Micro-RNAs: viral genome and robustness of gene expression in the host

BY JACQUES DEMONGEOT1,*, EMMANUEL DROUET2, ADRIEN ELENA1, ANDRÉS MOREIRA3, YASSINE RECHOUM2 AND SYLVAIN SENÉ4

1 Faculty of Medicine, University Joseph Fourier, Grenoble, TIMC-IMAG UMR UJF/CNRS 5525, 38700 La Tronche, France
2 Unit of Virus Host Cell Interactions, UVHCI, UMI 3265 UJF-EMBL-CNRS, 6 rue Jules Horowitz, B.P. 181 38042 Grenoble cedex 9, France
3 Instituto de Sistemas Complejos, Universidad Santa Maria, Artilleria 470, Cerro Artilleria, Valparaiso, Chile
4 LIRIS, INSA Lyon, Avenue Albert Einstein, 69100 Villeurbanne, France

For comparing RNA rings or hairpins with reference or random ring sequences, circular versions of distances and distributions like those of Hamming and Gumbel are needed. We define these circular versions and we apply these new tools to the comparison of RNA relics (such as micro-RNAs and tRNAs) with viral genomes that have coevolved with them. Then we show how robust are the regulation networks incorporating in their boundary micro-RNAs as sources or new feedback loops involving ubiquitous proteins like p53 (which is a micro-RNA transcription factor) or oligopeptides regulating protein translation. Eventually, we propose a new coevolution game between viral and host genomes.

Keywords: micro-RNAs; circular Hamming distance; circular Gumbel distribution; viral genome; robustness in regulatory networks

1. Introduction

A challenge 40 years ago was to give an objective score summarizing the genetic distance between a host (e.g. human) and an infectious agent (e.g. Haemophilus influenzae) in order to predict the latter’s pathogenicity or virulence. In the classical Gatlin diagram (Gatlin 1968), whose variables were DNA redundancy $R$ and GC per cent of genomes, the quadratic distance between the two genomes was a way to compare them, based on their global content in puric and pyrimidic base distribution (figure 1a). Now comparing genomes that come from infectious agents, hosts and vectors is always pertinent, and more sophisticated tools use entropy or circular distances based on the distribution of nucleic bases along DNA (Vinga & Almeida 2004) when all the sequences of their genomes can be used, even when these genomes have a complex architecture (in their information organization or in their topology). This is the case, for example, *

*Author for correspondence (jacques.demongeot@imag.fr).

One contribution of 17 to a Theme Issue ‘From biological and clinical experiments to mathematical models’.

This journal is © 2009 The Royal Society
for the circular DNA of the 4 600 755 bp length chromosome of *Yersinia pestis* (http://cmr.tigr.org/tigr-scripts/CMR/shared/CircularGenomeDisplay.cgi) or for the hepatitis D circular RNA (http://pathmicro.med.sc.edu/virol/hepatitis-virus.htm), also known as the delta agent, more similar to a plant viroid than to a complete virus (figure 2). The main difference with the historical approach performed in the 1960s is that now we can compare chain or ring sequences of RNA or DNA with reference sequences or random rings (figure 1b), with appropriate distances and distribution functions expressing the variability of these distances among a population of given chains or rings.

In figure 1a, the redundancy $R$ (that is, the ability of the genome to repeat pairs of bases) is defined as follows, where $(1 - p)$ denotes GC per cent:

$$R = 1 + p[P_{AU/AU} \log_2 P_{AU/AU} + (1 - P_{AU/AU}) \log_2(1 - P_{AU/AU})]$$

$$+ (1 - p)[P_{GC/AU} \log_2 P_{GC/AU} + (1 - P_{GC/AU}) \log_2(1 - P_{GC/AU})],$$

where $P_{AU/AU}$ (respectively, $P_{GC/AU}$) is the probability to have a base A or U after a base A or U (respectively, G or C).

In this paper, we give the essentials of the mathematical properties of these distances in the case of rings (Demongeot 1978; Demongeot & Besson 1983; Demongeot & Moreira 2007a,b), the work for chains having been extensively published already (e.g. Comet et al. 1999; Bacro & Comet 2000), and later make the comparison between genomes of some coevolving triplets (host, vector and infectious agent) in virology. For example, with data from recent studies (Jopling et al. 2005), we will show how some human (host) or mosquito (vector)
micro-RNAs coming from their UTR (untranslated) genomes fit with the genomes of some viruses (infectious agent), and we discuss a possible coevolution giving this fit as a result of a global game favouring the survival of the three interacting species, each winning (a 'win/win/win game').

2. Distances between rings and chains

If we wish to compare chains of dinucleotides, we could use classical distances between integer vectors, such as that defined by Hamming, but, in the case of rings, the vectors are considered as the same if one is a rotation of the other (Moreira 2003). Let us consider a finite alphabet $A$ and a fixed integer $n$ denoting the length of the rings, described from vectors in $A^n$. We first introduce a notation for the rotation: given $x \in A^n$, such that $x = (x_0, x_1, \ldots, x_{n-1})$, $\sigma^i(x) = (x_i, \ldots, x_{n-1}, x_0, \ldots, x_{i-1})$ is the $i$th circular left-permutation. It is evident that the following properties hold: $\sigma$ is invertible, $\sigma^i(\sigma^j(x)) = \sigma^{i+j}(x)$ and $\sigma^i(x) = \sigma^{i \mod n}(x)$. We define the notion of equivalence under rotation, denoted ‘$\equiv$’, for two vectors $x, y \in A^n$, by

$$x \equiv y \iff \exists k: x = \sigma^k(y).$$

It is easy to see that this is an equivalence relation. Our space of rings will hence be $A^n/\equiv$, the quotient composed of the equivalence classes of the vectors, and a ring will be described as $[x] \in A^n/\equiv$.

(a) Circular Hamming distance

The most usual way to compare vectors with values in a finite alphabet is through the Hamming distance. Given two vectors $x, y \in A^n$, the Hamming distance $d_H(x, y)$ is defined as the number of positions at which the corresponding elements are different. For rings, we consider two rings equivalent if one is a rotation of the other, and the distance between them is the minimum number of rotations needed to make them equivalent.

Figure 2. (a) Circular Y. pestis chromosome, (b) circular hepatitis D RNA and (c) various forms of RNA chains, rings and hairpins.
distance between them is
\[ d_H(x, y) = \# \{ i \in \{0, \ldots, n-1\} : x_i \neq y_i \}. \]

In other words, it is the number of positions in which the values of the vectors differ. The function \( d_H \) is a metric: it is non-negative, symmetric, satisfies the triangle inequality and a null distance implies identity of the vectors. It is also easy to see that, for all \( i \in \{0, \ldots, n-1\} \),
\[ d_H(x, y) = d_H(\sigma^i(x), \sigma^i(y)) \]
and hence
\[ d_H(x, \sigma^i(y)) = d_H(\sigma^{-i}(x), y). \]

Using this last property, we define the circular Hamming distance between two rings \([x]\) and \([y]\) as
\[ d_H^c([x], [y]) = \min d_H(x, \sigma^k(y)), \quad 0 \leq k \leq n-1. \]

In general, the minimum between two metrics is not necessarily a metric, but here it holds.

**Lemma 2.1.** The circular Hamming distance \( d_H^c \) is a metric on \( A^n/\equiv \).

**Proof.** If \( d_H^c([x], [y]) = 0 \), this implies that there exists a \( k \) such that \( d_H(x, \sigma^k(y)) = 0 \); hence
\[ x = \sigma^k(y) \quad \text{and} \quad [x] = [y]. \]

Let us now prove the symmetry:
\[ d_H^c([x], [y]) = \min d_H(x, \sigma^k(y)) = \min d_H(\sigma^{-k}(x), y) \]
\[ = \min d_H(y, \sigma^{-k}(x)) = d_H^c([y], [x]). \]

Let \([x], [y], [z] \in A^n/\equiv\). We must show that the triangle inequality is satisfied, i.e.
\[ d_H^c([x], [y]) \leq d_H^c([x], [z]) + d_H^c([z], [y]). \]

Let \( i \) and \( j \) be such that \( d_H^c([z], [x]) = d_H(z, \sigma^i(x)) \) and \( d_H^c([z], [y]) = d_H(z, \sigma^j(y)) \). In addition, we define
\[ a = \# \{ k : \sigma^i(x)_k \neq \sigma^j(y)_k = z_k \}, \]
\[ b = \# \{ k : \sigma^i(x)_k = \sigma^j(y)_k \neq z_k \}, \]
\[ c = \# \{ k : \sigma^j(y)_k \neq \sigma^i(x)_k = z_k \}, \]
\[ d = \# \{ k : \sigma^i(x)_k \neq \sigma^j(y)_k, \sigma^i(x)_k \neq z_k, \sigma^j(y)_k \neq z_k \}. \]

Then
\[ d_H(\sigma^i(x), \sigma^j(y)) = a + c + d \leq (a + b + d) + (b + c + d) \]
\[ = d_H(\sigma^i(x), z) + d_H(z, \sigma^j(y)), \]
and hence
\[ d_{\Pi}^c([x],[y]) \leq d_{\Pi}(\sigma^i(x),\sigma^j(y)) \leq d_{\Pi}(\sigma^i(x),z) + d_{\Pi}(z,\sigma^j(y)) = d_{\Pi}^c([x],[z]) + d_{\Pi}^c([z],[y]). \]

\[(b)\ Maxsubstrings\ distance\]

We now define another distance measure denoted by \(d_s\), which evaluates the existence of substrings shared by the rings; more precisely, we define \(d_s([x],[y])\) as the difference between \(n\) and the longest length of the substrings present in both rings:
\[ d_s([x],[y]) = n - \max_{i,j} \{ m \in \{0,\ldots,n\} : \sigma^i(x)_k = \sigma^j(y)_k \text{ for } 0 \leq k \leq m \}. \]

It is easy to see that \(d_s\) is a semi-metric in \(A^n/\equiv\). It is not a metric, since the triangle inequality may fail when substrings shared by \([z]\) with \([x]\) and \([y]\) have as intersection two disconnected subchains, i.e. when taken together, the shared substrings cover \(z\), and intersect each other in both of their extremities. For example, the triangle inequality may fail for \(d_s\) if
\[
[x] = [abcd], \quad d_s([x],[y]) = 3,
[y] = [cbdd], \quad d_s([y],[z]) = 1,
[z] = [cbed], \quad d_s([z],[x]) = 1
\]
and
\[ d_s([x],[y]) > d_s([x],[z]) + d_s([z],[y]). \]

**Lemma 2.2.** We have \(d_{\Pi}^c \leq d_s\).

**Proof.** Since \(d_{\Pi}(x,y) = \# \{ i : x_i \neq y_i \}\), we also have \(n - d_{\Pi}(x,y) = \# \{ i : x_i = y_i \}\), and hence we can write \(d_{\Pi}^c\) as
\[ d_{\Pi}^c([x],[y]) = \min_k [n - \# \{ i : x_i = \sigma^k(y)_i \}] = n - \max_k \# \{ i : x_i = \sigma^k(y)_i \}. \]

If the longest substring shared by \([x]\) and \([y]\) is of length \(m\), then we have
\[ \max_k \# \{ i : x_i = \sigma^k(y)_i \} \geq m, \]
and thus
\[ d_{\Pi}^c([x],[y]) = n - \max_k \# \{ i : x_i = \sigma^k(y)_i \} \leq n - m = d_s([x],[y]). \]

\[(c)\ Shuffle\ distance\]

Until now, the two ‘distances’ we have defined can measure some form of similarity between rings, but each of them has advantages as well as disadvantages. The circular Hamming distance \(d_{\Pi}^c\) measures similarities between rings, but ignores their order. If we apply a permutation to both rings, the distance would not change. Hence, in a scenario of rings cut into pieces, which are shuffled and then come together to build new rings, \(d_{\Pi}^c\) will not capture much of what happens with those substrings. On the other hand, \(n - d_s\) measures the size of the longest common substrings between rings, but does not tell us anything about the other sequences.
Hence, in order to capture another aspect of the idea of similarity in which we are interested, we introduce a third function, $d_t$. This function will be finite only for pairs of rings $[x], [y]$ that use the same amount of each kind of letter in $A$, i.e. such that $d_t(x, \alpha \ldots \alpha) = d_t(y, \alpha \ldots \alpha)$, for all $\alpha$ in $A$, where $\alpha \ldots \alpha$ is the sequence of $A^n$ made by the concatenation of $n \alpha$; it will be $\infty$ otherwise. In a finite case, we define $d_t([x],[y])$ as the minimum number of cuts to be made in $[x]$ so that, after reordering the resulting pieces, we may obtain $[y]$.

Lemma 2.3. The shuffle distance $d_t$ is a metric on $A^n/\equiv$.

Proof. If $d_t([x],[y]) = 0$, then no cut is necessary, and the rings must be identical. Symmetry is easy to see, since the pieces used to go in opposite directions are the same. Finally, for the triangle inequality, $d_t([x],[y]) \leq d_t([x],[z]) + d_t([z],[y])$, we cut $[x]$ in the optimal way to build $[z]$, and then in addition we make the cuts needed to build $[y]$ out of $[z]$. In this way, we pass from $[x]$ to $[y]$ with $d_t([x],[z]) + d_t([z],[y])$ cuts; this may not be the optimal way of going from $[x]$ to $[y]$, but it provides an upper bound for $d_t([x],[y])$, proving the inequality. The previous argument holds for the case where all values are finite; if the left-hand side of the inequality is infinite, then letter usage is different in $[x]$ and $[y]$, and since they cannot both share the letter usage of $[z]$, the right side will be infinite too. \[\blacksquare\]

(d) The semi-metric $d_t^*$

For speeding up the computation, the ‘distance’ we eventually propose will be an approximation of $d_t$. Given two rings $[x]$ and $[y]$, we remove from both of them one of the longest substrings they share, leaving two words $x'$ and $y'$. Then, we initialize two lists of words; let us denote by $x^{(k)}$ the set of (one or two) subwords left in the words of $x^{(k-1)}$ by removing one of the longest substrings common with the words of $y^{(k-1)}$; then these two lists are $P_x = \{x', x'', x^{(3)}, \ldots\}$ and $P_y = \{y', y'', y^{(3)}, \ldots\}$. At each time step, the lists contain a family of non-overlapping substrings of $[x]$ and $[y]$, respectively. More precisely, at each iteration $k$, the algorithm finds the longest substrings between two words, taken from each list at the same level $x^{(k)}$ and $y^{(k)}$ (i.e. maximizing over all possible pairs between these words from $x^{(k)}$ and $y^{(k)}$), removes one of these substrings from these words, and returns the remaining words to the respective lists. We define $d_t^*$ as the number $N$ of iterations of the algorithm until the word sets $x^{(N)}$ and $y^{(N)}$ are empty. It is easy to see why we call this function $d_t^*$: it represents the same idea as $d_t$, cutting the sequences into the required number of pieces in order to obtain one by reassembling the pieces of the other and reciprocally. The function $d_t^*$ is a semi-metric (the triangle inequality may fail).

3. Circular Hamming distribution and circular Gumbel distribution

If one of the sequences to compare is a fixed chain $x$, the other being a random ring $[y]$, both being of length $n$, let us denote by $M$ the random variable equal to the number of matches between them; we have $M = n - \min_{k=1, \ldots, n} d_t(x, \sigma^k(y))$, where $\sigma^k(y)$ is the chain obtained by opening $y$ at the letter of phase $k$. We will call circular Hamming distribution the probability law of $M$. The expected
number of matches $E(M)$ in the case of the comparison of an RNA chain with a reference RNA ring each having for example 22 bases is less than the maximum number of matches observed in the case of comparison with 22 independent chains of length 22, because a change in the origin of phases on the ring does not correspond strictly to a new chain tossing. Then we can write

$$P(M < k) > P\left(\bigcap_{i=1,...,22} (X_i < k)\right),$$

where the $X_i$ are independent and identically distributed (i.i.d.) random variables, having as common distribution the binomial law $B(22,1/4)$, i.e. the distribution of a binomial variable $X$ equal to the number of matches between the given RNA chain and a random reference RNA chain of the same length (we suppose that the occurrence of each base A, U, G, C has the probability 1/4). By exploiting the binomial histogram (figure 3), we obtain

$$P(M < 15) > P(X \leq 14)^{22} \approx 1,$$
$$P(M < 14) > P(X \leq 13)^{22} = (0.9999)^{22} = 0.998 \approx 1 - 22 \times 0.0001 = 0.998,$$
$$P(M < 13) > P(X \leq 12)^{22} = (0.9993)^{22} = 0.985 \approx 1 - 22 \times 0.0007 = 0.985,$$
$$P(M < 12) > P(X \leq 11)^{22} = (0.997)^{22} = 0.936 \approx 1 - 22 \times 0.003 = 0.934,$$
$$P(M < 11) > P(X \leq 10)^{22} = (0.99)^{22} = 0.802,$$
$$P(M < 10) > P(X \leq 9)^{22} = (0.97)^{22} = 0.512,$$
$$P(M < 9) > (P(X \leq 8))^{22} = (0.925)^{22} = 0.180,$$
$$P(M < 8) > (P(X \leq 7))^{22} = (0.839)^{22} = 0.021,$$
$$P(M < 7) > (P(X \leq 6))^{22} = (0.699)^{22} = 4 \times 10^{-4}.$$

Hence, we have

$$E(M) = \sum_{i=0,...,22} P(M \geq k) = \sum_{i=1,...,23} (1 - P(M < k))$$
$$= 23 - \sum_{i=1,...,23} P(M < k) < 23 - \sum_{i=0,...,22} P(X \leq k)^{22} \approx 9.6.$$
Let us note that this result is in agreement with the inequality whose proof is reported by Hill & Kertz (1981), which gives a majorant equal to 11. \( E(M) \) is also of course strictly larger than the expected number in the case of comparison with only one reference random chain, i.e. \( 22/4 = 5.5 \), hence \( E(M) \) lies in the interval \([6, 10]\).

The observed empirical mean (see §5) in the numerical experiments shows a value near 9.5, i.e. about the value of the expectation of the supremum of 22 binomial variables \( B(22,1/4) \). This observation suggests a conjecture: the distribution of \( M \) is in general a convex compromise between the binomial law of \( X \), the \( \sup_{i=1,...,n} X_i \) distribution and the Dirac distribution located on the singleton \{22\} (with weights to determine). The extremal distributions can be obtained in the following circumstances. If the length of the reference random ring is going to infinity, the length of the given RNA remaining finite equal to 22, \( E(M) \) tends to be equal to the binomial expectation 5.5. If, on the contrary, the length of the given RNA tends to infinity as the length of the reference random ring remains fixed to 22, the perfect fit is asymptotically observed and \( E(M) \) tends to 22. If both lengths remain the same, equal to \( n \) and if \( n \) tends to infinity, we observe the \( \sup_{i=1,...,n} X_i \) distribution, whose expectation is about 9.6, if \( n = 22 \). This last case is observed in our example. If \( n \) is small, the bias observed in simulations with respect to the \( \sup_{i=1,...,n} X_i \) distribution is because of the relatively weak number \( A_n \) of aperiodic rings (i.e. rings each of whose circular permutation is different from the others) among the \( R_n \) possible rings (Ruskey & Sawada 2000):

\[
A_n = \sum_{d \text{ prime number divisor of } n} \mu(n/d)4^{d/n}
\]

and

\[
R_n = \sum_{d \text{ prime number divisor of } n} \phi(n/d)4^{d/n},
\]

where \( \mu \) and \( \phi \) are, respectively, the Möbius and the Euler functions. For example, we have for rings of \( n \) nucleotides having only two states (puric and pyrimidic): \( A_8 = 30 \) and \( R_8 = 36 \), but \( A_{22} = 190\,557 \) and \( R_{22} = 190\,746 \), which shows the reduction of the bias when \( n \) increases.

We will call \textit{circular Gumbel distribution} the probability distribution of the random variable defined by \((M - E(M))/\sigma(M)\), where \( \sigma(M) \) is the standard deviation (s.d.) of \( M \). This quantity is random, but partially independent of the length (here 22) of the reference RNA ring. For a ‘circular’ \( Z \)-score it could play the same role as the ‘classical’ Gumbel distribution for the ‘classical’ \( Z \)-score (Gumbel 1958; Comet et al. 1999). By using an upper bound of the large deviations of this distribution given by the supremum of binomial variables, we can show for example the significance (at the threshold of 2.5%) of the fit between specific chains (200 siRNAs from http://www.rnainterference.org/HumanSequences.html) and a reference ring called AL (cf. §5). The circular Gumbel distribution can be estimated using a von Mises–Tychonov kernel (Shmalyi 2005).
4. RNA relics

The RNA relics (essentially tRNA loops, siRNAs and micro-RNAs) are made of short sequences (length of about 20 bases) having the same function in many realms (viral, bacterial, vegetal, animal) and a weak interspecific variability as for the genetic code, which is universal (Eigen 1971; Labouygues 1976; Hopfield 1978; Trifonov & Sussman 1980; Eigen et al. 1981; Figureau & Pouzet 1984; Hartman 1984; Swanson 1984; Hobish et al. 1995; Szathmary & Maynard Smith 1997; Trinquier & Sanejouand 1998; Yarus 2000; de Duve 2002; Hornos et al. 2004; Wang & Schultz 2005; Wang et al. 2006). This is for example the case for the tRNA loops, which are highly invariant between species and amino acids, and it has been recently discovered that it also holds for micro-RNAs, which are small sequences of mean length 22 (see §7), present in the non-coding regions of many known genomes (especially of plants and animals), whose maturation (figure 4) process allows the interaction with mRNAs, preventing in general their translation in ribosomes. These micro-RNAs are particularly useful as cancer biomarkers (Calin et al. 2004) and could also be used in infectious diseases for predicting the pathogenicity of the infectious agents.

During the first step of the maturation process, the micro-RNAs (miRs) have a hairpin structure (http://protein3d.ucifcrf.gov/shuyun/Web/talk/Talk04.pdf), and both bioinformatics approaches and direct cloning methods have identified many such miRs, including orthologues from various species: the repository miRBase (http://microrna.sanger.ac.uk) contains over 5000 annotated miRs, including numerous human miR genes. Many miRs are ubiquitously expressed, whereas others are expressed in a cell-type specific manner. Because a single miR can target transcripts from multiple genes and, conversely, several miRs can control a single target (Krek et al. 2005), the miRs and their targets function as a complex regulatory network. We take advantage of the complete sequencing of vectors like Anopheles gambiae (Holt et al. 2002; Hill et al. 2005) and Aedes aegypti (Nene et al. 2007) and also use the 5′-untranslated region (5′-UTR) part of viral RNAs, like a typical isolated mRNA of Hepacivirus, hepatitis C virus (HCV), a 341-nucleotide sequence containing an internal ribosome entry site (IRES) required for the initiation of translation. It is fully admitted that the 5′- and 3′-UTRs may play a role in the initiation of negative-strand synthesis of viral RNAs released from entering virions, switching from negative-strand synthesis to synthesis of progeny plus strand RNA at late times after infection, and finally in the initiation of translation and in the packaging of virus plus strand RNA into particles (Markoff 2004). Until recently very little was known about regulation of Flavivirus RNA replication and translation, in particular via the RNA interference machinery (Bartenschlager et al. 2004), but in Jopling et al. (2005) a human liver-specific miR (miR-122) enhances intracellular levels of HCV RNAs, and a recent work noted that this miR was likely to facilitate replication of the viral RNA (Appel & Bartenschlager 2006). By searching matches between miRs and viral genomes, we also discovered that a dozen miRs had a conserved coincidence in all four dengue virus subtypes, and also a dozen in all five HCV subtypes, with three miRNAs present in both, and from them only one, called Anopheles gambiae miRNA-281, was found with a coincidence in the same UTR (5′) and in the same sense (+) for dengue and HCV. Its matching with dengue virus is interesting: for the subtypes 1, 2 and 3,
Figure 4. (a) Micro-RNAs (denoted also as miRNAs) and mRNAs maturation and (b) RNA secondary structures of let-7 pre-miRNAs with the same final sequence (in red), in *Candida elegans*, *Drosophila melanogaster* and *Homo sapiens*.
Micro-RNAs: viral genome and robustness

(a) ...uuuu–ACAA...
    u–A
    u–G
    a–U
    u–A
    u–g
    a–u
    g–c
    a–u
    g–c
    a
    c
    g
    a

(b) 81
    uguugggcuauucugagcgu
miR-34

...uuuuauuagacgcaucucugAUGAACAAn...
| | | | | | | | | | | |
uauuucucucguuaagguac
miR-281

Figure 5. (a) Hairpin form in zone 5′-UTR of the dengue virus and (b) hybridization with miR-281 and miR-34.

(a) 78
...uuuuauuagacgcaucucugAUGAACAAn...
| | | | | | | | | | | |
gagagagagacgcaucucugacggu
(b) 78
...aggcgcttttt...
| | | | | | | | | | | |
uagcgcgaac
uguugggcuauucugagcgu

Figure 6. Matching human miR-518c (bottom) with (a) dengue and (b) HCV sequences.

it matches exactly the end of the 5′-UTR, right before the beginning of the first CDS (coding sequence). It turns out that this part, in the absence of the miR, has a high hairpin-building potential, hybridized in chain form if the miR is added (figure 5).

Concerning human miRs, if the virus requires something to ‘open up’ the 5′-end, then it should also happen with Homo sapiens miR-518c (cf. http://microrna.sanger.ac.uk/cgi-bin/sequences/mirna_entry.pl?acc=MI0003159; figure 6), in which the matching concerns the Watson–Crick pairing plus the G-U pairing, with two hydrogen bonds, which occurs fairly often in RNA (but rarely in DNA).

For each mature miR and each target sequence, we slide the Watson–Crick complement of the miR over the target sequence on all possible positions. Thus, for each position, we compare a sequence \( m_1, m_2, \ldots, m_L \) (the miR) with a segment of the target, \( s_i, s_{i+1}, \ldots, s_{i+L-1} \). We define \( v_j = 1 \) if \( m_j = s_{i+j-1} \) and \( v_j = -1 \) otherwise. We consider the segment [start, stop] a candidate match, if:

(i) \( v_{\text{start}} = v_{\text{stop}} = 1 \) and \( \sum_{\text{start} \leq j \leq \text{stop}} v_j \geq 7; \)
(ii) it is maximal, i.e. not contained in a larger segment verifying previous conditions.

When we analyse the mean match score (calculated for all miRs of species indicated in the legends of figures 7 and 8) along the viral 5′-UTR, we notice a best match for the hosts whose coevolution with the virus has been the closest (e.g. showing a better fit for Gallus gallus than Homo sapiens for West Nile virus and the inverse for HCV, the fit being identified as the integral of the mean match
Figure 7. Percentage of mean matches between miRs of various genomes (blue, *G. gallus*; violet, *H. sapiens*; and yellow, *A. gambiae*) and West Nile 5′-UTR.

Figure 8. Percentage of mean matches between miRs of various genomes (blue, *G. gallus*; violet, *H. sapiens*; and yellow, *A. gambiae*) and HCV 5′-UTR. Green, Jopling.

curve). If we focus on precise miRs (figure 9), we find good matches between some of them and the 5′-UTR, showing a better resistance of some hosts, similar to the human miR-122 at the beginning of the HCV 5′-UTR (cf. figure 8; Jopling *et al.* 2005) and dengue 5′-UTR (figure 9).
Micro-RNAs: viral genome and robustness

(a) miR-122/HCV
acaccattgacacctcca
acacactaggtacacctcca (HCV 7-25)

(b) miR-122/dengue
acaaacacca
acaaacacca (Dengue 10-19)

(c) miR-17_5p/dengue
gcactgtaagcactttg
gcaggtgaagctatg (Dengue 73-89)

Figure 9. Good matches of human miR-122 with (a) HCV and (b) dengue and of (c) miR-17_5p with dengue, the fit being localized on the viral 5′-UTR.

It is clear that the genomic congruences shown above are more pertinent than the proximities in the Gatlin diagram, but they are calculated in the same spirit. Complementary studies, namely of modelling and simulation, should be performed in order to understand well the effective role of miRs in the host and the vector regulatory networks during viral infection. A variational principle maximizing the benefit that each species (host, vector and virus) is getting in this three-player game has also to be found in order to explain why the coevolution has produced these fits between the three genomes. This evolutionary variational principle would involve only the three genomes and no exogenous information (with respect to the set of game players) coming, for example, from ancestral genomes. However, if we want to introduce an external reference in order to emphasize the internal homogeneity of a given genome with respect to the set of all possible genomes, we need to calculate distances to this referential set and show that they are smaller between the given genome and the referential set than between the given genome and a set of randomly chosen genomes.

5. Primitive genome and comparison with RNA relics

It has been shown (Demongeot & Besson 1996; Moreira 2003; Demongeot & Moreira 2007a,b) that specific RNA rings (e.g. the ring shown in figure 10, called AL for archetypal loop) could be selected as solutions of a variational principle: to be of minimal length favouring RNA naturation or renaturation after denaturation, as well as RNA replication processes (figure 10a,b) and to offer at least one reasonable affinity site for each amino acid (in the sense of the stereo-chemical theory of the genetic code, i.e. with electrostatic and/or van der Waals interactions). AL is represented in the Gatlin diagram (figure 1) and lies between the archaeabacteria and the human mitochondrial genome. The total number of all the rings (denoted alRNAs) selected under this variational principle is 29,520. They all have a length of 22 bases, with a hairpin secondary structure (figure 10b,c), and are close (for the distances of §1) to all known tRNA relics essentially made of a succession of tRNA loops (Moreira 2003; Demongeot & Moreira 2007a,b), which can come from ancestral hairpins (Di Giulio 1992, 1997, 2009; Fujishima et al. 2008) and be present in miRNAs conserved or not in humans (Bentwich et al. 2005). Explaining the proximity or identity in the case of some tRNAs, like Oenothera lamarckiana Gly-tRNA, comes from the fact that rings subsolutions of the variational problem present a tRNA-like structure (figure 11), creating stems as for the O. lamarckiana Gly-tRNA clover leaf.

Phil. Trans. R. Soc. A (2009)
The ring AL fits with a high significance (less than 2.5% in figure 12) with siRNAs and miRs involved in many important cell functions. The mean and s.d. of the mean matching score (for the 22-circular Hamming distance) between all alRNAs and all known miRs are 9.634 and 0.088 (blue curve in figure 13). If we compare all known miRs with randomized samples from the set of all RNA rings having a length of 22 bases (there are about $16 \times 10^{12}$ such rings) and presenting the same base composition as the 29520 alRNAs, these values become 9.558 and 0.11 (yellow curve in figure 13).

The comparison between all the known miRs with AL gives a mean of 9.77935, over $\mu + 1.645\sigma$ with respect to alRNAs, and slightly over $\mu + 2\sigma$ with respect to randomized rings. Then, the mean matching score is significantly ($p = 0.005$) higher with AL for the set of all the known miRs ($\mu = 9.78$, $\sigma = 0.09$) than for a set of miRNAs obtained by chance ($\mu = 9.56$, $\sigma = 0.11$). In the same way, the expectation of the maximal length $L$ of consecutive matches of known miRs is significantly higher ($p = 0.01$) with AL ($P(L \geq 6) = 0.093$) than with miRs obtained by chance with the same base composition ($P(L \geq 6) = 0.07$). The barycentre of all known miRs also has a significantly higher number of matches with AL: mean and s.d. are 4.131 and 0.041 for randomized rings with the same
Micro-RNAs: viral genome and robustness

Figure 11. Clover leaf structure fitting the loops of the *O. lamarckiana* Gly-tRNA. The primary sequence (top) has four main parts (a, b, c, d), which constitute the loops of the clover leaf secondary structure (bottom) made from a ring of length 75 bases containing all the 64 triplets.

base composition as alRNAs; 4.227 and 0.070 for the set of 29 520 alRNA rings; and 4.31781 for AL, more than 4.5 times the s.d. from the mean 4.131. A unilateral test of mean shows that the match with AL is significantly better than the match with randomized rings (*p* = 10⁻⁶) (figure 14).

In figure 15, we show a significantly lower maxsubstring distance between known miRs observed in at least two species (repmiRs) and randomized miRs (with the same base composition as repmiRs) than with AL. More generally, in figure 16, we summarize for the two different distances introduced in §1 and for the circular version of the classical edit distance (the edit distance between two strings of characters is the number of operations required to transform one of them into the other) the proximity diagram (generalization of the Gatlin diagram) between the set of all known tRNA loops, the set of ancestral ring solutions of the variational problem of §5 and the set of all known miRs. An explanation of this proximity could lie in the fact that these structures with a low interspecific variability (tRNAs loops and miRs, as well as siRNAs as shown in figure 12) come from the same primitive reservoir of RNA rings satisfying a variational principle and that the fitness to their function (protein building for tRNAs and translation control for miRs) has been from the beginning sufficiently high to ensure their survival. We hope in the future to find the same type of variational principle explaining the fitness we have observed in §4 between host, vector and virus genomes.
Figure 12. Matches between human small RNAs and AL showing a mean p-value less than 2.5 per cent between AL and 12 siRNAs randomly chosen from the database http://www.rnainterference.org/HumanSequences.html (p-value majored by using the supremum of binomial variables instead of the circular Gumbel law), between anti-AL and two human miRNAs, and between a tRNA ancestor, AL, the Lewin’s tRNA loops (Lewin 2007) and the human miRNA 448, close to AL (14/22 matches).
Figure 13. Distribution of the mean matching score (22-circular Hamming distance) between all known miRs and the 29520 solutions of the variational problem (blue), and with a sample of the same size from the $16 \times 10^{12}$ randomized RNA rings of length 22, having the same base composition as the solutions (yellow). The position of AL is indicated with a black arrow.

Figure 14. Distributions as in figure 13, but for the maxsubstrings length. Blue, all 29520 rings; pink, 29520 random rings with same base composition.
Figure 15. Distribution of the maxsubstrings length between known miRs observed in at least two species (repmiRs) and randomized miRs (with the same base content as repmiRs).

Figure 16. Set of the random small RNAs having a length of 22 bases (mean of the length of 319 human micro-RNAs on histogram bottom), with indication of the barycentres of three sets: all known tRNA loops (t), ancestral RNA rings (a) and all known miRs (μ).
Micro-RNAs: viral genome and robustness

Figure 17. Coexistence of a primitive genome $G_1$ and of an evolutive genome $G_2$.

Figure 18. Coexistence of three species: host, vector and virus.

6. Beyond a common fitness function between RNA relics, viral genome and primitive genome

(a) A first equilibrium between two genomes, a primitive and an evolutive

Let us suppose that a primitive RNA genome $G_1$ is able to appear well protected against denaturation by amino acids $AA_i$, having a great affinity with it. To evolve, $G_1$ needs a second RNA genome $G_2$ with which it has the following relationship, summarized in figure 17.

(i) $G_1$ is able to make peptides $P$ (by confining amino acids, these giving peptides because of their proximity and ability to create covalent bonds), and it exports $P$ as ‘capsule’ peptides to contribute to the protection of $G_2$ (which presents affinity only for some amino acids of $P$ like $AA_1$).

(ii) $G_2$ is able to duplicate and grow through the classical operations such as mutation, insertion, deletion, inversion and translocation (Faraut & Demongeot 2000) with nucleic material already present in the environment, for example,
nucleic bases synthesized from Miller’s type reaction (Johnson et al. 2008), and to export small RNA fragments to G1.

(iii) Finally, the coevolution of G1 and G2 allows a first equilibrium between two genomes, G1 able to capitalize on the evolutionary memory and G2 able to evolve and ensure possibilities of evolution to G1. The game with two players, G1 and G2, leads to an equilibrium with two winners, each of them transmitting to the other its main survival feature, i.e. peptide protection for G1 and ability to evolve/adapt for G2.

(b) A second equilibrium between three genomes, a host, a virus and a vector

We can consider that after the first stage of evolution with two genomes that we have described above, another game appeared and went to equilibrium (cf. figure 18). This game consists in exchanging proteins and RNA (or DNA) between three players. The host, like the primitive genome G1, capitalizes on the evolutionary memory and is able, if infected by the RNA (or DNA) of a virus, to replicate it and to make the proteins necessary for its protective capsule. The virus plays the same role as G2 by being able to evolve rapidly in a given environment. It continues to contribute to the evolution of G1 by incorporating and leaving a part of its RNA inside G1 (whose molecular form becomes more stable and adapted to a conservative replication by adopting the DNA configuration; this is also the case for certain viruses that have adopted this more stable form for their genome). To be more efficient, in particular, at passing through the defences of the host, the virus uses a third species, the vector (which can also be an intermediary host susceptible to start the multiplication of the virus) well adapted to transport the viral RNA inside the host cells. The game still leads to an equilibrium with only winners: the host and the vector are increasing their adaptability, and the virus ensures its survival and multiplication. Because this game corresponds to a coevolution over a long time, it is not surprising now to find common RNAs between host, vector and virus, as we have shown in the previous sections, these common sequences being just the traces of the exchanges between the three species. A computational implementation of this game is possible and will be presented and discussed in a further paper.

7. Robustness of the micro-RNA control

Micro-RNAs 17_5p or 34 (figures 5, 8, 9 and 12) match not only with viral genomes and the AL ring, but also with mRNAs of proteins controlling important functions. Among these functions, proliferation is ruled by the cell-cycle network, whose boundary elements act on the transcription factor E2F, which belongs to the core of the network, made of a double positive loop (figure 19). By fixing the states of these micro-RNAs or of p53 (a transcription factor of miR-34; Jouanneau & Larsen 2007) to the value 1 (corresponding to their expression state), four limit cycles occur in their dynamics, when all elements of the cell-cycle network are synchronously updated. These limit cycles are never present in the case of parallel updating, when the boundary genes are in state 0 (figure 19). More generally, a strongly connected subnetwork of size three like those containing E2F can present four different dynamical behaviours: Cy (respectively, Fi and Mi), in
Figure 19. (a) Cell-cycle network controlling the mitosis events. (b) Some of the attractors of the dynamics described with fixed boundary conditions. (c) Attractors in synchronous and parallel updating modes without fixed boundary conditions. The attraction basin relative size (ABRS) denotes the percentage of the initial states lying in the basin of the attractor and the average diameter (AD) denotes the mean length of transient trajectories. (d) Attractors with miRs in state 1 in the parallel updating mode; the nodes are represented in the following order: p27, Cdk2, pCyCE_Cdk2, CyCE_Cdk2, miRNA 159, pCycA_Cdk2, CycA_Cdk2, Rbp-E2F, Rb-E2F, E2F, Rbp, Rb.
Table 1. Percentages of observed dynamical behaviours of strongly connected networks of size three before (upper two lines) and after (lower two lines) the addition of a micro-RNA, in terms of four classes, Cy, Fi, Mi and Ev.

<table>
<thead>
<tr>
<th></th>
<th>Cy</th>
<th>Fi</th>
<th>Mi</th>
<th>Ev</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60846</td>
<td>49071</td>
<td>19457</td>
<td>33406</td>
<td>161780</td>
</tr>
<tr>
<td></td>
<td>37.4%</td>
<td>30.1%</td>
<td>12.0%</td>
<td>20.5%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>168925</td>
<td>192673</td>
<td>49959</td>
<td>76786</td>
<td>488340</td>
</tr>
<tr>
<td></td>
<td>34.6%</td>
<td>39.5%</td>
<td>10.2%</td>
<td>15.7%</td>
<td>100%</td>
</tr>
</tbody>
</table>

which attractors are only limit cycles (respectively, only fixed configurations, and at least one fixed configuration and one limit cycle) independently of the updating mode; and Ev, in which attractors are only fixed configurations or only limit cycles for certain updating modes (Elena & Demongeot 2008; Elena et al. 2008; Elena 2009). The addition of a micro-RNA to these strongly connected subnetworks of size three increases the percentage of class Fi and decreases that of class Ev, hence improving the robustness of networks like the cell-cycle network, whose class becomes independent of the updating mode (table 1). The cell-cycle network is then very sensitive to its boundary elements, especially to the miRs. Then the viral mRNAs hybridizing these miRs can play a direct role in important cell functions such as proliferation; for the other main functions, robustness has already been studied in many papers (Ben Amor et al. 2008, 2009; Demongeot et al. 2000, 2006, 2008, 2009a, b).

8. Conclusion

We have shown in this paper that, for some RNA relics (i.e. RNA sequences well conserved among species) like tRNA loops and micro-RNA sequences extracted from their hairpin form, we have significant similarities. Because the mean length of these sequences is low (about 22), we used this to prove its existence as an intermediary, a reference set made of RNA rings selected from a variational principle (minimization of their length and maximization of their amino acid affinity in the stereochimical theory of the genetic code), which provided rings of length 22 only. Other small RNAs (like siRNAs) have also been tested, showing the same similarity. From this perspective, we could address the problem of the systematic detection of micro-RNAs in non-coding parts of genomes and show that there is a correlation between the low interspecific variability of these structures and their fit with the archetypal genome, as well as with the viral genome, both fits satisfying variational principles due to common coevolution. Even if Gilbert’s hypothesis (Gilbert 1986) of a primordial RNA world is not yet proved (cf. Ertem 2004; Shapiro 2007), the intense period of research about RNAs over the past 20 years is a reality. It has not been a ‘revolution’, but we can say, following Mello & Conte (2004), that ‘considering the potential role of RNA as a primordial biopolymer of life, it is perhaps more apt to call it an RNA ‘revelation’. RNA is not taking over the cell—it has been in control all along’.

Phil. Trans. R. Soc. A (2009)
Micro-RNAs: viral genome and robustness

References


Phil. Trans. R. Soc. A (2009)


