SECONDARY STRUCTURE OF RNA

URŠA URŠIČ

Faculty of Mathematics and Physics
University of Ljubljana

Ribonucleic acids (RNAs) are complex biopolymers consisting of mainly four building blocks – nucleotides (adenine (A), cytosine (C), guanine (G) and uracil (U)). Nucleotides are prone to forming hydrogen bonds in pairs A-U and C-G, as well as in some other variations, depending on the RNA nucleotide sequence. Base pairing is the key phenomenon leading to RNA folding into a secondary structure, which gives rise to many different RNA functions. Ongoing research is investigating ways to predict secondary structure, which may improve both current understanding of biological systems as well as design of new bionanotechnological tools.

SEKUNDARNA STRUKTURA RNA

Ribonukleinske kisline (molekule RNA) so kompleksne biopolimerne molekule, zgrajene iz štirih glavnih gradnikov, imenovanih nukleotidi (adenin (A), citozin (C), gvanin (G) in uracil (U)). Nukleotidi med seboj tvorijo bazne pare z vodikovimi vezmi, ki so najmočnejše med A-U in C-G, lahko pa nastanejo med mnogo različnimi kombinacijami, glede na zaporedje nukleotidov na verigi RNA. Tvorjenje baznih parov je najpomembnejši dejavnik zvijanja RNA molekul v sekundarno strukturo, kar privede do mnogo različnih funkcij RNA molekul. Raziskave so usmerjene v napovedovanje sekundarne strukture, saj bi boljše razumevanje zvijanja lahko privedlo do napredka pri razumevanju delovanja bioloških sistemov, kot tudi odkrivanja novih možnosti za bionanotehnološka orodja.

1. Introduction to RNA

Ribonucleic acid (RNA) is a heterogeneous polymeric molecule crucial for numerous biological processes. It consists of four different building blocks – nucleotides – which contain ribose, phosphates and base. Phosphate groups link ribose sugar rings into a backbone and each ribose has one of four bases attached as a side group (Fig. 1). Phosphate groups connect the 3rd carbon in the sugar ring of one nucleotide and the 5th carbon of the following one, which imposes directionality of the backbone. The two ends are then referred to as 5' and 3' end, where the 5th and the 3rd carbon are unlinked, respectively. Bases in RNA are adenine (A), cytosine (C,), guanine (G) and uracil (U) in contrast to DNA (deoxyribonucleic acid), where uracil is replaced with thymine (T). From the chemical point of view, RNA and DNA are fairly similar. Along with U-T substitution there is only an H group in DNA instead of the OH in RNA in the ribose ring. However, from the structural point of view, RNA is much closer to proteins since it easily adopts a secondary or even tertiary structure due to its usual single-stranded flexible form. DNA is normally double-stranded with perfectly complementary sequences of the two strands, which provides a firm structure more resilient to damage. The two strands are connected via hydrogen bond in a base pair. In DNA, guanine always binds to a cytosine and adenine always binds to a thymine. Base

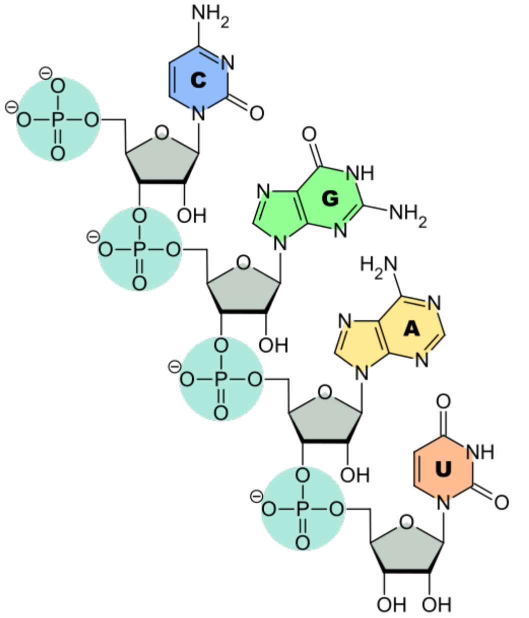

Figure 1. The backbone of RNA consists of ribose sugar rings (gray) linked with phosphate groups (light blue). Each ribose is attached to one of the four different bases (A, C, G, U).

pairs are formed in RNA as well, but in this case occur intra-molecularly – within a single strand – which leads to formation of helices and loops in the secondary structure. [1]

The sequence of nucleotides in RNA defines its structure and functions. Mistakenly, RNA used to be seen for a long time only as an intermediary between the DNA and the proteins. Numerous studies have since proven there is a vast number of different functions, many of them still unknown. [2]

2. RNA folding

The primary structure, which is defined as a sequence of nucleotides, folds into a more complex structure due to intra-molecular interactions between nucleotides. Complex spatial structure is primarily governed by bonding of nucleotides within a single RNA strand forming base pairs (see Fig. 2).

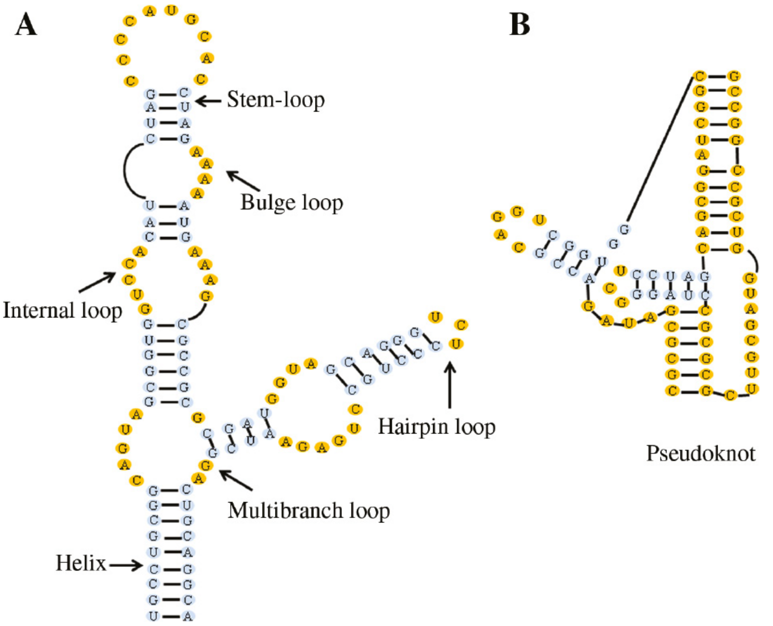

Figure 2. Possible representation of an RNA secondary structure, where letters indicate RNA sequence and gray nucleotides indicate paired bases, whereas yellow ones are unpaired. A) shows helices and different types of loops and B) illustrates a possible structure of a pseudoknot. [3]

2.1 Base pairs and elements of secondary structure

Chemical properties of nucleotides allow the formation of hydrogen bonds. These are electrostatic interactions between a hydrogen atom and a lone electron pair, crucial for secondary structure stability. Hydrogen bonds in RNA can be of two types: polar (between the atoms O:H) and apolar (between the atoms N:H). Two nucleotides connected with one or more hydrogen bonds form a base pair. Base pairs are generally characterized as either canonical or non-canonical, where canonical base pairs are also named Watson-Crick base pairs. They occur between C and G or A and U (T substitutes U in DNA). In DNA, Watson-Crick base pairs form a very strong and rigid structure. Similarly in RNA, these base pairs form the strongest hydrogen bonds and are therefore favorable for the minimization of its total free energy. On the other hand, all the other weaker base pairs are categorized as non-canonical. They occur between many nucleotide combinations – theoretically there are more than 150 non-canonical base pairs due to different kinds of bonding, such as cis/trans combination according to relative base positions and different hydrogen bond conformation in each pair, as there is more than one possibility for a hydrogen bond to be formed. While numerous kinds of base pairs are adopted, not all of the theoretical possibilities are energetically favorable. [4] A very common non-canonical base pair is formed by guanine (G) and uracil (U) and is named wobble pair.

Bases in secondary structure are generally divided into two different classes: those that are included in base pairs (also known as stems) and those that are not, i.e., unpaired bases. Therefore, the secondary structure can be seen as a set of different elements formed by paired and unpaired bases (see Fig. 2).

2.2 Tools for RNA secondary structure prediction

With discoveries of RNA sequences a far-from-trivial question arises; how RNA sequences fold into a more complex 3D structure, which can take on such a number of different functions. This is almost impossible to answer without a middle step, that is solving a simplified 2D structure also known as secondary structure. Secondary structure is defined as a pair (I, S), where I is an RNA sequence with a known set S of base pairs between its constituent nucleotides. An example of secondary structure is shown in Fig. 2.

The most common approach to the prediction of RNA secondary structure is **free energy minimization**, which will be described in Sec. 2.3 The key to this technique is to link possible secondary structures to corresponding free energies.

Another prediction model is **modeling kinetics**, where the basic principle is observation of the step-by-step folding mechanism where the completely unfolded chain follows a folding path by either formation or a switch of a single base pair. Every chain will eventually fold into a minimum free energy structure, therefore a characteristic folding time can be defined. However, some structures have long-lived metastable states where the structure is not in the minimal free energy state but the characteristic time of that structure is extremely long. Folding kinetic approach offers an insight into these structures. [5]

Loop decomposition is used to simplify the structure problem into smaller independent compounds – loops. A loop is characterized by its length and its degree. Degree of a loop is defined as the number of stems attached to the loop. For example (see Fig. 2A), a hairpin loop is attached to only one stem and thus is of degree 1, whereas the degree of both an internal loop and of a bulge loop is 2. Loops with degree > 2 are called multibranch loops. Loop decomposition is particularly useful in free energy minimization models for RNA secondary structure prediction, since the total free energy can be approximated by the sum of free energies of all the independent elements. Note that pseudoknots (Fig. 2B) prevent decomposition into independent structures.

2.3 Free energy model of RNA secondary structure

RNA sequence is written as a string of N letters indicating each base, where N is the total number of nucleotides. For example,

GGGCGUGUGCGUAG ... GUCACCA.

In theory, sequence I is written as a string of N elements, for example $I = x_1, x_2, ..., x_N$, where $x_i \in A, U, C, G$. If two elements x_i and x_j form a base pair, it is denoted as (i, j), where $1 \leq i < j \leq N$. Secondary structure is defined by the set S of base pairs, with three basic rules.

- Each base can be involved in only one base pair; further interactions are then considered only in the tertiary or quaternary structure.

- Pair bases (i, j) must be separated by at least 3 unpaired base pairs, $j > (i + 3)$, due to the RNA backbone flexibility.

- Pseudoknots (see Fig. 2B) are not allowed in the secondary structure. This means that for two pairs (i, j) and (k, l) either condition $i < k < l < j$ or $i < j < k < l$ must be met.

Note that the rules mentioned above can vary from a model to a model, but are most commonly used in this form. Pseudoknots are not only computationally complex, but also hard to observe experimentally. They prevent the simplification of the problem with loop decomposition approximation and are therefore often theoretically forbidden without a major toll on the accuracy of the model.

Every secondary structure S can be associated with a Gibbs free energy $G(S)$. Various models exist to determine the value of $G(S)$. One option is to sum over the energies of every base pair

$$G(S) = \sum_{(i,j) \in S} G\big((i, j)\big). \tag{1}$$

Note that base pairing lowers the free energy in comparison to an unfolded chain. This model does not take into account the entropy contribution of the paired bases, which can change the overall $G(S)$. More accurate model uses loop decomposition, where the complete secondary structure is divided into several loops as described in 2.2 Gibbs free energy with this model changes Eq. 1 to include factors corresponding to every loop, composed of three parts – a free energy contribution of the mismatched pairs in a stem, a contribution of the loop size and a special term, used for empirical correction of some unusually stable cases. The obtained free energy is of form

$$G(S) = \sum_{(i,j) \in S} G\big((i, j)\big) + \sum_{loop \in S} \left(G_{mismatch} + G_{size} + G_{special}\right). \tag{2}$$

Experimental results of many loop free energies are stored in publicly accessible databases and are further used for computational predictions. [6]

The accuracy of this technique is just over 70% for shorter RNA chains (up to 700 nucleotides) and even decreases for longer RNA chains. [7] The energy minimization approach is not perfectly accurate due to several reasons. First, interaction rules are incomplete, meaning that the interactions in the base pairs are often more complex than involving just the two bases and the energy of a non-coupled base depends on its position. Second reason lies in the process of folding itself, as some RNA structures depend on folding kinetics pathways. Here, folding kinetics models ought to be used for prediction improvement. Another complication occurs due to multiple possible secondary structures of the same RNA sequence depending on the surrounding medium. This property is very well used in nature, for example in regulatory systems. What is more, secondary structure in real life always exists in a water-based solvent with non-zero ionic concentration. This also contributes to the free energy variations, as the RNA chain contains also hydrophobic and charged sites which behave differently in solvent than *in vacuo*. Such systems cannot be analytically predicted, and prediction models have to consequently take into account empirically specified parameters. In recent years, studies have been searching for a way to minimize the use of empirically obtained data. [8] Many RNAs also have modified nucleotides like inosine and pseudouridine, for which there is not much knowledge on their influence on the stability of the loops. [7]

Regardless of the model used, one obtains free energies corresponding to different possible conformations. A structure with minimal free energy is presumed to be the one adopted in nature.

2.4 Boltzmann partition function

According to energy minimization models, RNA will fold into the secondary structure of the lowest possible free energy when in equilibrium. For most RNA sequences, there are plenty of different structures with free energies very close to the minimum free energy, as the Watson-Crick binding energy is of the same order of magnitude as thermal energy. This gives rise to the uncertainty of the predicted set of base pairs that can be quantified with the use of partition function Q, which is the sum over Boltzmann factors $e^{\frac{-G(S)}{RT}}$ corresponding to the probability of a given structure S.

$$Q = \sum_S e^{\frac{-G(S)}{RT}}, \tag{3}$$

where $G(S)$ is the Gibbs free energy of a certain structure determined relatively to the free energy of the unfolded structure, R is the gas constant and T is absolute temperature. The probability $p(S)$ of a given structure is

$$p(S) = \frac{e^{\frac{-G(S)}{RT}}}{Q}. \tag{4}$$

Given an RNA sequence, different computational algorithms are used to determine *in silico* structure, following models based on the basic idea explained above.

2.5 Energy landscapes

Very often, knowing the minimum free energy structure of an RNA is not enough to explain biological processes in the cell that the RNA is involved in. It is very common for an RNA to fluctuate around the minimum free energy state due to interactions with other molecules and performing fluctuation dependent functions. Therefore, energy landscapes are introduced to assist with this problem. Energy landscape of RNA structure is obtained by determining the free energies of all the possible structural conformations of an RNA chain and indicating their corresponding free energies. One possible representation of this would be a high-dimensional energy landscape with all degrees of freedom. However, this kind of representation is not visualizable; instead, a reduced but demonstrative energy landscape is used. Dependence of the free energy is reduced to two parameters and is defined as

$$F(n, nn) = -k_B T \ln Q(n, nn, T), \tag{5}$$

where $Q(n, nn, T)$ is the partition function which counts all the possible conformations containing n native contacts and nn non-native contacts. Native contact is any specific base pair in a given secondary structure which is the same as in the native structure – the minimum free energy state – and a non-native contact is when a particular base pair does not exist in the native structure. Helmholtz free energy F is used instead of Gibbs free energy G only for practical reasons as most papers work with F. An example of energy landscapes at different temperatures for a simple RNA chain is shown in Fig. 3.

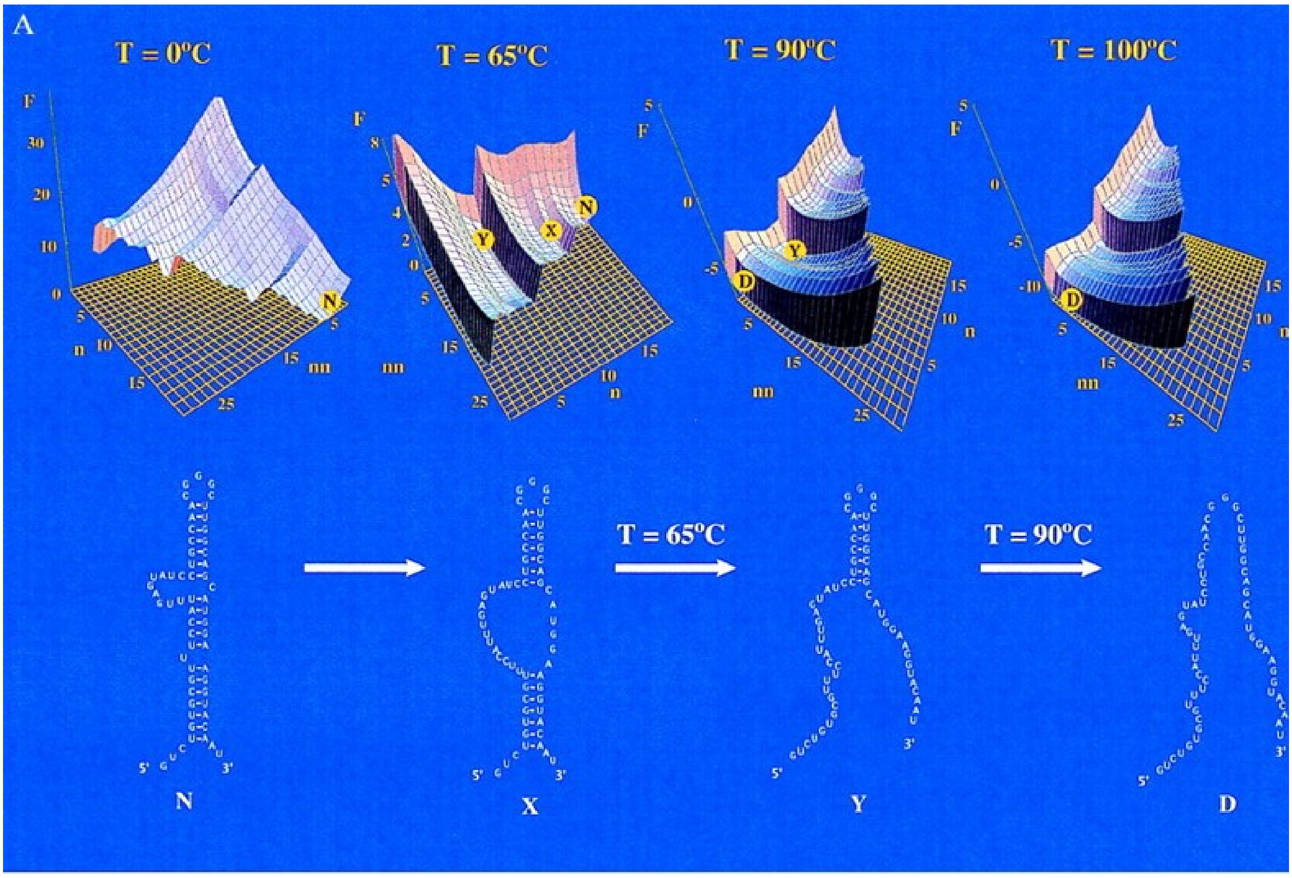

Figure 3. Energy landscapes representing free energy dependence of structures containing n native pairs and nn non-native pairs at four different temperatures. Letters indicate points in a n-nn plane and the corresponding secondary structures are shown at the bottom of the image. [9]

As the temperature changes, so does the energy landscape, since there is a competition between enthalpy and entropy. From that one can extract the information about folding pathways: how the RNA chain changes its secondary structure upon temperature changes. Does it follow the minimum energy state or not? In the minimum free energy model of RNA folding, an order parameter can be defined with the connection to the energy landscapes. Order parameter can be a structural or energetic measure describing the match between a given structure and the native one. [9]

3. Experimental techniques

Predictions have little value if they cannot be experimentally tested. Recently, many different experimental techniques have been developed to determine the RNA structure; the three main techniques will be briefly reviewed here.

3.1 X-ray crystallography

X-ray crystallography was the first technique to provide insightful images of RNA. [10] The process consists of four main steps in the acquisition of RNA secondary structure. First, a homogeneous single crystal of the observed RNA molecule is needed. Secondly, the crystal is immobilized in the intense X-ray beam and then rotated gradually. The result are diffracted X-ray beams at different angles and different intensities. The diffracted beams are measured and collected to form the diffraction pattern. This is used to determine the structure of Bravais lattice. The third step is called phasing. The second part of experiment only allows one to obtain the intensities of each diffracted beam. However, for the complete knowledge of the electron density in the unit cell one needs phase information as well. The phase can be measured in a few different procedures, for example heavy atom derivatization or molecular replacement with a known molecular structure. Using intensities and phases of each diffracted spot, an electron density can be calculated with the inverse Fourier transform of the structural factor. Last step in the experiment is refinement of the data obtained to get a clearer and usable structure. [11]

3.2 NMR spectroscopy

A powerful technique for determining structures of RNA chains using nuclear magnetic resonance is called NMR spectroscopy. RNA molecules are first labeled with NMR active elements (^{1}H, ^{13}C, ^{15}N, ^{31}P). These elements must have their nuclear spin not equal to zero. When these requirements are met, the nuclei in the strong magnetic field will have energy degeneration eliminated, therefore based on the type of nucleus and its interactions with its surroundings, the differences between energies of different states are unique. By applying electromagnetic radiation to the sample in the strong magnetic field certain wavelengths – those matching the energy differences between states – will be absorbed. In this way one can obtain the absorption spectrum and perform structure reconstruction by using various computational approaches. [12]

3.3 Cryo-electron microscopy

Technique that is attracting a lot of attention in recent years is based on electron microscopy. The main advantage over X-ray crystallography is that it does not require the crystallization, which in many cases does not occur. It also allows observation of molecules in close-to-native state, because samples are kept in a native-like solution. The solution is quench frozen with liquid nitrogen and enables water to amorphously solidify, preserving the density and avoiding water crystallization. Electron beam is focused on the sample, scatters on the RNA molecules and is then magnified and detected. The molecules' orientations are random, therefore one can obtain several 2D projections, which can be modeled into a 3D structure. A disadvantage of this technique is its relatively low resolution, providing valuable images only for larger structures. [13]

Other techniques exist but are less popular for determining RNA structures. They are mostly based on chemical or biological processes using structure-specific chemicals and enzymes. Another comparative technique is temperature denaturation for investigating the strength of intra-molecular forces responsible for preserving 3D structures. [14]

4. RNA nanotechnology

Apart from satisfying the pure curiosity of the human kind, there is more to gain from knowing the principles behind RNA secondary structure formation and its function. Artificially synthesized RNA molecules based on accurate prediction models can lead to wonderful bionanoparticles in various shapes (Fig. 4).

In nature, RNA forms a vast variety of 3D structures with specific functions. RNA is more flexible in contrast to DNA and can adopt much greater variety of structures due to the combinations of canonical and noncanonical base pairings. This gives huge opportunities for the design of novel RNA structures with a never-ending list of possible applications. Although early studies of RNA nanotechnology produced RNA nanostructures based on intuition and practical experiences,

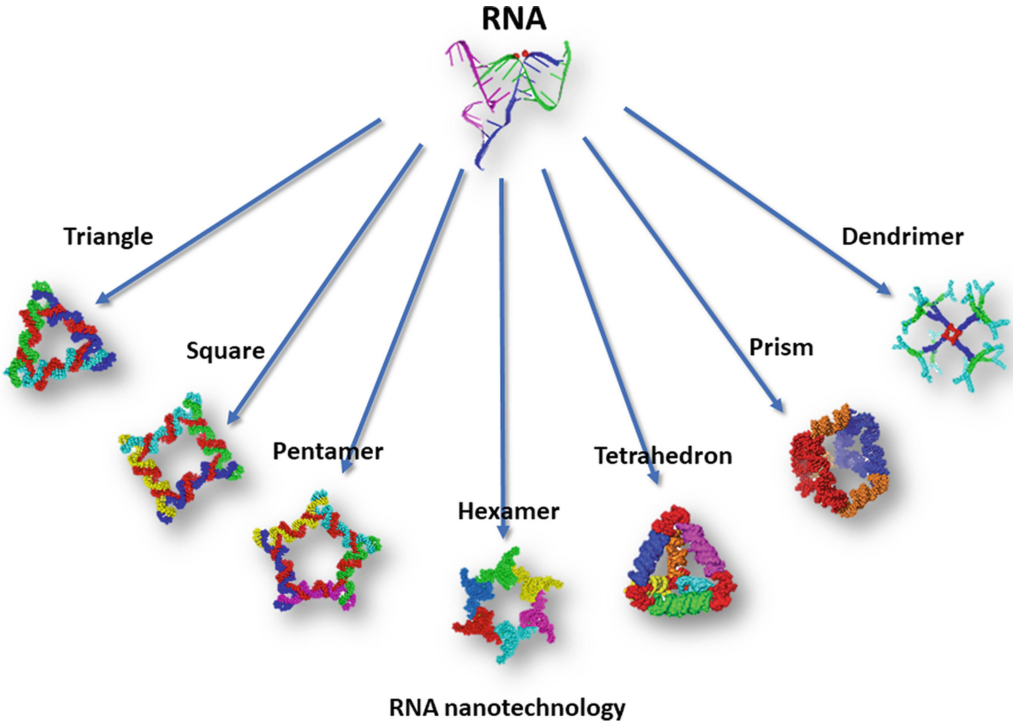

Figure 4. RNA can be synthesized in such a sequence that it adopts a specific structure. [15]

the value of computational methods for secondary structure prediction is increasing as it gives rise to more complex programmable structures.

The process by which RNA functional nanoparticles are discovered starts with the the concept: a specific structure should be linked to the desired function. The second step includes programming of the RNA sequences that will adopt the appropriate structure for building blocks of the complete nanoparticle. This is where the prediction models are needed. Thirdly, building blocks are synthesized and then appropriately assembled. There are two categories: templated and non-templated assembly. The first one requires external force, structure or spatial constraint for assembling the building blocks, whereas the second one self-assembles upon inter-molecular interactions. After obtaining the final structure it should be characterized by an appropriate observation technique. In the absence of complications, the medical or biotechnological applications can be developed.

The most desirable applications of RNA are of medical manner, for example targeted delivery of drugs, gene silencing and cell recognition and binding for diagnosis. Advantages of RNA for medical applications are its ability to trespass the nuclear membrane for genetic treatments and the absence of proteins as many proteins are recognized as pathogens by human immune system and are therefore degraded. RNA nanostructures are also used as mediators for self-assembly of other nanoparticles, such as palladium or cationic gold. An advancing field in bionanotechnology is also synthetic biological circuitry, where bionanomolecules are designed in order to perform functions directly comparable to logic gates. With that, the program written as a genetic sequence could be run in *in vivo* in cells. Biological circuits can be either used for studying cell processes or influencing them. Latter option offers possibilities for self regulatory medical treatment. [16]

5. Conclusions

Predictions of RNA secondary structure are only small, yet essential, steps on the way to completely understand different biological processes. RNA folding is a hierarchical process, meaning that the strongest interactions are hydrogen bonds in base pairs, governing the folding of an RNA chain first into a secondary structure and then weaker interactions (hydrophobicity, electrostatics) determine a tertiary structure. Knowing the concept behind RNA secondary structure offers explanations of biological processes and opportunities for *de novo* design. Nanotechnology using biomolecules, such as RNA, is a very promising but poorly known area, which requires well established 3D structure prediction. Hopefully in near future, we will be able to develop this field even further and us it to our advantage.

6. Acknowledgements

I sincerely appreciate all the assistance with preparation, writing and presentation of this work from my menthor Dr. Anže Rapoš Božič.

REFERENCES

[1] Higgs, P. G., Quarterly Reviews of Biophysics **33** (2000) 199.

[2] Smirnov, A., Schneider, C., Hör, J., and Vogel, J., Current Opinion in Microbiology **39** (2017) 152.

[3] Singh, N., Sharma, S. D., and Yennamalli, R. M., Bio-Algorithms and Med-Systems **13** (2017) 187.

[4] Lemieux, S. and Major, F., Nucleic Acids Research **30** (2002) 4250.

[5] Flamm, C., Fontana, W., Hofacker, I. L., and Schuster, P., RNA **6** (2000) 325.

[6] Turner, D. H. and Mathews, D. H., Nucleic Acids Research **38** (2010) D280.

[7] Mathews, D. H. and Turner, D. H., Current Opinion in Structural Biology **16** (2006) 270.

[8] Vendeix, F. A., Munoz, A. M., and Agris, P. F., RNA **15** (2009) 2278.

[9] Chen, S.-J. and Dill, K. A., Proceedings of the National Academy of Sciences **97** (2000) 646.

[10] Westhof, E., RNA **21** (2015) 486.

[11] Holbrook, S. R. and Kim, S.-H., Biopolymers: Original Research on Biomolecules **44** (1997) 3.

[12] Fürtig, B., Richter, C., Wöhnert, J., and Schwalbe, H., ChemBioChem **4** (2003) 936.

[13] Garmann, R. F. et al., RNA **21** (2015) 877.

[14] Felden, B., Current Opinion in Microbiology **10** (2007) 286.

[15] Yaradoddi, J. et al., Handbook of Ecomaterials: RNA Nanotechnology, page 3587, Springer Nature Switzerland, 2019.

[16] Guo, P., Nature Nanotechnology **5** (2010) 833.