MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **57** (2007) 135-142

ISSN 0340 - 6253

Emulating the Function of Introns in Pre-mRNA

Vladimir R. Rosenfeld

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

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

(Received June 30, 2006)

Abstract

We set forth some physical and mathematical views of the function of intronic loops in pre-mRNA (precursors of messenger ribonucleic acid) [1]. The physical reasoning is based on the oscillatory-wave model and leads to the modern ideas of quantum information science. The mathematical simulation employs difference and differential equations. Notice that an independent semigroup-theoretical approach to modeling closed loops in pre-mRNA and crystallized proteins was discussed in [2].

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

1 Physically emulating the function of introns

At any stage of our interdisciplinary discussion, the reader might raise many questions about physical models that could emulate the function of introns in pre-mRNA (precursors of messenger ribonucleic acid). We shall try to elaborate on this by carrying out a series of gedanken experiments.

First, let us imagine a string (or wire) with beads having the same diameter, but their masses may be different. Following our polynucleotide context, one can simply set the number s of sorts (or distinct masses) of beads to be equal to 4 while the four distinct masses m_a, m_c, m_g, and m_u (whose numerical values are not important) thus uniquely correspond to the four types

of nucleotides $\boldsymbol{a}$, $\boldsymbol{c}$, $\boldsymbol{g}$, and $\boldsymbol{u}$; but this restriction imposed on s is not necessary, in general. Let the string be (horizontally) stretched and have its ends fixed in some frame; just in case, we shall need an adequately long string, to avoid any possible boundary effects. Consider a standing oscillatory wave in our system supported by a balanced external action; such a wave, if it resembles a harmonic one, is an alternating sequence of local maxima and minima, where between nearest maxima and minima there are also neutral points whose (vertical) position is unchanged and like that of a nonoscillating string. Here, one should specially concentrate one's attention on the sequences of beads stretched between pairs of equivalent phase points; by definition, a phase distance between any pair of such points is equal to an integer number of periods of the wave. Taking into account that the fixed points of the string, at its ends, are both neutral points, it is very convenient to specify any pair of in-phase neutral points as a *canonical pair* of points. For a string with l bead sites, there are s^l different arrangements of beads (regarding all possible assortments of beads and their order in a sequence, but disregarding right-left symmetry). For ease of our consideration, we shall assume that one and the same external action (as above) will generate a sustained wave in each l-bead string (out of s^l).

Of special interest for us are all complete subsequences of beads that connect canonical pairs of points—that is, subsequences $a_1a_2\cdots a_p$ belonging to segments $\bullet a_1a_2\cdots a_p\bullet$, where the two $\bullet$s symbolize a canonical pair of points in the string. By analogy, call such subsequences between $\bullet$s *canonical substrings*. For further discussion, we need to drop any possible restrictions that might be imposed on the shape of the oscillatory wave. In doing this, one admits irregular distribution of neutral points along the string, which allows consideration of canonical substrings of varied lengths that fit real lengths of introns, in pre-mRNA. In the latter case, the entire string also corresponds to a certain coding word with (separated) cancelable factors, not necessarily of equal lengths [1, 2].

If the beads are threaded on a string in an absolutely random fashion, canonical subsequences can have no preferences in bead assortment and their arrangement, therein. The only restriction is their length which is due to a fixed external action generating a standing wave. That is why we shall consider only nonrandomly threaded strings (resembling linguistic sequences of pre-RNA). The latter case takes into account specific interactions of nucleotides, in pre-RNA, that necessarily imply structural preferences. The inspection of all the s^l beaded strings should give the set Y of all possible canonical substrings. However, in any case, the length of canonical

segments may vary.

Now let a wave propagate along the string. Under this, every canonical pair of points synchronously moves with the wave while the distance between these may alter, which is just an instance of algebraic description involving orthodox semigroups [1, 2]. Our alternative approach [1, 2] is targeted at a universal (semigroup-theoretical) description of all canonical substrings—both in a string with a standing wave and one with a moving wave. In other words, one and the same set Y above should obey the two cases at once. A biological analogy occurs in the case of shifting the frame in a polynucleotide sequence when the triplet code is changed, but the order of nucleotides remains the same.

Let our string with l beads twist and coil to allow it to make self-crossings and form closed loops. In case of nonoscillating string, the sizes of such loops are determined by the stiffness and resilience of the string only. However, if one additionally introduces a standing wave into our system (like that considered in a stretched string above) everything at once becomes rather definite. In particular, a necessary (but not sufficient, in general) condition for self-crossing of the string is that any two points which will thus contact with one another should be just in-phase points. Apparently, contacts can take place just between points of a string that are at the phase distance equaling an integer number of periods of wave oscillation. Thus, we come to a very important inference: Every closed-loop factor is a canonical substring or a consecutive sequence of these [1].

At last, notice that intronic factors have been represented in [1, 2] by cancelable factors of the words that encode pre-mRNA, as well as closed-loop factors in polypeptides. When cancelable factors (read: canonical substrings) correspond to idempotents of an orthodox monoid, where built-in idempotents are allowed, built-in closed loops (as well as consecutive closed loops) can also be represented. Hence, the reader can infer that the semigroup-theoretical modeling of closed loops is quite plausible; applied mathematical findings presented herein may be used for a better understanding of our biological context. But the main inference of our emulation is this: Introns, as spacers, provide a synchronized action of respective points of a genome! Cutting out an intron during splicing does not alter anything at any point of a string which is not (yet) removed; however, the removal of it decreases the number of synchronously acting ”phase” points of genomic sequence [1]

Here, we want to refer to a modern branch which is called *quantum information science*

[3]. Quantum information scientists are fathoming the relation between classical and quantum units of information, the novel ways that quantum information can be processed, and the pivotal importance of a quantum feature called *entanglement*, which entails peculiar connections between different objects. In the perspective, this is especially of interest from the point of view of bioinformatics. A key feature of it is the understanding that groups of two or more quantum objects can have states that are entangled. These entangled states have properties fundamentally unlike anything to be thought of as essentially a new type of physical resource that can be used to perform interesting tasks. The members of an entangled collection of objects do not have their own individual quantum states. Only the group as a whole has a well-defined state. Entangled objects behave as if they were connected with one another no matter how far apart they are—distance does not attenuate entanglement in the slightest. If something is entangled with other objects, a measurement of it simultaneously provides information about its partners.

Within our narrow context, entangled are parts of genome that are responsible for the work of the respective single gene. According to our model considerations, entangled 'in-phase' points of the genome are exactly those that are separated by an intronic closed-loop sequence in pre-mRNA. As well as the closed loops in crystallized proteins [1, 2] (see also [4–8]), intronic loops in pre-mRNA can algebraically been presented by cancelable factors of coding words. The very length of a cancelable factor that connects two 'entangled characters' thus does not matter; however, nonetheless, the inner characters of one cancelable factor may well be entangled, in their own right, with inner characters of another factor (or of a few at once). Therefore, adjusting the length of closed loops (during biosynthesis) in nature may selectively switch on or shut down respective genes. The entanglement exactly through a cancelable segment is, however, only one type thereof (namely, the strongest one). As it had already been discussed above, this type emulates in-phase points in the string with a sustained wave. But there can also be cases with fixing different phase shifts—not just the in-phase synchronization.

A pre-mRNA sequence may possess many classes of in-phase points. Nature knows how to economically edit such idealized sequences in order to obtain more informative ones. Its method, particularly, includes putting the locks between closed loops [4]. Just the locks considerably increase the number of the classes of in-phase genomic points that are entangled and act with the maximum synergism, within each class thereof. Accordingly, the very locks are nothing more than fragments of (incomplete) closed-loop sequences, whose inserting locally

shifts the phase (or frame). In our opinion, this idea is worth further investigation.

Above, we just considered the gedanken experiment where some signal or distant influence propagates along the string with self-crossings. Due to the closed loops attached to the linear part thereof, the signal has two possible ways, in lieu of the only one (as it must be in a purely linear chain, without loops therein), and this is tantamount to the fact that each closed loop acts as a phase shifter with constant phase shift and redistributor of the energy dissipated among different parts of a system. Accordingly, each closed loop of proper length serves as a transducer (resonator/attenuator, band-pass filter, etc.) for a wave that spreads along the linear system to which it is attached. Thus, the loops serve also for the processing and storage of information. The last is due to the induced phase shift that locally emerges in a dynamic information flux.

In order to demonstrate that the closed loops in question can indeed be band-pass filters we propose to consider below the following problem that can rigorously be proven using differential and difference equations [1].

2 The propagation of a wave along the string with attached closed loops

We shall determine the analytical form of the wave that does not alter after it has traversed a given string with attached closed loops [1]. This will also provide a means for solving the converse problem—determining the loop content of the system which a given wave should pass through unchanged.

Let $w(x)$ be a function of a complex (or real) variable x that describes the propagation of a certain wave along a string. The word "string" may denote a chain of atoms linked with chemical bonds and whatever else. We shall assume $w(x)$ to have all the derivatives up to $w^{(m)}(x)$, where m is a natural that will be defined later.

Now consider a chain of atoms with an attached closed loop which is formed due to linking previously unadjacent atoms therein. Let a wave spread from one end to the other. When it meets the first point where the loop is attached (this one is shared by a linear part of the system and by the loop), two further ways are possible: $(*)$ visiting the second contact point of the

loop and, then, traveling directly along the second linear part of the chain; and $(**)$ by-passing around the loop and, then, moving along the second linear segment, which begins from the second contact point of the loop, too. In the former case, consecutively visiting the two contact points needs in traversing just 1 chemical bond, whereas in the latter case, the number of bonds to be traversed is $c-1$, where c is the total number of atoms in the loop which also includes the two contact ones. Thus, the lag l of the second wave with respect to the first one equals $(c-1)-1=c-2$. The resulting wave that finally reaches the second end of the string is additively described by the function $\gamma_1 w(x)+\gamma_2 w(x-l)$, where the coefficients γ_1 and γ_2 take into account partial contributions of the two waves. Since we shall consider just the very principle of such a mathematical description, it will be assumed below that $\gamma_1=\gamma_2=1$; therefore, our model function for the resulting wave that has crossed the region with an attached closed loop will hereafter be described by the function

$$r(x)=w(x)+w(x-l). \tag{1}$$

Note that $w(x-l)=w(x)$ in every point of its domain iff (if and only if) w is a periodic function with the period l; thus, in general, $w(x)\not\equiv r(x)/2$.

Due to the differential-operator formalism, one can rewrite (1) as follows:

$$r(x)=\left[1+\exp\left(-l\frac{d}{dx}\right)\right]w(x), \tag{2}$$

where the operator $\exp\left(-l\frac{d}{dx}\right)$ shifts x by l (which appears with the sign minus).

In the case of s $(s\geq 1)$ consecutive loops formed in a similar way, the resulting function $r(x)$ is given by the following expression:

$$r(x)=\left(\prod_{j=1}^{s}D_j\right)w(x), \tag{3}$$

where l_j is the delay done by the j-th loop and $D_j=\left[1+\exp\left(-l_j\frac{d}{dx}\right)\right]$ $(1\leq j\leq s)$. Note that the order in which the loops follow between the ends of a string are thus inessential.

Here, we can come closer to performing our task concerning a specific situation when the $r(x)$ in (3) remains to be equal to the original function $w(x)$. In this case,

$$r(x)-w(x)=\left[\left(\prod_{j=1}^{s}D_j\right)-1\right]w(x)=\omega w(x)=0, \tag{4}$$

where $\omega=[\cdots]$.

After collecting terms, the differential equation $\omega w(x) = 0$ can be reduced as follows:

$$a_0 \exp\left[-n\frac{d}{dx}\right] w(x) + a_1 \exp\left[-(n-1)\frac{d}{dx}\right] w(x) + \cdots + a_m \exp\left[-(n-m)\frac{d}{dx}\right] w(x) = 0, \tag{5}$$

where $n = \sum_{j=1}^{s} l_j$, $m = n - \min(l_1, l_2, \ldots, l_s)$, and $a_0, a_1, \ldots, a_m$ are nonnegative integers while $a_0 = 1$ and $\sum_{t=0}^{m} a_t = 2^s$ (in general, these coefficients may be arbitrary numbers).

The difference equation equivalent to (5) is:

$$w(x-n) + a_1 w(x-n+1) + \cdots + a_m w(x-n+m) = 0, \tag{6}$$

where all the characters have the same meaning as in (5).

How to solve (6) head-on is not evident; however, it is so for (5). Namely, an elementary solution of (5) is $w(\lambda; x) = \lambda^x$, where λ is an arbitrary root of the polynomial

$$x^n(x^{-n} + a_1 x^{-(n-1)} + \cdots + a_m x^{-(n-m)}) = 1 + a_1 x + \cdots + a_m x^m = 0, \tag{7}$$

with the same coefficients as above.

Thus, in general, any wave which is described by the function

$$w(x) = \sum_{k=1}^{m} b_k \lambda_k^x \tag{8}$$

and only this will pass through our system without any alteration in it, whereas any different wave will necessarily alter. If one sets a different function $r(x)$ for the resulting wave that reaches the second end, a (rather) more complicated, inhomogeneous equation

$$\omega w(x) = r(x) - w(x) \tag{9}$$

has to be solved, which is as yet beyond the present scope.

As an exercise, one may make sure that the linear system with three attached closed loops of lengths 3, 4, and 6 (whence $l_1 = 3 - 2 = 1$, $l_2 = 4 - 2 = 2$, and $l_3 = 6 - 2 = 4$) does correspond to the polynomial (7) of the form

$$1 + x + x^2 + x^3 + x^4 + x^5 + x^6 = 0, \tag{10}$$

whose roots are $\exp\left(\frac{2\pi k i}{7}\right)$ $(1 \le k \le 6; i = \sqrt{-1})$.

Obviously, in general, if $r(x)$ takes the form of the R. H. S. of (8) and the polynomial (7) has all roots equal to respective roots of 1, then the loop content of the system can be determined analytically, moving back from (8) to (3).

Thus, we have demonstrated that intronic loops do operate as phase-shifters and band-pass filters in pre-mRNA. A similar part is also played by closed loops in globular proteins.

ACKNOWLEDGMENTS

We are thankful to our referees for all their remarks.

References

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

[2] Rosenfeld V. R., Using Semigroups in Modeling of Genomic Sequences, *MATCH Commun. Math. Comput. Chem.*, 2006, v. 56, p. 281–290.

[3] Nielsen M. A., Rules for a Complex Quantum World, *Scientific American*, 2002, v. 287, p. 49–57.

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

[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.*, 2000, v. 466, p. 283-286.

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

[7] Berezovsky I. N., Kirzhner V. M., Kirzhner A. Z., Rosenfeld V. R., and Trifonov E. N., Closed Loops: Persistence of the Protein Chain Returns, *Protein Eng.*, 2002, v. 15, p. 955–957.

[8] Berezovsky I. N., Kirzhner A. Z., Kirzhner V. M., Rosenfeld V. R., and Trifonov E. N., Protein Sequences Yield a Proteomic Code, *J. Biomol. Struct. Dynam.*, 2003, v. 21, p. 317–326.