J Math Chem (2015) 53:1488–1494
DOI 10.1007/s10910-015-0501-y

ORIGINAL PAPER

Equivalent genomic (proteomic) sequences and semigroups

Vladimir R. Rosenfeld[1,2]

Received: 18 February 2015 / Accepted: 30 March 2015 / Published online: 7 April 2015

Abstract We discuss some combinatorial properties of genomic and proteomic sequences and propose semigroup theory as a versatile algebraic method for their study. In particular, we consider biologically equivalent but not identical sequences and finding hidden regularities therein.

Keywords Genomic and proteomic sequences · Circular permutations · Associativity · Semigroup · Monoid · Group · Equality of words · Homomorphism · Idempotent · Normal subsemigroup · Hidden regularities

1 Introduction

Amidst the constantly changing environment, every organism tries to survive. This includes its striking conservatism to continue produce *almost* identical biopolymer molecules—DNA, RNA, and proteins. Here, the word "almost" needs in a special clarification. As it was generally observed, some segments of such biopolymers allow to substitute them with distinct ones, so that this does not alter the functions of an organism. On the other hand, the replacement of these with very similar segments may bring different biological disturbances. Taking into account both equally important facts, we state herein our mathematical task, as it will be done below.

✉ Vladimir R. Rosenfeld
vladimir_rosenfeld@yahoo.com; rosenfev@tamug.edu

1 Mathematical Chemistry Group, Department of Marine Sciences, Texas A&M University at Galveston, Galveston, TX 77553-1675, USA

2 Present Address: Department of Computer Science and Mathematics, Ariel University, 40700 Ariel, Israel

In a more general context, one may consider not only *substitutions* – but also *insertions* and *deletions* (which combined together give the term *indels*, used in bioinformatics) as well as *permutations* in a biopolymer chain (read also: *genomic* or *proteomic sequence*). Since our sequences are represented by *words* (*strings*) over the alphabet encoding nucleotides or amino acids, the problem of equivalent alternations of a protein or polynucleotide chain is nothing else than the problem of the *equivalence of words* (being *synonyms*). This equivalence, in particular, means the *equality* of respective words as products of factors, e.g., characters, in a certain *semigroup*. That is, equal words here represent equal elements of a semigroup (specially called a *monoid* if contains a unit). Just such a semigroup-theoretical approach is considered herein.

A *nonempty set* $X = \{x_1, x_2, \ldots, x_n\}$ is an arbitrary collection of elements without repetitions, of any nature, such as numbers, amino acids, *etc.*. The *cardinality*, or *size*, $|X|$ of the set X is the number n of elements thereof ($|X| = n$). A function $f(x, y)$ in two variables x and y (where in general $f(x, y) \neq f(y, x)$) implements a *binary operation* on the set X if $\forall x, y \in X$ the value $z = f(x, y)$ also belongs to $X (x, y, z \in X)$. Say, let X be the set of all even numbers and $f(x, y) := x \cdot y$; setting $x = 2$ and $y = 4$, we obtain $2 \cdot 4 = 8$, where 8 is also an even number and, hence, also belongs to X. They say that the set X is *closed* with respect to this operation $(\cdot)$.

A *groupoid* $(X; (\cdot))$ is a nonempty set X with a binary operation $(\cdot)$, which is conditionally called herein the *multiplication*, such that $\forall a, b \in X\, a \cdot b \in X$. Here, we use a simpler notation ab for $a \cdot b$, as usually done in algebra.

A *semigroup* $S = (S; (\cdot))$ is an *associative groupoid* such that $\forall a, b, c \in S\, abc = (ab)c = a(bc)$. An example is again the numbers; say, $3 \cdot 4 \cdot 5 = (3 \cdot 4) \cdot 5 = 3 \cdot (4 \cdot 5) = 12 \cdot 5 = 3 \cdot 20 = 60$. The associativity of S means the ability to arbitrarily partition (by parentheses) the product of elements into factors. They say that the product of elements, in S, does not depend on a distribution of parentheses. Another instance is matrices, when one considers the usual (scalar) product thereof. But the vector product of vectors is not associative! An example of an associative operation from calculus is the composition of two functions $f(x)$ and $\varphi(x)$ defined as $f\varphi \equiv f \cdot \varphi := f[\varphi(x)]$. In our context, the major part is played by the concatenation of words, say when a word $w = w_1 w_2 w_3$ can be obtained by either concatenating $w_1 w_2$ and w_3 or by concatenating w_1 and $w_2 w_3$, where w may represent a polynucleotide or polypeptide chain.

A *unit*, or *neutral element*, e in S is the element such that $\forall a \in S\, ea = ae = a$. In the case of the numbers, clearly, $e = 1$. A semigroup may be without a unit, but one may always deliberately add e to an arbitrary semigroup S, without it, and obtain a semigroup S^1 with a unit, which is specially called a *monoid*. Herein, we work with multiplicative monoids, but use also the more general term "semigroup" to reduce at times the overuse of the term "monoid" and when a statement is true for both words interchangeably. Anyway, a monoid is an instance of the semigroup, such as the monoid of all (res. integer, positive, even, odd, rational, complex) numbers by multiplication. Note that the most familiar specific case of a monoid is a *group* G, where, in particular, each element has its unique inverse, as is the monoid of all rational numbers without a zero (0); say, 2 has $(1/2)$ as its inverse (and *vice versa*), and $2 \cdot (1/2) = (1/2) \cdot 2 = 1 = e$. Recall symmetry groups having symmetry operations as

elements, and where a binary operation produces the result of consecutively performing two symmetry operations. In the case of words representing biological sequences (or whichever else), a role of a unit is played by an empty word (with no character therein).

A semigroup S is *commutative* (*noncommutative* or *anticommutative*), if $\forall a, b \in S\, ab = ba$ ($\exists a, b, c, d \in S$ such that $ab = ba$ but $cd \neq dc$ or $\forall a, b \in S\, ab \neq ba$, respectively). An infinite (anticommutative) semigroup F of words over a finite alphabet A is a *free semigroup*, if all products of characters from A are algebraically unequal therein.

A *word* (or *string*) is a sequence of characters over the alphabet $A = \{a_1, a_2, \ldots\}$ thereof. In the case of proteins, A represents all (characters denoting) amino acids, whose number is 20 or 21, with a triple of stop codons encoding the nonexistent 21st amino acid. Herein, we identify the alphabet A with a semigroup S but without its unit, if any. Since we earlier decided to write down products of elements of S without multiplication signs ($a \cdot b \cdot c \Rightarrow abc$), the products and words have the same algebraic meaning. We specially note the following. Two words are orthographically equally spelled if and only if these are identical (as "abc" and "abc" again), whereas, in an algebraic sense (of the product of elements of S), there may exist infinitely many equal words having distinct spellings (and also lengths).

Now, we turn to more practical actions.

2 The main part

We begin with the following result:

Lemma 1 *Let $M(3 \leq |M| = n)$ be an arbitrary finite monoid, and let w denote an arbitrary (nonempty) word over the alphabet $A = M\backslash\{e\}$. Then, there exist infinitely many words v over A such that vw and wv are orthographically distinct but algebraically equal: $vw = wv$.*

Proof For any pair ab and cd of orthographically distinct words of length 2, words $abcd$ and $cdab$ are also orthographically distinct ($a, b, c, d \in A$). Since there exist $(n-1)^2 > n-1$ nonempty two-character words, there is at least one pair of algebraically equal words, among them; and the same is also true for such words of any length $l \geq 2$. This completes the proof. □

Remark In Lemma 1, w is an arbitrary nonunit element of a finite monoid M. Obviously, in the case of a nonfree infinite monoid M_∞, there also exist infinitely many pairs v and w of orthographically distinct but algebraically equal words ($vw = wv$); however, we cannot *a priori* assert that such pairs exist for an arbitrary w, as in the case of a finite monoid M, in Lemma 1.

Corollary 1.1 *Let $M(3 \leq |M| = n)$ be an arbitrary finite monoid, and let w denote an arbitrary (nonempty) word over the alphabet $A = M\backslash\{e\}$. Also, let σ denote a circular permutation of a word $u = wv$. Then, there exist infinitely many pairs of words u and respective circular permutations σ thereof such that algebraically $\sigma u = u$.*

Proof By virtue of Lemma 1, there exist infinitely many words v such that $wv = vw$. But the last equality also notates a circular permutation of the word $u = wv$ by $|w|$ positions clockwise (where $|w|$ is the length of w). This affords the proof. □

Note that the possibility of equivalent circular permutations of a genomic sequence is by now an experimentally established fact [1–6]. This might also be predicted purely theoretically using Corollary 1.1 or directly Lemma 1. But there exists also a special case in which circular permutations of a nucleotide sequence do not alter a circular sequence of produced amino acids in the respective protein (translated from the same circular sequence of nucleotides with a shift of frame), as was rigorously studied in [7].

The next statement is a generalizing corollary of Lemma 1:

Lemma 2 *Let $M(3 \leq |M| = n)$ be an arbitrary finite monoid, and let w be and arbitrary (nonempty) word over the alphabet $A = M\backslash\{e\}$. Then, there exist infinitely many pairs of orthographically distinct but algebraically equal words $avwb$ and $awvb (avwb = awvb)$, where v is a word over A and $a, b \in M$.*

Proof By virtue of Lemma 1, $\exists\, v, w \in M$ such that $vw = wv$; and, due to the associativity of M, $\forall a, b \in M avwb = a(vw)b = a(wv)b = awvb$. Since the infiniteness of the number of such pairs $avwb$ and $awvb$ is apparent, we arrive at the overall proof. □

Lemma 2 in turn leads to a more general result, *viz.*:

Proposition 3 *Let $M(3 \leq |M| = n)$ be an arbitrary finite monoid. Let $w_j (j \geq 1)$ denote an arbitrary finite word over the alphabet $A = M\backslash\{e\}$. Then, there exist infinitely many pairs $\alpha = a_1 v_1 w_1 a_2 v_2 w_2 \cdots b$ and $\beta = a_1 v_1^* w_1^* a_2 v_2^* w_2^* \cdots b (\alpha = \beta)$ of orthographically distinct but algebraically equal in M words, where a_j, $b \in M$, and $v_j^* w_j^* = v_j w_j$.*

Proof It is due to a repetitive application of Lemma 2. □

This general case is also illustrated by an experimental data [8]. It is worth specially emphasizing that the discussed possibility of equivalent permutations of genomic sequences is a direct consequence of their combinatorial semigroup-theoretical nature.

The overall approach should include orthographically distinct but algebraically equal pairs of words, e.g.,

$$a_1 b_1 a_2 b_2 \cdots a_t b_t = a_1 b_1^* a_2 b_2^* \cdots a_t b_t^*, \tag{1}$$

where $a_j, b_j, b_j^* \in M, b_j \neq b_j^*$, and $j \in [1, t]$. Earlier, we studied [9] a special case with $b_j \in A$ and $b_j^* = e (j \geq 1)$, which corresponds to equivalent insertions/deletions (indels).

Lastly, what is more general than the equality of words is the *equivalence* thereof, according to a certain criterion. Such a criterion may be due to the *homomorphism* (one-valued mapping), if any, of the monoid M onto a group G; accordingly, equivalent elements and equal to them words are those which are preimages of one common

element $g \in G$. Here, say, all *idempotents* (elements a for which $a^2 = a$) of M are the preimages of the unit e of G; in a more general case, they may belong to a *normal subsemigroup N of M* (in notation: $N \trianglelefteq M$), which may locally be defined as a subsemigroup being the preimage of the unit e of G (under a certain homomorphism), while all idempotents of M belong to N. We want to note that, in algebraic sense, the words over the alphabet $B \in G\backslash\{e\}$ are all equal to respective homomorphic images of words over $A = M\backslash\{e\}$ and may also be of certain (independent) interest for researchers (see also below).

3 Discussion

Here, we make a more general theoretical discussion which may give an idea of further realistic applications of semigroups in the context of our paper, though such applications may have not yet been implemented in practice. We realize that the very idea of speaking the semigroup-theory language may often come later than practical approaches emerged in nonmathematical areas, such as chemistry and biology. Part of these approaches may also be semigroup-theoretic in their essence but they have not yet made use of the notion of semigroup.

Equivalent permutations and the other transformations of words are not, as such, the only "linguistic" example of an application of semigroups. As a continuation of the topic, one may consider the recognition of hidden regularities in a sequence. Here, we may carry out an elementary gedanken experiment consisting of two steps.

First, write down the entire alphabet as a word:

$$\underline{w} = abcdefghijklmnopqrstuvwxyz. \tag{2}$$

Which regularities may w have, besides the alphabetical order? For instance, let us imagine that the alphabet is partitioned into nonintersecting subsets of characters such that all characters in each subset have an assigned common value or meaning, in a certain (abstract) sense. Let such subsets be: $X_1 := \{a, e, i, m, q, u, y\}$; $X_2 := \{b, f, j, n, r, v, z\}$; $X_3 := \{c, g, k, o, s, w\}$; $X_4 := \{d, h, l, p, t, x\}$. Since the characters within each subset are equal for us, we rewrite our word using just the first elements of each subset and obtain:

$$\underline{w'} = abcdabcdabcdabcdabcdabcdab, \tag{3}$$

i.e., the substitution of a representative of each subset $X_j (j \in [1, 4])$ for all the other characters thereof, in $\underline{w}$, "developed" (like a photo) a possible hidden regularity in $\underline{w}$. This orthographic transformation of $\underline{w}$ to $\underline{w'}$ is done using mapping of the alphabet into itself, having just four images: a, b, c, d. Of course, this specific regularity devised here is just an imagined artificial trick to illustrate how "erasing differences" among certain characters may similarly develop a more real regularity in practice. Note that "erasing differences" between characters is a familiar method in bioinformatics [10–13] (which also allows using empty boxes for "wobble" characters). In particular, they so study evolutionary mutations of certain factors of genomic sequences, when

it is needed to find a common ancestor of factors whose transformation has made them considerably deviated from a common origin. Apparently, in general, one may consider the equivalence of factors of lengths ≥ 1 in a word. Recall that we considered above equal factors of words – such factors (composed from more than one character) might also be reduced to common spellings thereof, each of which is algebraically equal to any other in a common equivalence class. This may also develop a hidden regularity.

Second, one may consider the semigroup S of all mappings of consecutive characters of the word $\underline{w}$ into $\underline{w}$ itself, obeying a criterion of the preimage-image equivalence of factors (when an image factor equals its preimage one). Since all such mappings, if any, comprise by composition a semigroup, the last may in general contain normal subsemigroups. If at least one proper normal subsemigroup N exists, then there exists also the homomorphism of S onto a group G; and *vice versa*, the existence of such a homomorphism is always associated with the existence of N. Namely, G keeps (representations of) possible symmetry operations related to the hidden symmetry of $\underline{w}$. The task of practically describing S is not so easy in general but, fortunately, even without it, here we know that two nonunit elements of our G correspond to shifts by four characters to the right and left, along $\underline{w}'$, until there will be achieved a respective end thereof. Hence, we can "retrospectively" conclude that S does possess a proper normal subsemigroup. Thus, indeed, theory of semigroups may describe (or even find) hidden regularities, e.g., *periodicity* in a sequence (or 1-dimensional *quasicrystalline order*, in other cases).

As to the very existence of such regularities, the famous theorem of Van der Waerden, Ramsey, and Shirshov (see p. 1 in [14]) states that every infinite sequence of a finite number of symbols contains an arbitrary long periodic subsequence. It is not in general true for finite sequences. But another famous, Sampling Theorem of Whittaker, Nyquist, Shannon, and Kotel'nikov [15] (about the spectrum of a transmitted complex signal) gives, as an indirect corollary, an assertion that the probability to find an (almost) periodic subsequence in a finite sequence grows with the length thereof. Since biological applications presuppose very large semigroups and even infinite ones, knowing a type of such a semigroup may itself play a helpful role (see [9]) in the interdisciplinary research on genomic and proteomic sequences.

4 Conclusions

Using a reduced alphabet of symbols to denote nucleotides (amino acids) may be helpful in discovering regularities. Say, using three digits 1, 2, and 3 to denote fatty amino acids, polar amino acids, and serine, respectively, allowed to carry out a certain original research on the polypeptide sequence of the bacterium *Escherichia coli* [16]. In a broader context, sequences with imagined regularities may also be studied in a similar way [17].

Semigroups may be helpful in two cases. First, semigroups may algebraically describe the orthography of equivalent words (synonyms) and help to study the language to which the words pertain, when the equivalence of words is one important aspect of this. Second, considering a word merely as a "meaningless" sequence of

symbols, theory of semigroups may help to find hidden regularities in such a sequence and describe possible symmetry operations thereon.

Acknowledgments We acknowledge the support of the Welch Foundation of Houston, Texas (through grant BD–0894) and the Ministry of Absorption of the State Israel (through fellowship "Shapiro").

References

1. B.A. Cunningham, J.J. Hemperley, T.P. Hopp, G.M. Edelman, Favin versus concanavalin A: circularly permuted forms of amino acid sequences. Proc. Natl. Acad. Sci. USA **76**, 3215–3222 (1976)
2. M. Hahn, K. Piotukh, R. Borriss, U. Heinemann, Native-like in vivo folding of a circularly permuted jellyroll protein shown by crystal structure analysis. Proc. Natl. Acad. Sci. USA **91**(22), 10417–10421 (1994)
3. Y. Lindqvist, G. Schneider, Circular permutation of natural protein sequences: structural evidence. Curr. Opin. Struct. Biol. **7**(3), 422–427 (1997)
4. J. Ay, M. Hahn, K. Decanniere, K. Piotukh, R. Borriss, U. Heinemann, Crystal structures and properties of de novo circularly permuted 1; 3–1; 4-beta-glucanases. Proteins **30**(2), 155–167 (1998)
5. S. Uliel, A. Fliess, A. Amir, R. Unger, A simple algorithm for detecting circular permutations in proteins. Bioinformatics **15**(11), 930–936 (1999)
6. S. Uliel, A. Fliess, R. Unger, Naturally occurring circular permutations in proteins. Protein Eng. **14**(8), 533–544 (2001)
7. V.R. Rosenfeld, D.J. Klein, Cyclic nucleotide sequences codonically invariant under frame shifting. Studia Univ. Babes-Bolyai Chemia **55**(4), 177–182 (2010). (Available from the authors!)
8. E.A. Nalefski, J.J. Falke, The C2 domain calcium-binding motif: structural and functional diversity. Protein Sci. **5**, 2375–2390 (1996)
9. V.R. Rosenfeld, Using semigroups in modeling of genomic sequences. MATCH Commun. Math. Comput. Chem. **56**(2), 281–290 (2006)
10. Z.M. Frenkel, E.N. Trifonov, Evolutionary networks in the formatted protein sequence space. J. Comput. Biol. **14**(8), 1044–1057 (2007)
11. Z.M. Frenkel, E.N. Trifonov, Walking through the protein sequence space: towards new generation of the homology modeling. PROTEINS: Struct. Funct. Bioinform. **67**, 271–284 (2007)
12. Y. Sobolevsky, Z.M. Frenkel, E.N. Trifonov, Combinations of ancestral modules in proteins. J. Mol. Evol. **65**, 640–650 (2007)
13. Z.M. Frenkel, E.N. Trifonov, Walking through protein sequence space. J. Theor. Biol. **244**, 77–80 (2007)
14. J. Karhumäki, Ramsey theory and related topics, www.math.utu.fi/en/home/karhumak/Ramsey (Lecture Notes), University of Turku, (2004), 65 pp
15. C. Bissel, "The sampling theorem", Communications Engineer, July/July 2007, IET, UK, ISSN 1479–8352
16. V.R. Rosenfeld, Studying the polypeptide sequence (α-code) of Escherichia coli. J. Theor. Chem. (Hindawi), vol. 2013, Article ID 961378, 5 pp
17. V.R. Rosenfeld, Selfcomplementary, selfreverse cyclic nucleotide sequences codonically invariant under frame shifting. J. Math. Chem. **51**(10), 2644–2653 (2013)