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

JACQUES DEMONGEOT[1,*], EMMANUEL DROUET[2], ADRIEN ELENA[1], ANDRES MOREIRA[3], YASSINE RECHOUM[2] AND SYLVAIN SENE[4]

[1]University J. Fourier of Grenoble, TIMC-IMAG UMR UJF/CNRS 5525, Faculty of Medicine, 38700 La Tronche, France
E-mail: Jacques.Demongeot@imag.fr; E-mail: Adrien.Elena@imag.fr
[2]EMBL Grenoble Outstation, 6 rue Jules Horowitz, BP181, 38042 Grenoble, France
E-mail: drouet@embl-grenoble.fr; E-mail: rechoum@embl-grenoble.fr
[3]Instituto de Sistemas Complejos, Artilleria 470, Cerro Artilleria, Valparaiso, Chile
E-mail: amoreira@inf.utfsm.cl
[4] INSA Lyon, LIRIS, Avenue A. Einstein, 69100 Villeurbanne, France
E-mail: ssene@ens-lyon.org

By comparing RNA rings or hairpins to reference or random ring sequences, circular versions of distances and distributions like Hamming and Gumbel one's are needed. We define these circular versions and we apply these new tools to the comparison of RNA relics like micro-RNAs and tRNAs, to viral genomes having co-evolved with them. Then we show how robust are the regulation networks incorporating in their boundary micro-RNAs as gardens of Eden or in new feed-back loops involving ubiquitous proteins like p53 or oligopeptids regulating traduction. Eventually, we propose a new co-evolution 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 its pathogenicity or virulence. In the classical Gatlin diagram (Gatlin 1968), whose variables were DNA redundancy R and GC % of genomes, the quadratic distance between 2 genomes was a way to compare them, based on their global content in puric and pyrimidic bases distribution (Figure 1). Now comparing genomes coming from infectious agents, hosts and vectors is always pertinent, and more sophisticated tools using entropy or circular distances based on 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): it is for example the case of the circular DNA of the 4600755 bp length chromosome *Yersinia pestis* (http://cmr.tigr.org/tigr-scripts/CMR/shared/CircularGenomeDisplay.cgi) or of the Hepatitis D circular RNA (http://pathmicro.med.sc.edu/virol/hepatitis-virus.htm) also known as Delta agent, more similar to a plant viroid than to a complete virus (Figure 2). The main difference with the historical approach done in the sixties is that now we can compare chain or ring sequences of RNA or DNA to reference sequences or to random rings (Figure 1 right), with appropriate distances and distribution functions expressing the variability of these distances among a population of given chains or rings.

*Author for correspondence (Jacques.Demongeot@imag.fr)

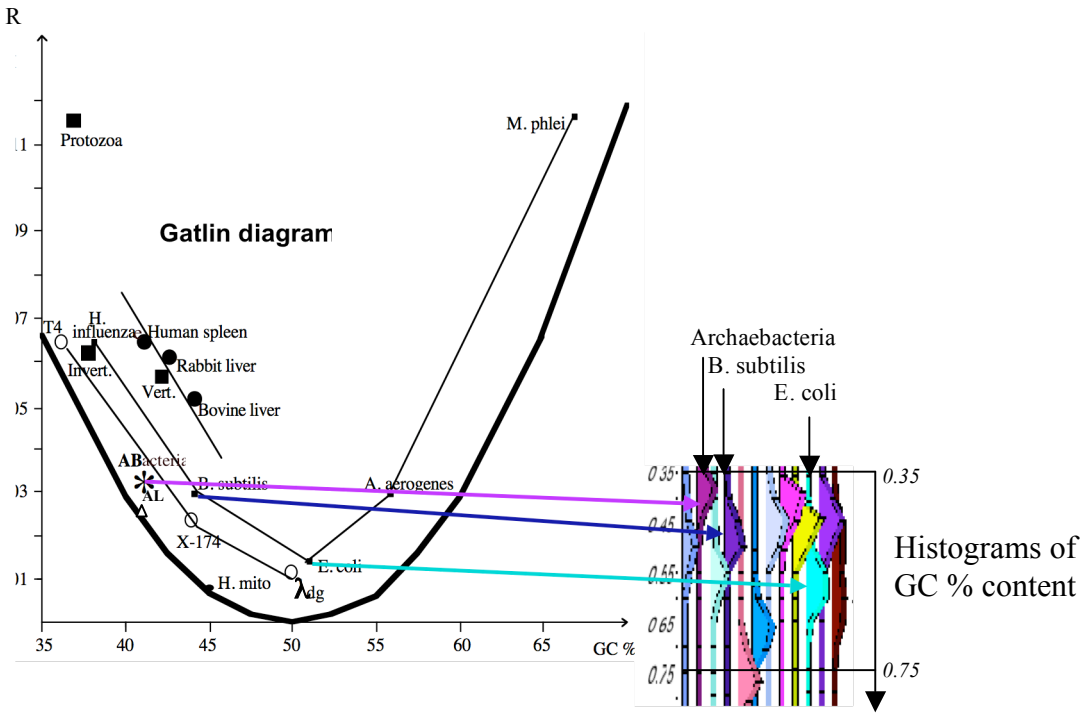


Figure 1. Genomes representation in the 2D Gatlin diagram, with redundancy R (y axis), GC % content in sequences (left), and histograms of GC % content of these genomes (right)

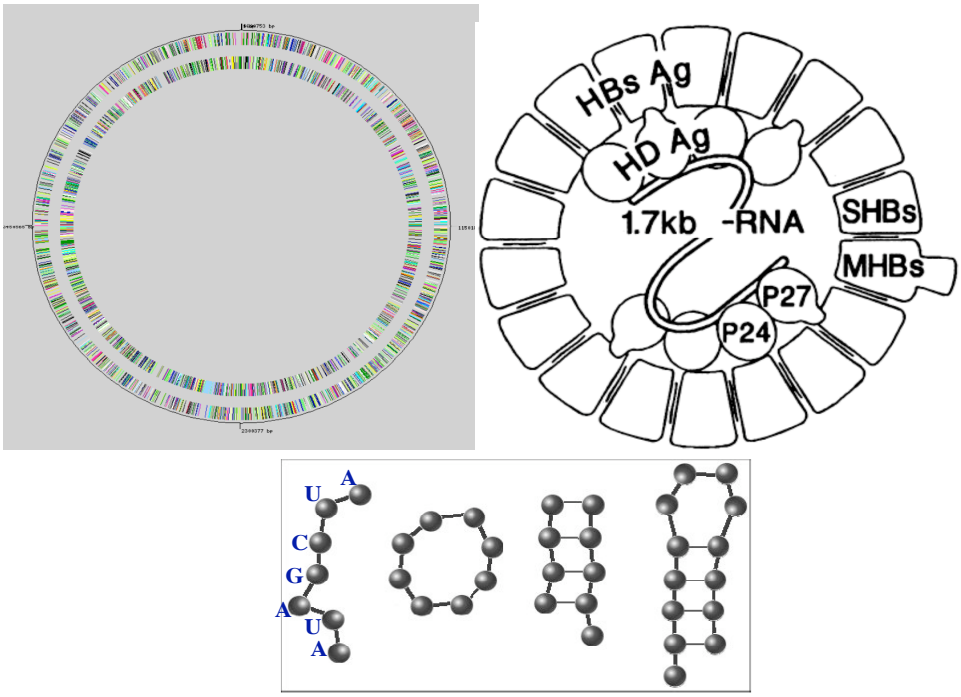


Figure 2. Circular *Yersina pestis* chromosome (left), circular Hepatitis D RNA (right) and various forms of RNA chains, rings and hairpins

In Figure 1, the redundancy R (that is the ability of the genome to repeat pairs of bases) is defined as follows, if p denotes GC %:

$R=1+p[P_{AU/AU}Log_2P_{AU/AU})+(1-P_{AU/AU})Log_2(1-P_{AU/AU})]+(1-p)[P_{GC/AU}Log_2P_{GC/AU})+(1-P_{GC/AU})Log_2(1-P_{GC/AU})]$,

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

In this paper we give the essential of the mathematical properties of these distances in the case of rings (Demongeot & Moreira 2007), the work for chains being extensively already published (cf. for example (Comet *et al.* 1999; Bacro & Comet 2000)) and after do the comparison between genomes of some co-evolving triplets (host, vector and infectious agent) in virology. For example, with data coming 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 argument about a possible co-evolution giving this fit as resulting from a global game favoring the survival of the three interacting species, each winning (a "win/win/win game").

2. Distances between rings and chains

If we would like to compare chains of dinucleotides, we could use classical distances between integer vectors, 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 introduce first a notation for the rotation: given $x \in A^n$, $\sigma(x) = (x_1,..., x_{n-1}, x_0)$ is the circular permutation. It is evident that following properties hold: σ is invertible, $\sigma^i(\sigma^j(x))=\sigma^{i+j}(x)$ and $\sigma^i(x)=\sigma^{i \pmod n}(x)$. We define the notion of equivalence under rotation, denoted "$\equiv$", for two vectors $x, y \in A^n$, by:

$$x \equiv y \Leftrightarrow \exists\, k : x = \sigma^k(y)$$

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 $[x] \in A^n/\equiv$.

2.1 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 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, it satisfies the triangle inequality and a null distance implies identity of the vectors. It is also easy to see that:

$\forall i \in \{0, \ldots, n - 1\}$, $d_H(x, y) = d_H(\sigma^i(x), \sigma^i(y))$, 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:

$$d^c_H([x], [y]) = \min_{0 \leq k \leq n-1} d_H(x, \sigma^k(y))$$

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

Lemma 1. *d^c_H is a metric on $A^n/\equiv$.*

*Proof.*1. If $d^c_H([x], [y]) = 0$, this implies that there exists k such that $d_H(x, \sigma^k(y)) = 0$; hence:

$$x = \sigma^k(y) \text{ and } [x] = [y]$$

2. Let us now prove the symmetry:

$$d^c_H([x], [y]) = \min_k d_H(x, \sigma^k(y)) = \min_k d_H(\sigma^{-k}(x), y) = \min_k d_H(y, \sigma^{-k}(x)) = d^c_H([y], [x])$$

3. Let $[x], [y], [z] \in A^n/\equiv$. We must show that the triangular inequality is satisfied, i.e., that:

$$d^c_H([x], [y]) \leq d^c_H([x], [z]) + d^c_H([z], [y])$$

Let i, j be such that: $d^c_H([z], [x]) = d_H(z, \sigma^i(x))$, $d^c_H([z], [y]) = d_H(z, \sigma^j(y))$

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:
$d^c_H([x], [y]) \leq d_H(\sigma^i(x), \sigma^j(y)) \leq d_H(\sigma^i(x), z) + d_H(z, \sigma^j(y)) = d^c_H([x], [z]) + d^c_H([z], [y])$ ■

2.2 Maxsubstrings distance
We define now 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 diffence 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, triangle inequality may fail for d_s if:
[x] = [abcd], $d_s([x], [y]) = 3$
[y] = [cbdd], $d_s([y], [z]) = 1$
[z] = [cbcd], $d_s([z], [x]) = 1$
and $d_s([x], [y]) > d_s([x], [z]) + d_s([z], [y])$

Lemma 2. *We have: $d^c_H \leq d_s$.*
Proof. Since $d_H(x, y) = \#\{i: x_i \neq y_i\}$, we have also $n - d_H(x, y) = \#\{i: x_i = y_i\}$, and hence we can write d^c_H as:

$$d^c_H([x], [y]) = \min_k [n - \#\{i: x_i = \sigma^k(y)_i\}] = n - \max_k \#\{i: x_i = \sigma^k(y)_i\}.$$

If the longest substring shared by [x] and [y] is of length m, then we have:

$$\max_k \#\{i: x_i = \sigma^k(y)_i\} \geq m, \text{ and thus:}$$
$$d^c_H([x], [y]) = n - \max_k \#\{i: x_i = \sigma^k(y)^i\} \leq n - m = d_s([x], [y])$$ ■

2.3 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. Circular Hamming distance d^c_H 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^c_H 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] which use the same amount of each kind of letter in A, i.e. such that $d_H(x, \alpha...\alpha) = d_H(y, \alpha...\alpha)$, for all α in A, where $\alpha...\alpha$ is the sequence of A^n made of the concatenation of n α; it will be ∞ 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 3. *d_t is a metric on $A^n/\equiv$.*
Proof. If $d_s([x], [y]) = 0$, 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 we do in addition 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; it 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 share both the letter usage of [z], the right side will be infinite too ■

2.4 The semi-metric d^*_t

For speeding 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'. With them we initialize 2 lists of words; let us denote $x^{(k)}$ the set of (1 or 2) subwords leaved in the words of $x^{(k-1)}$ by removing one of the longest substrings common with the words of $y^{(k-1)}$; then these 2 lists are $P_x=\{x', x'', x^{(3)},...\}$ and $P_y=\{y', y'', y^{(3)},...\}$. 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 2 words, taken from each list at the same level $x^{(k)}$ and $y^{(k)}$ (i.e. maximizing over all possible pairing 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 words set $x^{(N)}$ 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 in the required number of pieces in order to obtain one by reassembling the pieces of the other and reciprocally. 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: $M=n-\min_{k=1,...,n} d_H(x, \sigma^k(y))$, 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 a RNA chain with a reference RNA ring having for example each 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 of the origin of phases on the ring does not correspond strictly to a new chain tossing. Then we can write: $P(M<k)>P(\cap_{i=1,...,22}(X_i<k))$, where the X_i's are independent 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 ¼). By exploiting the binomial histogram (Figure 3), we obtain:

$P(M<15) > P(X\leq 14)^{22}\approx 1$
$P(M<14) > P(X\leq 13)^{22}=(0.9999)^{22}=0.998\approx 1-22x0.0001=0.998$
$P(M<13) > P(X\leq 12)^{22}=(0.9993)^{22}=0.985\approx 1-22x0.0007=0.985$
$P(M<12) > P(X\leq 11)^{22}=(0.997)^{22}=0.936\approx 1-22x0.003=0.934$
$P(M<11) > P(X\leq 10)^{22}=(0.99)^{22}=0.802$
$P(M<10) > P(X\leq 9)^{22}=(0.97)^{22}=0.512$
$P(M<9) > (P(X\leq 8))^{22}=(0.925)^{22}=0.180$
$P(M<8) > (P(X\leq 7))^{22}=(0.839)^{22}=0.021$
$P(M<7) > (P(X\leq 6))^{22}=(0.699)^{22}=4\ 10^{-4}$

Hence, we have:

$E(M)=\Sigma_{i=0,\ldots,22}P(M\geq k)=\Sigma_{i=1,\ldots,23}(1-P(M<k))=23-\Sigma_{i=1,\ldots,23}P(M<k) < 23-\Sigma_{i=0,\ldots,22}P(X\leq k)^{22}\approx 9.6$

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[.

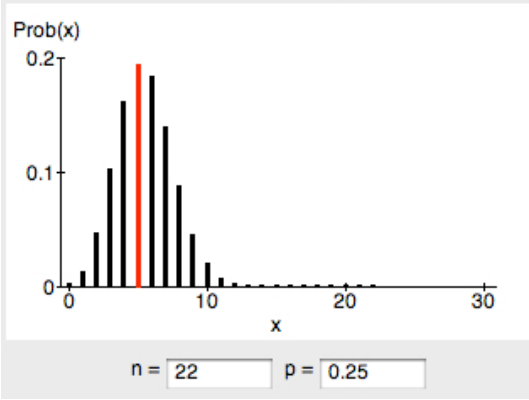


Figure 3. Binomial histogram B(22,1/4): P_1=Prob(X=1)=0.0131, P_5=0.1933, P_{14}=0.0001

The observed empirical mean (Figure 16) 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 $\sup_{i=1,\ldots,n}X_i$ 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 $\sup_{i=1,\ldots,n}X_i$ 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 $\sup_{i=1,\ldots,n}X_i$ distribution is due to the relatively weak number A_n of aperiodic rings (i.e. rings whose each circular permutation is different from the others) among the R_n possible rings (Ruskey & Sawada 2000):
$A_n = \Sigma_{d \text{ prime number divisor of } n}\ \mu(n/d)4^d/n$ and $R_n = \Sigma_{d \text{ prime number divisor of } n}\ \phi(n/d)4^d/n$, where μ and ϕ 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}=190557 and R_{22}=190746, which shows the reduction of the bias when n increases.

We will call "circular" Gumbel distribution the probability distribution of the random variable defined by: $(M-E(M))/\sigma(M)$, where $\sigma(M)$ is the standard deviation of M.

This quantity is random, but partially independent of the length (here 22) of the reference RNA ring. It could play for a "circular" Z-score 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 large deviations of this distribution given by the supremum of binomial variables, we can show for example the significativity (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. Section 5). The circular Gumbel distribution can be estimated by using a von Mises-Tychonov kernel (Shmaliy 2005).

4. RNA relics

The RNA relics (essentially tRNA loops, si-ARNs 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. It is for example the case of 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 (Figure 19), present in the non coding regions of many known genomes (specially 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 be also used in infectious diseases for predicting the pathogenicity of the infectious agents.

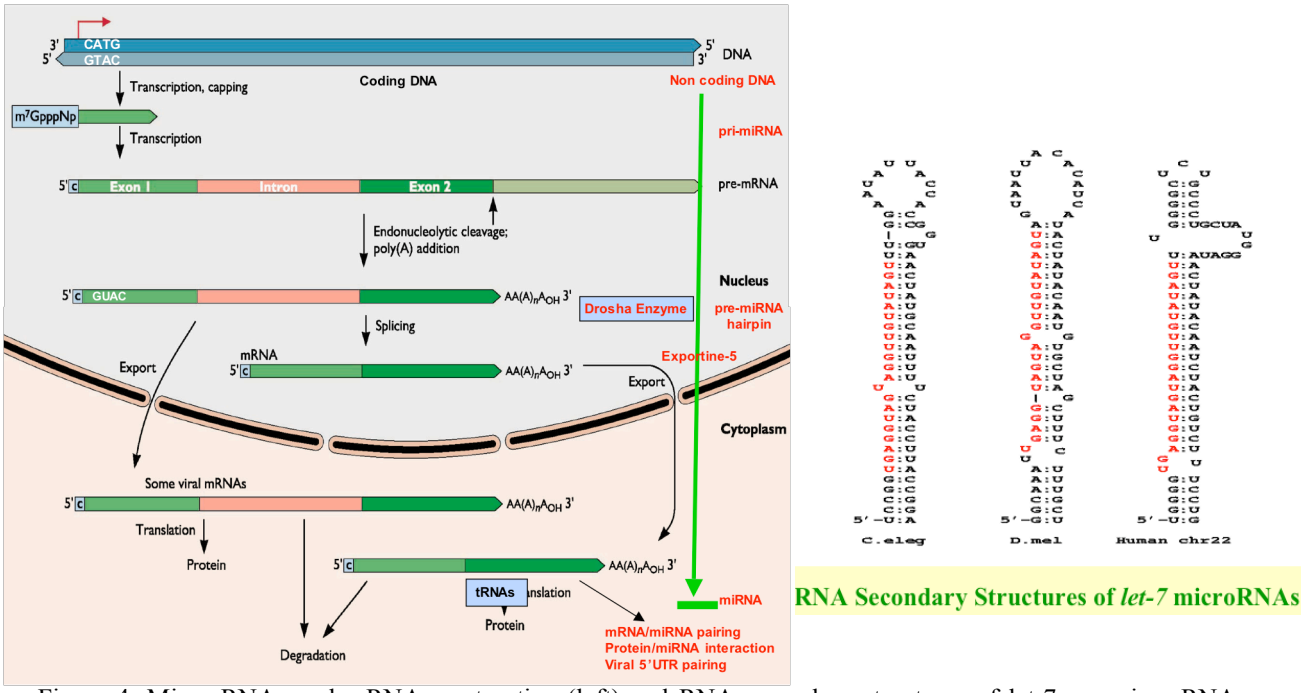


Figure 4. Micro-RNAs and mRNAs maturation (left) and RNA secondary structures of let-7 pre-micro-RNAs with the same final sequence (in red), in *C. elegans*, *D. melanogaster* and *H. sapiens* (right)

During the first step of the maturation process, the micro-ARNs (miRs) have a hairpin structure (http://protein3d.ncifcrf.gov/shuyun/Web/talk/Talk04.pdf), and both bioinformatic approaches and direct cloning methods have identified many such miRs, including orthologs 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 egypti* (Nene *et al*. 2007) and used also the 5' UnTranslated Region (5'UTR) part of viral RNAs, like a typical isolate mRNA of the *Hepacivirus*, Hepatitis C virus (HCV), a 341 nucleotides sequence containing an Internal Ribosome Entry Site (IRES) required for the translation initiation. It is fully admitted that the 5' and 3' UTRs

may play a role in the initiation of negative-strand synthesis of virus 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 note that this miR was likely to facilitate replication of the viral RNA (Appel & Bartenschlager 2006). By searching matches between miRs and viral genomes, we discovered also that a dozen of miRs had a conserved coincidence in all four Dengue virus subtypes, and also a dozen in all five HCV subtypes, with 3 miRNAs present in both, and from them only one, called *Anopheles gambiae* miRNA-281, with a coincidence in the same UTR (5') and in same sense (+) for Dengue and HCV. Its matching with Dengue virus is interesting: for the subtypes 1, 2 and 3, 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).

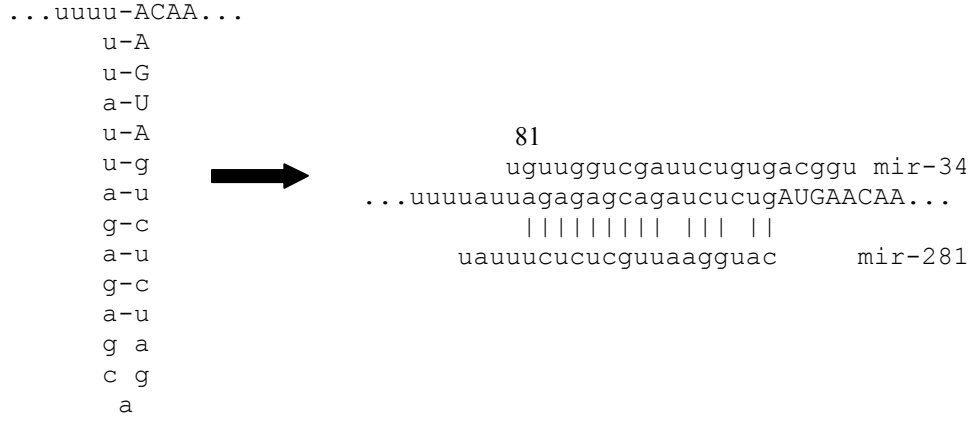


Figure 5. Hairpin form in zone 5' UTR of the Dengue virus (left) and hybridization with miRs 281 and 34 (right)

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 and 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).

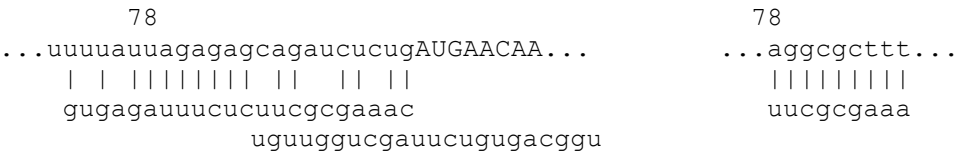


Figure 6. Matching human mir-518c (bottom) with Dengue (left) and HCV (right) sequences

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 $m_1m_2\ldots m_L$ (the miR) with a segment of the target, $s_is_{i+1}\ldots s_{i+L-1}$. 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:

- $v_{start} = v_{stop} = 1$ and $\sum_{start \le j \le stop} v_j \ge 7$
- 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 legend of Figures 7 and 8) along the viral 5' UTR, we can notice a best match for the hosts whose co-

evolution 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 Dengue 1 virus, the fit being identified as the integral of the mean match curve). If we focus on precise miRs (Figure 9), we can find good matches between some of them and the 5' UTR, showing a better resistance of some hosts, like for the human miR-122 at the beginning of the HCV 5' UTR (Jopling *et al*. 2005; Demongeot *et al.* 2008 b) and Dengue 5' UTR (Figures 8 & 9).

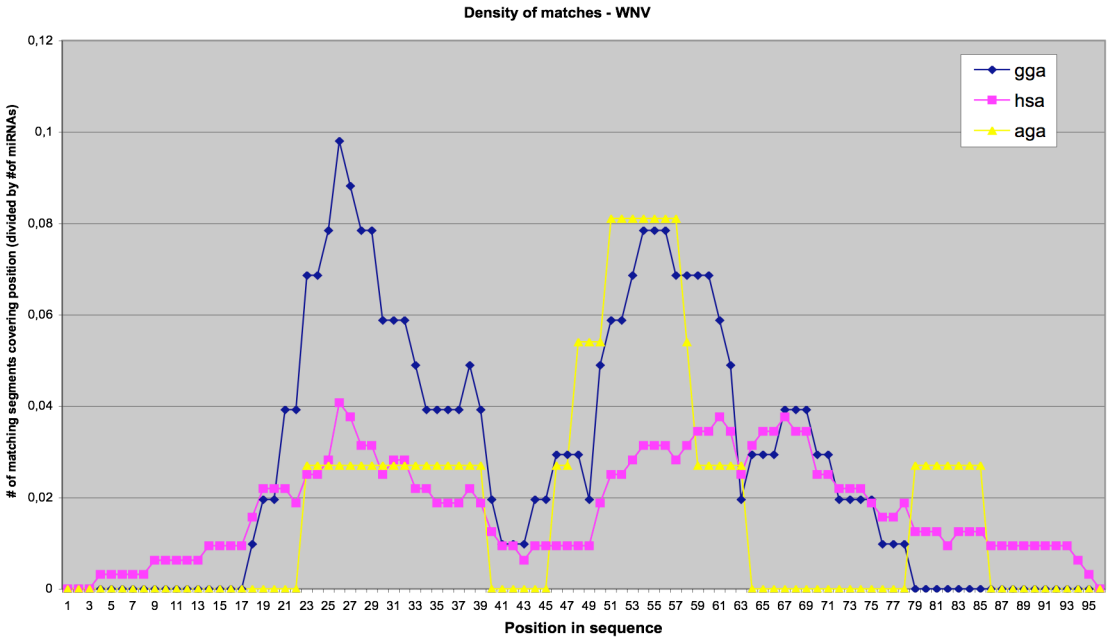


Figure 7. % of mean matches between miRs of various genomes (blue *Gallus gallus*, violine *Homo sapiens* and yellow *Anopheles gambiae*) and West Nile 5' UTR

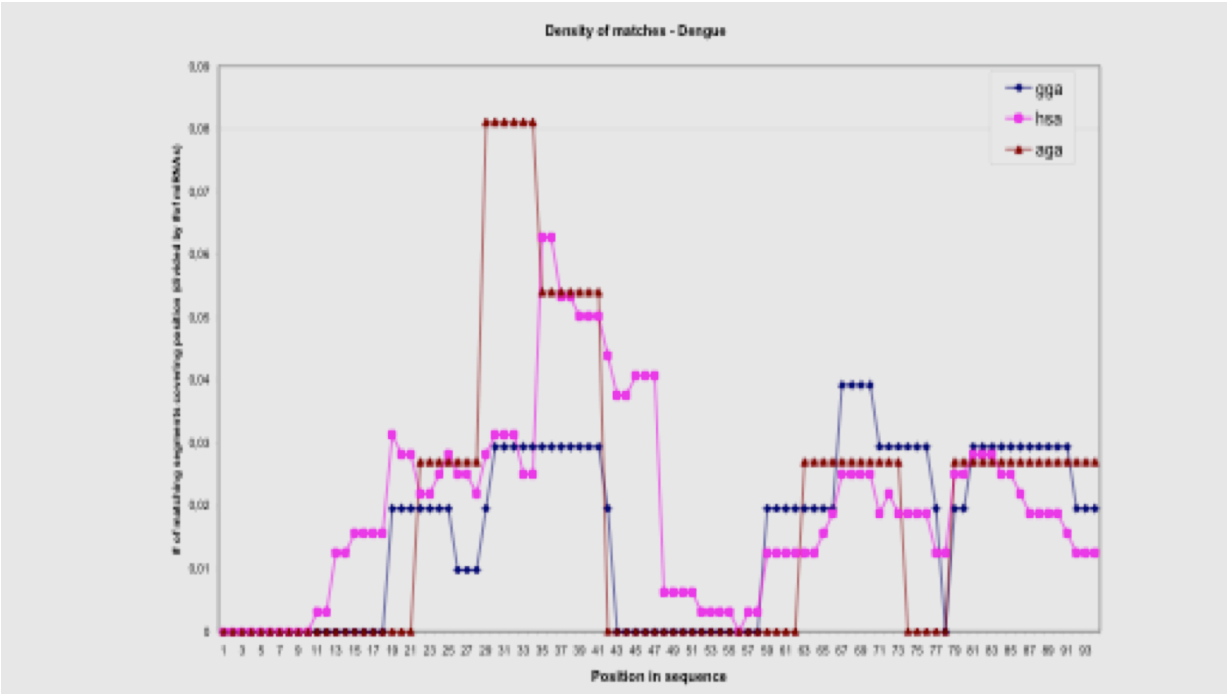


Figure 8. % of mean matches between miRs of various genomes (blue *Gallus gallus*, violine *Homo sapiens* and brown *Anopheles gambiae*) and Dengue 1 virus 5' UTR

miR-122/HCV	acaccattgtcacactcca acacactaggtacactcca (HCV 7-25)
miR-122/Dengue	acaaacacca acaaacacca (Dengue 10-19)
miR-17_5p/Dengue	gcactgtaagcactttg gcacggtaagagctatg (Dengue 73-89)

Figure 9. Good matches of human miR-122 with HCV and Dengue (top), and of miR17_5p with Dengue (bottom), 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 well understand the effective role of miRs in the host and the vector regulatory networks during viral infection. A variational principle maximizing the benefit each species (host, vector and virus) is getting in this 3 players game has also to be found in order to explain why the co-evolution has produced these fits between the three genomes. This evolutionary variational principle would involve only the three genomes and no exogeneous information (with respect to the game players set) information coming for example from ancestral genomes. However, if we want to introduce an external referential 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, than between the given genome and a set of randomly chosen genome.

5. Primitive genome and comparison with RNA relics

It has been shown in (Demongeot & Besson 1996; Moreira 2003; Demongeot & Moreira 2007) that specific RNA rings (e.g. the ring shown on 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 10 top) 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 electro-static and/or van der Waals interactions). AL is represented in the Gatlin diagram (Figure 1) and lies between the Archae Bacteria and the human mitochondrial genome. All selected rings under this variational principle, denoted alRNAs, are 29520, have all a length of 22, can present a hairpin secondary structure (Figure 10 bottom), and are narrow for the distances of Section 1 to all known tRNA relics essentially made of succession of tRNA loops (Moreira 2003; Demongeot & Moreira 2007). Explaining the proximity or identity in the case of some tRNAs, like *Oenothera lamarckiana* Gly-tRNA, comes from the fact that rings sub-solutions of the variational problem present a tRNA-like structure (Figure 11), creating stems as for the *O. lamarckiana* Gly-tRNA clover leaf.

The ring AL fits with a high significativity (less than 2.5 % in Figure 12) with siRNAs and miRs involved in many important cell functions. The mean and standard deviation of the mean matching score (for the 22-circular Hamming distance) between all alRNAs and all known miRs are $\mu=9.634$ and $\sigma=0.088$ (blue curve on Figure 13). If we compare all known miRs to randomized samples from the set of all RNA rings having a length of 22 bases (there are about 16 10^{12} such rings) and presenting the same base composition as the 29520 alRNAs, these values become $\mu'=9.558$ and $\sigma'=0.11$ (yellow curve on Figure 13).

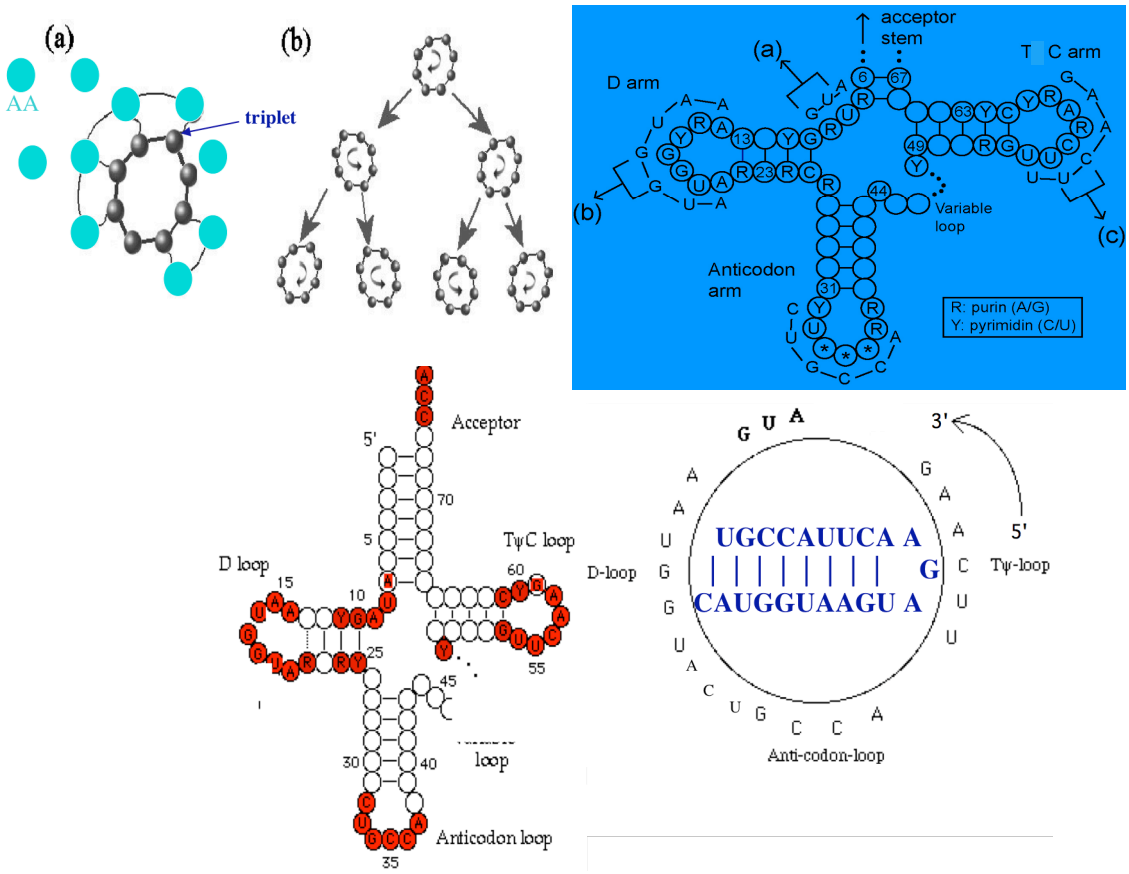


Figure 10. Selection of a ring called AL satisfying a variational principle for amino acids affinity and for renaturation and replication processes optimization (top left), made of the succession of overlapped codons of all amino-acids (like the archetypal Lewin's tRNA, top right); best fit of the AL ring with a specific tRNA, the *Oenothera lamarckiana* Gly-tRNA (bottom left), and hairpin inside the ring form of AL (bottom right)

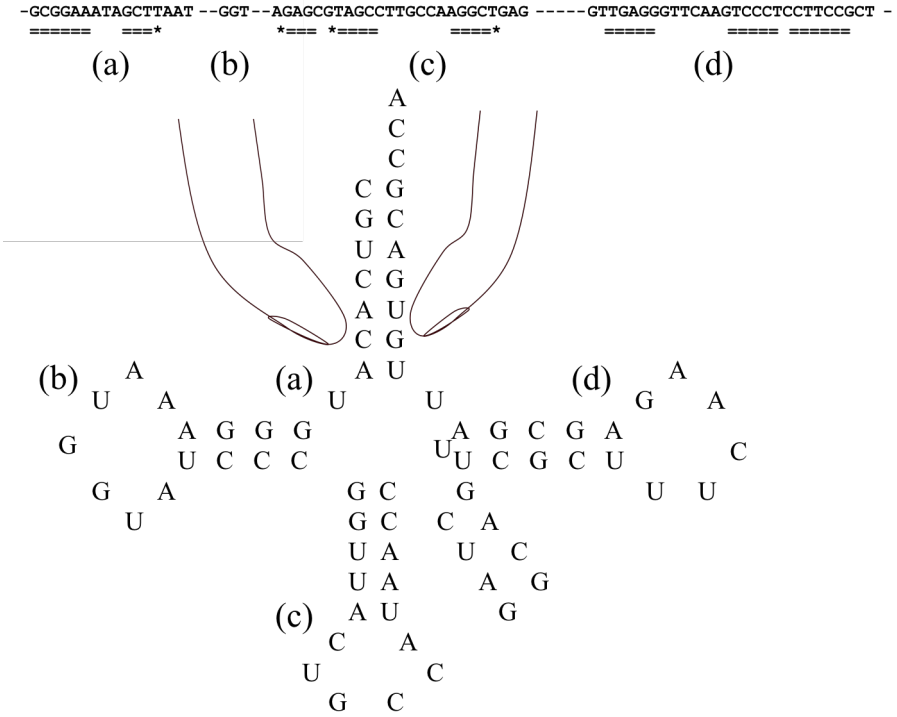


Figure 11. Clover leaf structure fitting the loops of the *Oenothera lamarckiana* Gly-tRNA (top) from a ring of length 75 containing all the 64 triplets (bottom)

NM_005572 Hs.491359 LMNA Lamin A/C
5'- ACTGGACTTCCAGAAGAAC -3'
AL UGCCAUUCAAGAUGAAUGGUAC nb of AL matches: 13/19 p-value=.002

NM_001904.2 Hs.476018 CTNNB1 Catenin (cadherin-associated protein), beta 1, 88kDa
5'-CUAUCAGGAUGACGCGG-3'
UGCCAUUCAAGAUGAAUGGUAC 10/17 .05

NM_033360.2 Hs.505033 KRAS2 V-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog
5'-TTCAAGAGACTACGCCA-3'
UGCCAUUCAAGAUGAAUGGUAC 11/17 .001

NM_001641.2 Hs.73722 APEX1 APEX nuclease (multifunctional DNA repair enzyme) 1
5'-AACCTGCCACACTCAAGATC-3'
UGCCAUUCAAGAUGAAUGGUAC 12/20 .02

NM_006839.1 Hs.148559 IMMT Inner membrane protein, mitochondrial (mitofilin)
5'-AAUUGCUGGAGCUGGCCUUTT-3'
UGCCAUUCAAGAUGAAUGGUAC 12/21 0.035

NM_016485.3 Hs.431367 C6ORF55 Chromosome 6 open reading frame 55
5'-GAATGAAGATCGATAGTAA-3'
UGCCAUUCAAGAUGAAUGGUAC 13/19 0.001

NM_016485.3 Hs.431367 C6ORF55 Chromosome 6 open reading frame 55
5'-GCAGTGCTTTGCAGTATGA-3'
UGCCAUUCAAGAUGAAUGGUAC 12/19 0.01

NM_016410.2 Hs.415534 SNF7DC2 SNF7 domain containing 2
5'-GAGAGGGTCCTGCAAAGAA-3'
UGCCAUUCAAGAUGAAUGGUAC 11/19 0.043

NM_004827.1 Hs.480218 ABCG2 ATP-binding cassette, sub-family G (WHITE), member 2
5'-AAGATGATTGTTCGTCCCTGCT-3'
UGCCAUUCAAGAUGAAUGGUAC 13/22 0.015

NM_212535.1 Hs.460355 PRKCB1 Protein kinase C, beta 1
5'-AAGCGCTGCGTCATGAATGTT-3'
UGCCAUUCAAGAUGAAUGGUAC 12/21 0.035

NM_004068.2 Hs.518460 AP2M1 Adaptor-related protein complex 2, mu 1 subunit
5'-AAGGUCCAGU-CAUUCCAAAUG-3'
UGCCAUUCAAGAUGAAUGGUAC 12/21 0.035

NM_002940.1 Hs.12013 ABCE1 ATP-binding cassette, sub-family E (OABP), member 1
5'-AGAGTTGTCCTGTAGTTCG-3'
UGCCAUUCAAGAUGAAUGGUAC 11/19 0.043

3'-UGUUGGUCGAUUCUGUGACGGU-5':hsa-miR-34a
GUUCUACUUACCAUGACGGUAA 11/23 anti-AL + 3 GA or UC matches $3\ 10^{-4}$

3'-UGAUGGACGUGACAUUCGUGAAA-5':hsa-miR-17_5p
GUUCUACUUACCAUGACGGUAA 11/23 anti-AL + 1 GA or UC matches $5\ 10^{-4}$

Figure 12. Matches between human small RNAs and AL showing a mean p-value less than 2.5% between AL and 12 siRNAs randomly chosen in the data base http://www.rnainterference.org/HumanSequences.html (p-value majored by using the supremum of binomial variables instead of the circular Gumbel law) and between antiAL and two human miRs

The comparison between all known miRs with AL gives a mean matching score $\mu''=9.78$ ($\sigma''=0.09$) over $\mu+1.645\sigma$ with respect to the distribution of this score among alRNAs, and slightly over $\mu'+2\sigma'$ with respect to randomized rings with same base composition as alRNAs. In the same way, all known miRs have a mean of the maximal length L of consecutive matches with AL $\nu''=4.32$ ($\sigma''=0.06$) over $\nu+1.645\sigma$ with respect to the distribution of L among alRNAsrings ($\nu=4.22$; $\sigma=0.06$) and over $\nu'+2\sigma'$ with respect to randomized rings ($\nu'=4.1$; $\sigma'=0.09$) (Figure 14). Then the mean of the matching score and of L are significantly ($p=0.05$) higher between AL and all known miRs than for a set of miRs obtained by chance.

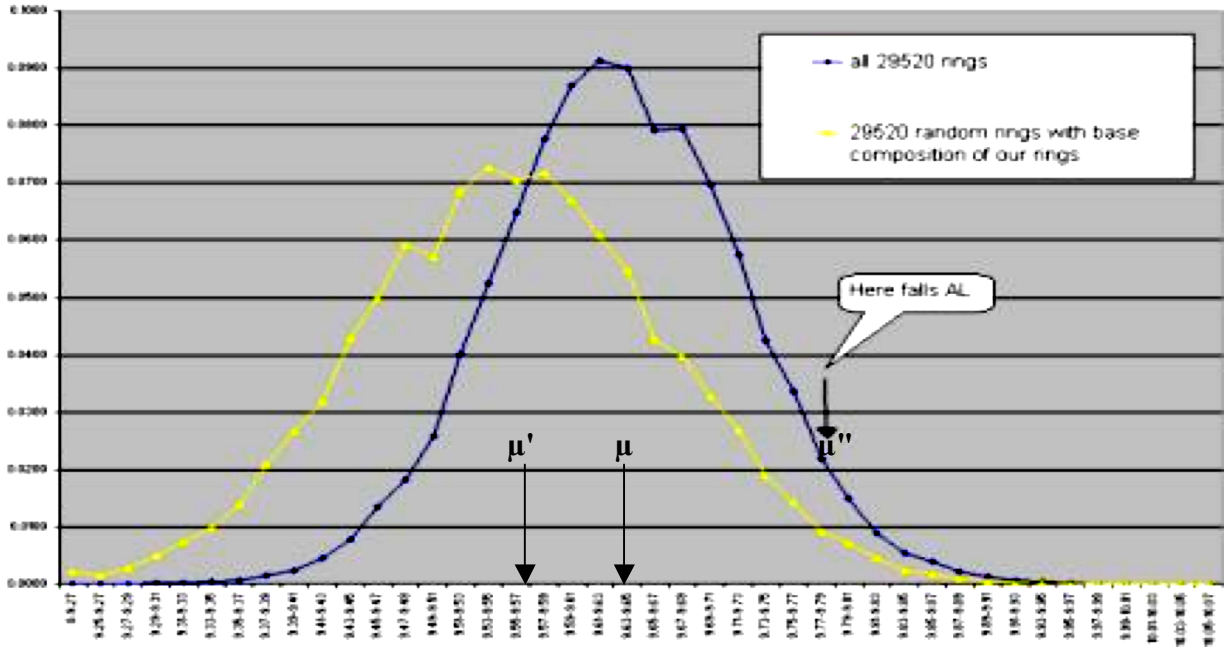


Figure 13. The distribution of the mean matching score (22-circular Hamming distance) between all known miRs and the 29520 solutions of the variational problem (blue), and with a sample of the same size from the 16 10^{12} randomized RNA rings of length 22, having the same base composition as the solutions (yellow). The position of AL is indicated with a black arrow.

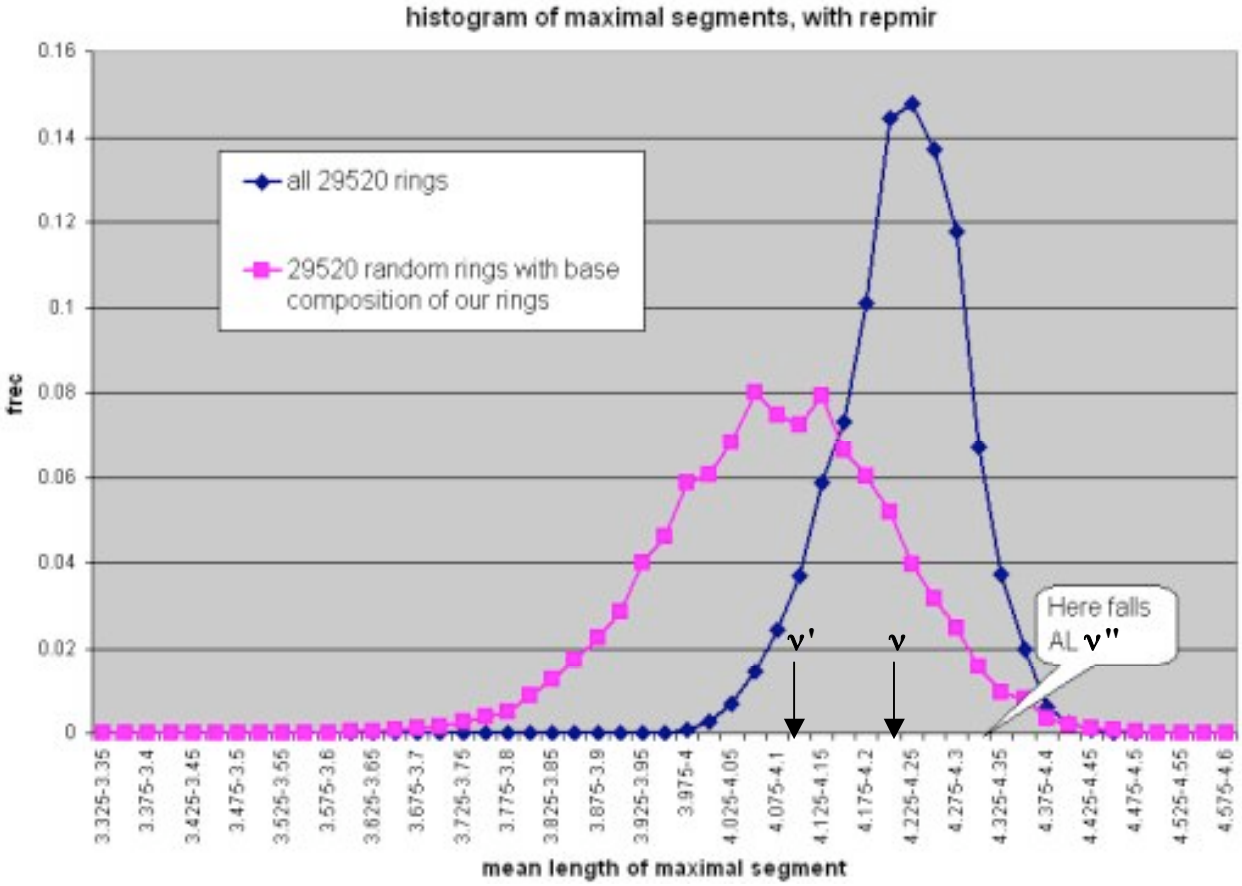


Figure 14. Distributions like in Figure 13, but for the maxsubstrings length L

We can generalize the Gatlin diagram (Figure 1) in a proximity diagram between the set of all known tRNA loops, the set of the ancestral rings solutions of the variational problem and the set of all known miRs based on calculations using the distances introduced in Section 2 (Demongeot *et al.* 2008 a & b). 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) are coming from the same primitive reservoir of RNA rings satisfying the 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. In the future, we hope to find the same type of variational principle explaining the fitness between host, vector and virus genomes.

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

6.1 A first equilibrium between a primitive and an evolutive genome

Let us suppose that a primitive RNA genome G_1 appeared, well protected against denaturation by amino acids AA_i having with it a great affinity. For evolving, G_1 needs a second RNA genome G_2 with which it has the following relationship, summarized in Figure 15:

- G_1 favors the formation of peptides P (by confining amino acids, these ones giving peptides because of their proximity and ability to create covalent peptidic bonds), and exports P as "capsid" peptides to contribute to the protection of G_2 (which presents affinity only for some amino acids of P like AA_1)
- G_2 is able to duplicate and growth (by using the classical operators of mutation, insertion,...) and can export small RNA fragments able to be inserted in G_1.

Finally the co-evolution of G_1 and G_2 allows a first equilibrium between 2 genomes, G_1 able to capitalise the evolution memory and G_2 able to evolve and ensure possibilities of evolution to G_1. The game with 2 players, G_1 and G_2, leads to an equilibrium with 2 winners, each of them transmitting to the other its main survival feature, i.e. peptide protection for G_1 and evolutility/adaptability for G_2.

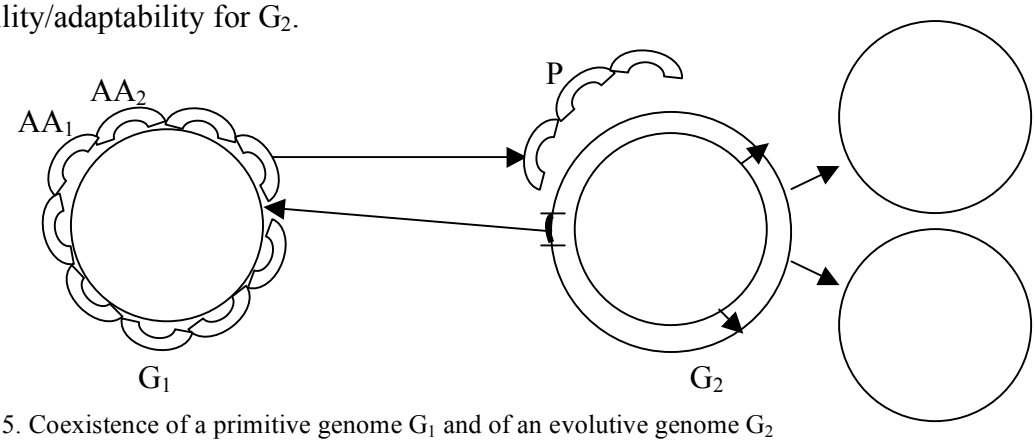


Figure 15. Coexistence of a primitive genome G_1 and of an evolutive genome G_2

6.2 A second equilibrium between 3 genomes, a host, a virus and a vector

We can consider that after a first stage of evolution with 2 genomes as described in 6.1, an other game appeared and went to equilibrium (Figure 16). This game consists in exchanging proteins and RNA (or DNA) between 3 players, a host, which like the primitive genome G_1 capitalises the evolution memory and is able, if infected by the RNA (or DNA) of a virus, to replicate it and to build the proteins necessary to its protective capsid. The virus plays the same role than G_2 by being able to evolve and adapt rapidly in a given environment. It

contributes to the evolution of G_1 by incorporating a part of its RNA inside G_1, whose molecular form became more stable and adapted to a conservative replication, by adopting the DNA configuration (it has also been the case for certain viruses which have adopted this more stable form for their genome). For being more efficient, in particular to pass through the host defences, viruses use a third species, a vector, which can also be an intermediary host susceptible to start the multiplication of the virus, well adapted to the transport of the viral RNA inside the host cells. The game is still leading to an equilibrium with 3 winners: the host and the vector are increasing their adaptability, and the virus ensures its survival and multiplication. Because this game corresponds to a co-evolution during a long time, it is not surprizing to find now common RNAs between host, vector and virus, as we have shown in the previous Sections, these common sequences being just the traces of past exchanges between the 3 species. An informatic implementation of this game is possible and will be presented and discussed in a further paper.

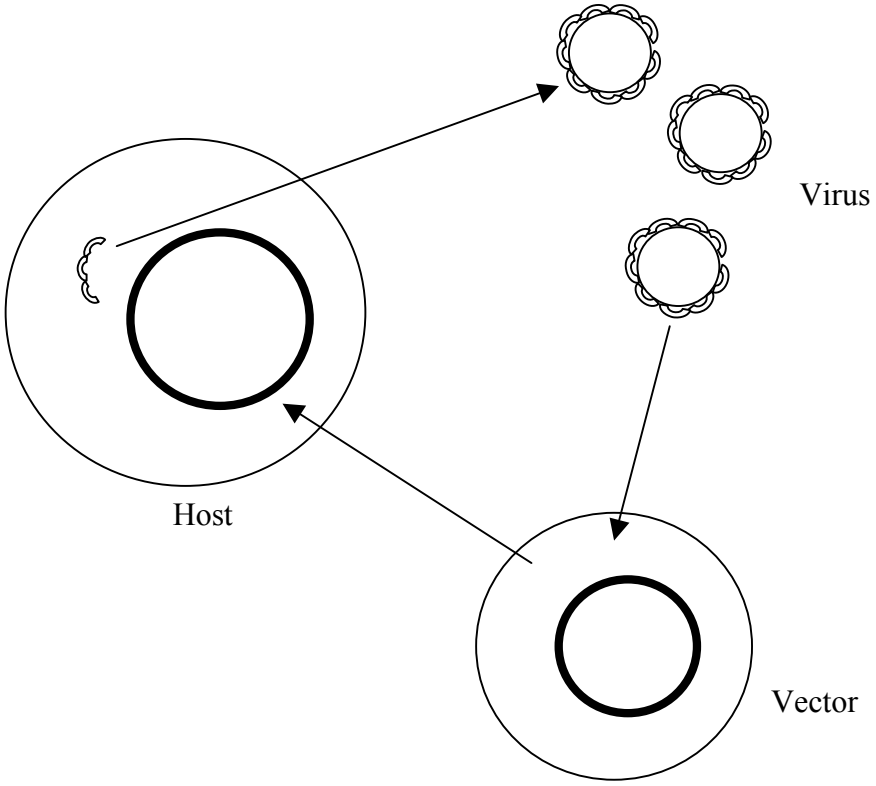
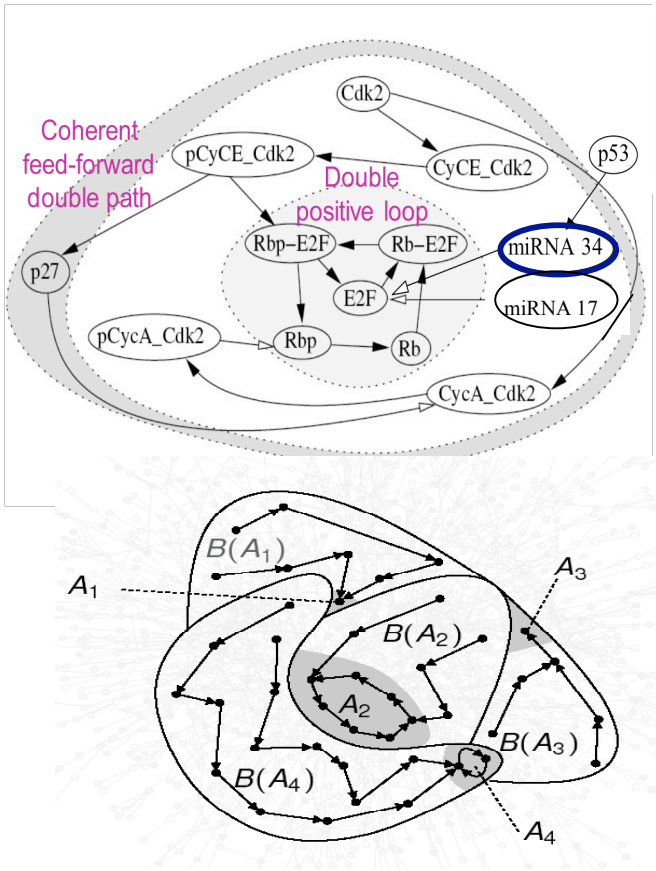


Figure 16. Coexistence of 3 species, host, vector and virus

7. Robustness of the micro-RNA control

MicroRNAs 17_5p or 34 (Figures 5, 8, 9, 12) are matching with viral genomes and AL ring, but also with mRNAs of proteins controlling important functions like those of the cell cycle network as boundary elements (http://microrna.sanger.ac.uk) acting on the transcription factor E2F which belongs to a core made of a double positive loop (Figure 17 left). By fixing the state of these miRs or of the p53 (transcrition factor of the miR-34) to the value 1 (corresponding to their state of expression) and by updating parallelly all the genes of the network, four limit cycles occur in its dynamics, never observed when the boundary states were free in the parallel case (Figure 17 right).



	Sequential updating		Parallel updating		
Nature	Attractor	ABRS	Attractor	ABRS	AD
Fixed point 1	000000000000	6.25%	000000000000	0.5%	1.45
Fixed point 2	000000011111	56.25%	000000011111	99.5%	5.40
Limit cycle 1	000000001000 000000010111	37.5%	None		

	Nature	Attractor	ABRS
A_3	Fixed point 1	000010000000	0.12%
A_1	Fixed point 2	000010011011	0.54%
A_2	Limit cycle 1	000010001010 00001001000 1	10.6%
A_4	Limit cycle 2	000010010011 000010001011 000010011001 000010011010	26.6%
	Limit cycle 3	000010010010 000010000011 000010001001 000010011000	32.4%
	Limit cycle 4	000010000010 000010000001 000010001000 000010010000	23.7%

Figure 17. Cell cycle network (top left) controlling the mitosis events. Attractors of the dynamics (bottom left) described without fixed boundary conditions in synchronous and parallel updating modes (top right), ABRS denoting the percentage of the initial states lying in the basin of the attractor, and with miRs in state 1 in the parallel updating mode (bottom right); 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

Cy	Fi	Mi	Ev	Total
60846	49071	19457	33406	162780
37, 4%	30, 1%	12, 0%	20, 5%	100%
Cy	Fi	Mi	Ev	Total
168925	192673	49959	76783	488340
34, 6%	39, 5%	10, 2%	15, 7%	100%

Figure 18. Percentages of observed dynamical behaviors without (top) and with (bottom) miRs at the boundary of all regulatory networks having 3 genes

Cy	Fi	Mi	Ev	Total
7, 38	4, 25	2, 33	4, 88	5, 32
Cy	Fi	Mi	Ev	Total
7, 53	4, 94	2, 83	5, 14	5, 61

Figure 19. Mean diameter (for the Hamming distance) of the smallest attraction basin without (top) and with (bottom) miRs at the boundary of all regulatory networks having 3 genes

The observed dynamical behaviors of a regulatory network can be dispatched into four classes: those having for all updating modes only limit cycles *Cy*, those having only fixed configurations *Fi*, and those having at least one fixed configuration and one limit cycle *Mi*; the last class is made of behaviors presenting either only fixed configurations for certain updating modes or occurrence of additional limit cycles for the other updating modes (Elena 2009). If we simulate the dynamical behavior of all regulatory neworks having 3 genes, we observe that the presence of additional miRs at their boundary reinforces the class *Fi* (Figure 18) and also increases the stability of the attractors, by augmenting the mean diameter (for the Hamming distance) of their smallest attraction basin (Figure 19).

The cell cycle network and particularly its 3 genes core Rbp-E2F/Rb-E2F/E2F, the *Arabidopsis thaliana* flowering network (Sené 2008), as well as statistically all the 3 genes regulatory networks, are very sensitive to their boundary elements, especially to the miRs action. Then the viral mRNAs hybridizing these miRs can play a direct role (by displacing the miRs hybridization with cellular mRNAs) on important cell functions like the proliferation.

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-RNAs sequences we had significant similarities. The mean length of these sequences is low (about 22), and we used to prove the existence of these similarities an intermediary reference set made of RNA rings selected from a variational principle (minimization of their length and maximization of their amino-acids affinity, in the framework of the stereo-chemical theory of the genetic code), which provided only rings of length 22. Other small RNAs (like siRNAs) have also been tested showing the same similarity. In perspectives we could address the problem of the systematic detection of micro-RNAs in non coding parts of the genomes and show that there could be a correlation between the low interspecific variability of these structures and their fit with the archetypal genome, as well as with viral genomes, due to a common co-evolution. Even if the Gilbert's hypothesis (Gilbert 1986) of a primordial RNA world is not yet proved (Ertem 2004; Shapiro 2007), the intense period of research about RNA's since 20 years is a reality. It has not been a "revolution", but we can say following (Mello and 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."

References

Appel, N. & Bartenschlager, R. 2006 A novel function for a miR: Negative regulators can do positive for the hepatitis C virus. *Hepatology* **43**, 612-615.

Bacro, J.N. & Comet, J.P. 2000 Sequence alignment : an approximation law for the Z-value with applications to databank scanning. *Computers & Chemistry* **25**, 401-410.

Bartenschlager, R., Frese, M. & Pietschmann, T. 2004 Novel insights into hepatitis C virus replication and persistence. *Advances in Virus Research* **63**, 71-180.

Ben Amor, H., Demongeot, J., Elena, A. & Sené, S. 2008 Structural Sensitivity of Neural and Genetic Networks. *LNCS* **5317**, 973-986.

Ben Amor, H., Cadau, S., Elena, A., Dhouailly, D. & Demongeot, J. accepted Regulatory networks analysis: robustness in biological regulatory networks. In: *AINA' 09 & BLSMC' 09*. Piscataway: IEEE Proceedings.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y. & Bentwich, Z. 2005 Identification of hundreds of conserved and non conserved human micro-RNAs. *Nature Genetics* **17**, 766-770.

Blalock, J. E. & Bost, K. 1986 Binding of peptides that are specified by complementary RNAs. *Biochem. J.* **234**, 679-683.

Bosnacki, D., ten Eikelder, H. & Hilbers, P. 2003 Genetic code as a Gray code revisited. In: *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences METMBS'03*, pp. 447-456. Athens Georgia: CSREA Press.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyogova, S., Illarionova, A., Fushan, A., Vinogradova, T. & Sverdlov, E. 2003 The human genome contains many types of chimeric retrogenes generated through in vivo recombination. *Nucleic Acids Res*. **31**, 4385-4390.

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M. & Croce, C.M. 2004 Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. USA* **101**, 2999-3004.

Chang, J. & Taylor, J. M. 2003 Susceptibility of Human Hepatitis Delta Virus RNAs to Small Interfering RNA Action. *J. Virology* 77, 9728-9731.

Cognet, J. 2006 Personal communication.

Comet, J. P., Aude, J. C., Glémet, E., Hénaut, A., Risler, J. L., Slonimski, P. P., & Codani, J. J. 1999 Significance of Z-value statistic of Smith-Waterman scores for protein alignments. *Computers and Chemistry* **23**, 317–331.

Dale, T., Smith, R. & Serra, M. J. 2000 A test of the model to predict unusually stable RNA hairpin loop stability. *RNA* **6**, 608-615.

de Duve, C. 2002 *Life Evolving*. Oxford UK: Oxford University Press.

Demongeot, J. 1978 Sur la possibilité de considérer le code génétique comme un code à enchaînement. *Revue de Biomaths* **62**, 61-66.

Demongeot, J. & Besson, J. 1983 Code génétique et codes à enchaînement I. *C. R. Acad. Sc.* **296**, 807-810.

Demongeot, J. & Besson, J. 1996 Genetic code and cyclic codes II' *C. R. Acad. Sc.* **319**, 520-528.

Demongeot, J., Aracena, J., Ben Lamine, S., Mermet, M. A. & Cohen, O. 2000 Hot spots in chromosomal breakage: from description to etiology. In: *Comparative Genomics*, (ed. D. Sankoff & J.H. Nadeau), pp. 71-85. Amsterdam: Kluwer.

Demongeot, J., Elena, A. & Weil, G. 2006 Potential automata. Application to the genetic code III. *Comptes Rendus Biologies* **329**, 953-962.

Demongeot, J. & Moreira, A. 2007 A circular Hamming distance, circular Gumbel distribution, RNA relics and primitive genome. In: *AINA' 07*, pp. 719-726. Piscataway: IEEE Proceedings.

Demongeot, J. & Moreira, A. 2007 A circular RNA at the origin of life. *J. Theor. Biol.* **249**, 314-324.

Demongeot, J., Morvan, M. & S. Sené, S. 2008 Robustness of Dynamical Systems Attraction Basins Against State Perturbations: Theoretical Protocol and Application in Systems Biology. In: *IEEE ARES-CISIS' 08 & IIBM' 08*, pp. 675-681. Piscataway: IEEE Proceedings.

Demongeot, J., Glade, N. & Moreira, A. 2008 Evolution and RNA Relics. A Systems Biology View. *Acta Biotheoretica* **56**, 5-25.

Demongeot, J., Goles, E. & Sené, S. (accepted) Regulatory networks analysis: robustness in artificial regulatory networks. In: *AINA' 09 & BLSMC' 09*. Piscataway: IEEE Proceedings.

di Giulio, M. 1992 On the origin of the tRNA molecule. *J. Theor. Biol.* **159**, 199-214.

di Giulio, M. 1997 On the origin of the genetic code. *J. Theor. Biol.* **187**, pp. 573-581.

Eigen, M. 1971 Molekuläre Selbstorganisation und Evolution. *Naturwissenschaften* **58**, 465-523.

Eigen, M., Gardiner, W., Schuster, P. and Winkler-Oswatitsch, R. 1981 The origin of genetic information. *Sci. Am*. **244**, 88-92.

Elena, A., Ben-Amor, H., Glade, N. & Demongeot , J. 2008 Motifs in regulatory networks and their structural robustness. In: *IEEE BIBE' 08*, pp. 234-242. Piscataway: IEEE Proceedings.

Elena, A. & Demongeot, J. 2008 Interaction motifs in regulatory networks and structural robustness. In: *IEEE ARES-CISIS' 08 & IIBM' 08*, pp. 682-686. Piscataway: IEEE Proceedings.

Elena, A. 2009 *Robustesse des réseaux d'automates booléens à seuil aux modes d'itération*. PhD Thesis Grenoble: University Joseph Fourier.

Ertem, G. 2004 Montmorillonite, Oligonucleotides, RNA and Origin of Life. *Orig. Life* **34**, 549-570.

Faraut, T. & Demongeot, J. 2000 Benefits of a model of segregation for the understanding of chromosomal evolution. In : *Comparative Genomics* (ed. D. Sankoff & J.H. Nadeau), pp. 13-17. Amsterdam: Kluwer.

Figureau, A. & Pouzet, M. 1984 Genetic code and optimal resistance to the effects of mutations *Orig. Life* **14**, 579-588.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. 1986 Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* **83**, 9373-9377.

Gatlin, L. L. 1968 The information content of DNA. II. *J. Theor. Biol.* **18** 181-194.

Gilbert, W. 1986 The RNA world. *Nature*, **319**, 618.

Gilis, D., Massar, S., Cerf, N. and Rooman, M. 2001 Optimality of the genetic code withrespect to protein stability and amino-acids frequencies. *Genome Biology* **2**, 1-12.

Glansdorff, P. & Prigogine, I. 1971 *Structure, Stabilité et Fluctuations*. Paris: Masson.

Gottesman, S. 2005 μ for microbes: noncoding regulatory RNA in bacteria. *Trends in Gen*. **21**, 399-404.

Gumbel, E. J. 1958 *Statistics of extremes*. Columbia: Columbia University Press.

Hartman, H. 1984 Speculations on the evolution of the genetic code III. *Orig. Life* **14**, 643-648.

Hayes, B. 1998 The invention of the genetic code. *American Scientist* **86**, 8-14.

He, M., Petoukhov, S. & Ricci, P. 2004 Genetic code, Hamming distance and stochastic matrices. *Bull. Math. Biology* **66**, 1405-1421.

Hill, C. A., Kafatos, F., Stansfield, S. K. & Collins, F. H. 2005 Arthropod-borne diseases: vector control in the genomics era. *Nature Reviews Microbiology* **3**, 262-268.

Hill, T. P. & Kertz, R. P. 1981 Additive comparisons of stop rule and supremum expectations of uniformly bounded independent random variables. *Proc. AMS* **83**, 582-585.

Hobish, M. K., Wickramasinghe, N. S. M. D. & Ponnamperuma, C. 1995 Direct interaction between amino-acids and nucleotides as a possible physico-chemical basis for the origin of the genetic code. *Advances in Space Research* **15**, 365-375.

Holt, R. A. *et al.* 2002 The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129-149.

Hopfield, J. 1978 Origin of the genetic code: A testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. *Proc. Natl. Acad. Sci. USA* **75**, 4334-4338.

Hornos, J. E. M., Braggion, L., Magini, M. & Forger, M. 2004 Symmetry preservation in the evolution of the genetic code. *Life* **56**, 125-130.

Hughes, J. & Coffin, J. 2004 Human endogeneous retrovirus K solo-LTR formation and insertional polymorphism: implications for human and viral evolution. *Proc. Natl. Acad. Sci. USA* **101**, 1668-1672.

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M. & Sarnow, P. 2005 Modulation of hepatitis C virus RNA abundance by a liver-specific miR. *Science* **309**, 1577-1581.

Jouanneau, J. & Larsen C. J. 2007 Les microARN : un « bras armé » du suppresseur de tumeur P53. *Bull. Cancer* **94**, 634-635.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus K. C., Stoffel, M. & Rajewsky, N. 2005 Combinatorial microRNA target predictions. *Nature Genetics* **37**, 495-500.

Knight, R. & Landweber, L. 1998 Rhyme and reason: RNA-arginine interactions and the genetic code. *Chemistry & Biology* **5**, 215-220.

Labouygues, J. M. 1976 New mathematical model of genetic code with passive resistance to mutations or buccion and complementary dynamic protective mechanisms against noise at genome level. *Agressologie* **17**, 329-335.

Laurent, M. 1996 Prion diseases and the «protein only» hypothesis: a theoretical dynamic study. *Biochem. J.* **318**, 35-39.

Lewin, B. 2007 *Genes*. Boston: Jones & Bartlett Pub..

Magini, M. & Hornos, J. E. M. 2003 A dynamical system for the algebraic approach to thegenetic code. *Brazilian J. Physics* **33**, 825-830.

Majerfeld, I., Puthenvedu, D. & Yarus, M. 2005 RNA affinity for molecular L-histidine; Genetic code origins. *J. Mol. Evol*. **61**, 226-235.

Markoff, L. 2004 5' and 3' noncoding regions in flavivirus RNA. *Adv. in Virus Research* **59**, 177-228.
Mello, C. G. & Conte, D. C. 2004 Revealing the world of RNA interference. *Nature*, **431**, 338-342
Mitaku, S., Hirokawa, T. & Tsuji, T. 2002 Amphiphilicity index of polar amino-acids as an aid in the characterization of AA preference at membrane-water interfaces. *Bioinformatics* **18**, 608-616.
Moreira, A. 2003 *PhD Thesis*. Santiago de Chile: Universidad de Chili.
Moreira, A. 2004 Genetic algorithms for the imitation of genomic styles in protein backtranslation. *Theoretical Computer Science* **322**, 297-312.
Nene, V. *et al.* 2007 Genome Sequence of *Aedes aegypti*, a Major Arbovirus Vector. *Science* **316**, 1703-1704.
Oliva, R., Cavallo, L. & Tramontano, A. 2006 Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in tRNA tertiary interactions. *Nucl. Acids Research* **34**, 865-879.
Pasqual, N., Gallagher, M., Aude-Garcia, C., Loiodice, M., Thuderoz, F., Demongeot, J., Ceredig, R., Marche, P. & Jouvin-Marche, E. 2002 Quantitative and Qualitative Changes in ADV-AJ Rearrangements During Mouse Thymocytes Differentiation: Implication For a Limited TCR ALPHA Chain Repertoire. *J. Exper. Medicine* **196**, 1163-1174.
Poole, A., Jeffares, D. & Penny, D. 1998 The path from the RNA world. *J. Mol. Evol.* **46**, 1-17.
Rodin, S., Ohno, S. & Rodin, A. 1993 tRNAs with complementary anticodons: Could they reflect early evolution of discriminative genetic code adaptors? *Proc. Natl. Acad. Sci. USA* **90**, 4723-4727.
Ruskey, F. & Sawada, J. 2000 Generating necklaces and strings with forbidden substrings. *Lecture Notes in Computer Sciences* **1858**, 330-339.
Sciarrino, A. 2003 A mathematical model accounting for the organization in multiplets of the genetic code. *BioSystems* **69**, 1-13.
Sené, S. 2008 *Influence des conditions de bord dans les réseaux d'automates booléens à seuil et application à la biologie*. PhD thesis Grenoble: University Joseph Fourier.
Shapiro, R. 2007 A Simpler Origin of Life. *Scientific American* **296**, 46-53.
Shimizu, M. 1995 Specific aminoacetylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. *J. Biochem. (Tokyo)* **117**, 23-26.
Shmaliy, Y. S. 2005 Von Mises/Tikhonov-based distributions for systems with differential phase measurement. *Signal Processing* **85**, 693-703.
Swanson, R. 1984 A Unifying Concept for the Amino Acid Code. *Bull. Math. Biology* **46**, 187-203.
Szathmary, E. & Maynard Smith, J. 1997 From replicators to reproducers: the first major transitions leading to life. *J. Theor. Biol.* **187**, 555-571.
Thom, R. 1972 *Stabilité structurelle et Morphogenèse*. Benjamin : New York.
Toulokhonov, I., Artsimovitch, I. & Landick, R. 2001 Allosteric control of RNA-polymerase by a site that contacts nascent RNA hairpins. *Science* **292**, 730-733.
Trifonov, E. & Sussman, J. 1980 The pitch of chromatin DNA is reflected in nucleotide sequence. *Proc. Natl. Acad. Sci. USA* **77**, 3816-3820.
Trifonov, E. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139-151.
Trinquier, G. & Sanejouand, Y. H. 1998 Which effective property of amino-acids is best preserved by the genetic code? *Protein Engineering* **11**, 153-169.
Vinga, S. & Almeida, J. S. 2004 Rényi continuous entropy of DNA sequences. *J. Theor. Biol.* **231**, 377-388.
Wang, L. & Schultz, P. 2005 Expanding the genetic code. *Angew. Chem. Int. Ed*. **44**, 34-66.
Wang, Y. L., Bao, J., Sun Y. & Yang, J. 2006 Energy and structural analysis of double nucleic acid triplets. *J. Theor. Biol*. **238**, 85-103.
Yarus, M. 2000 RNA-ligand chemistry: a testable source for the genetic code. *RNA* **6**, 475-484.
Zhang, B. H., Pan, X. P., Wang, Q. L., Cobb, G. & Anderson, T. 2005 Identification and characterization of new plant μRNAs using EST analysis. *Cell Research* **15**, 336-360.