doi:10.2533/chimia.2014.629 Chimia 68 (2014) 629–632

Predicting the Genes Regulated by MicroRNAs *via* Binding Sites in the 3' Untranslated and Coding Regions

Jiří Vaníček*

Abstract: MicroRNAs form one of the groups of small noncoding RNA molecules that have completely changed our understanding of gene regulatory networks. Because microRNAs have been discovered only relatively recently, most of their functions remain unknown, providing a challenge to both experiment and theory. I review several computational approaches pursued in our group to answer this challenge. In particular, I show that a few rather simple ideas can go a long way in predicting accurately genes regulated by microRNAs *via* binding sites both in the coding and 3' untranslated regions (3'UTRs). Finally, I mention briefly several applications, including two collaborations with experimental groups, which have shed new light on the latency and reactivation of herpesviruses, and on the maturation of red blood cells.

Keywords: Conservation · MicroRNA · MicroRNA target · Target prediction algorithm · RNA secondary structure

1. Introduction

MicroRNAs (miRNAs) are small, 20–23 nucleotides (nt) long noncoding RNA molecules that function as negative gene regulators by binding to target messenger RNAs (mRNAs) and either degrading them or repressing their translation into protein.[1] MicroRNAs play an important role in cell differentiation, development, cancer, and other biological processes in species ranging from viruses to humans.[2] While more than thousand miRNAs are encoded in the human genome, most of their target genes remain unknown, especially since direct validation of the functionality of individual miRNA-target pairs is rather tedious and expensive.

Fortunately, several high-throughput techniques can be used for indirect validation, including high-throughput proteomics methods,[3,4] in which protein levels of thousands of genes are monitored upon overexpression of a miRNA of interest, various methods for quantifying mRNA expression (microarrays, RNA-Seq, *etc.*), or high-throughput cross-linking immunoprecipitation (CLIP),[5,6] which provides coordinates of thousands of mRNA regions bound by the Argonaute-miRNA ribonucleoprotein complexes. However, since these techniques are indirect and even more expensive than direct validation, they are usually combined with computational approaches to focus the search for functional miRNA targets, and this is, in fact, one of the research interests in our group.

2. Predicting MicroRNA Targets in the 3'Untranslated Region (3'UTR)

The challenge faced by miRNA target prediction algorithms becomes obvious by comparing miRNAs with closely related small interfering RNAs (siRNAs): Predicting targets of siRNAs is relatively straightforward since these ~22 nt long RNA molecules must be fully complementary to their target RNA sequences; one can simply scan the human genome for such complementary sequences and be rather confident that those found will be functional. Indeed, assuming that each of the four (A, C, G, T) nucleotides is equally likely and ignoring genomic repeats, one finds that a specific sequence of 22 nt would appear by chance on average only once in a sequence of $4^{22} \approx 1.8 \times 10^{13}$ nt. In other words, it would rarely show up by chance in the human genome whose length is approximately 3×10^9 nt.

Searching for targets of miRNAs, which are also ~22 nt long, is more difficult since they require complementarity only within a so-called 'seed' region, *i.e.* a region of six to seven consecutive nucleotides starting at position 2 of the miRNA (see Fig. 1). Assuming a typical human 3'UTR length of about 1000 nt, a simple back-of-the-envelope calculation shows that the naïve algorithm requiring complementarity to the 6-mer seed would predict that the 3'UTR of every fourth gene is regulated by each miRNA. The algorithm that works so well for siRNAs fails for miRNAs; it is therefore necessary to use additional criteria to predict miRNA targets precisely.

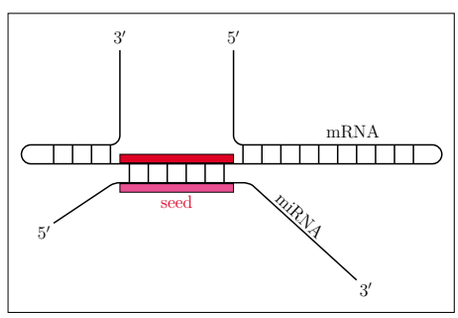

Fig. 1. Simplified representation of the binding between a miRNA and its mRNA target. The most important interaction occurs along a seed region of six to seven nucleotides.

Among the many criteria used to improve upon the naïve algorithm are conservation and accessibility of the binding site, hybridization energy between the miRNA and target mRNA, and various empirical rules based on training sets constructed

*Correspondence: Prof. Dr. J. Vaníček
Laboratory of Theoretical Physical Chemistry
Institut des Sciences et Ingénierie Chimiques
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Avenue Forel, CH-1015 Lausanne
Tel.: +41 21 693 47 36
E-mail: jiri.vanicek@epfl.ch

from experimentally validated miRNA-target pairs. It is not surprising that conservation of the binding site among several species (Fig. 2a) has been used successfully in many miRNA target prediction algorithms.[7–13] The justification is simple – sequences that have been conserved during evolution are much more likely to be functional. Another criterion uses the accessibility of the binding site and is based on the assumption that miRNAs are more likely to bind to accessible segments of mRNA, *i.e.* single-stranded regions of the secondary RNA structure (Fig. 2b). However, it is not necessary that the full 'seed match' (*i.e.* mRNA sequence complementary to the seed) be accessible at all times; four nucleotides are often sufficient to nucleate the binding (see Section 4 for more details).[14] Interestingly, in one of our papers,[15] we showed that some 'obviously good' criteria such as the hybridization energy between the miRNA and mRNA are poor for ranking predictions. One may object that the free energy of the interaction should also include the 'opening energy', *i.e.* the energy required to unwind the secondary mRNA structure to make it accessible for binding by the miRNA. Yet, we showed that even the 'total' free energy, computed as a signed sum of the opening and hybridization energies does not significantly improve the precision of target predictions beyond the simple algorithm requiring complementarity to a 7-mer seed.[15]

As a consequence, thermodynamic criteria are not used for ranking predictions in our algorithms.[13,15,16] Instead, we rank the predictions according to the over-representation of the seed match compared to a random background.[12] This is motivated by the assumption that functionally interacting miRNA-mRNA pairs have co-evolved, implying that the number of complementary sites present in regulated genes should be higher than the corresponding number appearing by chance in unregulated genes. In the 3'UTR, which does not contain as much information as the coding region, we model this random background by the second-order Markov model based on the nucleotide and dinucleotide composition of the given 3'UTR sequence.[12] The dinucleotide composition is important since adjacent CpG (-cytosine-phosphate-guanine-) pairs are under-represented in comparison with what would be expected from the C and G content alone.

In addition to this ranking criterion, we also use conservation and accessibility of the binding site. However, in contrast to algorithms that rank their predictions according to the extent of conservation and accessibility, these two criteria are in our algorithms only used as filters – we require a minimum amount of conservation and accessibility, but then rank the predictions according to over-representation.

To summarize, for each miRNA-3'UTR pair, let us denote by c the number of conserved and accessible n-mer seed matches in the 3'UTR and by l the total number of conserved and accessible n-mers in the 3'UTR. We first use the Markov model to compute the probability p to find an n-mer seed match by chance at any particular position of a random 3'UTR,[12,16] and then evaluate the probability P_{SH} (the single-hypothesis P-value) to find, by chance, at least c conserved and accessible seed matches in a random 3'UTR containing l conserved and accessible n-mers:

$$P_{SH} = \sum_{i=c}^{l-n+1} \binom{l-n+1}{i} p^i (1-p)^{l-n+1-i} \quad (1)$$

Lower P_{SH} values (*i.e.* stronger over-representation) imply a higher likelihood of co-evolution between the miRNA and the seed match, and hence, a higher likelihood of biological functionality. (In the definition of P_{SH}, 'accessible site' stands, more precisely, for what we call a 'partially accessible site', *i.e.* an n-mer containing at least one 4-mer that appears in the single-stranded region of at least 20% of secondary structures from the Boltzmann ensemble of secondary structures of the given 3'UTR.[15] Similarly, 'conserved n-mer' typically stands for an n-mer that is conserved in the human, chimp, rhesus, and mouse.[13])

There are several accurate prediction algorithms for predicting conserved targets of conserved miRNAs; these include, *e.g.* DIANA-microT,[7] PicTar,[8] TargetScan,[9,10] ElMMo,[11] the algorithm of Robins and Press,[12] and our algorithm PACCMIT.[13] However, we have shown[15] that as soon as the conservation requirement is dropped, *e.g.* when one is interested in species-specific targets, the precision of existing algorithms decreases drastically. Yet, we have shown that our simple algorithm based on accessibility and over-representation (PACMIT)[15] for predicting nonconserved targets remains accurate in such situations.[13] The algorithm was validated[13,15] on four very different and complementary datasets: the high-throughput (1) mRNA expression and (2) proteomics datasets of Selbach *et al.*[3] and Baek *et al.*,[4] in which protein and RNA levels of several thousand genes were monitored upon overexpression (or knockdown) of several miRNAs, (3) the high-throughput cross-linking immunoprecipitation (CLIP) datasets of Hafner *et al.*[5] and Chi *et al.*,[6] providing coordinates of thousands of mRNA regions bound by the Argonaute-miRNA ribonucleoprotein complexes, and (4) a dataset of Kertesz *et al.*[17] using luciferase reporter assays.

3. Predicting MicroRNA Targets in the Coding Region

Predicting miRNA targets in the coding region is much more difficult than in the 3'UTR for an obvious reason – the coding region contains another, biologically

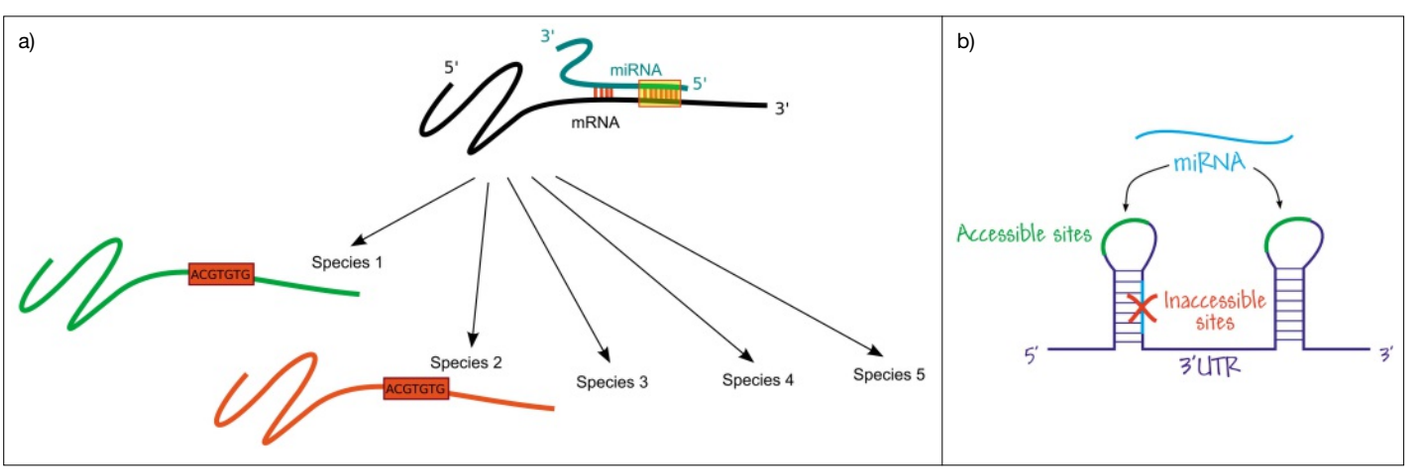

Fig. 2. (a) Conservation of miRNA binding site among several species. (b) Accessibility of miRNA binding site in the target mRNA.

much more important signal – the code for the protein. In other words, the signal that we seek (miRNA binding sites) is much weaker than the background (code for the protein); it is like looking for a needle in a haystack.

To predict miRNA targets in the coding region, we employ the same main idea as before – we look for binding sites that are over-represented compared to a random background, but now the background is constructed differently. The random background should not destroy the most important function of the coding region – the code for the protein. Fortunately, due to the redundancy of the genetic code such a background exists: since there are 4^3 = 64 codons (nucleotide 3-mers) and only 20 amino acids, there exist synonymous codons encoding the same amino acid. The desired background is constructed by replacing each codon in the real sequence with a randomly chosen synonymous codon. A less known fact adds another constraint to the choice of an optimal background. Namely, codon usage differs in different species and sometimes even in different genes in a single species, in order to fine-tune translational efficiency. In order to construct a background preserving only the codon usage, one would shuffle the codons of each coding sequence; the background satisfying the two constraints simultaneously, *i.e.* preserving both the amino acid sequence and codon usage, is constructed by shuffling only the synonymous codons (see Fig. 3).[18,19]

To summarize, let c again denote the number of seed matches of a given miRNA in the coding sequence of a given mRNA. In our PACCMIT-CDS algorithm,[20] the miRNA-mRNA pairs are ranked according to the probability that at least c seed matches would appear in the randomly generated coding sequence, in practice computed as the single-hypothesis P-value

$$P_{\mathrm{SH}} = \frac{N_c}{N_{\mathrm{total}}}, \qquad (2)$$

where N_c is the number of random sequences with at least c seed matches and N_{total} is the total number of random sequences. The random sequences are generated with the Durstenfeld modification of the Fisher-Yates algorithm for generating random permutations ('shuffles') of an array with N elements.[21] While the P-values of the top predictions are below 10^{-8}, evaluating all P-values for the whole genome with this resolution would take hundreds of years on a supercomputer, and, moreover, is unnecessary. Instead, the calculation was enormously accelerated by a gradual refinement of P-values: the high P-values were only evaluated with resolution 10^{-3},

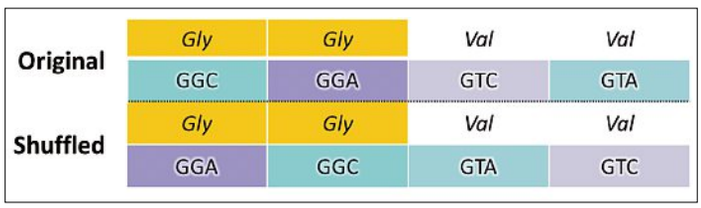

Fig. 3. Construction of a random background sequence required for predicting miRNA targets in the coding region with PACCMIT-CDS. The background is obtained by shuffling synonymous codons in order to preserve both the amino acid sequence encoded by the gene and the codon usage required for efficient translation.

and lower P-values for more significant miRNA-mRNA pairs with higher resolution, as needed.

By analyzing the coding sequences of the human genome, we demonstrated that the background preserving both the amino acid sequence and codon usage reduces the noise in target predictions more than backgrounds preserving none or only one of the two constraints. Moreover, we showed that considering conservation of the seed matches among related species increases enormously the signal-to-noise of the predictions. Note that the algorithm as well as the above analysis, including estimates of the signal-to-noise ratio, require only the well-known coding sequences, but – unlike other algorithms – PACCMIT-CDS does not rely on any training sets using proteomics, mRNA expression, or other experimental data.

We have, nevertheless, again validated the predictions of PACCMIT-CDS independently, in two different ways, using the large throughput (1) proteomics dataset of Selbach *et al.*[3] and (2) cross-linking immunoprecipitation (CLIP) dataset of Hafner *et al.*[5] In order to separate the well-established miRNA function *via* binding within the 3'UTR, we had to construct datasets with binding sites with seed matches only within the coding region. The proteomics and cross-linking immunoprecipitation datasets confirmed that the top predictions of PACCMIT-CDS are genes with more downregulated protein levels and genes with more sites bound by the Argonaute-miRNA ribonucleoprotein complexes within the coding region, respectively.

4. Nucleation of MicroRNA–Target Binding

It is agreed that the seed region is the most important determinant of the binding between the miRNA and its target. However, it is not clear how this seed binding is nucleated. Does the binding start from the 3' or 5' end of the miRNA? (Fig. 4). Ray Marín, a former Ph.D. student in my group, answered[22] this question *in silico* by a careful analysis of the high-throughput cross-linking immunoprecipitation datasets.[5,6] In particular, he compared the accessibility of RNA segments of length 1 to 7 nucleotides at various locations within bound seed matches with the accessibility of the corresponding regions in unbound seed matches. The conclusion of this analysis was that in contrast to unbound sites, in bound sites the 3' end of the seed match is much more accessible than the 5' end, a result that was recently confirmed experimentally by Wan *et al.*[23] The miRNA-mRNA binding is therefore much more likely to start from the 5' end of the miRNA seed, or, equivalently, from the 3' end of the seed match within the target mRNA.[22]

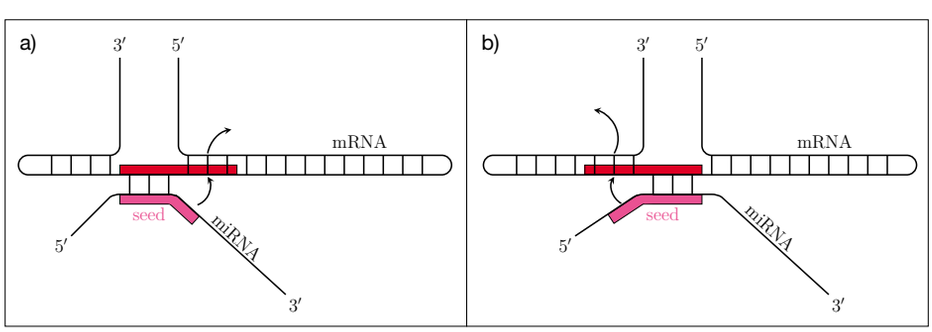

Fig. 4. Two alternative hypotheses about the nucleation of the miRNA-mRNA binding. (a) Binding starts at the 5' end of the seed (corresponding to the 3' end of the seed match). (b) Binding starts at the 3' end of the seed (corresponding to the 5' end of the seed match). By analyzing the accessibility of binding sites detected in cross-linking immunoprecipitation (CLIP) experiments, we showed that mechanism (a) is much more likely.

5. MicroRNA Functions in Herpesvirus Infection and in the Maturation of Red Blood Cells

In 2004 miRNAs were discovered in herpesviruses. This was very exciting because viruses, despite their very small genomes, still elude our understanding of their complicated life cycle, and because miRNAs, thanks to their small genomic size, are perfect candidates for storing regulatory information in the small viral genomes. (I should note that among viruses, herpesviruses are giants since their genomes contain 100–200 protein coding genes, which is, however, still two orders of magnitude smaller than the number of genes encoded in the human genome.) Moreover, herpesviruses have a very interesting life cycle, consisting of lytic and latent infections, which are very familiar in the case of the best known representative of human herpesviruses, the herpes simplex virus 1 (HSV-1): In the lytic infection, the virus replicates, causing the cold sores, and most of the viral genes are expressed; as a result the virus becomes visible to the immune system, which eventually eliminates the virus, except for a few virion particles that escape to the trigeminal ganglia, where they establish latent infection. In the latent infection, which can last for many years, the virus is dormant and all but invisible to the immune system. When the host becomes exposed to biological stress, such as heat shock, sunburn, infection, *etc.*, the virus, which depends on the well-being of its host, decides that it is time to find another host, and reactivates, starting the lytic infection again. In the case of varicella zoster virus (VZV), the primary infection – varicella (chicken pox), was even for a long time believed to be caused by a different pathogen than the reactivated infection – zoster (shingles).

Interestingly, in all herpesviruses there is one or at most a handful of so-called *immediate-early genes*, expression of which alone can cause reactivation. In the lytic expression cascade, these immediate-early genes activate *early genes*, responsible for DNA replication, which turns on the expression of the *late*, structural *genes*, comprising the majority of the genome.

Using our miRNA target prediction algorithms, we predicted that a viral miRNA hcmv-miR-112-1 targets the immediate early protein 1 (IE1) mRNA of the human cytomegalovirus,[16] and that human hsa-miR-200 miRNA family members target the 3'UTR of the immediate early protein 2 (IE2) of this virus,[24] thus helping either to maintain latency or to enter latency from the lytic infection. These predictions were confirmed experimentally by Eain Murphy, Tom Shenk, and Christine O'Connor from Princeton University and from the Lerner Research Institute.[16,24] Although the repression by miRNAs is only partial, targeting the immediate-early genes appears to be the optimal way to repress the lytic expression cascade.

In another collaboration, with Didier Trono's Laboratory of Virology and Genetics at EPFL, we identified a group of genes playing a critical role in the maturation of red blood cells (erythropoiesis). Building on preliminary miRNA and mRNA expression data obtained in Didier Trono's laboratory, we predicted theoretically which candidate genes were the most likely to be repressed by specific miRNAs. These predictions included genes mediating mitophagy – a process involving the elimination of mitochondria from red blood cells and maximizing the cells' ability to carry hemoglobin – and were confirmed experimentally in Trono's laboratory.[25]

6. Conclusion

In conclusion, I have described several simple, yet effective algorithms for predicting targets of both conserved[13,16,20,22] and species-specific[13,15,20,22] miRNAs, with binding sites both in the coding[20] and 3' untranslated regions.[13,15,16,22] In all of our algorithms we attempt to avoid empirical parameters and training sets; the only indispensable parameters are the length and location of the seed region (7 nt starting at position 2 of the miRNA), which are widely accepted parameters in most precise algorithms. We first select only the seed matches that are partially accessible and/or conserved in several species; the predictions with binding sites both in the coding and 3' untranslated regions are then ranked according to the over-representation of the conserved and/or accessible seed matches relative to an appropriate random background. At present, we work on removing the remaining parameters, such as the thresholds for conservation and accessibility, and on extending the algorithms to include the information from high throughput RNA expression experiments.

Acknowledgments

I would like to thank Ray Marín and Miroslav Šulc, who worked in my group on the theoretical and computational aspects of the projects discussed here, Harlan Robins from Fred Hutchinson Cancer Research Center and Arnold Levine from the Institute for Advanced Study for introducing me to the miRNAs and herpesviruses, our experimental collaborators Christine O'Connor and Eain Murphy from the Lerner Research Institute at the Cleveland Clinic, Tom Shenk from Princeton University, and Isabelle Barde and Didier Trono from EPFL. This work was supported by EPFL. Finally I thank Ray Marín, Miroslav Šulc, and Eduardo Zambrano for providing the figures.

Received: July 13, 2014

[1] B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **2005**, *120*, 15.
[2] W. Filipowicz, S. N. Bhattacharyya, N. Sonenberg, *Nat. Rev. Genet.* **2008**, *9*, 102.
[3] M. Selbach, B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, N. Rajewsky, *Nature* **2008**, *455*, 58.
[4] D. Baek, J. Villen, C. Shin, F. Camargo, S. Gygi, D. Bartel, *Nature* **2008**, *455*, 64.
[5] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, *Cell* **2010**, *141*, 129.
[6] S. W. Chi, J. B. Zang, A. Mele, R. B. Darnell, *Nature* **2009**, *460*, 479.
[7] M. Maragkakis, P. Alexiou, G. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, A. Hatzigeorgiou, *BMC Bioinformatics* **2009**, *10*, 295.
[8] A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, N. Rajewsky, *Nat. Genet.* **2005**, *37*, 495.
[9] R. C. Friedman, K. K.-H. Farh, C. B. Burge, D. P. Bartel, *Genome Res.* **2009**, *19*, 92.
[10] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, D. P. Bartel, *Mol. Cell* **2007**, *27*, 91.
[11] D. Gaidatzis, E. van Nimwegen, J. Hausser, M. Zavolan, *BMC Bioinformatics* **2007**, *8*, 69.
[12] H. Robins, W. H. Press, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15557.
[13] R. M. Marín, J. Vaníček, *PloS one* **2012**, *7*, e32208.
[14] H. Robins, Y. Li, R. W. Padgett, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4006.
[15] R. M. Marín, J. Vaníček, *Nucl. Acids Res.* **2011**, *39*, 19.
[16] E. Murphy, J. Vaníček, H. Robins, T. Shenk, A. J. Levine, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5453.
[17] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, E. Segal, *Nat. Genet.* **2007**, *39*, 1278.
[18] A. Fuglsang, *Biochem. Bioph. Res. Co*. **2004**, *316*, 755.
[19] H. Robins, M. Krasnitz, H. Barak, A. J. Levine, *J. Bacteriol.* **2005**, *187*, 8370.
[20] R. M. Marín, M. Sulc, J. Vaníček, *RNA* **2013**, *19*, 467.
[21] D. E. Knuth, 'The Art of Computer Programming', Vol. 2: 'Seminumerical algorithms', 3rd ed., Addison-Wesley: Boston, **1997**.
[22] R. M. Marín, F. Voellmy, T. von Erlach, J. Vaníček, *RNA* **2012**, *18*, 1760.
[23] Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, H. Y. Chang, *Nature* **2014**, *505*, 706.
[24] C. M. O'Connor, J. Vaníček, E. A. Murphy, *J. Virology* **2014**, *88*, 5524.
[25] I. Barde, B. Rauwel, R. M. Marín-Florez, A. Corsinotti, E. Laurenti, S. Verp, S. Offner, J. Marquis, A. Kapopoulou, J. Vaníček, D. Trono, *Science* **2013**, *340*, 350.