps of mRNA transcription and subsequent translation via tRNA. (By definition, codons occur only in mRNA: the equivalent three-letter sequences in the DNA sense strand are designated as *triplets*. Hereafter in this discussion, we adopt the National Center for Biotechnology Information (NCBI) bioinformatic convention and use a DNA triplet alphabet.)

Any dsDNA molecule may be read from six potential starting points, designated as *reading frames (RFs),* which are three-base windows that commence at the first, second, or third base from the 5' end of one strand, after which each frame repeats; or from the 5' end of the other strand starting at the opposite end of the molecule. Full-length DNA sequences of several hundred to more than a thousand bases that specify protein sequences that are hundreds of amino acids long are expected to show that only one of these RFs is an Open Reading Frame (ORF); that is, that it does not include a stop triplet over the required length of the polypeptide. As three out of 64 triplets are stops (TAA, TAG, and TGA), the five alternative RFs are expected to include multiple random stops at expected intervals of about 20 triplets: the first occurrence of a stop closes the RF. We designate this the *5&1 condition*. Commercial DNA software programs perform this process as a matter of routine, either from novel data or data mined from online resources such as GenBank.

3 ALGORITHMIC AND PROGRAMMING CONSIDERATIONS

An introduction to the theory of data mining for such ORFs typically begins with the propounding of short dsDNA sequence exemplars of length $L = 15 \sim 25$ base pairs that are constrained by the 5&1 condition. A practical algorithmic generator of such

exemplars must be able to access the entire space of dsDNA sequences that satisfy the 5&1 condition for a specified L, sample that space in an at least approximately random manner, and be efficient in terms of both central processing unit (CPU) run time and required memory space. We developed two such algorithms (Carr *et al*., 2014a), the first based on a two-level recursive search that generates a dsDNA skeleton with at least one stop codon in each of five frames, and then completes the remainder of the dsDNA sequence by adding bases at random to the skeleton so as to produce an ORF exemplar in which the 5&1 condition is maintained. An app that generates dsDNA sequence exemplars that satisfy the 5&1 condition for $L \leq 100$ is available at http://www.ucs.mun.ca/~donald/orf/biocomp/. We provide a more complete discussion of the pedagogical use of the web application in a previous study (Carr *et al*., 2014b). The second algorithm used an exhaustive search that enumerated all those dsDNA sequences of length L that satisfied the 5&1 condition without storing the results as exemplars.

The recursive and exhaustive algorithms show that there are no solutions for $L = 5 \sim 10$, and 96 for $L = 11$ (Figure 13.2, after Carr *et al*., 2014a). Enumerations from

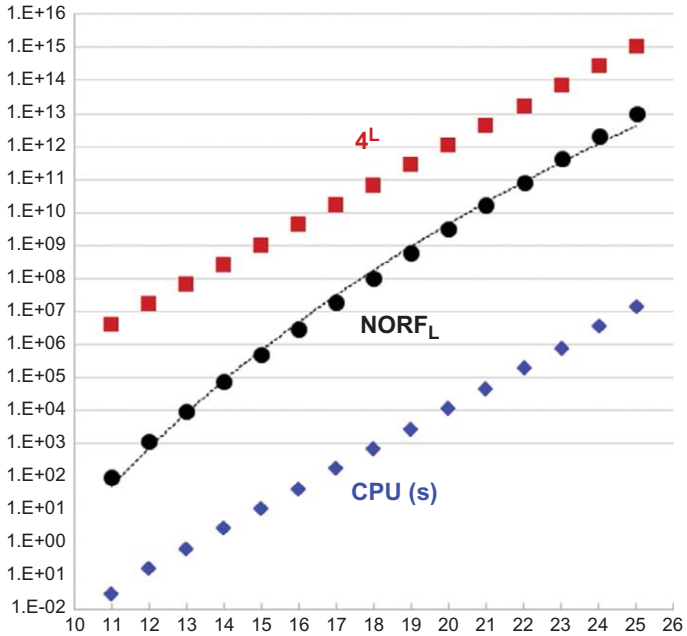


FIGURE 13.2

Semilogarithmic plot of the enumerated number of ORF exemplars of length L ($NORF_L$) for $L = 11 \sim 25$. The total number of possible dsDNA sequences of length L is 4^L (■). Required CPU time for the exhaustive algorithm is given in seconds (◆); CPU is log-linear with respect to L, as $CPU = 0.613(\log_{10} L) - 11.736$ ($r^2 = 0.9998$). After Figure 3 in Carr *et al.* (2014a).

the two methods agree for $11 \leq L \leq 19$, at which point the recursive algorithm succumbs to memory limitations. For $L < 22$, CPU usage for the exhaustive algorithm was measured on a single, quad-core PC. For $L \geq 22$, CPU usage was measured over a network of such machines: by $L = 25$, exact CPU usage is obscured by competing demands from other users on the same network. Calculation of the number of 5&1 solutions for $L > 25$ with the resources available to us would require several days.

4 ANALYTICAL AND RANDOM SAMPLING SOLUTIONS TO L > 25 SEQUENCES: TRIPLET-BASED APPROXIMATIONS

Given these limitations, we have developed a simplified analytical formulation of the 5&1 problem, in which the 64 triplets in the universal genetic code comprise $C = 61$ coding and $S = (64 - C) = 3$ stop triplets. If we disregard the actual nucleotide composition of coding and noncoding triplets and the overlapping nature of the six RFs, the probability that any given triplet is a coding triplet is $C = 61/64$. Next, the probability that a string of T triplets will be an ORF is simply calclulated as

$$p(\text{ORFT}) = C^T, \tag{13.1}$$

and the probability that such a string will include at least one stop is calculated as

$$p(\text{stop}) = 1 - C^T. \tag{13.2}$$

Then, an approximation of the probability that a string of triplets satisfies the 5&1 condition p(NORFT) is the joint probability that RF1 is open *and* RFs 2-5 are all closed, *or* that any of RF2, RF3, … RF6 is open and the other five RFs closed. Thus,

$$p(5\&1T) = (6)\left(C^T\right)\left(1 - C^T\right)^5 \tag{13.3}$$

Figure 13.3 shows a simultaneous plot of Eqs. (13.1), (13.2), and (13.3). Where Eq. (13.3) has a constant factor $K = 6$ and p(stop) enters the function as its fifth power, p(5&1 T)] initially tracks p(stop) toward the enumerable limit of $L = 25$ as observed, but the function maximizes at $T = 37$ ($L = 111$) at $p = 0.4$ of a 5&1 solution.

Thus, and counterintuitively, the scarcity of 5&1 solutions for smaller values of T ($T < 37$, $L < 111$) is determined by the low probability of exactly five simultaneously stopped frames $(1 - C^T)^5$, rather than the relative scarcity of ORFs $(C^T/4^L)$. For larger $T >> 37$, any given ORF is almost certainly accompanied by five frames with multiple stops.

We evaluated Eq. (13.3) as an estimator of p(5&1) by sampling for each of $L = 3 \sim 450$ (mod 3) a set 10^6 random dsDNA sequences, and ascertaining the fraction that satisfied the 5&1 condition under the universal genetic code. Figure 13.4 shows that Eq. (13.3) very slightly overestimates the proportion of 5&1 solutions in the Monte Carlo simulation for $L < 37$. This is to be expected given the absence of constraints in the triplet approximation (triplet assignments, overlapping RFs, etc.), and otherwise the equation provides a close upper bound.

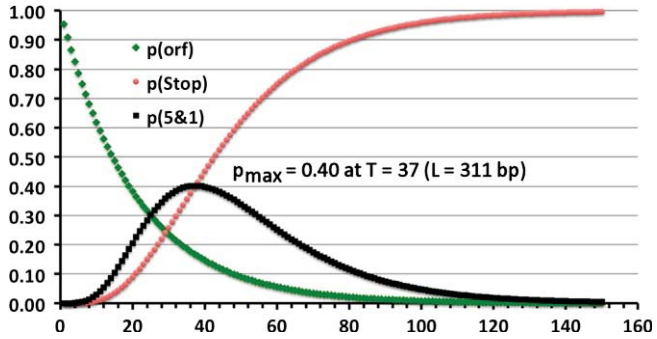


FIGURE 13.3

Plot of the three components of the triplet approximation of the 5&1 solution in Eq. (13.3) {p (5&1)}, for T = 1…150 (L = 3…450). Note that p(ORF) as a simple exponential (C^T) starts high but declines log-linearly toward zero as T increases, and p(stop) = $(1 - C^T)$ starts low but converges on 1.

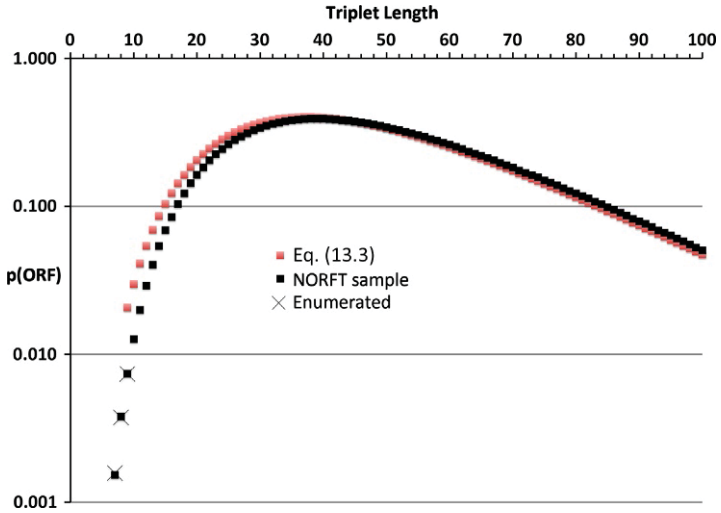
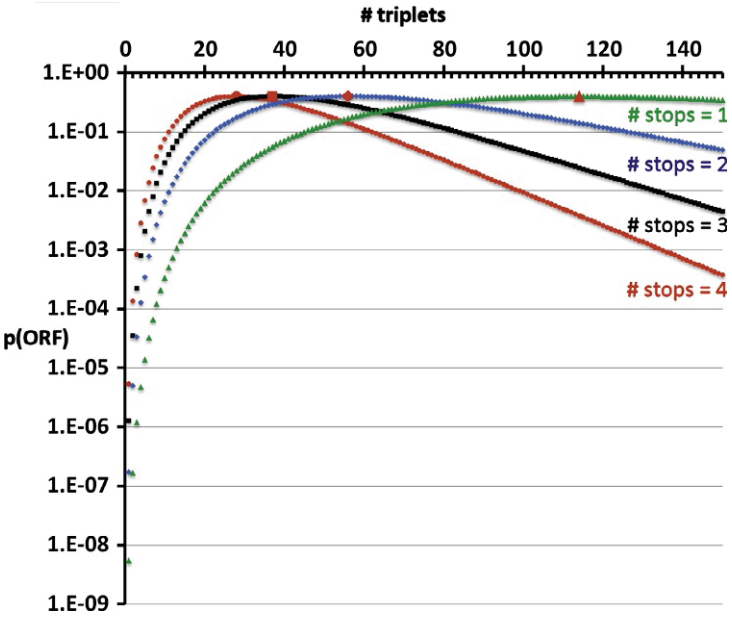


FIGURE 13.4

Probability of a 5&1 solution in triplet strings of length T = 6…100, as estimated by the approximation in Eq. (13.3) (red) and by random sampling of 10^6 dsDNA sequences of length L = 3 T (black). Crosses indicate exact enumerations for T = 6, 7, and 8, corresponding to L = 18, 21, and 24 in Figure 13.2.

5 ALTERNATIVE GENETIC CODES

Besides the universal code with three stop triplets (Figure 13.1), there are several variant codes with one, two, or four stops (Itzkovitz and Alon, 2007). Figure 13.5 shows simultaneous plots of Eq. (13.3) with S = 1, 2, or 4, such that C = 63/64, 62/64, and 60/64, respectively. All variants have the same $p_{max} = 0.40$ as the three-stop code, at L = 342, 168, and 84, respectively. This maximum arises as the zero (horizontal) slope of the first derivative of Eq. (13.3) when $C^T = 1/6$. Substituting this back into Eq. (13.3) gives p = 0.40. Because C is a constant for any one model of the code and co-occurs with T only in the form C^T, the derivative



(a)

T	$p_{low} = 0.10$	$p_{max} = 0.40$	$p_{high} = 0.10$
S = 4	11	28	62
S = 3	15	37	83
S = 2	22	56	126
S = 1	23	114	254

(b)

FIGURE 13.5

(a) Probability function of p(5&1T) for alternative genetic codes with different numbers of stop codons. The universal code has three stops (black), for which a sequence T = 37 triplets (L = 111) has the highest probability ($p_{max} = 0.4$) of providing a 5&1 solution. (b) Upper and lower bounds for p(5&1T) = 0.1, and $p_{max} = 0.4$, for alternative genetic codes with S = 4, 3, 2, and 1 stop triplets, respectively.

of p(NORFT) is necessarily identical for different values of C. The equation for p_{max} can then be rearranged and solved to predict T, as

$$T = \log(1/6) / (\log C) = -0.7782 / (\log C) \quad (13.4)$$

The upper bound on a 10% cutoff for p(5&1) increases rapidly as the number of stops decreases: for example, at the upper 10% probability bound, there are more than three times as many solutions with a one-stop (L = 254, T = 762) as with a three-stop code (L = 83, T = 249).

6 IMPLICATIONS FOR THE EVOLUTION OF ORF SIZE

Broad conservation of the universal code in nuclear genomes indicates that a three-stop code optimizes some selective advantage (Itzkovitz and Alon, 2007), whereas retention of an unstopped TGA in the common ancestor of all Metazoan mitochondrial DNA (mtDNA) codes suggests that there is some advantage for a two-stop code, and the relatively recent evolution of the four-stop code in Chordata offers some advantage over a three-stop intermediate (Cannaozzi and Schneider, 2012). We have shown here that short random DNA sequences have a high probability of including single ORF over certain size ranges, and that this probability is inversely proportional to the number of stops in the genetic codes used. Might size variation of ORF coding sequences across genetic codes be subject to natural selection?

A recent model of stop codon evolution (Johnson *et al.*, 2011) proposes that multi-stop codes provide a backstop against readthrough, balanced against an increased probability of random stop mutations. Like ours, the model predicts an inverse relationship between the number of stop triplets and the length of coding sequences. Consistent with this, their sampling of NCBI data shows a marked (though nonsignificant) relationship between longer coding sequence and fewer stops, for pairs of genomes in the same taxon alternatively decoded with one- versus two-, one- versus three-, or two- versus three-stop codes. There is no such trend for the two- versus four-stop Chordata comparison. Johnson *et al.* (2011) note a previous suggestion that reassignment of TGA from sense to stop has occurred frequently in association with the evolutionary reduction of genome size in mtDNA genomes, in apparent contradiction to the predicted direction. However, a phylogenetic perspective on the various mtDNA codes shows that this reassignment has occurred only once, in the shared ancestral code of all Animalia and Yeast (Ophisthokonta); this will be considered elsewhere.

In their data, coding sequences for genomes with three-stop codes are in the range of 250 ~ 400 bp, with animal mtDNA at about 300 bp: these are rather longer than our optimal of 111 or 84 bp for S = 3 or 4, respectively, but they are well within the range of reasonable probability (Figure 13.5). A longer coding sequence might also be assembled from several shorter single-ORF fragments, so long as the individual ORFs were assembled in the same RF. Recall that fragments shorter than the optimum are more likely to have multiple ORFs. Selection could then act to modify the

function of the corresponding polypeptide product while maintaining a single ORF. DNA sequences 5' or 3' to the ORF region can be added easily, since there is a high probability that any 3-bp sequence added in the open frame also will be open [$\sim(61/64)^3 = 0.87$], while the other five frames are already stopped. The fewer the stops, the longer the likely candidate sequences are. For example, in a four-stop code, there is less than 1% chance that an approximately 300-bp DNA will contain one and only one ORF, whereas for a one-stop code, there is a far greater than 10% chance that a sequence of many hundreds of base pairs will do so.

Are short, random DNA sequences with single ORFs of utility in evolution? It has recently been demonstrated that some free-living bacteria can take up *ex vivo*, fragmented DNA from the environment and incorporate it into their genomes by replication-dependent transformation (Overballe-Petersen *et al*., 2013). Fragments of $20 \sim 100$ bp were most efficiently transformed at higher rates than larger fragments. We have shown that random fragments of just this size are most likely to include a single ORF, which might mediate the success of any such horizontal transfer and its incorporation into the host genome as a functional coding sequence. Overballe-Petersen *et al*., 2013 hypothesize that "rates of molecular evolution in naturally transformable species may be influenced by the diversity of free environmental DNA." Our results suggest that one type of evolutionary diversity in random DNA may be the varying high probability that small fragments of various lengths will include unique ORFs subject to modification by natural selection.

ACKNOWLEDGMENTS

S. M. Carr, H. D. Marshall, and T. Wareham were supported by NSERC Discovery Grants during the preparation of this chapter. We thank K. Tahlan for pointing out the implications for lateral gene transfer in bacteria. We thank H. Arabnia for his leadership of the BioComp conferences, and his support for the preparation of this manuscript. SMC dedicates this chapter to Professor William D. Stansfield of California Polytechnic State University, San Luis Obispo, in recognition of his long service in genetics education.

REFERENCES

Cannaozzi, G., Schneider, A., 2012. Codon Evolution, Mechanisms and Models. Oxford University Press, Oxford.

Carr, S.M., Wareham, H.T., Craig, D., 2014a. An algorithmic and computational approach to open reading frames in short dsDNA sequences: evaluation of "Carr's Conjecture". In: Arabnia, H.R., Tran, Q.-N., Yang, M.Q. (Eds.), Proceedings of the International Conference on Bioinformatics and Computational Biology, pp. 37–43.

Carr, S.M., Craig, D., Wareham, H.T., 2014b. A web application for generation of DNA sequence exemplars with open and closed reading frames in genetics and bioinformatics education. CBE Life Sci. Educ. 13, 373–374.

Crick, F.H.C., 1966. The genetic code, yesterday, today, and tomorrow. Cold Spring Harbor Symp. Quant. Biol. 31, 3–9.

Itzkovitz, S., Alon, U., 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 17, 405–412.

Johnson, L., Cotton, J., Lichtenstein, C., Elgar, G., Nichols, R., Polly, D., Le Comber, S., 2011. Stops making sense: translational trade-offs and stop codon reassignment. BMC Evol. Biol. 11, 227.

Judson, H., 1996. The Eighth Day of Creation, second ed. Cold Spring Laboratories, Cold Spring Harbor, New York.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., O'Neal, C., 1965. RNA codewords and protein synthesis VII. On the general nature of the RNA code. Proc. Natl. Acad. Sci. U. S. A. 53, 1161–1168.

Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., Anderson, F., 1966. The RNA code and protein synthesis. Cold Spring Harbor Symp. Quant. Biol. 31, 11–24.

Overballe-Petersen, S., Harms, K., Orlando, L., Mayar, J., Rasmussen, S., Dahl, T., Rosing, M., Poole, A., Sicheritz-Ponten, T., Brunak, S., Inselmann, S., de Vries, J., Wackernagel, W., Pybus, O.G., Nielsen, R., Johnsen, P., Nielsen, K., Willerslev, E., 2013. Bacterial natural transformation by highly fragmented and damaged DNA. Proc. Natl. Acad. Sci. U. S. A. 110, 19860–19865.