Secondary structures in the coding regions of mRNAs: literature survey and comparison of prediction methods

Aleksandra Vasileva[3], Michael Kiening[1], Dmitrij Frishman[1,2,3]

[1]Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany, d.frishman@wzw.tum.de

[2]Institute for Bioinformatics and Systems Biology, HMGU German Research Center for Environmental Health, Neuherberg, Germany

[3]Department of Bioinformatics, St. Petersburg Polytechnic University, St. Petersburg, 195251, Russia, aleksandra.vasileva.spbpu@gmail.com

mRNA molecules contain three main domains – 5'UTR (untranslated region), CDS (coding region), and 3'UTR. RNA secondary structures formed both in UTRs and CDS are involved in a variety of regulatory functions [1]. However, while a large variety of structural elements in UTRs have been thoroughly documented, structural properties of coding regions remain poorly understood. For example, IRESs (internal ribosome entry sites) provide cap-independent or internal translation initiation by recruiting ribosomes for protein synthesis, thus obviating the requirement for the 7-methyl-guanosine moiety at the 5' terminus of the mRNA [2]. We found that out of 149 viral and 183 eukaryotic IRES structures mentioned in the literature only 11 and 16, respectively, are located in coding regions.

As long as the experimental determination of RNA structures remains difficult computational prediction methods, in spite of their insufficient accuracy, remain the main option for elucidating the structure of RNA molecules. Existing prediction algorithms are generally trained on a set of non-coding RNAs and their performance on coding regions is largely unknown.

We sought to evaluate the accuracy of prediction algorithms on a carefully curated set of RNA structures located in mRNA coding region. We included in our dataset only those structural elements that have been experimentally confirmed by structure probing (for example, by the 2'-hydroxyl acylation and primer extension approach (SHAPE) [3]) and possess experimentally documented function. Currently only 7 IRESs and 1 CRE element (cis-acting replication element) [4] fulfil these requirements.

We next used this small dataset to assess the performance of several widely used computational techniques, including those based on the Zuker-Stiegler algorithm for computing the minimal free energy (MFE) (RNAfold [5], mfold [6]) as well as those employing pairwise co-variation in multiple alignments (RNAalifold [5], RNAz[7]).

The crucial step is to compare predicted structures with the real ones. One of the options in this case is to compute the distance between two RNA structures. A popular algorithm to perform this task based on the edit distance between trees representing RNA secondary structure elements, RNA distance [5], is included in Vienna RNA package. Recently several new algorithms for computing the distance between structures appeared [8]. Work is in progress to assess the performance of these tools using our dataset.

Another possibility would be to use abstract shapes [9]. The main idea is to partition the folding space of structures into different classes of structures called shapes. These shapes can be easily obtained from the commonly used dot-bracket notation, as shown in Fig. 1.

(a) ..(((.((..(((....))).(((.....)))))))).

(b) _[_[_[_]_[_]]]_

Fig. 1. Two alternative representations of an RNA secondary structure. a. Dot-bracket representation; dots symbolize unpaired bases and matching parentheses symbolize base pairs. b. Shape representation of the same structure; underscores symbolize unpaired regions and pairs of square brackets symbolize stacking regions.

Our goal is thus to find a more adequate method for comparing RNA structures.

According to our visual analyses of predicted and real structures all predictors perform poorly (Fig. 2 and 3).

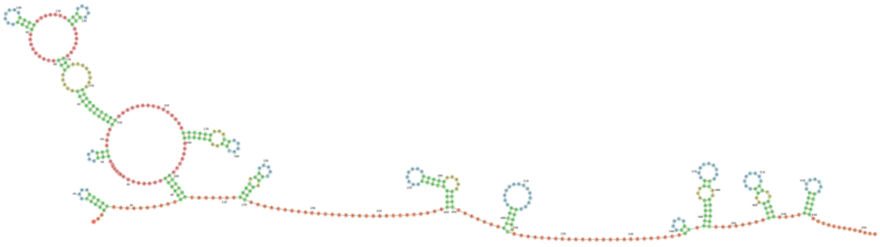

Fig. 2. Experimentally confirmed structure of the HIV-1 IRES.

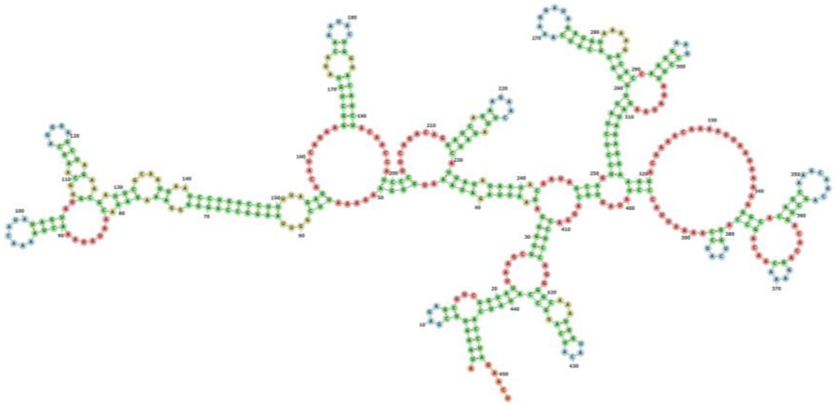

Fig. 3. Structure of the HIV-1 IRES predicted by RNAz.

RNAz usually provides completely wrong structures while RNAfold and mfold often correctly reconstruct the most energetically stable parts of the structure. Interestingly, mfold produced an essentially perfect prediction of the IRES in the 5’ UTR of the fibroblast growth factor 1 (FGF1) (Fig. 4). This IRES is a rather small and simple structure, which seems to be very stable.

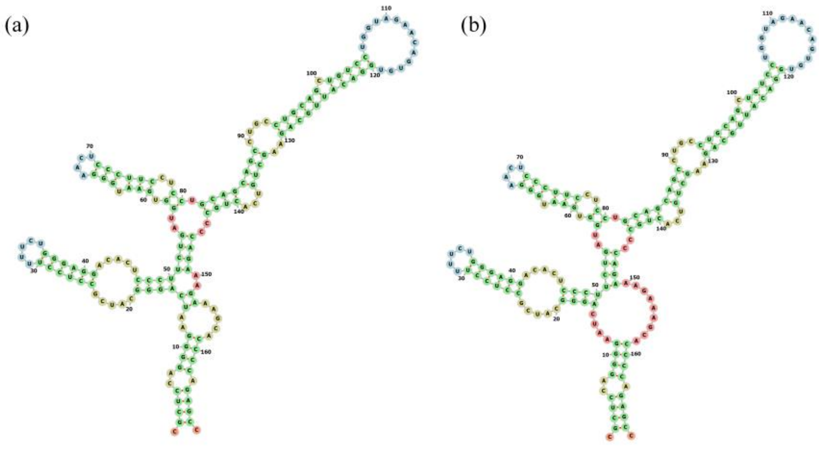

Fig. 4. IRES structures in the 5’UTR of FGF1. a. Experimentally verified structure. b. mfold prediction.

We are currently extending the dataset of experimentally verified coding RNA structures by further literature analyses. In collaboration with the University of Vienna these data will be used to retrain the RNAz method for predicting structurally conserved and thermodynamically stable RNA secondary structures both in the coding and non-coding regions of viruses and cellular organisms.

1. P.C.Bevilacqua, J.M.Blose (2008) Structures, kinetics, thermodynamics, and biological functions of RNA hairpins, *Annual Review of Physical Chemistry*, **59:** 79–103.

2. C.U.Hellen, P.Sarnow (2001) Internal ribosome entry sites in eukaryotic mRNA molecules, *Genes & development*, **15:** 1593–1612.

3. E.J.Merino et al. (2005) RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE), *Journal of the American Chemical Society*, **127:** 4223–4231.

4. S.Marton et al. (2013) RNA aptamer-mediated interference of HCV replication by targeting the CRE-5BSL3. 2 domain, *Journal of viral hepatitis*, **20:** 103–112.

5. R.Lorenz et al. (2011) ViennaRNA Package 2.0., *Algorithms for Molecular Biology*, **6:** 26.

6. M.Zuker, P.Stiegler (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, Nucleic acids research, **9:** 133–148.

7. I.Hofacker, P.F.Stadler (2010). RNAz 2.0: improved noncoding RNA detection, *Pacific Symposium on Biocomputing*, **15:** 69–79.

8. T.Ivry et al. (2009) An image processing approach to computing distances between RNA secondary structures dot plots, *Algorithms for Molecular Biology*, **4:** 1–19.

9. R.Giegerich et al. (2004) Abstract shapes of RNA, *Nucleic acids research*, **32:** 4843-4851.