STUDIA UBB. CHEMIA, LV, 4, 2010

CYCLIC NUCLEOTIDE SEQUENCES CODONICALLY INVARIANT UNDER FRAME SHIFTING

VLADIMIR R. ROSENFELD*, DOUGLAS J. KLEIN*

ABSTRACT. A shift of the frame in a polynucleotide sequence typically alters the codon content of the sequence. This provokes a question as to what sequence might be unaltered after shifting the frame. In fact, a linear sequence cannot exactly be so conserved – but there might be a possibility if it is a cyclic code subjected to a circular permutation, as we consider here. The solution is strikingly simple: A cyclic sequence of different nucleotides conserves a circular order of its codons under any shift of its frame if it has a length λ not divisible by 3 and is consecutively read κ times, or it is composed of κ repeated copies of a factor h of length λ, where κ is divisible by 3, while λ is not. For example, the sequence atcgatcgatcg has a factor atcg of length λ = 4 is repeated *κ* = 3 times. Translating this code without any shift gives isoleucine, aspartic acid, arginine, and serine, consecutively, or IDRS for short. The circular shift by 1 position results in SIDR, by 2 positions if produces RSID, and (here) at last, the circular shift by 3 positions gives DRSI. Apparently, all four translated codes of amino acids are the same relative to cyclic permutation. We conclude here discussing the *cyclically invariant* codes by noting that these can easily be enumerated using the famous Pólya's theorem.

Keywords: *nucleotide sequence, codonic, frame shift, cyclically invariant, permutable*

INTRODUCTION

Nucleotide sequences of DNA (desoxyribonucleic acid) and RNA (ribonucleic acid) are constructed from four types of nucleotides denoted by the characters A, C, G, and either T or U, where options T or U are used in cases of DNA or RNA, respectively. According to the complementarity of two strands in DNA, these four characters comprise two complementary pairs: A & T (or U) and C & G. See, *e. g.*, Ch. 13 in [1].

A cyclic shift of the frame in an RNA polynucleotide sequence, in general, alters the resulted sequenced codon content, by which we mean the net number of codons of each different type. This provokes questions as to whether there are codes unaltered after shifting the frame, and, if so, then what codes. Typically, a linear sequence of codons cannot exactly be

* *Mathematical Chemistry Group, Department of Marine Sciences, Texas A&M University at Galveston, 200 Seawolf Parkway, Galveston, TX 77553-1675, USA, rosenfev@tamug.edu, vladimir_rosenfeld@yahoo.com, kleind@tamug.edu*

so conserved – except possibly in certain circumstances. As announced in the title, cyclic sequences of nucleotides having this property do exist. Cutting such a cyclic sequence at an arbitrary position produces a linear factor f which then might be read as a sequence of codons. But with an alternative cut, a shift of the frame by one or two nucleotide positions can under suitable circumstances give the same codon content (*i. e.*, the same counts of acid type of codons -- and thence of each type of translated amino acid). Still one might imagine another scenario where a cyclic RNA is read without cutting, with the reading going round repeatedly – and this under different circumstances can again lead to codon conservation. That is, our considerations are connected with potential ways in which nature might create a kind of 'selfcorrecting code' for amino-acid content, or even codon sequences (up to cyclic permutation), such as to conserve the construction of proteins which are synthesized through translation of codons to amino acids. That is, regardless of the starting point, or reference frame choice for codon translation, the same result would be realized for a codonically invariant cyclic sequence. But also, instead of a cyclic RNA, one may also imagine a linear one having a similarly constructed, periodic factor whose frame shift produces the same circular shift of codons therein and thereby assures the same (apparently circular) order of a translated amino acid sequence.

RESULTS AND DISCUSSION

We begin with a selfevident statement:

Lemma 0. *A periodic cyclic sequence, of the length* $\geq 3,$ *obtained by repetition of just one nucleotide conserves a fixed codonic content which does not alter under any shift of frame.*

Note that both of DNA and RNA normally contain (long) stretches of mononucleotide repeats; besides the conservation of codon content, they may play an important role in base composition and genetic stability of a gene and gene functions, *etc.*. However, it is not yet properly understood -- how nature keeps a fixed-frame reading of general-type codons to reproduce many times the same polypeptide molecules, in organisms. Here, we apply some combinatorial reasoning to comprehend certain details of this complex natural phenomenon.

The first result of this paper is the following statement:

Lemma 1. *Let* f *be a cyclic sequence of nucleotides with a length* $|f|$ *not divisible by* 3 *and with not all nucleotides being the same. Then, there is conservation of a circular order of codons under any shift of frame if* f *is consecutively read* κ *times, where* κ *is divisible by* $3.$

Proof: First, we address the case where the length of nucleotide sequence is not a multiple of 3, say $3k \pm 1$, with k being a positive integer. First, choose a cyclic sequence f of length $|f| = 3k+1$. Starting from an arbitrary fixed point of the cycle, we can traverse $3k$ characters, or k complete codons, and have yet in reserve one spare nucleotide. Continuing cyclically, we utilize this remnant nucleotide as the first. Whence, codons in the second portion of k codons are all passed with the shift of nucleotides by one position to the left, with respect to the distribution into codons in the first $3k$ -nucleotide string. Now, we have two remnant nucleotides, from the right "end" of which now constitute the first two nucleotides of the next codon. Making a third tour now of k more codons along the same sequence of nucleotides produces a sequence of codons which stops at the same point where it was originally begun. That is, we have overall traversed a sequence of $3k+1$ complete codons where all the three possible shifts of the frame have been realized – meaning that the shifts have been made in a circular direction. Apparently, much the same holds true for a factor of length $3k-1$. This completes the proof.

The next statement is related to the preceding one:

Lemma 2. *Let f be a cyclic nucleotide sequence obtained by the κ-fold repetition of a factor h of a length $|h|$, let the nucleotides not all be the same, and let the whole sequence be read just once. Then, there is conservation of a circular order of codons under any shift of frame if κ is a multiple of 3*, while $|h|$ is not.

Proof: First, take a sequence f which is the κ-fold repetition of a factor h of a length $|h|$ not divisible by 3 and with κ being a multiple of three, as in conditions of Lemma 1. Since the tour around such a κ-fold cycle is tantamount to the κ-fold rotation along a cycle of length h (obtained by cyclically closing a factor h), the application of Lemma 1 gives here the proof of the statement.

It is convenient to merge both lemmas to state the following:

Proposition 3. *Let f be cyclic sequence of nucleotides. Then, f conserves a circular order of its codons under any shift of its frame if (0) all the nucleotides are the same, (1) f has a length λ not divisible by 3 and is consecutively read κ times, with κ a multiple of 3, or (2) f is composed of κ repeated copies of a factor h of length λ, where κ is divisible by 3, while λ is not.*

As a case in point, consider the sequence ***atcgatcgatcg***; here, the factor ***atcg*** of length $\lambda = 4$ is repeated three times. Translating this code without any shift gives isoleucine, aspartic acid, arginine, and serine, consecutively, or **IDRS** for short. The circular shift by 1 position results in **SIDR**, by 2 positions produces **RSID**, and (here) at last, the circular shift by 3 positions gives **DRSI**.

Apparently, all the four translated codes of amino acids are the same relative to some circular permutation. Besides codon conservation, the circumstances of Proposition 3 lead to a further consequence:

Proposition 4. *Let f be a cyclic sequence of different nucleotides satisfying one of the conditions of Proposition 3. Then, the cyclic sequence q of amino acids so translated from f is conserved under any cyclic shift of f (with q defined only relative to circular order).*

The Propositions 3 and 4 allow to conclude that a minimal linear factor g of a nucleotide sequence which guarantees to produce, upon translation, the respective factor of the amino acid 'with accuracy to a circular permutation' takes the form $g = acccb,$ where c is a factor of length $|c| = \lambda$ not divisible by $3;$ and prefix a & suffix b factors of a total length $|a|+|b|=2$ $(0=a;b=2)$ correspond to the last and first, consecutive nucleotides of *c*, respectively. The adjective "minimal" stands here to allow circular shifts by 1 or 2 positions. If a (res. b) is a longer factor of c and $|a|+|b| \geq \lambda,$ then g allows a (not minimal) number $|a|+|b|$ of circular permutations of the translated factor $g.$ Accordingly, one or two 'sparse' nucleotides form codons with 2 or 1 external nucleotides, respectively. Codonic nucleotides of the factor g encode a (periodic) factor of the respective amino acid sequence containing a not necessary integer number of repeated translates of $ccc.$ *Combinatorially, just this controls producing circular permutations in a protein domain.*

One might also consider the enumeration of the types of sequences of our Lemma 2, say, with the enumeration at fixed κ & $\lambda.$ That is, we seek the number of equivalence classes of cyclic sequences, where two such sequences are *equivalent* if one can be obtained from the other via a cyclic permutation (*i. e*., a power of the permutation which cycles the members one unit along the sequence, with the last member permuted to the first). We let $\#_{\kappa,\lambda}$ be the number of such equivalence classes consisting of κ segments each of length $\lambda,$ with κ divisible by 3 and λ not. Then:

Proposition 5. *Let $\#_{\kappa,\lambda}$ be the number of equivalence classes of cyclic (nucleotide) sequences having κ segments of length $\lambda.$ Then, $\#_{\kappa,\lambda} = \#_{1,\lambda}$.*

Proof: Each circular shift of an arbitrary (κ,λ)-sequence by one position is equivalent to asynchronous circular shift of every factor of length λ. Such a factor, if considered in a cyclic fashion, represents a $(1,\lambda)$-sequence, so that the number of distinct circular arrangements of nucleotides in both (fixed) (κ,λ)- and $(1,\lambda)$-sequences is the same. Since this is true for any (κ,λ)-sequence separately, it holds true for the entire set S of all circularly nonequivalent (κ,λ)-sequences with the set F of their representative λ-

factors (which are all distinct). With this one-to-one correspondence between the two sets, the proof is completed.

But now we note that $\#_{1,\lambda}$ is solved by Pólya's enumeration theory [2]. Indeed, a problem somewhat like this is a standard enumeration in many combinatorics texts: one ordinarily enumerates equivalence classes of beads on a necklace, with equivalence being determined by the
dihedral group, rather than the cyclic group as here. The additional "reflective" permutations of the dihedral group are absent in our case, since our nucleotide "beads" have a direction (or orientation) along the sequence.

But further, we might clarify a point concerning sequences of types (κ,λ) and (κ',λ') with κ & κ' each divisible by 3 while λ & λ' not. In particular, it can turn out that some sequences can be of both types when there is a fixed total number of nucleotides $\#_{\kappa,\lambda} = \#_{\kappa',\lambda'} \equiv N$. In particular, if *N* has a maximum power $p > 1$ of 3 as a divisor, then $N \equiv 3^p \lambda''$ with λ'' not divisible by 3, and sequences of type (κ'',λ'') (with $\kappa'' = 3^p$, as both κ & κ' are divisible by κ''). Indeed, all the sequences of a type (κ,λ) are again counted in those of type (κ',λ') if κ is a divisor of κ' Thus, we might introduce the count $\#_{\kappa,\lambda}$ of cyclic sequences such that this includes no cyclic sequences of other types. By virtue of the Proposition 5, we can reduce this count to the enumeration of all cyclic $(1,\lambda)$-sequences that are a repetition of no block of length λ' being a divisor of λ. The calculation of $\#'_{\kappa,\lambda}$ first includes determining the numbers $\#_{1,\lambda'}$ for all distinct divisors λ' of λ; and, then, the general inclusion-exclusion procedure applies [2]. Now, one can in fact obtain those counts in terms of the classical Möbius functions [2]. In particular, the number-theoretic Möbius function $\mu(n)$ is defined as follows:

$$\mu(n) := \begin{cases} 0 & \text{if } n \text{ is not square-free;} \\ (-1)^k & \text{if } n = \text{product of } k \text{ distinct primes.} \end{cases} \tag{1}$$

Using the inclusion-exclusion principle, we state the following:

Proposition 6. *Let $\#'_{\kappa,\lambda}$ be the number of cyclic (κ,λ)-sequences such that includes no cyclic sequences of other types. Then,*

$$\#'_{\kappa,\lambda} = \sum_{d|\lambda} \mu(\lambda/d)\#_{(\lambda\kappa/d),d} = \sum_{d|\lambda} \mu(\lambda/d)\#_{1,d} \tag{2}$$

Where the d summation is over divisors of λ.

Proof: The first equality in (2) follows from the general inclusion-exclusion principle applying the number-theoretic Möbius function $\mu(n)$, as this given by (1). The second equality follows from the Proposition 5, which completes the overall proof.

Preceding experimental observations [3–9] of the last 30 years have unequivocally demonstrated the existence of naturally occurring cyclic permutations of the amino acid sequence of a protein. Our present Propositions 3 and 4 determine a sufficient combinatorial condition imposed on respective factors of a nucleotide sequence to guarantee the practical occurrence of this phenomenon.

Concluding, we also mention that in a wider context, which includes also an algebraic simulation of *alternative splicing*, two other cyclic invariances of nucleotide sequences were earlier considered by Propositions 1 and 2 in [10], which do not directly take into account the distribution of nucleotides into codons. Besides, in nature, there are cases of biologically tolerated shuffling of factors of a nucleotide sequence which conserves the inventory of translated amino acids, together with all multiplicities thereof [11]. In other words, there exist also noncircular permutations of a nucleotide sequence conserving the ratios of codonically encoded amino acids (and, maybe, the assortment of codons themselves, without equivalent replacements thereof), whereas a circular order in which they (would normally) follow may be altered. Here, "would" is also used to anticipate a possible perspective of gene engineering which might apply such a principle. Presumably, this may give a new impetus to further interdisciplinary studies of invariant permutable codes, including those which are not cyclically invariant.

ACKNOWLEDGMENTS

The authors acknowledge the support (via grant BD–0894) from the Welch Foundation of Houston, Texas.

REFERENCES

1. J.D.D. Watson, "Molecular Biology of the Gene", 3rd ed., W.A. Benjamin Inc., New York, **1976**.
2. G.H. Hardy and E.M. Wright, An introduction to the Theory of Numbers, Oxford University Press, London, 1938; the 5th edition, 1979.
3. B.A. Cunningham, J.J., Hemperley T.P. Hopp, G.M. Edelman, *Proc. Natl. Acad. Sci. USA*, **1976**, *76*, 3215.
4. M. Hahn, K. Piotukh, R. Borriss, and U. Heinemann, *Proc. Natl. Acad. Sci. USA*, **1994,** *91* (*22*), 10417.
5. Y. Lindqvist and G. Schneider, *Curr. Opin. Struct. Biol., **1997***, ***7*** (*3*), 422.
6. J. Av, M. Hahn, K. Decanniere, K. Piotukh, R. Borriss, and U. Heinemann, *Proteins*, **1998**, *30* (*2*), 155.
7. S. Uliel, A. Fliess, A. Amir, and R. Unger*, Bioinformatics*, **1999**, *15* (*11*), 930.
8. S. Uliel, A. Fliess, and R. Unger, *Protein Engineering*, **2001**, *14* (*8*), 533.
9. J. Weiner 3rd, G. Thomas, and E. Bornberg-Baurer, *Bioinformatics,* **2005**, *21* (*7*), 932.
10. V.R. Rosenfeld, *MATCH Commun. Math. Comput. Chem.*, **2006**, *56* (*2*), 281.
11. E.A. Nalefski and J.J. Falke, *Protein Sci.,* **1996**, *5*, 2375.