MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **56** (2006) 281-290

ISSN 0340 - 6253

Using Semigroups in Modeling of Genomic Sequences

Vladimir R. Rosenfeld

Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

E-mail: vladimir@research.haifa.ac.il

(Received May 26, 2006)

Abstract

We discuss a unified semigroup-theoretical approach to modeling polynucleotide sequences of RNA (ribonucleic acid) and polypeptide sequences of proteins. Primary attention is given to the existence of closed loops in these biopolymers. By way of illustration, some mathematical facts are established and biological interpretation thereof is proposed.

The paper is addressed to biologists, chemists, physicists, and mathematicians that are engaged in interdisciplinary research on RNA and proteins.

1 Introduction

Evolution creates new, diversified genes from ready fragments of already existent ones. Were this done not under selection, the creation of new genes might look like random concatenation of respective genomic subsequences. The concatenation of subsequences is an associative process which is algebraically described by a *semigroup* [1–4]; but only special types of semigroups can simulate the governing selection rules. In order to elaborate a reasonable semigroup model of the process in question, we are going to take here into account known constructional features of pre-mRNA (precursors of messenger RNA) and crystallized proteins [5-8].

Above all, we must consider the existence of intronic loops in pre-mRNA (assuming the shape of a hairpin, clover- or maple-leaf) and like closed loops in crystallized proteins (single, successive, and composite loops containing smaller built-in ones) [5]. Under the semigroup approach that we shall consider below, all these loops will be modeled by (locally) cancelable factors of the strings that encode respective genomic sequences. Of particular interest are factors of the strings which represent *idempotents* of a modeling semigroup. In such a special case, just the presence of complex loops with built-in smaller ones (as well as successive loops) in both pre-mRNA and crystallized proteins unambiguously determines the type of this semigroup—it must be just an *orthodox monoid* [2, 3]. Moreover, the same conclusion can be drawn from a more general fact that genomic sequences possess fractal properties [9].

The next section will introduce some necessarily basic information from the theory of semigroups that will be used later on.

2 Preliminaries

This section epitomizes some known facts from [1–4; 10–15], which will be used by us for deriving special corollaries later on.

A *semigroup* $(S, \cdot)$ is a nonempty set with a binary operation, denoted by $(\cdot)$, satisfying the associativity law: $x\cdot(y\cdot z) = (x\cdot y)\cdot z$ for all $x, y, z \in S$. A semigroup S has an *identity* if there exists an element $1 \in S$ such that $1 \cdot x = x \cdot 1 = x$ for every $x \in S$; the fact that $1 \in S$ is often indicated in notation by writing S^1 in lieu of S. Here, note that 1 can be added to (elements of) any semigroup S, even though S did not originally contain it; usually (see [1], 1.1, p. 4 or [4], Ch. 1, Def. 1.3, p. 2), they use, in lieu of S^1, the notation $S \cup \{1\}$ to stress that S does not contain 1. Regardless of its genesis, a semigroup S^1 is specially termed a *monoid*.

An *idempotent* is an element $e \in S$ such that $e \cdot e = e$; for every element g of a finite semigroup S there is the minimum exponent m, such that g^m is an idempotent.

Since all semigroups that must be considered below are multiplicative ones, we shall hereafter adopt everywhere a simpler dotless notation ab instead of the formerly used $a \cdot b$.

The *order* $|S|$ of a finite semigroup S is the number n of its elements ($|S| = n$). Given a subset A of a semigroup S, there exists the minimum subsemigroup of S containing A. It consists of all finite products of elements of A and is denoted by $\langle A\rangle$ or A^+ if the latter notation is more convenient (see below). In case A has only one element x, we write $\langle x\rangle$ and call this a *cyclic* (or *monogenic*) subsemigroup of S. Any subset A of S such that $\langle A\rangle = S$ is called a *set* (or *system*) *of generators* of S. There exists a minimal set (or system) of relations among elements of S that assures uniquely reconstructing S from A; these are called *generating relations* (or *relators*, for short, to rhyme with ”generators”). Generators and relators represent the *genetic code* of a semigroup S.

Owing to the homonymy above, the molecular biologist may recall that the ‘genetic code’ in his/her own scientific area also provides the set of generators (which are in fact 4 characters coding nucleotides) and generating relations (among the characters) that completely control producing correct protein sequences in nature.

The *order* of an element $g \in S$ is the order $|\langle g\rangle| = q$ of the cyclic subsemigroup $\langle g\rangle$. Here, $\langle g\rangle = \{g^1, g^2, \ldots, g^q\}$ is the set of the first q powers of g, which are all distinct elements of S, whereas every subsequent power g^t $(t > q)$ of g necessarily coincides with the respective element of $\langle g\rangle$. Thus, for an element $g \in S$ of finite order there exists the minimum number r, such that $g^r = g^s$ for some $s > r$; the number r is called the *index* of an element g. Also, there exists the minimum number p such that $g^r = g^{r+p}$; the number p is called the *period* of an element g. The element $e = g^m$ $(m = 0 \,(\mathrm{mod}\ p); m \geq r)$ is an idempotent in the cyclic subsemigroup $\langle g\rangle$. Note that a subset $G = \{g^r, g^{r+1}, \ldots, g^q\} \subseteq \langle g\rangle$ $(q = r + p - 1)$ is a cyclic subgroup of S of order p, with the same idempotent g^m as an identity. The pair (r, p) of the numbers is called the *type* of an element g. For any two naturals r and p there exists a monogenic semigroup of type (r, p). Two finite monogenic semigroups are isomorphic iff (if and only if) their types coincide. Thus, from the standpoint of isomorphism, for every type there is the only finite monogenic semigroup having this type. The latter circumstance is of crucial importance for modeling different processes and objects (see the Introduction, [14] and, especially [4, 11]), when a model should be analogous to what is simulated with it.

An element $a \in S$ is called *regular* if $a \in aSa$, i.e., if there is $x \in S$ such that $a = axa$. Hence, it follows that the elements $e = ax$ and $f = xa$ are idempotents and, moreover, the element e (element f) serves as a *left* (*right*) *unit* for a; if also $e = f$, then a is a *group element* (belonging to a certain subgroup of the semigroup S). Conversely, if an element a has a left (right) unit belonging to the set aS (set Sa), then a is, obviously, regular. An element a is regular iff the main left ideal $\{a\} \cup Sa$ (main right ideal $\{a\} \cup aS$) is generated by some idempotent.

Elements a and b are called *inverse* for one another (or, as a noun, *inverses*) if $aba = a$ and $bab = b$. Every regular element a has at least one inverse A^*. A semigroup is called *regular* if all its elements are regular.

A regular semigroup S is called *inverse* if for each $s \in S$, there is a unique inverse element $t \in S$ such that $sts = s$ and $tst = t$. We shall write this element t as s^* (like in the general case of regular semigroups) or as s^{-1} (like in a more specific case of groups).

A regular semigroup S is called *orthodox* [2, 3] if the set $E(S)$ of all its idempotents is a semigroup. Let $V(a)$ denote the set of all inverses of an element $a \in S$, in S; then for any two elements $a, b \in S$ either $V(a) = V(b)$ or $V(a) \cap V(b) = \emptyset$. This determines an equivalence relation $\sim$ on the elements of S: for $a, b \in S$ $a \sim b$ iff $V(a) = V(b)$. Thus, any orthodox semigroup S can be distributed into a direct sum of all its disjoint subsets $V(a)$. Let further $s \in S$ and $e \in E(S)$, then $ses^* \in E(s)$, one can thus derive nested (built-in) idempotent words from any idempotent ones of S (see Theorem 1.1.9 on p. 9 in [2]).

Of special interest are regular semigroups without skew pairs of idempotents (see [10]). Let S be a regular semigroup with the set of idempotents $E(S)$. Given $x, y \in S$, we say that (x, y) is a *skew pair* if $xy \notin E(S)$ whereas $yx \in E(S)$. A regular semigroup that contains a skew pair of idempotents is not orthodox (see [10], below Def. 1 on p. 265). We shall use just the orthodox monoid as a case of the monoid without skew pairs (see Proposition 2, below).

At this point, we should note that the theory of semigroups has intensively been used in the theory of automata and mathematical linguistics [4, 11]. Part of the terminology in the last two theories came from physicists, logicians, and other specialists [12–14]; this gives a curious mixture sometimes, but it is rather convenient in practice; and we want to introduce some useful notions from it.

An *alphabet* A is a finite set, whose elements are characters; $|A| = d$. Evidently, a case in point can be the 4 characters above that code nucleotides. Throughout this text, A will simultaneously perform the function of the set of generators of a certain finite semigroup $S = \langle A \rangle)$; we shall denote by M the monoid obtained by adding 1 to S ($M = S^1$).

A *word* (over the alphabet A) is a finite sequence $(a_1, a_2, \dots, a_l)$ of letters of A; the integer l is the *length* of the word. In practice, the notation $(a_1, a_2, \dots, a_l)$ is shortened to $a_1 a_2 \cdots a_l$. The *empty* word, which is the unique word of length 0, is denoted by 1 (which is also consistent with the identity above). One says that a word u is a *factor* of a word w if $w = aub$, where a

and b are not necessarily nonempty words.

The set of all words over a finite alphabet A that also includes the empty word is denoted by A^*. (Note that a^* above and A^* here is the usual allowed connotation.) Equipped with the concatenation of words, A^* is the free (multiplicative) monoid on the set A, with the empty word as an identity. A *language* L of A^* is a set of words over A $(L \subset A^*)$. The following property is a necessary attribute of the language describing genomic sequences. Namely, a language L is called *factorial* [12] if $L = F(L)$, where $F(L)$ denotes the set of all the factors of all words of L; evidently, in this case, $M = M(L) = \langle L \rangle \cup \{1\}$.

By definition, a *cancelable* word $ufv = a_1 a_2 \cdots a_k$ $(u, f, v \in L; a_1, \ldots, a_k \in A)$, with a (*locally*) *cancelable* factor $f = a_s a_{s+1} \cdots a_{s+t}$ $(1 < s \leq s + t \leq k)$, is a word for which $ufv = uv$ in M; here, "locally" merely reminds us that, in general, there may also exist other pairs u' and v' $(u', v' \in M)$ for which $u'fv' \neq u'v'$.

Employing the monoid M, in this paper, is needed to us because we shall propose an algebraic model in which all factors of L that encode closed loops in pre-RNA (resp. crystallized proteins) are cancelable, as the respective products of elements in M, factors. This allows us to deduce some 'orthographic rules' for the genomic-sequence language, which can practically be used in the analysis of natural polynucleotide and polypeptide sequences.

Since biological data are insufficient for the exact characterization of M, some model approximations have to be employed. Here, we should recall that the closed loops in pre-RNA and crystallized proteins may be singular and multiple, following one another along the chain; there exist also bigger composite loops containing smaller built-in ones [5] (see [6–8]). Evidently, our model should take into account this whole variety of loops. Since closed-loop subsequences comprise the major part of a genomic sequence, we thereby describe the fractal properties of such a sequence, which was comprehensively studied in [9]. One can readily reformulate the notion of the fractality of a genomic sequence (or just of a closed-loop subsequence) in algebraic terms as follows. Let $u = u_1 u_2$ and $v = v_1 v_2$ be two arbitrary (locally) cancelable factors representing closed loops. Consider four derivative words $w_1 = uv$, $w_2 = vu$, $w_3 = u_1 v u_2$, and $w_4 = v_1 u v_2$. Obviously, according to the biological background, it must be claimed, in this case, that all of the four derivative words w_1, w_2, w_3, and w_4 should be (locally) cancelable factors of respective words, of L, in their own right. Here, we note that the only type of regular monoid that has these same properties for its arbitrary idempotent words, u and v $(u^2 = u$ and

$v^2 = v$ in M), is the orthodox monoid M (see Theorem 1.1.9 on p. 9 in [2] or [3]). Thus, to our model, the orthodox monoid M is of special use. On a wider scale, the derivation of longer (not necessary idempotent) factors of a genomic sequence from shorter ones, in the same fractal manner as we described deriving w_3 and w_4 above, was experimentally confirmed in [9].

We can also discuss additional evidence for M being an orthodox monoid. Let M^k $(k \geq 1)$ locally denote the family of all the n^k k-ary products $u_1 u_2 \cdots u_k$ of elements of M. Since all elements of M^k can be reduced to respective elements of M $(M^k \subseteq M)$, one can introduce the parameter $\mu_k(u)$, which is equal to the share of an element $u \in M$ in M^k. Let M describe some evolutionary process and let the consecutive powers $M^1 = M, M^2, M^3$, *et seq.* of M simulate its successive (elementary) stages; in our context, such stages can be imitated, for example, by reproductive events of the polynucleotide synthesis. In every specific situation, the type of the monoid M can play the principal role. If, for the sake of a gedanken experiment, we assume that M is a group G, we obtain $\forall u, k$ $(u \in G, k \geq 1)$ that $\mu_k(u) = \frac{1}{|G|} = \text{const}$; in other words, the share of any element u, of a group G, in G^k does not depend on k. Apparently, it is impossible to use a group as the model for any evolutionary process; on the contrary, groups are known to be the best language for describing equilibria and invariant actions (such as symmetry operations). Thus, the monoid M cannot be a group. On the other hand, M should be kindred enough to a group because our model simulating long-term evolution should also assure the invariability of any kind of living organisms, for a (large) number of generations. The monoid which is not a group, but which is the most kindred to it, is an inverse monoid. Since any inverse semigroup is an orthodox semigroup, one can legitimately assert that M is, indeed, an orthodox monoid, which also corroborates the reasoning above.

At last, let $w = u_1 f_1 u_2 f_2 \cdots u_t f_t u_{t+1}$ $(u_i, f_j \in L; 1 \leq j < i \leq t+1)$ be a word (string) representing the polypeptide sequence of a crystallized protein (resp. a polynucleotide sequence of a pre-mRNA), where factors $f_1, f_2, \ldots, f_t$ correspond to (all) closed loops of a crystallized protein (resp. pre-mRNA). We are briefly going to employ an algebraic model in which every factor f_j $(1 \leq j \leq t)$ representing a closed loop is a (locally) cancelable factor of w, and conversely. In this model, removing a cancelable factor f of a cancelable word ufv, which produces an algebraically equal abbreviated word uv, simulates a fundamental biological process called *splicing* (*i.e.*, converting pre-mRNA into mRNA). The validity of the 'abbreviating' equality $ufv = uv$ is confirmed by the well-known basic fact that both original pre-mRNA and its successor mRNA store up the same reproductive genetic information, in organisms. It will enable us to determine some properties of the genomic language L, to which we shall turn below.

3 Main results

We begin this section with the following statement.

Proposition 1. *Let a word* $f^t = (a_1a_2\cdots a_s)^t$ $(s \geq 2; \forall t \in \mathbf{N})$ *be a cancelable factor of a longer word* uf^tv *(*$uf^tv = uv$ *in* M*). Moreover, let* $g = a_{\pi 1}a_{\pi 2}\cdots a_{\pi s}$ *be (another) word obtained by an arbitrary circular permutation* π *of characters* $a_1, a_2, \ldots, a_s$ *in* f*. Then* $\forall t \in \mathbf{N}$ g^k $(1 \leq k \leq t-1)$ *is also a cancelable factor of the word* uf^tv*, in* M*.*

Proof. First, recall that both the multiplication in M and concatenation of factors are associative operations. Therefore, for any k, t $(1 \leq k \leq t-1)$ and π we have $uf^tv = ux(g^k)yv = uv$. Here, taking into account the periodicity of the factor f and the definition of the factor g, we obtain that $xy = f^{t-k}$ $(t-k \geq 1)$. Then, according to the conditions of Proposition 1, $uxyv = uf^{t-k}v = uv$. Hence, it follows that $ux(g^k)yv = uxyv$ $(t-k \geq 1)$; consequently, g^k is indeed a cancelable factor of the word uf^tv, which completes the proof. □

Evidently, the number of all orthographically distinct factors g in the conditions of Proposition 1 equals the number of all pairwise incongruous cyclic shifts of a string $a_1a_2\cdots a_s$. In addition, it is clear that all (not necessarily distinct) factors g^k $(1 \leq k \leq t-1)$ can be canceled out in the factor f^t, of uf^tv, in $[s(t-k)+1]$ algebraically equivalent ways.

Apparently, Proposition 1 may be helpful in the context of alternative splicing when one considers introns lying in longer periodic factors of a genomic sequence; in particular, this may propose the first practical rule for compiling the genomic-sequence language L. It is rather easier to detect just one cyclic realization of some loop sequence and purely theoretically deduce from it all the other ”rotated” ones than to find every possible case by chance. Note that he possibility of circular permutations in genomic sequences was experimentally studied in [16].

Now recall that closed loops comprise the major part of a genomic sequence (see [5–8]). Therefore, speaking of the fractality of genomic sequences [9], we legitimately raise a question about the fractality of closed-loop sequences (*i.e.*, cancelable factors). Here, mention that the general model of closed loops above allows one to describe the fractality of cancelable factors provided that M is an orthodox monoid and cancelable factors correspond to idempotents of M. In this connection, we propose also the following result, borrowed from [17].

Proposition 2. *Let a word $e = a_1 a_2 \cdots a_c$ represent an idempotent of an orthodox monoid M. Let further $h = a_{\pi 1} a_{\pi 2} \cdots a_{\pi c}$ be (another) word obtained by an arbitrary circular permutation π of characters $a_1, a_2, \ldots, a_c$ in e. Then h also represents an idempotent of M.*

Proof. As known (see [10], below Def. 1 on p. 265), the statement is true for $c = 2$, when $e = xy$. Consider the next case $c = 3$, when $e = uvw$. Using the associativity of the product uvw, in M, and denoting vw by b, we can reduce this case to the precedent one; therefore, $bu = vwu$ is an idempotent as well. Similarly, denoting uv by a, we obtain that $wa = wuv$ is also an idempotent. Thus, our statement holds good for $c = 3$, too. Since the procedure can further inductively be extended to the case $c = 4$ *et seq*, we arrive at the general proof. □

In addition, note that inverted repeats can be redefined in terms of the involutory operation ($*$), given in the definition of the inverse semigroup above. Indeed, one can set $\mathbf{a} = \mathbf{u}^*$ ($\mathbf{u} = \mathbf{a}^*$) and $\mathbf{c} = \mathbf{g}^*$ ($\mathbf{g} = \mathbf{c}^*$) because $\mathbf{a}$ and $\mathbf{u}$, as well as $\mathbf{c}$ and $\mathbf{g}$, are strongly complementary characters of the alphabet B ($|B| = 4$) of nucleotides, as it takes place in decoding the polynucleotide sequence of DNA. Accordingly, any *inverted repeat* b takes the following form:

$$b = b_1 b_2 \cdots b_s (b_1 b_2 \cdots b_s)^* = b_1 b_2 \cdots b_s b_s^* \cdots b_2^* b_1^*. \tag{1}$$

Were (1) specifically written down for an element of an inverse semigroup, one could also rewrite it as follows

$$b = b_1 b_2 \cdots b_s (b_1 b_2 \cdots b_s)^{-1} = b_1 b_2 \cdots b_s b_s^{-1} \cdots b_2^{-1} b_1^{-1}. \tag{2}$$

However, then, the sequence b in (1) and (2) is an idempotent one, in every inverse (sub-)semigroup to which it belongs; and we do presuppose that M contains an inverse submonoid $M_{\div}$ ($\{1\} \subseteq M_{\div} \subseteq M$). Here, it is necessary to recall that the inverted repeats normally enter into all intronic loops in pre-mRNA. From the semigroup-theoretical point of view, these loops would be treated in perfect analogy to closed loops in crystallized proteins ([5–8]) and to their abiological counterparts, studied in the general polymer chemistry ([18]). This is particularly favorable in the case of interdisciplinary research.

Acknowledgments

I am cordially thankful to Prof. Francisco Torrens (València) and my anonymous reviewers for their constructive remarks.

References

[1] Clifford A. H. and Preston G. B., *The Algebraic Theory of Semigroups*, 2 vols., 2nd ed., Am. Math. Soc., Providence, 1967.

[2] Higgins P. M., *Techniques of Semigroup Theory*, Oxford Univ. Press, Oxford, 1992.

[3] Shevrin L. N., Semigroups, in: *General Algebra* (L. A. Skornyakov, ed.), v. 2, Nauka, Moscow, 1991, p. 11–191 (Russian).

[4] Lallement G., *Semigroups and Combinatorial Applications*, Wiley, New York, 1979.

[5] Berezovsky I. N., Grosberg A. Y., and Trifonov E. N., Closed Loops of Nearly Standard Size: Common Basic Element of Protein Structure, *FEBS Lett.* **466** (2000) 283–286.

[6] Berezovsky I. N. and Trifonov E. N., Van der Waals Locks; Loop-n-Lock Structure of Globular Proteins, *J. Mol. Biol.* **307** (2001) 1419–1426.

[7] Berezovsky I. N. and Trifonov E. N., Loop Fold Nature of Globular Proteins, *Protein Eng.* **14** (2001) 403–407.

[8] Berezovsky I. N., Kirzhner V. M., Kirzhner A., and Trifonov E. N., Protein Folding: Looping from Hydrophobic Nuclei, *Proteins: Struct. Funct. Genet.* **45** (2001) 346–350.

[9] Almirantis Y. and Provata A., An Evolutionary Model for the Origin of Non-Randomness, Long-Range Order and Fractality in the Genome, *BioEssays* **23** (2001) 647–656.

[10] Blyth T. S. and Almeida Santos M. H., Regular Semigroups with Skew Pairs of Idempotents, *Semigroup Forum* **65** (2002) 264–274.

[11] Pin J. E., Finite Semigroups and Recognizable Languages, in: *Semigroups, Formal Languages and Groups* (J. Fountain, ed.), Kluwer, Dordrecht, 1995.

[12] De Luca A. and Varricchio S., Factorial Languages whose Growth Function is Quadratically upper Bounded, *Inf. Process. Lett.* **30** (1989) 283–288.

[13] De Luca A. and Varricchio S., Some Combinatorial Properties of the Thue-Morse Sequence and a Problem in Semigroups, *Theor. Comput. Sci.* **63** (1989) 333–348.

[14] De Luca A. and Varricchio S., A Combinatorial Theorem on p-Power-Free Words and an Application to Semigroups, *RAIRO Inf. Théor.* **24** (1990) 205–228.

[15] Lothaire M., *Combinatorics on Words*, Ser.: Encyclopedia of Mathematics and Its Applications, v. 17, Addison-Wesley, Reading, 1983.

[16] Uliel S., Fliess A., Amir A., and Unger R., A Simple Algorithm for Detecting Circular Permutations in Proteins, *Bioinformatics* **15** (1999) 930–936.

[17] Rosenfeld V. R., Function-Related Linguistics of Noncoding Sequences, *Doctoral Thesis*, The University of Haifa, 2003.

[18] Rosenfeld Vladimir R. and Rosenfeld Victor R., Groupoids and Classification of Polymerization Reactions, in: *The Use of Computers in Spectroscopy and Chemical Research, Novosibirsk, 1983, Theses of the All-Union Conference*, Novosibirsk, 1983, pp. 195–196 (Russian).