TECHNISCHE UNIVERSITÄT MÜNCHEN

Fachgebiet für Bioinformatik

Sequence-structure relationships in mRNAs

Andrey Chursov

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung un Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. A. Gierl
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. D. Frischmann
	2. Univ.-Prof. Dr. B. Rost

Die Dissertation wurde am 26.08.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung un Umwelt am 12.11.2013 angenommen.

Contents

List of Figures **III**

Summary **6**

Zusammenfassung **8**

1 Introduction **11**

1.1 RNA Structure 12
1.2 Functions of RNA 18
1.3 Experimental Techniques 24
1.4 Theoretical Prediction Methods 31
1.5 Thesis Motivation and Outline of the Work 39

2 Results **41**

2.1 Sequence-structure Relationships in Yeast mRNAs 41
2.2 Specific Temperature-induced Perturbations of Secondary mRNA Structures are Associated with the Cold-adapted Temperature-sensitive Phenotype of Influenza A Virus 50
2.3 RNAtips: Analysis of Temperature-induced Changes of RNA Secondary Structure 61
2.4 Conservation of mRNA Secondary Structures May Filter Out Mutations in *Escherichia coli* Evolution 70

3 Discussion **79**

4 Conclusions and outlook **84**

Bibliography **86**

Acknowledgements **107**

Curriculum Vitae **109**

Glossary/Abbreviations **111**

List of Figures

1.1 Hierarchy of RNA structure 13
1.2 Secondary structural motifs 15
1.3 Graphical representations of RNA secondary structures 16

Summary

Essentially, all models are wrong, but some are useful.
—George E. P. Box

The ability to measure and predict the structure in which an RNA molecule will fold is crucial to our understanding of biological processes in a cell. Recently, many new functions of RNA, which are performed owing to specific conformation of particular RNAs, have been discovered. At the same time, ongoing nucleotide substitutions may influence the ability of an RNA molecule to fold in a proper structure and may cause both negative and positive impacts on the organism. Hence, such mutations are of great interest for biologists, virologists, drug designers, etc. The present work contributes to our understanding of sequence-structure relationships in messenger RNAs, which recently demonstrated having other regulatory functions in addition to encoding protein sequences.

To begin, we performed the first comprehensive analysis of sequence-structure relationships in the coding regions of yeast mRNAs. This work was done by analyzing the experimentally measured data produced by the first high-throughput approach for genome-wide probing of RNA structures termed PARS. Our results demonstrated that only those pairs of sequences which have a very high level of sequence identity (greater than 85-90%) have similar PARS profiles. However, the structures of pairs of sequences with lower sequence identities seem to be completely unrelated to each other. The latter fact was also demonstrated with theoretically predicted probabilities of nucleotides to be in a double-stranded conformation for orthologous mRNAs.

Arguably, the most efficient anti-influenza vaccines ever used have been cold-adapted (ca), temperature-sensitive (ts) live attenuated influenza vaccines (LAIV). The ca/ts phenotype leads to impaired growth at an elevated temperature of approximately

39°C, while permitting viral growth at lower temperatures. Thus, cold-adapted, temperature-sensitive (ca/ts) strains can be produced at the factory, but cannot cause significant harm to a patient. For more than fifty years, scientists were trying to understand why ca/ts mutants of the influenza virus used as vaccines possessed temperature sensitivity different from their wild type counterparts. Despite significant effort devoted to explaining temperature sensitivity, the molecular mechanism(s) causing the ca/ts phenotype in influenza A viruses remain unclear. We have demonstrated that influenza mRNAs of ca/ts vaccine strains contain clusters of nucleotides, which would undergo temperature-induced structural perturbations differently than corresponding non-vaccine strains. Thus, ca/ts phenomena can have the temperature-induced change of RNA structures as an underlining mechanism. These conclusions are further supported by the fact that clusters of temperature-sensitive positions specific for ca/ts strains do appear in response to mutations causing ca/ts phenotype, but not in sets of computer-generated RNA sequences containing the same number of random mutations. To the best of our knowledge, our approach is the first attempt to explain ca/ts phenomena through perturbations of RNA structures.

Based on the algorithm used in solving the ca/ts phenomena, the web server RNAtips.org has been implemented. Before this server, there had been no convenient way for a researcher to investigate how mutations in RNA (both coding and non-coding) may cause changes in RNA folding upon temperature elevation. It is the first widely available tool for such analysis which may be of interest to scientists studying temperature effects on different organisms.

Finally, we demonstrated that mutations disrupting mRNA secondary structure might be filtered out in the course of bacterial evolution. To test this hypothesis, the influence of single nucleotide polymorphisms (occurring during the "long-term evolution experiment" organized by Richard Lenski) on the changes in minimum free energy values were analyzed. Nucleotide substitutions occurring between the first and the 40,000$^{\text{th}}$ generations were investigated; and, changes in folding energies resulting from the mutations between essential and nonessential genes were compared. The statistical tests that were performed clearly indicated that preservation of the secondary structure of messenger RNAs might serve as a previously unknown mechanism of bacterial evolution.

Zusammenfassung

Die Fähigkeit, die Struktur eines RNA Moleküls zu bestimmen und vorherzusagen, ist entscheidend für unser Verständnis von biologischen Prozessen in der Zelle. In jüngster Vergangenheit wurden viele neue Funktionen von RNA entdeckt, die durch spezifische Konformationen dieser Moleküle ermöglicht werden. Zugleich können Nukleotidsubstitutionen die Fähigkeit eines RNA-Moleküls, sich korrekt zu falten, beeinflussen und damit sowohl negative als auch positive Effekte auf den Organismus haben. Daher sind diese Mutationen von großem Interesse für Biologen, Virologen, Wirkstoff-Entwickler, etc. Die vorliegende Arbeit ist ein Beitrag zu unserem Verständnis von Sequenz-Struktur-Beziehungen in messenger RNA, die, wie kürzlich gezeigt wurde, zusätzliche regulatorische Funktionen besitzt, die über die Kodierung von Proteinsequenzen hinausgehen.

Zunächst wurde die erste umfassende Untersuchung von Sequenz-Struktur-Beziehungen in den kodierenden Regionen der Hefe mRNA durchgeführt. Dazu wurden experimentelle Daten genutzt, die durch die erste Hochdurchsatz-Methode für die genomweite Analyse von RNA-Struktur (PARS) gewonnen wurden. Unsere Ergebnisse zeigen, dass nur Sequenzpaare mit einem sehr hohen Grad an Sequenzähnlichkeit (mehr als 85-90%) vergleichbare PARS-Profile aufweisen. Die Strukturen von Sequenzpaaren mit geringerer Sequenzähnlichkeit hingegen, schienen völlig verschieden zu sein. Diese letzte Beobachtung wurde zusätzlich durch theoretisch vorhergesagte Wahrscheinlichkeiten für das Vorliegen von Nukleotiden in einer Doppelstrang-Konformation in orthologen mRNAs bestätigt.

Die bislang wirksamsten Impfstoffe gegen Influenza sind Präparate mit kälteadaptierten (ca), temperatursensitiven (ts) attenuierten Influenza Lebendimpfstoffen (live attenuated influenza vaccines, LAIV). Der ca/ts Phänotyp führt zu einer Verminderung des viralen Wachstums bei erhöhten Temperaturen von etwa 39°C, während virales Wachstum bei niedrigeren Temperaturen ermöglicht wird. So können kälteadaptierte

temperatursensitive (ca/ts) Stämme industriell hergestellt werden, ohne wesentlichen Schaden in den Patienten zu verursachen. Seit mehr als 50 Jahren versuchen Wissenschaftler zu verstehen, warum die ca/ts Mutanten, die für Impfungen verwendet werden, eine andere Temperatursensitivität aufweisen als der zugehörige Wildtyp. Obwohl großer Arbeitsaufwand in die Erklärung der Temperatursensitivität investiert wurde, sind die molekularen Mechanismen, die den ca/ts Phänotyp von Influenza A Viren verursachen, bislang unklar. Wir haben gezeigt, dass die mRNAs der ca/ts Impfstämme Nukleotid-Cluster enthalten, welche andere temperaturinduzierten Strukturperturbationen durchlaufen als mRNAs entsprechender Wildtyp-Stämme. Es kann daher vermutet werden, dass diese temperaturinduzierten Änderungen der RNA-Struktur grundlegender Mechanismus des ca/ts Phänomens ist. Diese Schlussfolgerung wird zudem durch die Tatsache gestützt, dass Cluster von temperatursensitiven Positionen, die spezifisch für ca/ts-Stämme sind, als Reaktion auf ca/ts-Phänotyp verursachende Mutationen auftreten, jedoch nicht in computergenerierten RNA Sequenzen, welche die gleiche Anzahl an zufälligen Mutationen enthalten. Nach unserem Kenntnisstand ist unsere Herangehensweise der erste Versuch, den ca/ts Phänotyp durch Veränderungen in der RNA Struktur zu erklären.

Basierend auf dem Algorithmus für die Lösung des ca/ts Phänomens wurde der Webserver RNAtips.org entwickelt. Bisher gab es für Wissenschaftler keine geeignete Möglichkeit zu untersuchen, in welcher Weise RNA Mutationen (sowohl in kodierenden als auch in nicht-kodierenden Regionen) zu Änderungen der RNA-Faltung in Folge von Temperaturerhöhung führen. Es handelt sich hierbei um das erste weit verfügbare Tool für diese Art von Analysen, welche für Forscher, die Einflüsse von Temperatur auf verschiedene Organismen untersuchen, relevant sein dürften.

Zuletzt haben wir gezeigt, dass Mutationen, welche die Sekundärstruktur von mRNA zerstören, im Verlauf der bakteriellen Evolution herausgefiltert werden. Um diese Hypothese zu testen, wurde der Effekt von single nucleotide polymorphisms (aufgetreten während des “long-term evolution experiment” von Richard Lenski) auf die Änderungen der minimalen freien Energie untersucht. Nukleotidsubstitutionen, die zwischen der ersten und 40.000sten Generation auftraten, wurden analysiert und Änderungen der Faltungsenergie in Folge von Mutationen wurden zwischen essentiellen und nicht-essentiellen Genen verglichen. Die durchgeführten statistischen Tests zeigten deutlich, dass die Erhaltung der Sekundärstruktur von mRNA ein bislang unbekannter Mechanismus in der bakteriellen Evolution sein kann.

Chapter 1

Introduction

After the discovery of RNA, it was assumed that in all organisms, except RNA viruses, RNA was solely a mediator between DNA and protein (Crick, 1958; Crick et al., 1970). Then, transfer RNA (tRNA) was discovered (Crick, 1962); and later in 1960^s, it was suggested that, at the early stage of evolution, RNA could both be genetic material and play a catalyst role that promotes its own replication (Crick, 1968; Orgel, 1968; Woese, 1967). The later idea obtained support only after experimental discoveries of RNA with catalytic activity (Kruger et al., 1982; Guerrier-Takada et al., 1983; Cech, 1986), which led to Nobel laureate Walter Gilbert in 1986 introducing the concept RNA World to emphasize a world of free-living RNA molecules (Gilbert, 1986). The main feature of the RNA World is existence of a molecule functioning as an RNA-dependent RNA polymerase that is able to produce complementary RNAs from itself or copies of itself, and subsequently, produce additional copies of itself from the complementary RNAs produced during the previous step (Robertson and Joyce, 2012).

Problems of the origin of the RNA World and the early evolution of life are still far from being solved. It is also difficult to say how the first RNA replicase ribozyme arose (Robertson and Joyce, 2012). Nonetheless, the discovery of catalytic RNA clearly demonstrated that RNA might have a lot of different unknown functions. Since then, many new classes of functional non-coding RNAs, as well as other functions of messenger RNA apart from encoding amino acid sequence, have become known. We now know of many roles that RNA plays inside a cell which scientists did not even think about several decades ago; yet, there is still much to be discovered about RNA functions.

Similar to proteins, the functional roles of RNA are determined by the structure it folds. Hence, the necessity to understand those functions leads to the need to investigate what conformation the RNA molecule folds, what relationships between a structure and a sequence is, what evolutionary constraints on the structure are, and so on. These questions do not have univocal answers. Instead, they have to be considered from the different angles; this work presents some of the potential answers.

1.1 RNA Structure

A ribonucleic acid, or RNA, is a single-stranded polymer composed of four nucleotide subunits: adenine (A), cytosine (C), guanine (G), and uracil (U). The length of the RNA molecules, observed in nature, lies in a wide range from about 20 to thousands of nucleotides (Bevilacqua and Blose, 2008). Each nucleotide consists of a base, a ribose, and a phosphate (Tinoco and Bustamante, 1999). Similar to DNA (deoxyribonucleic acid), nucleotides in RNA may form hydrogen bonds with each other. The standard or canonical base pairs, similar to those discovered by Watson and Crick in DNAs (Watson and Crick, 1953a,b), are G-C and A-U based on three hydrogen and two hydrogen bonds respectively. As G-C pairs are based on three hydrogen bonds, they are more energetically stable than A-U base pairs. Non-canonical interactions are also possible (the most common is G-U), but such base pairs are highly unstable and hence occur very rarely. Therefore, hereinafter only canonical base pairs will be considered and non-canonical pairs are not discussed. A set of such interactions between nucleotides determines the structure in which an RNA molecule folds. Despite the fact that RNA and DNA are very similar, the structure of RNA is very different from those of DNA. DNA contains two complementary sequences, which form a famous double helix. The RNA molecule is usually presented as a single strand and folds to itself forming intra-molecular short helices (Higgs, 2000).

One can divide an RNA structure into four different levels: primary, secondary, tertiary, and quaternary. The primary structure of an RNA molecule is just its sequence of nucleotides describing the RNA molecule (Figure 1.1a). A secondary structure of RNA can be thought of as a two-dimensional folding containing a set

a) GCCCGGAUAGCUCAGUCGGUAGAGCAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCGAUUCCGAGUCCGGGCACCA

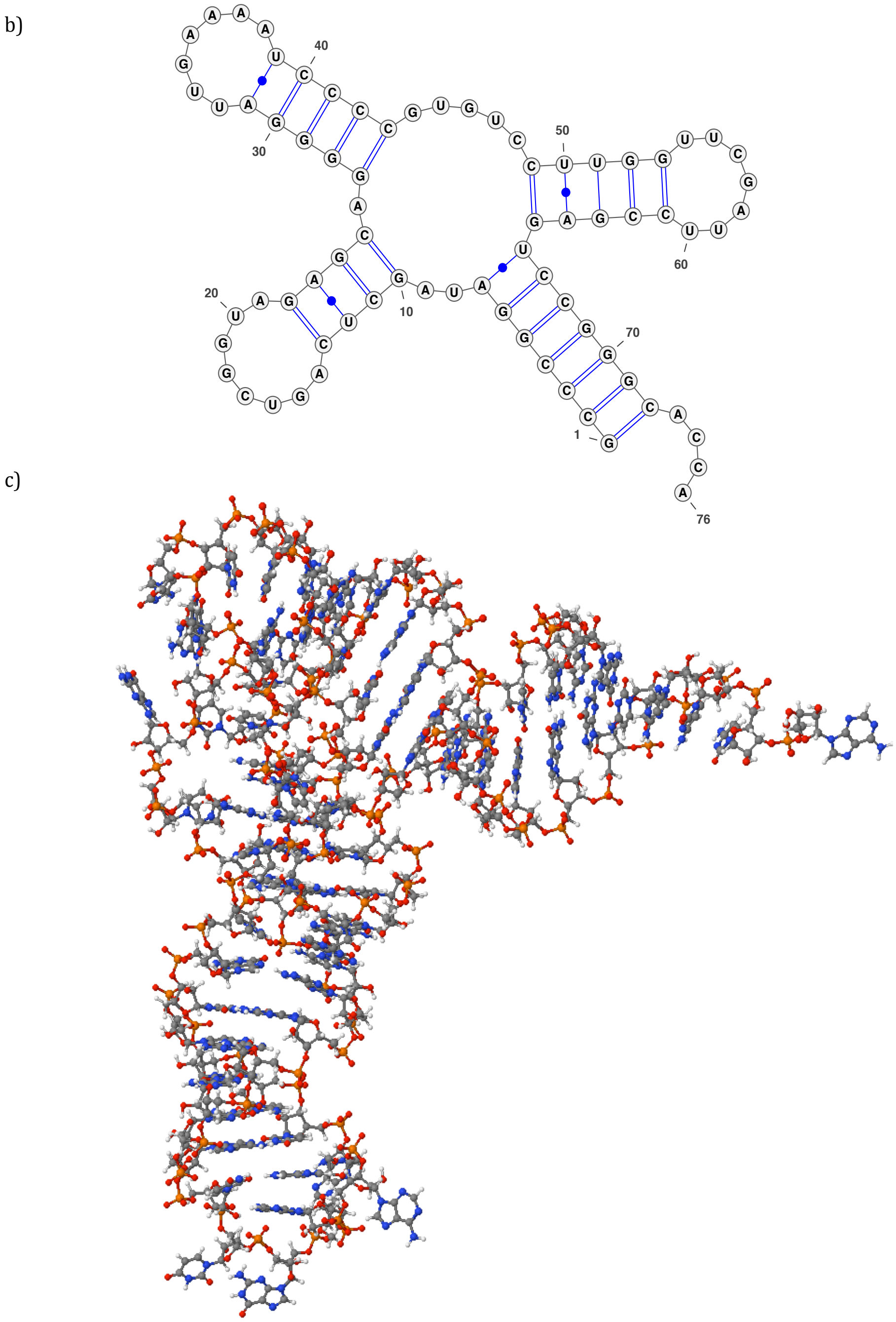


Figure 1.1: ***RNA structure of phenylalanine tRNA from Escherichia coli. a)*** *Primary structure.* ***b)*** *Secondary structure (created by VARNA visualization tool (Darty et al., 2009)).* ***c)*** *Tertiary structure (created by RNA structure 3D modeling server RNAComposer (Popenda et al., 2012)).*

of base pairs that are formed (Figure 1.1b). The tertiary structure is what form the molecule has in a three-dimensional space with a description of the location of every atom. Figure 1.1c shows the tertiary structure for a phenylalanine transfer RNA (tRNA) of Escherichia coli. Quaternary structure is generally considered as a set of complex interactions of the folded nucleic acid with other molecules.

The diverse roles of RNA are mainly determined via its spatial structures that bind small regulatory RNAs, large protein ligands, or with machineries responsible for translation initiation and splice site selection (Weeks, 2010). However, due to its complexity, the RNA tertiary structure is a less studied subject compared to the secondary structure. Therefore, hereinafter we consider only secondary structures of RNAs if another is not directly indicated.

Due to the ability of RNA molecules to fold into a native conformation in very short periods of time, it is thought that RNA folding is a hierarchical process (Pyle and Green, 1995; Brion and Westhof, 1997; Tinoco and Bustamante, 1999). For a long RNA, the number of potential conformations is large. Thus, it would require a lot of time to search through all of the states for a native structure, additionally, some intermediate organization should exist to lead the folding pathway (Brion and Westhof, 1997). The hierarchical nature of RNA folding means that secondary structural elements are determined completely by primary sequence, independently of tertiary structure, and are likely to form before tertiary interactions, which do not change secondary structure significantly (Pyle and Green, 1995). Furthermore, it is accepted that energies of secondary interactions are much larger than those of tertiary ones; hence, secondary conformations are more stable than tertiary folding. It is obvious that the assumption of hierarchy in RNA folding is only a simplification of the real processes going on in the cell. In some cases in which a complex pseudoknot (Figure 1.2) has to form, organizations of secondary and tertiary structures are not separable from each other (Gluick and Draper, 1994). However, in other cases, hierarchical folding was even observed experimentally (Greenleaf et al., 2008). Therefore, it is generally believed that in most cases this model describes the folding process well enough and can be applied to investigate functions of RNA (Tinoco and Bustamante, 1999).

A complex secondary structure of any RNA can be considered as a set of different

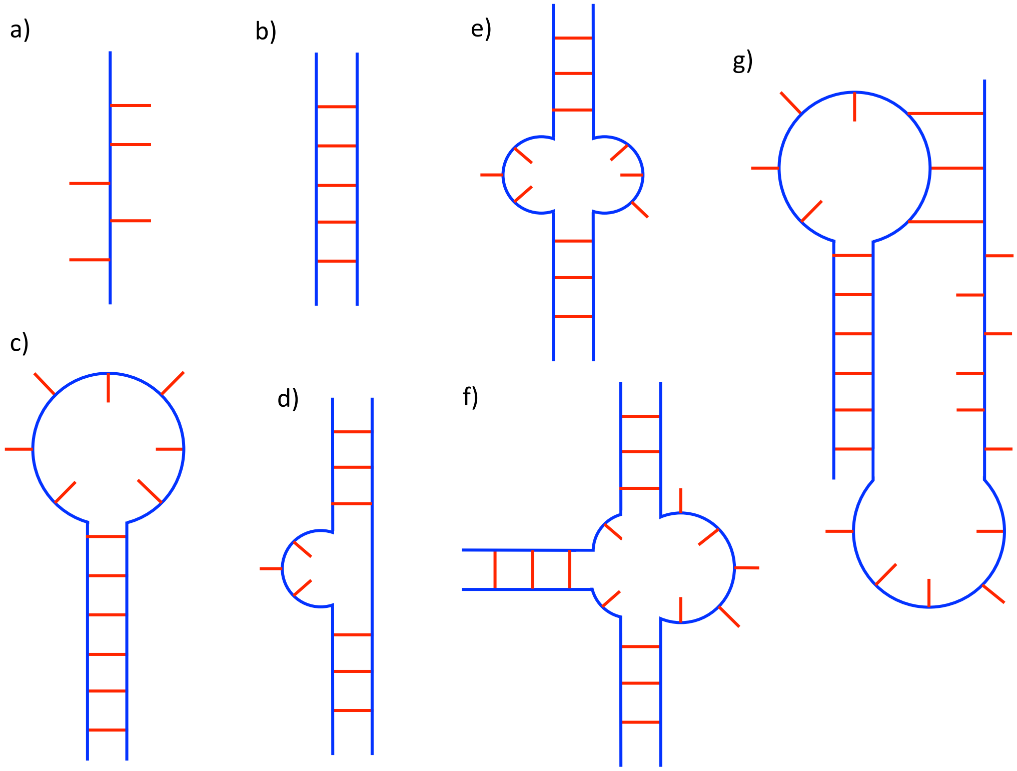


Figure 1.2: ***Secondary structural motifs in RNA. The RNA backbone is blue, and both unpaired and paired bases are red.*** ***a)*** *Single-stranded RNA.* ***b)*** *Double-stranded RNA helix.* ***c)*** *Hairpin consisting of stem and loop.* ***d)*** *Bulge loop.* ***e)*** *Interior loop.* ***f)*** *Multiloop (junction).* ***g)*** *Pseudoknot.*

smaller elements interconnected with each other, as depicted in Figure 1.2. Those elements are single stranded region, base-paired double helical segment, loops (hairpin loop, bulge loop, internal loop, multiloop or junction), and pseudoknot. Thus, secondary structure presents a single polynucleotide chain folding back upon itself with the formation of double helices and looped-out regions. Hairpin, or stem-loop structure, is the most frequent element of RNA secondary structure. It consists of a stem, base-paired double helical region, and a hairpin loop, which has to be minimum of three nucleotides to avoid steric hindrance with base pair in the stem (Bevilacqua and Blose, 2008). Bulge loops form when one or more bases on one strand cannot form base pairs with nucleotides on the other strand. Interior loops contain unpaired bases on both strands and may be either symmetric or asymmetric depending on whether the numbers of unpaired nucleotides on each side are equal or not. Multiloop, or junction, is an area of connection of three or more double helices separated by single-strand regions of zero or more nucleotides (Hendrix et al., 2005). And last but not least is a pseudoknot which is a secondary structural motif with non-nested base pairings.

Graphically base pairings can be presented in different ways (Figure 1.3). One of the most common approaches is to draw a simple planar plot similar to the one shown on Figure 1.1b. Other ways include: dot-bracket notation (A string of dots

a)

```
GCCCGGAUAGCUCAGUCGGUAGAGCAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCGAUUCCGAGUCCGGGCACCA
(((((((..((((........)))).(((((.......))))).....(((((.......))))))))))))....
```

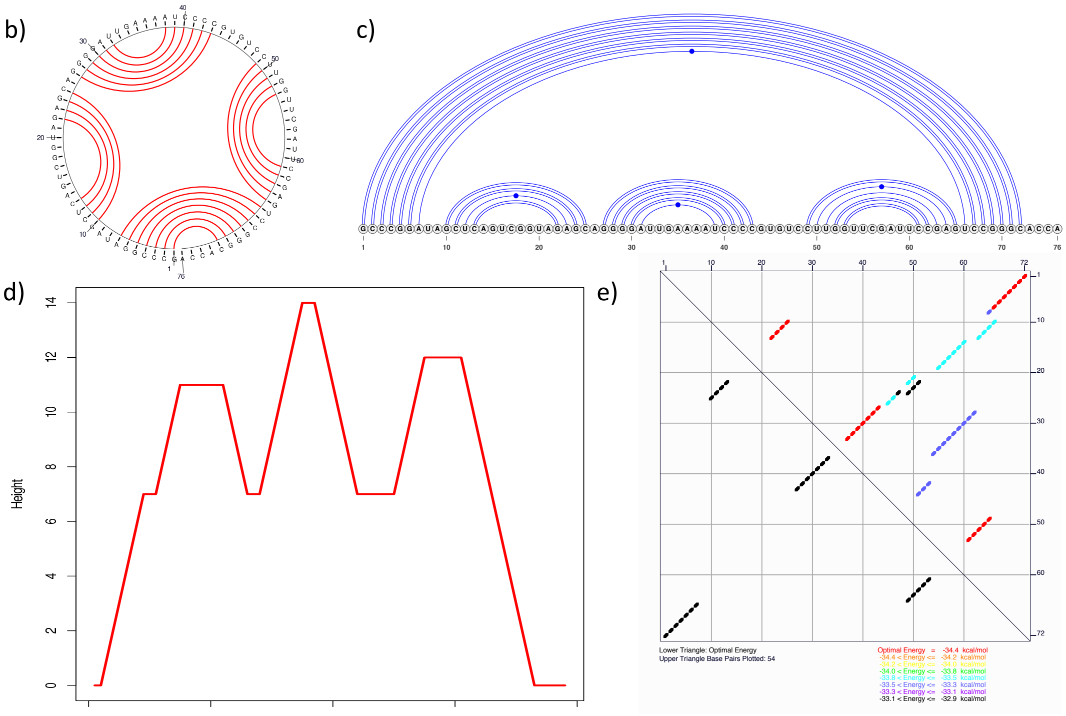

b)
c)
d)
e)
Height
0
2
4
6
8
10
12
14
Lower Triangle: Optimal Energy
Upper Triangle Base Pairs Plotted: 54
Optimal Energy = -34.4 kcal/mol
-34.4 < Energy <= -34.2 kcal/mol
-34.2 < Energy <= -34.0 kcal/mol
-34.0 < Energy <= -33.8 kcal/mol
-33.8 < Energy <= -33.5 kcal/mol
-33.5 < Energy <= -33.3 kcal/mol
-33.3 < Energy <= -33.1 kcal/mol
-33.1 < Energy <= -32.9 kcal/mol

Figure 1.3: ***Different ways to represent graphically RNA secondary structure include:*** ***a)*** *Dot-bracket notation.* ***b)*** *Circular representation (created by sfold (Ding et al., 2004)).* ***c)*** *Linear representation (created by VARNA visualization tool (Darty et al., 2009)).* ***d)*** *Mountain plot (created by RNAfold (Gruber et al., 2008)).* ***e)*** *Energy dot plot (created by mfold (Zuker, 2003)).*

and parenthesis of the same length as the corresponding sequence. A dot at position i means that i^{th} nucleotide is unpaired. In the case i^{th} and j^{th} bases are paired, it is depicted with an open bracket at i^{th} position and closed bracket at j^{th} position.); circular representation, where RNA sequence is presented as a circle and every paired nucleotides are demonstrated with an arc; linear representation, which is similar to circular plot, but the RNA sequence is shown as a line; diagonal dot plot (This presentation is a two dimensional matrix divided into two parts. Usually, the upper right triangle demonstrates probabilities of nucleotides to be paired; namely, the bigger the dot at position (i, j), the higher the probability of i^{th} and j^{th} bases to be in a double-stranded conformation. Similarly, the lower left triangle demonstrates those base pairs that correspond to a conformation with the lowest free energy.); and, mountain plot (It is simpler to consider this type of plot as a graphical representation of a dot-bracket notation. '(', ')', and '.'are represented with a line going up, down, and horizontally, respectively. Thus, the symmetric slopes represent a helix region and plateaus represent single-stranded regions).

It is also a consideration that some sort of hierarchy exists between different secondary structural elements, namely that short-range interactions should form faster than the long-term interactions (Higgs, 2000). Additionally, synthesis of RNA molecules starts at their 5´-end; hence, theoretically it is possible that structures at that end may form before the synthesis of the complete molecule finishes. However, according to thermodynamics, an RNA molecule likely folds in a most thermodynamically stable conformation, in other words, in a structure with minimum free energy that ignores all intermediate states occurring during the folding process. Therefore, it is generally assumed that during the process in which the RNA folds toward the structure with the lowest free energy, some temporal base pairs might form, but those interactions will be rearranged later (Tinoco and Bustamante, 1999).

Many researches, both experimental and computational, have aimed to investigate the stability of RNA structures. For example, it was shown that real tRNAs have thermodynamically more stable structure than random sequences of the same length and base composition (Higgs, 1993, 1995). It is likely that the high stability of tRNA structure compared to alternative conformations is crucial for the function of the molecule (Higgs, 2000). Lower free energies (resulting to higher stability) of real sequences compared to the random ones were also reported for mRNAs (Seffens and

Digby, 1999). However, if shuffling is made with preserving dinucleotide frequencies, then the difference between the minimum free energy (MFE) values of real and random sequences is lower (Workman and Krogh, 1999). Researches of other long RNAs, including rRNA, rRNase P and introns, showed the same inference about higher stability of the natural sequences (Schultes et al., 1999). These facts suggest the existence of thermodynamic constraints in the course of evolution.

RNA sequence can potentially fold into many RNA structures (Wan et al., 2011). Thus, another interesting set of experiments was aimed to study the number of possible suboptimal conformations of one sequence as a function of folding energy (Higgs, 1993). Studying of tRNA demonstrated that there is an energy gap between MFE structure and suboptimal conformations (Wuchty et al., 1999; Chen and Dill, 2000). However, several researchers demonstrated that large RNAs fold via rugged energy landscape, and that during the folding process a molecule can get one of the metastable conformations (Higgs, 2000; Weeks, 2010). Moreover, Höbartner and Micura experimentally demonstrated that even very short RNA sequences in thermodynamic equilibrium could fold into different structures with some ratio of frequencies between them (Höbartner and Micura, 2003). From a theoretical point of view, it resulted in the wide application of partition function to calculating the probabilities of nucleotides to be in a double-stranded conformation. As was mentioned above, at the equilibrium state an RNA molecule should be folded in a conformation with minimum free energy. However, if there are several possible conformations with very close values of free energy, then, according to statistical mechanics, the probability that an RNA molecule folds in each of them is proportional to the Boltzmann factor of the conformation. Nevertheless, the process of transition from one structure to another is not currently well understood.

1.2 Functions of RNA

During the last couple of decades, many articles have been published, which reveal different aspects of RNA functioning both as a protein-coding intermediate and as a regulatory agent. Such dual functionality of RNA does not seem surprising from the viewpoint of the RNA World hypothesis (Kloc et al., 2011). The list of known roles of RNAs has grown up substantially. A large number of non-coding RNAs (ncRNAs) have been discovered, which are transcribed and carry out essential structural, catalytic or regulatory functions within a cell, but do not encode for amino acid sequences (Eddy, 2001). Those functions are influenced by RNA secondary and tertiary conformations according to which RNA molecules can interact with other RNAs, ligands and RNA-binding proteins (RBPs) (Wan et al., 2011). Also, the existence of functional structural elements in untranslated and coding regions of messenger RNAs has been presented in recent experimental and computational papers (Bevilacqua and Blose, 2008; Wan et al., 2011). Plenty of excellent reviews on the different aspects of various functions that RNAs perform have been published; therefore, only some examples will be given here.

The three most known types of ribonucleic acids are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). All of them participate in protein synthesis, and their roles in this process are well documented. Messenger RNAs are copies of genes, which are translated into their corresponding protein; tRNAs are adaptor molecules, one end of which can read the triplet code in the mRNA, and an amino acid is attached to another end; finally, rRNAs move sequentially along the mRNA and catalyze the attachment of amino acids to the growing peptide chain. The described role of tRNA in protein synthesis is the most known function of transfer RNA. Meanwhile, tRNA is also involved in many other biological processes apart from protein synthesis, such as regulating the transcription of mRNA for enzymes associated with biosynthesis of its amino acid both in prokaryotic and eukaryotic systems, participation in making a DNA copy of the viral RNA, and acting as an enzyme inhibitor (Rich and RajBhandary, 1976).

After the first cases of discovering catalytic activity of RNA molecules (namely that the RNA moiety of ribonuclease P from *Escherichia coli* cleaves precursors of the transfer RNA molecules (Guerrier-Takada et al., 1983; Guerrier-Takada and Altman, 1984), and that ribosomal RNA of Tetrahymena whose ribosomal RNA contains a self-splicing exon (Kruger et al., 1982; Zaug et al., 1983, 1984)), it became clear that other catalytic functions might be performed by RNA as well. The range of catalytic activity currently known to be fulfilled by RNAs is rather wide (Tarasow and Eaton, 1998). For instance, priming of DNA synthesis in mitochondria (Wong and Clayton, 1986), priming reverse transcription (Kikuchi et al., 1986), and others (Greider and Blackburn, 1985; Fire et al., 1998).

Since the discovery of RNA catalytic functions, many cases in which RNA molecules carry out other essential functions within a cell, apart from encoding proteins, have been discovered. Most of those RNAs are newly identified non-coding RNAs that participate in different biological processes, and the number of which constantly grows (Eddy, 2001). In particular, there are several interesting cases in which an RNA molecule performs its function almost without folding into any conformation. For instance, microRNAs (miRNAs) have been discovered that bind to complementary regions of target messenger RNA and, thus, inhibit the translation of corresponding protein (Matzke and Birchler, 2005; Bevilacqua and Blose, 2008). Another example is small interfering RNAs (siRNAs) that form base pairs with complementary mRNA and target it for degradation (Matzke and Birchler, 2005). Such mechanism of targeting mRNAs for cleavage is called RNA interference (RNAi) (Matzke and Birchler, 2005). The degradation of mRNA in this case occurs by RNA-induced silencing complex (RISC), the core component of which is either Argonaute protein or RNA-dependent RNA polymerase (Matzke and Birchler, 2005).

Although there are several RNAs, which act without forming specific RNA structure, most RNA molecules have to fold into a particular shape or contain a certain structural element, such as hairpin or pseudoknot, to perform their function (Bevilacqua and Blose, 2008). Regulatory RNAs termed riboswitches have been found in bacterial messenger RNAs and appear to control gene expression (Nudler and Mironov, 2004; Serganov and Patel, 2007; Montange and Batey, 2008). Small RNAs (sRNAs), identified in a wide range of bacteria, have been shown to regulate the expression of proteins, and to modulate the activity and stability of messenger RNAs (Gottesman and Storz, 2011). RNAs also play an important role in mechanisms of phage defense in bacteria (Marraffini and Sontheimer, 2010).

An extremely interesting area of research is the investigation of roles that different types of RNAs can play in different diseases. miRNAs, long non-coding RNAs (lncRNAs), small nuclear RNAs (snoRNAs), large intergenic non-coding RNAs (lincRNAs), and others have been associated with different disorders, which makes them potential new therapeutic targets (Taft et al., 2010; Esteller, 2011). miRNAs have been associated with cancer, neurological disorders, cardiovascular disorders, and other diseases; snoRNAs have been associated with cancer (Esteller, 2011). Many long non-coding RNAs (lncRNAs) have also been associated with different diseases including cancer (Kung et al., 2013). For example, numerous lncRNAs demonstrate different expression levels between normal and cancer cells (Huarte and Rinn, 2010). Several lncRNAs have demonstrated inhibiting tumor suppressor functions in cancer cells (Gutschner and Diederichs, 2012). In addition, long non-coding RNAs have been noted to play roles in metastasis formation (Gutschner and Diederichs, 2012). Obviously, we are only beginning to understand all these molecular mechanisms and there is still plenty to learn.

The biology of viruses and how it is affected by their RNA conformations is another area of great significance. Numerous studies on viruses have revealed the importance of RNA structures for virus assembly (Larson and McPherson, 2001; Schneemann, 2006; Hutchinson et al., 2010). One RNA of HIV-1 virus folds into two different conformations promoting different functions (Lu et al., 2011). Locations of the most stable predicted structural motifs correlate with regions were structures are thought to play an important role in the genomes of the RNA picornavirus (Palmenberg and Sgro, 1997). Translation initiation can occur in the middle of an mRNA sequence owing to the existence of internal ribosome entry sites (IRES) (Jackson and Kaminski, 1995). Additionally, pseudoknots regulating gene expression and genome replication have been identified in many viruses (Brierley et al., 2007). All these examples clearly demonstrate how significant RNA structures are for a diverse range of viral activities.

For many years, it was thought that only non-synonymous nucleotide substitutions in messenger RNAs, which lead to change in the function of the protein, might be deleterious for an organism. Yet, recent studies have demonstrated that mRNAs, both in eukaryotes and prokaryotes, have other hidden non-coding functions which are performed by secondary structural elements and are unrelated to encoding for correspondent proteins (Ulveling et al., 2011). Therefore, those mutations, which result in a change in mRNA secondary structure, might affect some essential biological processes due their affect on the ability of mRNA to interact with other RNAs and proteins in a proper way.

Another generally accepted belief was that pre-mRNA is solely a passive transcript, from which messages are produced by spliceosome, a large protein-RNA complex that removes non-coding introns. Now, there are many examples that RNA structures are actively involved in this regulation and can either inhibit or aid splicing (Warf and Berglund, 2010). Splicing can be regulated by structures in pre-mRNA directly, or by proteins or small molecules, which bind RNA structural elements. One genome-wide study also reported that the existence of predicted secondary structure elements with low free energy near splice sites decreases the efficiency of splicing compared to introns, which have structural elements with higher free energy (Shepard and Hertel, 2008).

Structural elements in messenger RNAs can regulate gene expression by controlling the efficiency of translation initiation. RNA structure can influence accessibility of the start codon or other signals that should be recognized by a ribosome (McCarthy and Gualerzi, 1990). A pseudoknot in retroviral RNA is responsible for ribosomal frameshifting, and results in the production of two proteins at particular ratios, which in turn are required for viral propagation (Chamorro et al., 1992). Frameshifting has been observed in other viruses and species as well (Giedroc et al., 2000), including Rous sarcoma virus (Jacks and Varmus, 1985; Jacks et al., 1988), coronaviruses (Bredenbeek et al., 1990), and bacteria (Tsuchihashi and Kornberg, 1990; Flower and McHenry, 1990).

One of the methods for regulating gene expression is promoting or preventing the degradation of target mRNA (Emory et al., 1992). The range of mRNA half-lives within a single cell is rather wide. For instance, in *E. coli* it can vary from seconds up to nearly one hour (Emory et al., 1992). At the same time, in eukaryotic organisms mRNAs usually live longer than in prokaryotic species (Belasco and Chen, 1988). In many cases, the degradation rate of messenger RNAs is determined by the stability of some hairpin in the structure of this molecule. For example, particular hairpin structure near the 3´-end of the transcript was experimentally demonstrated protecting mRNA from degradation in *Rhodobacter capsulatus* (Belasco and Chen, 1988). A stem-loop element in the 5´-untranslated region *E. coli ompA* mRNA slows down its degradation, although the sequence in this region is relatively unimportant (Emory et al., 1992). However, hairpins that may promote mRNA degradation have been observed as well. For instance, it was shown that in *Escherichia coli*, during the processing of the mRNA for ribosomal protein S20, a hairpin structure helps to localize a cleavage site for a single-strand specific endonuclease RNase E (Mackie and Genereaux, 1993; Bevilacqua and Blose, 2008).

Dual functionality of coding RNAs, both in eukaryotic and prokaryotic organisms, has been recently observed in several cases (Kloc et al., 2011; Ulveling et al., 2011). One very interesting example is a so called transfer-messenger RNA (tmRNA), initially known as 10Sa RNA and expressed from the *SsrA* gene, which combines features of both mRNA and tRNA (Jentsch, 1996; Atkins and Gesteland, 1996). In all eubacteria, tmRNA directs a process called trans-translation aimed to release stalled ribosomes from a defective mRNA (lacking a stop codon) (Keiler et al., 1996; Gillet and Felden, 2001). The structure of tmRNA from *E. coli* was determined experimentally (Felden et al., 1997). It contains two domains: an mRNA-like and a tRNA-like, which can be charged with alanine that will link to the truncated polypeptide chain. Transfer-messenger RNA engages the stalled ribosome and mRNA-like domain replaces the defective mRNA bound. Hence, translation switches to tmRNA from the broken mRNA. The following translation of the tmRNA adds a peptide tag to the nascent protein, which targets the polypeptide for rapid degradation (Keiler, 2008). Thus, trans-translation is a quality control mechanism, which ensures that synthesized proteins are correct (Keiler, 2008).

Another interesting case is the *p53* tumor suppressor, mutations of which are found in approximately 50% of human tumors (Soussi and Wiman, 2007). Due to this fact, *p53* is a fascinating research target for clinicians and researchers. Usually *p53* protein is persistently degraded, but, upon stress, its activation eliminates tumor cells (Farnebo et al., 2010; Ulveling et al., 2011). *p53* tumor suppressor activity is mainly regulated by the E3 ubiquitin ligase *Mdm2*, which binds *p53* protein and targets it for degradation (Candeias et al., 2008; Farnebo et al., 2010). Recently, it has been demonstrated that messenger RNA of *p53* gene interacts directly with *Mdm2* protein and thus restrains the *Mdm2* activity of promoting *p53* degradation (Candeias et al., 2008). Additionally, it was shown that synonymous single nucleotide polymorphisms in the *p53* mRNA can impair its interaction with *Mdm2*, and, as a consequence, activity of *p53* in this case will be decreased (Candeias et al., 2008).

Other known cases of messenger RNAs that can also fulfill some structural role include: *VegT* mRNA from *Xenopus*, which is involved in the cytokeratin network of primordial germ; *oskar* mRNA from *Drosophila melanogaster*, which was shown being responsible for the early oogenesis; and others (Kloc et al., 2011; Ulveling et al., 2011).

Undoubtedly, functions of molecules within a cell depend on their three-dimensional structures. Therefore, to understand the functions that RNA molecules may perform, it is crucial to be able to determine atomic coordinates nucleotides in a 3D conformation and what residues are base-paired. Nevertheless, neither present experimental techniques nor theoretical predictions allow analyzing high-throughput data on tertiary structures of RNAs. Moreover, it is extremely difficult to predict three-dimensional folding. As a result, the first step toward determining and studying the RNA tertiary conformation is to identify its secondary structure (base pairing interactions within a molecule).

1.3 Experimental Techniques

The first ever experimentally determined structure was a conformation of transfer RNA (tRNA), which was crucial to the understanding of molecular mechanisms of protein synthesis and other biological functions of tRNA (Sigler, 1975). Initially an X-ray diffraction analysis was applied to yeast phenylalanine tRNA to measure its three-dimensional folding (Kim et al., 1971, 1972, 1973). The electron density maps from the experiments showed double helix regions connected to each other with weaker regions of electron density. These were interpreted as a confirmation of the idea first proposed by Holley and his collaborators (Holley et al., 1965) who sequenced alanine transfer RNA from yeast and suggested that tRNAs could be folded into a secondary structure widely known as cloverleaf (Figure 1b). Later, that method was improved and it allowed measuring the structure first at 3 angstrom resolution (Suddath et al., 1974; Kim et al., 1974; Robertus et al., 1974b), and then at 2.5 angstrom resolution (Quigley et al., 1975; Ladner et al., 1975). Further research enabled reciprocal space refinement and the ability to model precise atomic coordinates of the entire tRNA molecule (Sussman et al., 1978; Hingerty et al., 1978). Subsequently, the X-ray crystallography was used to determine the structure of another elongator transfer RNA (Giege et al., 1977) and yeast aspartic acid tRNA (Westhof et al., 1988b), confirming the proposed general tRNA structure. To an RNA, other than tRNA, the X-ray crystallography was first applied in 1994, when a structure of hammerhead ribozyme was determined at 2.6 angstrom resolution (Pley et al., 1994). Further development of crystallographic methods is described in the review by Holbrook (Holbrook and Kim, 1997).

X-ray diffraction from crystals gives the highest accuracy, but does not say whether the structure in the crystal is the same as the structure in solution, hence, other methods had to be adopted to answer this question (Rich and RajBhandary, 1976; Holbrook and Kim, 1997). The high-resolution NMR (nuclear magnetic resonance) spectroscopy study was first applied to determine the structure of yeast phenylalanine tRNA (Kearns and Shulman, 1974). Next, high-resolution NMR spectra of several other purified tRNAs was examined by different groups (Reid and Robillard, 1975; Reid et al., 1975; Daniel and Cohn, 1975; Wong et al., 1975; Bolton and Kearns, 1975). Results of those NMR studies supported the conclusion that the structure of the tRNA molecule in solution is identical to the 3D conformation determined in the crystal (Rich and RajBhandary, 1976). Such knowledge of the three-dimensional structure of tRNA led to the better understanding of chemistry and the role of transfer RNA in different biological processes (protein synthesis, transcription of messenger RNA, reverse transcription, etc.) (Rich and RajBhandary, 1976). The size of the RNA that can be analyzed at atomic resolution by NMR is continually increasing, but slowly. For many years the upper limit was approximately 100 nt (Tinoco and Bustamante, 1999). Only recently, a nuclear magnetic resonance approach, which enables detection of structural elements within longer sequences, has been developed and applied to investigation of HIV-1 5´-leader RNA (Lu et al., 2011). In addition, both X-ray approach and NMR studies require of large amounts of highly purified material (Ehresmann et al., 1987).

Another experimental approach to probe structure is the use of chemical modifications to test the reactivity of every nucleotide. An RNA of interest is modified somehow by treating it with a specific chemical reagent in such a way that any two modification events are independent from each other (Weeks, 2010). Some bases will be reactive while others will react at a much slower rate. Such reactivity of bases identifies which nucleotides are unpaired and which ones are paired. Two approaches of determining modified nucleotides are using end-labeled RNA molecules, which allows the detection of scissions in the RNA chain, and primer extension, which detects stops of transcription at modified sites (Ehresmann et al., 1987). What nucleotides should react depends upon the reagent used (Rich and RajBhandary, 1976). The list of reagents, which have been applied to examine the secondary and tertiary structures of transfer ribonucleic acid, includes: β-ethoxy-α-ketobutyraldehyde (kethoxal), which reacts with guanine (Litt, 1969); methoxyamine (Cashmore et al., 1971; Robertus et al., 1974a; Chang, 1973) and hydrogen sulfide (Miura et al., 1982), which react with cytosine; 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (CMCT), which used to map unpaired uridines and guanosines (Robertus et al., 1974a; Chang, 1973); ethylnitrosourea (ENU), a reagent ethylating phosphates in nucleic acids (Vlassov et al., 1981, 1983; Garret et al., 1984a; Romby et al., 1985); dimethyl sulfate (DMS), which reacts with the N1 of adenine, the N3 of cytosine and the N7 of guanine (Peattie and Gilbert, 1980; de Bruijn and Klug, 1983; Garret et al., 1984a; Romby et al., 1987); diethyl pyrocarbonate (DEPC), which reacts with the N7 of adenosines (Peattie and Gilbert, 1980; de Bruijn and Klug, 1983; Garret et al., 1984a; Romby et al., 1987); and others (Igo-Kemenes and Zachau, 1969, 1971; Vary and Vournakis, 1984a; Brunel and Romby, 2000; Rocca-Serra et al., 2011).

Following tRNA, different parts of ribosomal RNA became the subject of interest and were studied with different chemical probes. For example, mouse 5S ribosomal RNA was tested with kethoxal (Miura et al., 1983b) and hydrogen sulfide (Miura et al., 1983a), *E. coli* 16S ribosomal RNA was analyzed with diethyl pyrocarbonate (Van Stolk and Noller, 1984); investigation of the interaction of ribosomal protein S4 with *E. coli* 16S rRNA was done with using of kethoxal and DMS (Stern et al., 1986), binding of ribosomal protein S8 to 16S ribosomal RNA was studied with DMS, CMCT, DEPC and ethylnitrosourea (Mougel et al., 1986); bisulfite, which converts unpaired cytosine to uridine, was used to probe the RNA structure of 5S rRNA from *Spinacea oleracea* (Pieler et al., 1983); and many others.

Another group of methods is similar to chemical probing and is based on using a structure-specific enzymatic probe, which cleaves RNA at single- or double-stranded regions. In fact, many groups combined data from chemical and enzymatic structure probes (Ehresmann et al., 1987). As with chemical reagents, there are several enzymes, most of which cut the RNA within unpaired regions. First studies of RNA base pairing were performed applying single-stranded-specific RNase T1, which cuts unpaired guanosines, and S1 nuclease, which cleaves preferentially all single-stranded nucleotides, to digest transfer RNAs from different organisms and at different environment conditions (Wurst et al., 1978; Wrede et al., 1979b; Wrede and Rich, 1979; Wrede et al., 1979a). The obtained results were consistent with previously determined three-dimensional folding.

Other enzymatic reagents that are actively used to probe RNA structure include RNase U2, which cuts unpaired adenines (Mougel et al., 1986; Baudin et al., 1987); RNase Cl3, which cleaves unpaired cytidines, adenosines and uridines, but for the latter two it requires longer incubation time and high concentration of enzyme (Florentz et al., 1982); RNase T2, which cuts single-stranded adenosine residues (Vary and Vournakis, 1984b; Kean and Draper, 1985; Romaniuk, 1985; Christiansen et al., 1987). There are also other enzymes used to probe RNA structure, which similar to nuclease S1 cleave single-stranded RNA regions without being specific to a particular nucleotide. For example, RNase ONE, which cleaves all unpaired bases, was used in studying structure elements of umbravirus and panicovirus (Wang et al., 2009); RNase J1 from *Bacillus subtilis* was used to solve the structure of the *hbs* mRNA (Daou-Chabo and Condon, 2009); Neurospora crassa nuclease was used to study interactions between beef tryptophan transfer RNA and avian myeloblastosis reverse transcriptase (Garret et al., 1984b).

RNase V1 from cobra venom is the only enzyme that cuts preferentially double-stranded regions (Wan et al., 2011). This ribonuclease specifically cleaves RNA in regions that are helical, indicating where the RNA is base paired. It was first applied to probe the structure of yeast phenylalanine and *E. coli* methionine tRNAs, and results demonstrated that the V1 nuclease also recognizes non-canonical base pairs and tertiary interactions, in addition to usual secondary helices (Lockard and Kumar, 1981). Owing to this uncommon specificity, it is a widely used enzyme for probing RNA structure (Favorova et al., 1981; Troutt et al., 1982; Lowman and Draper, 1986).

The use of chemical modifications for testing RNA structure is very time consuming and requires a lot of effort (Weeks, 2010). Although chemical probing with a variety of structure-specific probes provides comprehensive information at the nucleotide level, data obtained solely from chemical probing techniques do not show which nucleotides are base pairing with each other (Ehresmann et al., 1987). However, such data can be directly incorporated as folding constraints into dynamic programming algorithms for secondary structure prediction (Mathews et al., 2004).

The current state of art in chemical probing techniques is the one termed Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) and based on the discovery that the nucleophilic reactivity of a ribose 2'-hydroxyl group is gated by local nucleotide flexibility (Merino et al., 2005; Wilkinson et al., 2005, 2006). It was first applied to reproduce the well-studied structure of aspartic acid transfer RNA in yeast (Westhof et al., 1985, 1988b,a; Perret et al., 1990). Single nucleotide resolution SHAPE chemistry computes nucleotide flexibility at all four ribonucleotides and differentiates paired residues from flexible ones. Knowledge of such local flexibilities of nucleotide positions allows determining the RNA secondary structure. However, SHAPE chemistry is a rather slow technique. Original protocol required two days to complete for RNA with only 100-200 nucleotides (Wilkinson et al., 2006). Later, a new faster-acting reagent was designed to improve the SHAPE chemistry (Mortimer and Weeks, 2007); nonetheless, the entire procedure still required a lot of time.

Further development of SHAPE technology allowed analyzing long RNAs in a single experiment and measuring flexibility at more than 99% of the bases (Wilkinson et al., 2008; Watts et al., 2009). Thus, SHAPE measurements yield comprehensive information about what nucleotides are paired and what nucleotides are unpaired in the RNA structure. Such improved SHAPE technology was used to assess the RNA secondary structure of a complete HIV-1 genome and revealed many previously unrecognized structural elements and long-range interactions (Wilkinson et al., 2008; Watts et al., 2009). Additionally, as with other chemical probing techniques, experimental data produced by the SHAPE analysis can be coupled with computational prediction methods. This feature has been implemented in the RNAstructure program (Mathews et al., 2004) and increases the accuracy of secondary structure predictions dramatically (Deigan et al., 2009; Low and Weeks, 2010). For example, taking into account SHAPE reactivity information in benchmarking was performed on 16S ribosomal RNA of *Escherichia coli*, the crystal structure of which was earlier solved at 3 angstrom resolution (Wimberly et al., 2000). The accuracy of the structure, based solely on a thermodynamic model, was less than 50%; however, the SHAPE-directed structure modeling of *E. coli* 16S rRNA demonstrated higher than 95% accuracy (Deigan et al., 2009).

The two main disadvantages of all the methods described above are that they are limited to probing one RNA molecule at a time, and only a few hundred bases can be examined per experiment. Recently, several high-throughput methods of determination of the RNA structures have been suggested. The first technique, termed Parallel Analysis of RNA Structures (PARS), was successfully applied to measure the secondary structures of the messenger RNAs for over 3,000 distinct transcripts of the *Saccharomyces cerevisiae* (Kertesz et al., 2010). The method is based on treating mRNAs separately with S1 nuclease and RNase V1, nucleases, which cleave single- and double-stranded RNA, respectively. Then, RNA is converted into a cDNA library. High-throughput sequencing of cDNA library enables identification of the cleavage sites. Those nucleotides, whose RNase V1 cleavage number is higher than S1 nuclease cleavage number, are considered base-paired. And the other way round, the nucleotides, whose RNase V1 cleavage number is lower than S1 nuclease cleavage number, are considered unpaired. The enzymatic footprinting takes about five days to complete and subsequent sequencing and analysis requires six to eight days (Wan et al., 2013).

Analysis of yeast structural profiles measured by PARS revealed that nucleotides in the coding regions of mRNAs are prone to appear in double-stranded conformations more often than nucleotides in the untranslated regions (UTRs) (Kertesz et al., 2010; Mauger and Weeks, 2010). A similar finding was reported for the HIV-1 virus genome (Watts et al., 2009). Another detail demonstrated by PARS was that the efficiency of mRNA translation is anti-correlated to the probability of nucleotides near the translation start site to be in a double-stranded conformation (Kertesz et al., 2010; Mauger and Weeks, 2010). Lately, a new approach similar to PARS analysis, termed Parallel Analysis of RNA structures with Temperature Elevation (PARTE), was suggested by the same group (Wan et al., 2012). It was applied to probe yeast RNA structures and different temperatures and it helped to identify thousands of putative RNA thermometers (Wan et al., 2012).

Zheng et al. combined nuclease-based mapping with high-throughput sequencing and applied this technique, which they called dsRNA-seq, to a genome-wide analysis of Arabidopsis (Zheng et al., 2010). The approach was based on treating RNA samples with a single-strand specific RNase One. It allowed them to identify highly stable regions of secondary structure, and also to identify many new small RNAs. Later, this approach evolved. A double-strand specific RNase V1 was added to this methodology; and, it was applied to the analysis of *Drosophila melanogaster* and

Caenorhabditis elegans transcriptomes (Li et al., 2012). Interestingly, they found that nucleotides both in the 5´- and 3´-untranslated regions have higher propensity to be in a double stranded conformation than nucleotides in the coding regions. This finding was interpreted as the existence of many regulatory signals or interaction sites for RNA-binding proteins (Li et al., 2012).

Finally, an alternative technology, termed fragmentation sequencing (Frag-seq), was used to probe for single-stranded regions of mouse transcriptome (Underwood et al., 2010). The method relies on high-throughput sequencing of RNA fragments that resulted from treating RNAs in solution with P1 endonuclease, which cleaves the RNA of interest at single-stranded regions. A high number of cleaves at a particular position indicates that this nucleotide is unpaired. Through this method, known structured regions in noncoding RNAs were validated and new, previously unprobed RNAs, were tested.

Recently, SHAPE technology has also been paired with deep sequencing and was termed SHAPE-Seq (Lucks et al., 2011). Compared with other high-throughput methods of probing RNA structure, which use large nucleases, SHAPE-Seq uses a small chemical probe. As a result, it has considerably higher accuracy of measurement. This method was applied to probe the structure of the highly conserved *Bacillus subtilis* RNase P and to identify changes in the structure resulting from single nucleotide polymorphisms (SNPs). The method also can be further extended to determine how the structure changes due to RNA-RNA or RNA-protein interactions (Lucks et al., 2011).

It is interesting that PARS, dsRNA-seq and Frag-seq appeared almost simultaneously in research. This shows that high-throughput methods of RNA structure mapping are of great interest and rapid further development of such techniques is very likely. However, the biological importance of RNA and the long absence of experimental techniques for measuring RNA structure have resulted in a fast growing number of works that are analyzing RNA functions based solely on theoretical predictions.

1.4 Theoretical Prediction Methods

To better understand of the biological functions of RNA molecules within a cell, it is crucial to know their structures. Despite the fact that RNA structures play important roles in different biological processes, the experimental techniques to probe RNA structure by high-throughput sequencing are only beginning to appear. Therefore, the majority of researches connected to RNA structure is based on theoretical predictions of secondary structures.

There are several problems that make the theoretical prediction of RNA structures very complicated. First of all, RNA structures are dynamic, which means that RNA conformation depends on surrounding conditions (such as temperature, salt concentrations, etc.) and on the functional role that an RNA molecule is supposed to perform at the particular biological state (Weeks, 2010; Wan et al., 2011). Thus, many of the processes that influence the conformation into which RNA folds (e.g. folding kinetics, higher-order interactions, etc.), are too complex to be taken into account to produce high accuracy results. Another problem is that the number of theoretically possible conformations for an RNA sequence increases exponentially with the length of the sequence, N (Zuker and Sankoff, 1984; Mathews, 2006):

$$Number\ of\ secondary\ structures \approx (1.8)^N$$

From the experimental results, it is also becoming clear that RNA can fold into many stable states with energy somehow different from the global minimum (Höbartner and Micura, 2003; Weeks, 2010). Therefore, the longer the sequence, the worse the quality of prediction is. Several studies were aimed to assess the accuracy of theoretical predictions. For instance, Higgs demonstrated that 85% of tRNA structures were correctly predicted (Higgs, 1995). However, the accuracy of predicting longer sequences drops significantly (Zuker and Jacobson, 1995; Konings and Gutell, 1995; Fields and Gutell, 1996; Doshi et al., 2004). For instance, even the most accurate dynamic programming algorithm predicts correctly less than 50% of the base pairs for 16S ribosomal RNA of *E. coli* (Deigan et al., 2009; Weeks, 2010).

Since the conformation of RNA with the lowest possible value of free energy is considered the most thermodynamically stable, one of the most common methods of secondary structure prediction is based on searching for the MFE structure. But, to incorporate such energy parameters into a prediction algorithm, they have to be experimentally measured first. Thus, one potential explanations of poor accuracy is incorrectness of the experimental energy parameters for base pairings (Doshi et al., 2004). This factor can be especially important in the case that several alternative structures, with similar values of free energy, do exist. Another reason is that more than one structure may exist at equilibrium (Tinoco and Bustamante, 1999). Thus, predicting only one structure for a long sequence may not show the entire picture.

A possible way to increase the quality of RNA secondary structure predictions, especially for large RNA molecules, is to use data from experimental probing as constraints into the prediction algorithms. Several works have demonstrated that incorporating chemical probing data into a dynamic programming algorithm improves the accuracy of predictions dramatically (Mathews et al., 2004). For instance, indicating those bases, which demonstrated high reactivity towards chemical probes, as certainly unpaired helped to increase the accuracy of prediction from 50% to 72% for 16S ribosomal RNA (Weeks, 2010).

Computer modeling of RNA molecules and computing of atomic coordinates, when taking into account electrostatic interactions, are still extremely difficult tasks (Auffinger and Westhof, 1998). Therefore, in those cases when it is absolutely necessary to know atomic coordinates, the common solution is to use X-ray diffraction (Holbrook and Kim, 1997). However, in most cases, we would like to know the structure well enough to be able to understand the function it performs, instead of highest possible resolution (Tinoco and Bustamante, 1999). Thus, predicting RNA structures can be very useful either in interpreting experimental data concerning a particular RNA function, or in suggesting new RNA regions that may be functionally important and testing them experimentally (Seetin and Mathews, 2012).

It is generally accepted that the approach termed comparative sequence or covariation analysis is the most reliable method of determining a secondary structure of an RNA molecule (James et al., 1989; Pace et al., 1999; Weeks, 2010). The underlying assumption of this technique is that we would intuitively expect that if there is a functionally important element of secondary structure then all the available sequences must have this element of the structure (in other words, that structure should be more conserved by evolution than sequence). Therefore, the main goal that resulted from this idea is to find a base pairing pattern that fits all the sequences (i.e. if a nucleotide substitution occurs on one side of a helix, which disrupts the structure, a compensatory substitution should occur on the other side of the helix). In this case, we will see a high covariance or mutual information between those two positions. The main advantage of this algorithm is that it predicts both secondary and tertiary structures. However, this approach requires the existence of a large number of homologous sequences, which makes it impossible to apply in most cases. Another disadvantage of constructing a covariations model from a multiple alignment is that it may require significant effort from the researcher (Low and Weeks, 2010; Seetin and Mathews, 2012).

This method was first applied to the analysis of transfer RNA sequences, which demonstrated the existence of correlation between mutations occurring in the positions that are base paired according to the cloverleaf model (Madison et al., 1966; Levitt, 1969). Usually comparative analysis is performed on the sequences of the same RNA molecule from different species, but homologs from the same organism can also be investigated (Seetin and Mathews, 2012). The physical model underlying the covariation method is the following. Let us consider two nucleotides paired with each other, and, to simplify the description, let us assume that these bases are G and C. Mutation rates are usually rather low; hence, it is considered that two mutations cannot occur simultaneously and the compensatory substitutions represent a two-step process (Higgs, 2000). If a mutation happens in one of these two bases and it changes, for example, to U, but that base pair was important to the structure stability, then a second, compensatory mutation occurs later in the other base which will form a new AU base pair. Thus, homologous sequences used in the analysis may have a low level of sequence identity, but their helical regions will be perfectly aligned (Seetin and Mathews, 2012). The analysis of Drosophila rRNA clearly demonstrated that compensatory mutations usually occur through intermediate GU base pairs (Rousset et al., 1991).

Unfortunately, in the majority of cases, there are not enough homologous sequences to apply covariation analysis. Thus, in those cases, it is crucial to be able predict an RNA secondary structure from a single sequence; the most common method for this is based on free energy minimization (Mathews and Turner, 2006; Shapiro et al., 2007).

First attempts to estimate stability of RNA secondary structure by minimizing folding energy were done more than 40 years ago by Tinoco et al. (Tinoco et al., 1971, 1973). According to thermodynamics, at equilibrium state the structure with the lowest free energy should dominate (Turner et al., 1988). Algorithms based on this thermodynamic model usually find many possible structures and estimate a free energy value for each of them. The main assumption is that the total energy of a conformation is just a sum over energies of separate local structural components, like stems and loops (Tinoco and Bustamante, 1999). Free energies of base paired regions are negative, hence, more favorable; while, loops are usually taken into account with free energy penalties because loops do not make the structure more stable and are considered as unfavorable elements (SantaLucia and Turner, 1997). Additionally, the energy of a double-stranded region depends on the sequence. Namely, the energy of a helix depends not only on the type of base pairs in the helix, but also on the order of base pairs, so called base stacking. Stability of loops depends on the sequence as well (Mathews et al., 1999). Moreover, accuracy of free energy associated with a loop is much lower than accuracy of helix parameters (SantaLucia and Turner, 1997). At the same time, the energy of a base pair is considered dependent only on the types of adjacent base pairs. This is termed the nearest neighbor model (Tinoco and Bustamante, 1999). Thus, to assess the free energy of a particular structure, it is usually divided into elementary parts (energies of which are known or reasonably estimated) and then energies of those simple parts are combined (Higgs, 2000). The lower the free energy of a structure, the more stable this structure is considered.

The results of predictions made by free energy minimization algorithms strongly depend on experimental thermodynamic data. As the accuracy of measuring free energies of different interactions progresses, the quality of predictions improves as well (Mathews et al., 1999). Usually optical melting studies are used to experimentally determine energy parameters for the nearest neighbor model (Xia et al., 1998; Mathews and Turner, 2002b; Mathews et al., 2004). The Nearest Neighbor Database was created recently to summarize those parameters (Turner and Mathews, 2010). However, there are different methods of measuring the energy and different models for the stacking free energy in helices (SantaLucia and Turner, 1997). For example, there are pure computational approaches to estimate energy parameters (Andronescu et al., 2007) and theoretical optimization of experimentally measured parameters (Mathews et al., 1999).

In 1978, Nussinov et al. presented the first dynamic programming algorithm to find a particular folded form with the largest number of base pairs (Nussinov et al., 1978). The biological underground idea is that every base pairing contributes to the stability of the structure. Therefore, the more base pairings the structure has, the more stable an RNA molecule is. Thus, the algorithm maximizing the number of base pairings was developed. This algorithm included an important simplification, however, it did not take pseudoknots into account. If we number all the nucleotides in a sequence from 1 to N, then nucleotides i^{th} and j^{th} can form a base pair only if they are complementary and at least three other bases exist between them. Now, let us assume that i^{th} nucleotide is paired with j^{th} and k^{th} nucleotide is paired with l^{th}. There are three possible variants of their mutual location: (i) one of the base pairs is located aside of the other one ($i < j < k < l$); (ii) one pair is within the other ($i < k < l < j$), so called nested base pairs; (iii) they are intersecting ($i < k < j < l$). The latter case is called pseudoknot (Higgs, 2000). Eliminating pseudoknots resulted in the fact that the original algorithm has $O(n^3)$ time complexity and requires $O(n^2)$ memory; and hence, can be applied in most cases. Later, several attempts were also made to accelerate this folding algorithm by using graphics processing units (GPU) (Chang et al., 2010; Stojanovski et al., 2012; Su et al., 2013).

Usually, the structure predicted by the original Nussinov's algorithm is very different from a conformation into which a real RNA folds. This is because the algorithm takes into account only the number of possible base pairs and maximizes this number, which is not the best model from a thermodynamics point of view. It also does not take into account different energy parameters for different base pairs, and there are no penalties for the loops. However, a new version of their algorithm for RNA structure predictions with incorporated energy parameters for base pairs was presented two years later (Nussinov and Jacobson, 1980).

Nevertheless, energy rules for base stacking and destabilizing regions cannot be incorporated into the Nussinov's algorithm. This problem was solved by Zuker and Stiegler, who designed a new dynamic programming algorithm that allows taking into account such energy parameters (Zuker and Stiegler, 1981). In many cases, the structure with the minimum free energy was not consistent with those measured experimentally, but could be observed among the structures with free energy close to the minimum. Therefore, Zuker developed a new version of the algorithm, which allows finding not only the MFE structure, but all suboptimal structures within a particular range of free energy (Zuker, 1989). Theoretically, suboptimal structures correspond to less stable conformations. Yet, they also can be considered as highly probable alternate structures because of simplifications used in the algorithm and errors in the energy parameters measurements. This algorithm became a basis for a popular software package termed mfold (Zuker, 2003).

In 1990 McCaskill proposed a novel application of dynamic programming, namely to calculate partition function instead of just the MFE structure (McCaskill, 1990). The partition function describes the entire ensemble of secondary structures in thermodynamic equilibrium and is defined as a sum of Boltzmann factors over all the possible conformations of a particular sequence:

$$Z = \sum_{q_i} e^{\frac{-\Delta E(q_i)}{k_B T}}$$

where $-\Delta E(q_i)$ represents the difference in free energies between a particular conformation, q_i, and an unfolded state; k_B is the Boltzmann constant; and T is the temperature in kelvins. According to statistical physics, the probability of a given conformation q_i in the equilibrium can be assessed as:

$$P_i = \frac{e^{\frac{-\Delta E(q_i)}{k_B T}}}{Z}$$

From this formula it is obvious that the MFE structure of an RNA molecule, as well as the weights of different conformations in the partition function, depends on the temperature. Therefore, this approach enables investigating how an ensemble of alternative structures alters upon the temperature change instead of studying the most probable conformation. In addition, from the partition function, it is possible to compute the probability for any two bases to be paired (McCaskill, 1990). Thus, calculating a partition function reveals important information about the complete ensemble of possible alternative conformations and enables assessing the power of prediction algorithms (McCaskill, 1990). For instance, it was demonstrated that the base pairs, which have high predicted probability to be paired, have higher chances to be present in a real structure (Mathews et al., 2004). Also, it was shown that the probabilities of nucleotides to be paired are less sensitive to errors in free energy parameters (Layton and Bundschuh, 2005).

Later, a dynamic programming algorithm by Zuker for free energy minimization and computation of partition function was combined into a package termed the Vienna RNA Package (Hofacker et al., 1994). This implementation was demonstrated to be much faster than the original versions and became rather popular (Hofacker et al., 1994).

Since RNAs, especially long ones, may fold into different co-existing conformations, the partition function approach was extended by Ding and Lawrence to select a statistical representation of structures from the ensemble (Ding and Lawrence, 2003; Ding, 2006). This approach allows for easily estimating the probabilities of particular secondary structural elements, instead of individual base pairs. Also, it was demonstrated to be very useful in rational design of RNAs, which have to fold into a particular structure (Ding, 2006).

The methods described above do not take into account the kinetic aspects of RNA folding, rather, they try to estimate an equilibrium state. However, there is also a class of algorithms that try to predict RNA structure by simulating the folding process instead of simply optimizing free energy (Abrahams et al., 1990; WuJu and JiaJin, 1998). For instance, specially adapted forms of Monte Carlo simulations were applied (Fernández, 1992; Schmitz and Steger, 1996). Also, several groups proposed using genetic algorithms to solve the optimization problem of kinetic folding (van Batenburg et al., 1995; Benedetti and Morosetti, 1995; Shapiro and Wu, 1996; Shapiro et al., 2001). Nevertheless, all those algorithms have not achieved prevalence.

Since comparative analysis is considered the most accurate method, there have been many attempts to automate this process and to combine comparative and thermodynamics methods. Usually, such covariation models are built with using stochastic context-free grammars (SCFG), which use formal language theory to develop a grammar to describe RNA secondary structure, or hidden Markov models (Durbin, 1998; Dowell and Eddy, 2004; Do et al., 2006; Shapiro et al., 2007; Jossinet et al., 2007). The first type of such techniques begins with constructing a multiple sequence alignment and then predicting the consensus structure for the alignment (Hofacker et al., 2002; Bernhart et al., 2008; Bernhart and Hofacker, 2009; Mathews et al., 2010). This approach is very fast, but the quality of the consensus structure depends a lot on the quality of the alignment. Such methods, for example, were applied to analyze genomes of a wide range of viruses, including HIV-1, hepatitis C virus, hantavirus, and others (Lück et al., 1996; Tabaska et al., 1998; Hofacker et al., 1998; Hofacker and Stadler, 1999). The second technique does the opposite. They predict a set of suboptimal structures, which have a free energy value close to the minimum, and then search for a structure, which is common to all sequences in the alignment. This paradigm is implemented in the RNAshapes software tool (Steffen et al., 2006), which is very fast (Mathews et al., 2010; Seetin and Mathews, 2012). However, authors of this approach try to find a common topology, termed abstract shape (Giegerich et al., 2004), instead of analyzing real structures (Reeder and Giegerich, 2005). Finally, the last approach is to fold and align the sequences simultaneously (Gorodkin et al., 1997b,a; Mathews and Turner, 2002a). One of the first such algorithms was proposed by Sankoff (Sankoff, 1985), but for the S input sequences, it has a complexity of $O(n^{3S})$ which makes it impractical for the majority of cases. The algorithm suggested by Gorodkin et al. has a complexity of $O(n^4)$, but does not take into account multi-branched loops (Gorodkin et al., 1997b). Accounting for multi-branched loops increases the complexity to $O(n^6)$ and also makes it impractical for most applications (Gorodkin et al., 1997b). As this is the most expensive approach, it is usually applicable only to two sequences (Mathews et al., 2010).

As was mentioned earlier, there are experimental evidences that pseudoknots can be functionally important. Moreover, the number of known pseudoknots has been constantly increasing and has resulted to creating a pseudoknot database (Taufer et al., 2009). However, most of the developed algorithms for structure prediction do not contain pseudoknots (Lyngsø and Pedersen, 2000). Predicting pseudoknots remains a huge challenge and the accuracy of currently existing algorithms is still rather low. The main complication is that RNA secondary structure prediction with pseudoknots was proved being NP-hard, which means that prediction of the structure for a particular sequence can be performed computationally in a reasonable amount of time only for very short RNAs (Lyngsø, 2004). Thus, some simplifications are used in attempts to predict pseudoknots.

Another problem of predicting pseudoknots, apart from the high complexity of algorithms, is that although the melting behavior and thermodynamics data for some types of pseudoknots have been obtained (Wyatt et al., 1990; Gregorian and Crothers, 1995; Qiu et al., 1996; Theimer et al., 1998; Theimer and Giedroc, 1999; Gonzalez Jr and Tinoco Jr, 1999), there is still limited experimental information about energy parameters of pseudoknots. As a result, there have even been attempts to estimate thermodynamic parameters for pseudoknots only from theory (Gultyaev et al., 1999).

There are several groups, which have been working on developing dynamic programming algorithms for predicting pseudoknots. For example, Rivas and Eddy suggested a new dynamic programming algorithm that takes into account simple topologies of pseudoknots (Rivas and Eddy, 1999). However, the algorithm complexity is $O(n^6)$, which makes it impractical for applying to long sequences. A dynamic algorithm with $O(n^4)$ complexity was suggested by Akutsu (Akutsu, 2000). However, this algorithm is similar to the original algorithm by Nussinov in the sense that it just maximizes the number of base pairs in the structure and does not take into account energy parameters. Stochastic modeling (Cai et al., 2003; Xayaphoummine et al., 2003), heuristic algorithms (Ruan et al., 2004; Ren et al., 2005), and other approaches also have been suggested to solve this problem. However, the accuracy of those approaches is not very high yet.

1.5 Thesis Motivation and Outline of the Work

For many years it was considered that proteins and protein interactions control practically every biological process. More recently, however, an increasing number of scientific papers have been published that describe a growing list of examples for when RNA structures play important roles in cellular regulatory functioning. Therefore, a deeper understanding of how RNA folds is required, including the kinetic aspect of RNA folding, alternative conformations, etc. At the same time, indisputably, mutations occur in RNA molecules. Hence, some very interesting questions are: What are the relationships between sequence and structure? How big is the conformational change of RNA resulting from occurred mutations? Can structure conservation be a filter to mutations disrupting the structure? Can modifications in RNA structure play a role in different diseases and/or infections? Answering these and many other questions about RNAs will help us better understand the processes within a cell, will result in producing better vaccines and medicine in the future, and will assist in investigating predecessors of the RNA World and of the origin of life.

The following chapter presents the results that have been achieved from our contributing efforts to investigate these questions. Four articles are presented. The first paper describes an analysis of sequence-structure relationships in yeast mRNAs based on the first-ever published genome-wide measurements of base pairing propensities in mRNA structures. The second article presents an attempt to explain a molecular mechanism of the rather famous cold-adapted, temperature-sensitive phenotype of influenza virus. It shows that alterations in secondary structures of viral mRNAs upon temperature change may be a potential factor affecting the cold-adapted, temperature-sensitive phenotype. To demonstrate this fact, we developed a new computational method of determining highly temperature-sensitive regions of RNA structure. Based on this methodology, we implemented a web server described in the third paper. The fourth publication is aimed at assessing the importance of mRNA secondary structures in bacteria and how those structures may be a potential factor of filtering out nucleotide substitutions occurring in *E. coli* during the Long-term evolutionary experiment by Richard Lenski.

Finally, the last chapter briefly presents some conclusions and possible applications of our findings.

Chapter 2

Results

This chapter presents the results of our work which have been published as four papers in peer-reviewed scientific journals.

2.1 Sequence-structure Relationships in Yeast mRNAs

Andrey Chursov, Mathias C. Walter, Thorsten Schmidt, Andrei Mironov, Alexander Shneider and Dmitrij Frishman
Nucleic Acids Res., 40(3):956-962, 2012

Structural bioinformatics of mRNAs is still in its infancy due to the absence of experimentally known (rather than computationally predicted) structures. We provided the first ever analysis of sequence-structure relationships in eukaryotic mRNAs based on the experimental measurements of base pairing propensities published in 2010 by Kertesz and colleagues in Nature (Kertesz et al., 2010). Our main finding is that the relationship between sequence and structure divergence in mRNA molecules is much weaker than in small RNAs, implying a high degree of evolutionary neutrality.

On a more general vein, the objective of our work was to analyze global structural arrangements and their similarity as a function of sequence identity, similar in spirit to the original work of Chothia and Lesk (Chothia and Lesk, 1986). In this work, we focused on the comparison of experimentally determined as well as predicted secondary structures for yeast mRNA sequences that encode paralogous proteins. We considered potential correlations between sequences and mRNA structures. In order to detect such correlations, we used large-scale experimental data on yeast transcriptome RNA structure (Kertesz et al., 2010). We combined these data with theoretical predictions and compared structural and sequence similarities for a number of yeast paralogous genes. The results demonstrated correlations for relatively highly similar sequences (higher than 85-90%), and their absence for sequences with lower similarity.

The result we obtained was not anticipated. To our surprise, we found that only extremely similar sequences are folded into similar structures, while quite similar sequences, sharing as much as 80-85% identity, fold differently. Thanks to the Kertesz et al. dataset, which was the first large-scale measurement of mRNA structures ever published, we could then derive for the first time the quantitative dependence between sequence and structure divergence in mRNAs. Such dependence was not previously known.

The next obvious step was to compare mRNA structures for orthologous genes with a similar distribution of nucleotide sequence identities to the distribution of similarities between paralogs in *S. cerevisiae*. Does the structure diverge faster in paralogs than in orthologs with comparable sequence similarity? For such a comparison, we chose *Candida glabrata*, the closest organism with a completely sequenced genome, and compared predicted structures in *S. cerevisiae* with *C. glabrata*.

In addition, we have made all sequence alignments, together with experimentally determined and predicted structures, in FASTA format available as Supplementary Files.

The research was designed by Dmitrij Frishman and me. I did the programming and performed the research. The resulting data were analyzed by all the authors. The paper was written by myself, Andrei Mironov and Dmitrij Frishman.

956–962 *Nucleic Acids Research, 2012, Vol. 40, No. 3*
doi:10.1093/nar/gkr790

Published online 27 September 2011

Sequence–structure relationships in yeast mRNAs

Andrey Chursov[1], Mathias C. Walter[2], Thorsten Schmidt[2], Andrei Mironov[3,4], Alexander Shneider[5] and Dmitrij Frishman[1,2,*]

[1]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, Maximus-von-Imhof-Forum 3, D-85354, Freising, [2]Helmholtz Center Munich – German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany, [3]Department of Bioengineering and Bioinformatics, Moscow State University, Leninskie Gory, GSP-1, 119991, [4]Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny pereulok 19, 127994, Moscow, Russia and [5]Cure Lab, Inc., 43 Rybury Hillway, Needham, MA 02492, USA

Received July 3, 2011; Revised August 31, 2011; Accepted September 8, 2011

ABSTRACT

It is generally accepted that functionally important RNA structure is more conserved than sequence due to compensatory mutations that may alter the sequence without disrupting the structure. For small RNA molecules sequence–structure relationships are relatively well understood. However, structural bioinformatics of mRNAs is still in its infancy due to a virtual absence of experimental data. This report presents the first quantitative assessment of sequence–structure divergence in the coding regions of mRNA molecules based on recently published transcriptome-wide experimental determination of their base paring patterns. Structural resemblance in paralogous mRNA pairs quickly drops as sequence identity decreases from 100% to 85–90%. Structures of mRNAs sharing sequence identity below roughly 85% are essentially uncorrelated. This outcome is in dramatic contrast to small functional non-coding RNAs where sequence and structure divergence are correlated at very low levels of sequence similarity. The fact that very similar mRNA sequences can have vastly different secondary structures may imply that the particular global shape of base paired elements in coding regions does not play a major role in modulating gene expression and translation efficiency. Apparently, the need to maintain stable three-dimensional structures of encoded proteins places a much higher evolutionary pressure on mRNA sequences than on their RNA structures.

INTRODUCTION

Secondary structure elements both in the untranslated (UTR) and coding (CDS) regions of mRNAs have been implicated in a variety of regulatory functions (1). For example, riboswitches modulate gene expression through conformational changes in response to various stimuli (2). In addition, translation initiation, elongation, termination and translation efficiency all depend on higher order mRNA secondary structures in non-coding regions (3,4). Coding region hairpins have also been suggested to play a role in the regulation of translation (5). The relationship between RNA structure and gene expression in the coding regions of mRNAs has been demonstrated both computationally and experimentally (6–10). In particular, reduced mRNA stability near the start codon has been observed in a wide range of species, probably as a mechanism to facilitate ribosome binding or start codon recognition by initiator tRNA (11). Computational studies show that native mRNA sequences have lower folding energies and hence more stable structure than codon-randomized ones (5). The three mRNA functional domains—5′-UTR, CDS and 3′-UTR—form largely independent folding units, with base pairing across domain borders being rare (12). Evolutionary conserved local secondary structures have been identified in the CDS regions (13,14) and shown to be functional (15).

There is a selective pressure toward maintaining both stable RNA structures of coding regions and the three-dimensional folds of their encoded proteins (16). It has been argued that the redundancy of the genetic code plays an important role in satisfying these selection requirements (12). In general, however, sequence–structure relationships in mRNA-coding regions remain elusive; and, their spatial structure is unknown. While hundreds

---

*To whom correspondence should be addressed. Tel: +49 179 538 2799; Fax: +49 8161 712 186; Email: d.frishman@wzw.tum.de

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

of atomic resolution structures have been determined for smaller RNA molecules, most notably tRNAs, experimental structures of large RNAs are still rare (17). Until recently, direct experimental determination of mRNA structure has been impossible on a large scale. Furthermore, most insights into the evolutionary constraints acting on them arose from correlating predicted base paring patterns with the effects of site-directed mutagenesis on mRNA expression and degradation, as well as on the expression levels and activity of encoded protein products.

Significant progress has been made in predicting RNA secondary structure from sequence based on free-energy minimization (18), probabilistic models (19) and evolutionary information (20). However, the accuracy of current algorithms is still insufficient to model large molecules, primarily because the number of theoretically possible RNA secondary structures grows exponentially with the length of the sequence (21). Also, the free folding energy of millions of suboptimal structures is very close to the most stable structure. Lowest energy structures may not necessarily reflect folding *in vivo* (22) due to kinetic processes and protein–RNA interactions. Additionally, it is hard to model pseudoknots and unstructured regions (23).

More accurate prediction of RNA secondary structure can be achieved by using experimental constraints obtained from oligonucleotide data to guide free-energy minimization (24). Moreover, experimental methods have been developed that allow comprehensive monitoring of RNA structure at single nucleotide resolution. One such method, fragmentation sequencing, allows for reconstructing RNA structures by sequencing fragments of single-stranded RNA resulting from nuclease digestion. Another method, known as selective 2′-hydroxyl acylation and primer extension (SHAPE) (25), exploits the sensitivity of selective acetylation of the ribose 2′-hydroxyl position to local nucleotide flexibility, thereby allowing identification of those nucleotides that are conformationally constrained by base pairing. Accurate SHAPE-directed RNA structure determination has been reported for several types of RNA molecules, including *Escherichia coli* 16S RNA and yeast tRNA^{asp} (26), as well as for the entire HIV-1 genome (27). This latter work highlighted the intricate relationship between RNA sequences and protein structure of the encoded proteins. In particular, it was found that flexible loops in protein structures correspond to highly structured RNA elements, implying a functional role of mRNA structure in the modulation of ribosome processivity at domain boundaries.

In recent work, Kertesz and colleagues (28) reported the first transcriptome-wide experimental analysis of mRNA structures using the novel technology called parallel analysis of RNA structure (PARS). PARS enables the determination of base pairing probabilities at single nucleotide resolution by refolding RNAs *in vivo*, treating them with structure-specific enzymes and then sequencing the resulting fragments. Structural profiles were obtained for more than 3000 transcripts from the budding yeast *Saccharomyces cerevisiae*. The work of Kertesz *et al.* revealed higher degree of structuredness in the mRNA-coding regions compared with the 3′- and 5′-untranslated regions, implying a functional role of RNA structure in coding regions in regulating gene expression. The global data set of PARS profiles represents a true treasure trove for investigating sequence–structure and structure–function relationships in mRNAs.

This report provides the first comprehensive analysis of sequence–structure relationships in the coding regions of yeast mRNAs based on base pairing propensities measured by the PARS technology. It was found that PARS profiles of paralogous mRNAs show very strong, essentially linear, correlation sequence for identity levels upwards of 85–90%. Yet, pairs of more distantly related yeast transcripts secondary structure appear to be unrelated. Interestingly, predicted secondary structures of yeast paralogs display a similar behavior with respect to sequence identity; and, there is a significant correlation between experimental and theoretical structures, as noted previously (28). Theoretical structures of orthologous mRNA pairs from yeast and *Candida glabrata* are also uncorrelated for low sequence identity levels while for highly similar sequences no conclusion could be made due to lack of data.

MATERIALS AND METHODS

Experimental data on yeast mRNA secondary structure

Secondary structure profiles of 3000 transcripts from the budding yeast *S. cerevisiae* have recently been determined using a novel experimental strategy called PARS (28). For each individual nucleotide position of mRNAs, a PARS score reflects its likelihood to be in a double-stranded conformation. PARS scores for yeast transcripts were downloaded from http://genie.weizmann.ac.il/pubs/PARS10. 5′- and 3′-UTR regions were identified by sequence comparison with yeast amino acid sequences, and then excluded from consideration. In the following, a vector of PARS scores for a given transcript is referred to as its experimental structure.

Yeast paralogs

Data on paralogous yeast proteins were kindly provided by Martin Münsterkötter and Ulrich Güldner from the fungal genomics group at the Institute for Bioinformatics and Systems Biology (German Research Center for Environmental Health, Munich). A list of protein pairs sharing significant similarity (identity at the amino acid level >50%) was extracted from the SIMAP database (29). Additionally, the putative paralogs were required to have not >10% difference in sequence length. In total, 243 paralog pairs involving 409 different yeast genes satisfied these conditions.

Amino acid sequences of paralogous yeast proteins were globally aligned using the ggsearch program from the FASTA software suite (30). Amino acid sequence alignments were subsequently converted into mRNA sequence alignments; and, the percent identity between each pair of coding regions was calculated by dividing the number of identical nucleotides by the length of the alignment.

Orthologs from *C. glabrata*

Sequence data for *C. glabrata* were downloaded from the PEDANT genome database (31). A list of orthologous protein pairs between *S. cerevisiae* and *C. glabrata* was extracted from the eggNOG database (32). In total, we obtained 2327 ortholog pairs. The alignment procedure was the same as for paralogs, see above.

PARS score distances between yeast paralogs

To assess global structural similarity between pairs of aligned mRNA sequences, root mean square deviations (RMSDs) between vectors of PARS scores were calculated for all alignment positions that did not contain gaps. Additionally, for each transcript pair, profiles of local structural similarity were obtained by calculating RMSDs between PARS scores in non-gapped alignment positions within a sliding window of varying length, typically between 100 and 1000 nt.

Prediction of mRNA secondary structures

For each nucleotide position of transcript sequences, the theoretical probability to be in double-stranded conformation was calculated using the RNAfold method from the Vienna RNA package (33). As done similarly for experimental PARS scores (see above), RNAfold probability values were used to calculate global and local measures of structural similarity between aligned coding regions of mRNAs based on RMSD. For brevity, a vector of predicted probabilities of RNA bases in double-stranded conformation for a given transcript is further referred to as its theoretical structure.

Data availability

All sequence alignments together with experimentally determined and predicted structures are available in Supplementary Data.

RESULTS

By illustrating the data used in this study on a concrete example, the research results can be readily presented. Two yeast mRNA sequences, YBR092C and YBR093C, share 86.5% sequence identity, and their partial alignment is depicted in the top part of Figure 1. The position-dependent PARS scores for both sequences are shown in the middle part of Figure 1. Both graphs display a rather high degree or correlation, albeit not perfect. In the bottom part of Figure 1, theoretical structures (probabilities for individual bases to be paired) are drawn along the sequence. Figure 2 shows how distances between experimental and theoretical structures of YBR092C and YBR093C vary along the mRNA sequence dependent on sequence identity in a local sequence window. As

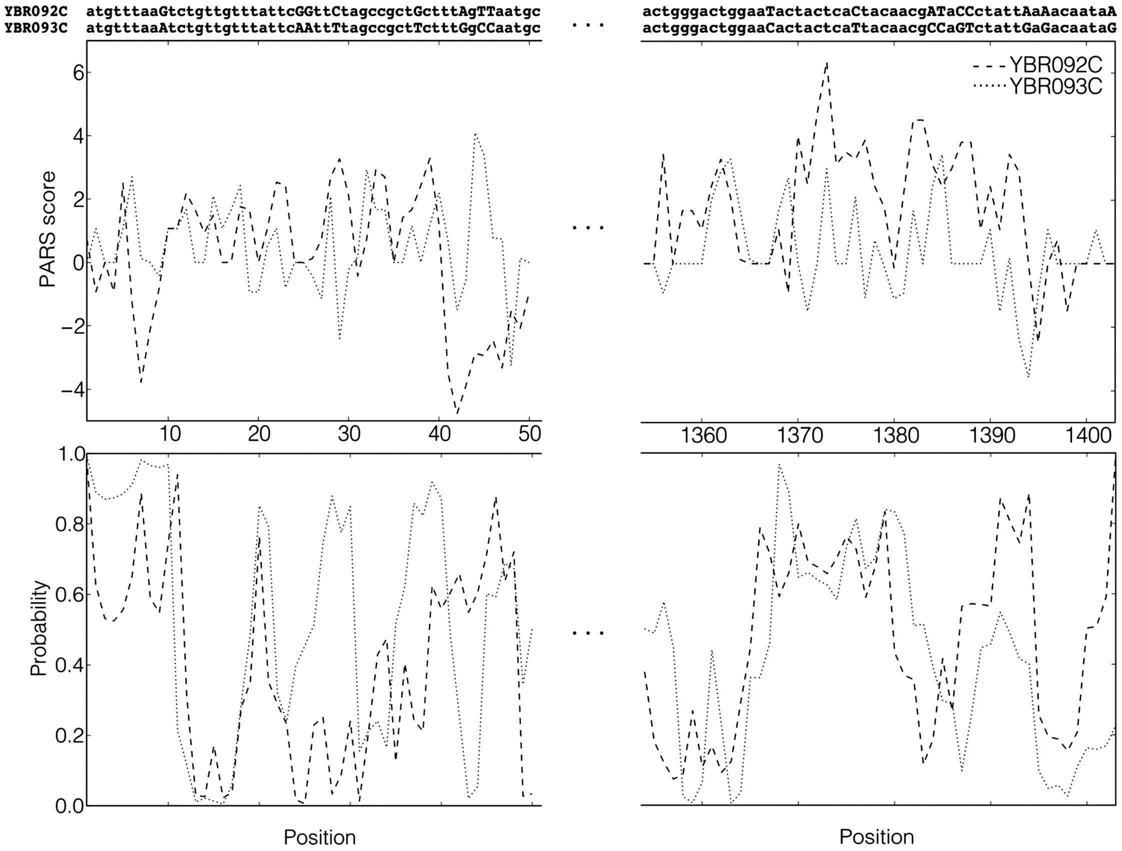


Figure 1. Sequence alignment, experimental and theoretical structures of the first and last 50 nt for the pair of yeast mRNA sequences YBR092C (dashed lines) and YBR093C (dotted lines).

expected, highly similar regions generally correspond to more similar structures.

Calculations exemplified in Figures 1 and 2 were performed for all pairs of paralogous mRNA sequences in our data set. Table 1 summarizes pair-wise correlations between the three evolutionary measures considered in this work for different ranges of sequence identities. Figure 3a shows how the difference between experimental structures depends on sequence similarity. PARS scores appear to be entirely uncorrelated for identity levels of up to ~85–90%. In this sequence identity range, the median RMSD between PARS score vectors does not differ from the median calculated for randomly selected mRNA pairs (dashed horizontal line in Figure 3a). For sequence identity levels over 85–90%, the distance between experimental structures shows essentially a linear dependence from sequence similarity (Supplementary Figure S1).

Upon conducting the same experiment with pairs of theoretical structures of yeast mRNAs, it was found that the distance between the structures also begins to depend on sequence similarity upward of roughly 85–90% identity (Figure 3b). For pairs with identity between sequences within the range from 97.5% to 100%, the median distance between theoretical structures constitutes 38% of the random level. Yet, for experimental structures, it is lower at 29%. The link between sequence and structure is thus stronger when experimental structures are considered. The distance between theoretical structures also shows a linear dependence from sequence similarity for sequence identity levels over 85–90% (Supplementary Figure S2).

Therefore, what is the significance of the sequence–structure dependence shown in Figure 3; and, how would it appear for codon-randomized mRNA sequences? Since experimental PARS scores are not available for randomly generated sequences, this issue could only be assessed for theoretical structures. For each pair of paralogs, one sequence was kept unchanged. In the second mRNA, however, mutations were randomly distributed along the sequence, keeping the encoded amino acid sequence, the codon usage and the total number of mutations between the paralogs unchanged. Overall, the divergence of structures between codon-randomized paralogs displays virtually the same dependence on sequence similarity as for native sequences (Supplementary Figure S3).

We also compared predicted structures between orthologous mRNAs from *S. cerevisiae* and the pathogenic yeast *C. glabrata* (Figure 4). Although *C. glabrata* is the most closely related organism to *S. cerevisiae* with a completely sequenced genome (34), no pair of orthologous mRNAs between these two organisms shares sequence identity >95% and thus no conclusion about structure divergence for very similar sequences could be made. However, for lower identity levels theoretical structures of orthologs are uncorrelated and thus behave the same way as paralogous structures.

DISCUSSION

In some sense, the current situation in RNA bioinformatics is reminiscent of the early days of structural

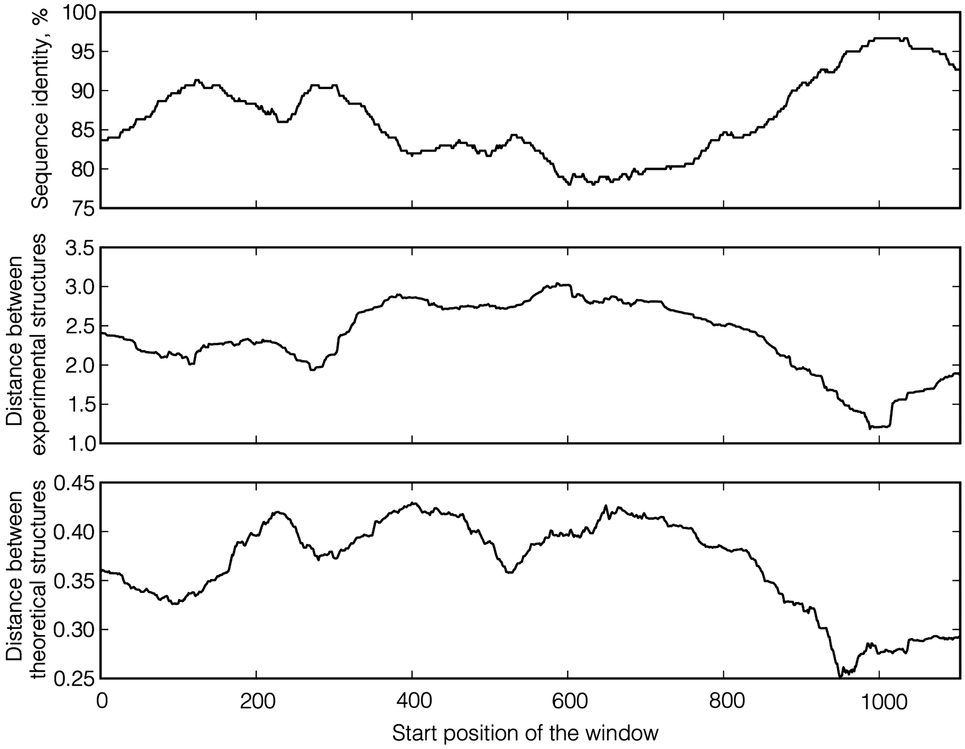


Figure 2. The profile of local structural similarity versus local sequence identity for the pair of yeast mRNA sequences YBR092C and YBR093C. The length of the sliding window is 300. The global sequence identity between these two sequences is 86.5%.

Table 1. Correlation coefficients and *P*-values for different ranges of sequence identity

Sequence identity range (%)	Sequence identity versus RMSD between experimental structures		Sequence identity versus RMSD between theoretical structures		RMSD between experimental structures versus RMSD between theoretical structures	
	Correlation coefficient	*P*-value	Correlation coefficient	*P*-value	Correlation coefficient	*P*-value
50–60	0.12	0.39	−0.07	0.62	0.14	0.31
60–70	0.14	0.22	−0.10	0.37	−0.02	0.87
70–80	−0.08	0.67	−0.08	0.67	−0.24	0.21
80–90	0.01	0.91	−0.14	0.40	0.04	0.79
90–100	−0.92	$5.66e^{-27}$	−0.75	$1.24e^{-12}$	0.69	$3.56e^{-10}$

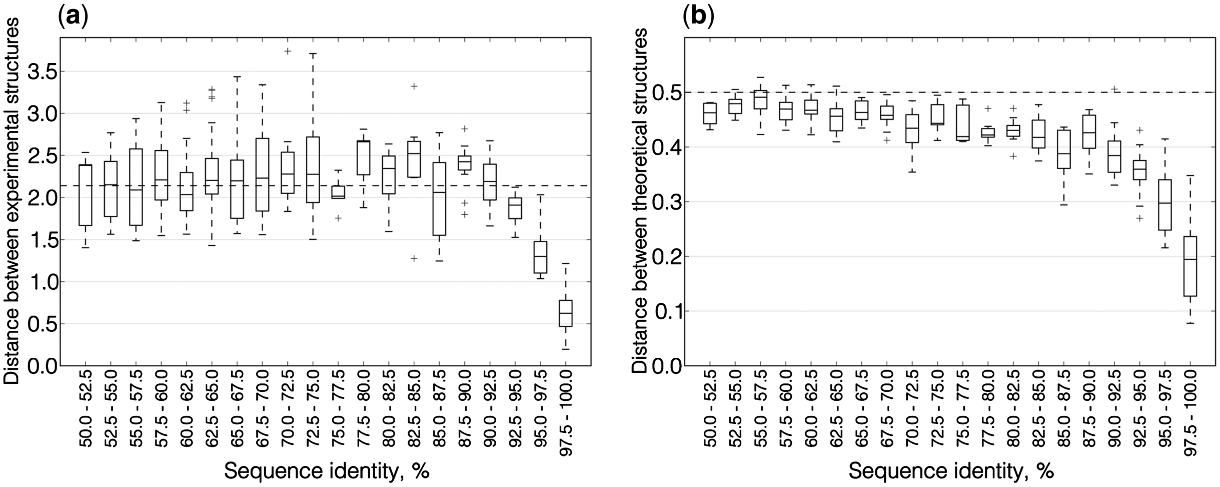


Figure 3. Boxplots of distances between structures of aligned paralogous mRNAs in different ranges of sequence similarity. Each box corresponds to the range of similarity 2.5%. The box extends from the lower to the upper quartile values, with a horizontal line at the median value. Whiskers demonstrate the entire range of the data. Crosses show outliers. (**a**) Distances between experimental structures. The average level of PARS score distances for alignments of random sequence pairs is 2.14 (dashed line). (**b**) Distances between theoretical structures. The average level of probability distance for alignments of random sequence pairs is 0.5 (dashed line).

bioinformatics of proteins, when the availability of a sufficiently large data set of X-ray structures allowed for the first comprehensive analysis of the relation between the divergence of sequence and structure in proteins (35). Until recently, studies of the evolutionary conservation of RNA structures were based on *in silico* predictions and largely limited to non-coding RNA. In the first large-scale study, Schudoma *et al.* (36) determined that in short RNA loops with known three-dimensional structures sequence identity >75% implies significant structural similarity. The most comprehensive investigation of sequence–structure relationships in RNA molecules to date is based on all-against-all pair-wise structural comparison of non-coding RNAs (tRNAs, rRNAs, riboswitches and riboswitches) with known spatial architectures (37). Assessment of evolutionary divergence revealed that the correlation between sequence and secondary structure conservation is highly significant for sequence identity levels in the range between just a few percentage points up to roughly 60% where this relationship saturates. Further increase of sequence similarity (60–100%) does not lead to an appreciable growth of secondary structure similarity. None of the studies mentioned above considered mRNAs because no mRNA structures are currently known at atomic resolution.

The principal finding of this research is that the correlation between sequence and structure in the coding regions of yeast mRNAs is much weaker than in small non-coding RNAs. Up to ∼85–90% sequence identity, the similarity of both experimental and theoretical base pairing propensities between paralogous yeast mRNAs is at random level; while, for more similar sequence pairs, sequence and structure are strongly correlated. This may imply that mRNAs do not experience a strong selective pressure to preserve a certain degree of structuredness. The fact that codon-randomized sequences display a similar behavior also indicates that there is no appreciable evolutionary pressure to preserve a particular RNA structure as long as the encoded protein remains unchanged. Taken together, these results underscore a high degree of evolutionary neutrality in yeast mRNA molecules, both at the level of primary (third codon position) and secondary (extent of base paring) structure.

On one hand, our findings are in strong contrast to many non-coding RNAs and *cis*-acting regulatory elements of mRNAs whose biological function is primarily

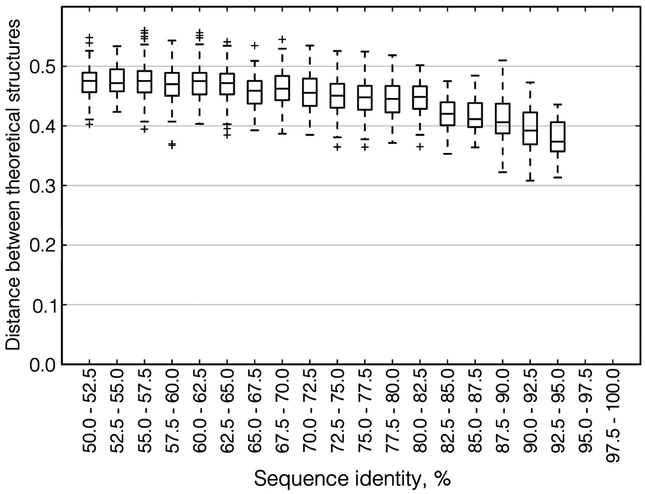


Figure 4. Boxplot of distances between theoretical structures of aligned orthologous mRNAs in different ranges of sequence similarity. Notation as in Figure 3.

mediated by their spatial architecture (38) stabilized by tertiary interactions, modified bases and interactions with proteins and small ligands. On the other hand, sequence–structure relationships observed in this work are compatible with the notion that, in general, RNA molecules do not have a single global structure. Instead, they exist as a highly dynamic ensemble of alternative conformations (39,40) that are often capable of performing different functions (41). The extent of base pairing may play a role in the regulation of pre-mRNA splicing, translation and mRNA degradation. Both experimentally determined PARS scores and computationally derived partition functions analyzed in this work are statistical measures that reflect the propensity of each nucleotide to form a base pair across a large number of metastable structures.

This analysis has several important limitations. First, PARS probes RNA structures *in vitro* rather than in the living cell and may not always reproduce functional RNA structures (42). Second, even if the base paring information obtained by the PARS technology were perfectly correct, it still merely represents a one-dimensional profile of structural propensities, a far cry from knowing the actual RNA secondary structure, let alone spatial architecture, for each individual molecule at any moment of time. Third, the findings do not rule out much stronger sequence–structure correlations in certain local structural elements of coding regions, such as reprogrammed genetic-decoding signals (43) or mRNA localization signals. We also cannot rule out the possibility that the degree of mRNA structuredness does have an important functional role in spite of quick erosion of structural similarity between paralogs with diminishing sequence similarity, and that this erosion reflects functional differentiation. However, we consider such explanation unlikely because the same behavior is observed between orthologous mRNAs. Finally, only a small subset of the PARS data constituted by pairs of sequence similar yeast mRNAs (paralogs) was explored. As a next step, it will be exciting to conduct comparative analyses of mRNA structuromes [the term coined by Westhof and Romby (44)], focusing on orthologous sequences from multiple organisms and taking into account important genomic variables, such as expression level and evolutionary rate. Given the current pace of high-throughput RNA analysis technologies there is no doubt that such data will become available in the near future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary figures S1–S3.

ACKNOWLEDGEMENTS

We would like to thank Dmitry Ivankov and Natalya Bogatyreva for helpful discussions and Janusz Bujnicki for illuminating comments on the article.

FUNDING

The DFG International Research Training Group ‘Regulation and Evolution of Cellular Systems’ (GRK 1563) and by the Russian Foundation for Basic Research (RFBR 09-04-92742). Funding for open access charge: German Research Center for Environmental Health, Munich.

Conflict of interest statement. None declared.

REFERENCES

1. Bevilacqua,P.C. and Blose,J.M. (2008) Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.*, **59**, 79–103.
2. Serganov,A. and Patel,D.J. (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.*, **8**, 776–790.
3. Gray,N.K. and Hentze,M.W. (1994) Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.*, **19**, 195–200.
4. Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.
5. Katz,L. and Burge,C.B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.*, **13**, 2042–2051.
6. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science*, **324**, 255–258.
7. Duan,J., Wainwright,M.S., Comeron,J.M., Saitou,N., Sanders,A.R., Gelernter,J. and Gejman,P.V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.*, **12**, 205–216.
8. Ilyinskii,P.O., Schmidt,T., Lukashev,D., Meriin,A.B., Thoidis,G., Frishman,D. and Shneider,A.M. (2009) Importance of mRNA secondary structural elements for the expression of influenza virus genes. *OMICS*, **13**, 421–430.
9. Carlini,D.B., Chen,Y. and Stephan,W. (2001) The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics*, **159**, 623–633.
10. Nackley,A.G., Shabalina,S.A., Tchivileva,I.E., Satterfield,K., Korchynskyi,O., Makarov,S.S., Maixner,W. and Diatchenko,L. (2006) Human catechol-O-methyltransferase haplotypes modulate

protein expression by altering mRNA secondary structure. *Science*, **314**, 1930–1933.

11. Gu,W., Zhou,T. and Wilke,C.O. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **6**, e1000664.
12. Shabalina,S.A., Ogurtsov,A.Y. and Spiridonov,N.A. (2006) A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.*, **34**, 2428–2437.
13. Meyer,I.M. and Miklós,I. (2005) Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.*, **33**, 6338–6348.
14. Steigele,S., Huber,W., Stocsits,C., Stadler,P.F. and Nieselt,K. (2007) Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol.*, **5**, 25.
15. Olivier,C., Poirier,G., Gendron,P., Boisgontier,A., Major,F. and Chartrand,P. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol.*, **25**, 4752–4766.
16. White,H.B. III, Laux,B.E. and Dennis,D. (1972) Messenger RNA structure: compatibility of hairpin loops with protein sequence. *Science*, **175**, 1264–1266.
17. Holbrook,S.R. (2008) Structural principles from large RNAs. *Annu. Rev. Biophys.*, **37**, 445–464.
18. Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.
19. Dowell,R.D. and Eddy,S.R. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 400.
20. Bernhart,S.H. and Hofacker,I.L. (2009) From consensus structure prediction to RNA gene finding. *Brief. Funct. Genomic Proteomic*, **8**, 461–471.
21. Meyer,I.M. and Miklós,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.
22. Mahen,E.M., Watson,P.Y., Cottrell,J.W. and Fedor,M.J. (2010) mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLoS Biol.*, **8**, e1000307.
23. Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.
24. Duan,S., Mathews,D.H. and Turner,D.H. (2006) Interpreting oligonucleotide microarray data to determine RNA secondary structure: application to the 3′ end of Bombyx mori R2 RNA. *Biochemistry*, **45**, 9819–9832.
25. Merino,E.J., Wilkinson,K.A., Coughlan,J.L. and Weeks,K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, **127**, 4223–4231.
26. Low,J.T. and Weeks,K.M. (2010) SHAPE-directed RNA secondary structure prediction. *Methods*, **52**, 150–158.
27. Watts,J.M., Dang,K.K., Gorelick,R.J., Leonard,C.W., Bess,J.W. Jr, Swanstrom,R., Burch,C.L. and Weeks,K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.
28. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
29. Rattei,T., Tischler,P., Götz,S., Jehl,M.A., Hoser,J., Arnold,R., Conesa,A. and Mewes,H.W. (2010) SIMAP–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res.*, **38**, D223–D226.
30. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.
31. Walter,M.C., Rattei,T., Arnold,R., Güldener,U., Münsterkötter,M., Nenova,K., Kastenmüller,G., Tischler,P., Wölling,A., Volz,A. *et al.* (2009) PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res.*, **37**, D408–D411.
32. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.
33. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neuböck,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.
34. Dujon,B. (2010) Yeast evolutionary genomics. *Nat. Rev. Genet.*, **11**, 512–524.
35. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
36. Schudoma,C., May,P., Nikiforova,V. and Walther,D. (2010) Sequence-structure relationships in RNA loops: establishing the basis for loop homology modeling. *Nucleic Acids Res.*, **38**, 970–980.
37. Capriotti,E. and Marti-Renom,M.A. (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, **11**, 322.
38. Gruber,A.R., Bernhart,S.H., Hofacker,I.L. and Washietl,S. (2008) Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.
39. Mironov,A.A., Dyakonova,L.P. and Kister,A.E. (1985) A kinetic approach to the prediction of RNA secondary structures. *J. Biomol. Struct. Dyn.*, **2**, 953–962.
40. Danilova,L.V., Pervouchine,D.D., Favorov,A.V. and Mironov,A.A. (2006) RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol.*, **4**, 589–596.
41. Zhao,P., Zhang,W.B. and Chen,S.J. (2010) Predicting secondary structural folding kinetics for nucleic acids. *Biophys. J.*, **98**, 1617–1625.
42. Mauger,D.M. and Weeks,K.M. (2010) Toward global RNA structure analysis. *Nat. Biotechnol.*, **28**, 1178–1179.
43. Namy,O., Rousset,J.P., Napthine,S. and Brierley,I. (2004) Reprogrammed genetic decoding in cellular gene expression. *Mol. Cell*, **13**, 157–168.
44. Westhof,E. and Romby,P. (2010) The RNA structurome: high-throughput probing. *Nat. Methods*, **7**, 965–967.

2.2 Specific Temperature-induced Perturbations of Secondary mRNA Structures are Associated with the Cold-adapted Temperature-sensitive Phenotype of Influenza A Virus

Andrey Chursov, Sebastian J. Kopetzky, Ignaty Leshchiner, Ivan Kondofersky, Fabian J. Theis, Dmitrij Frishman and Alexander Shneider

RNA biology, 9(10):1266-1274, 2012

Influenza A viruses constitute a serious threat to human health and vaccines against these viruses are of crucial importance. There are two types of vaccines against influenza A viruses: inactivated viruses, which are widespread in Europe; and live attenuated vaccines, which were developed in the USSR and are used in the USA. Live attenuated vaccines are cold-adapted, temperature-sensitive (ca/ts) mutant influenza A viruses. Mutations in the three proteins constituting the viral polymerase complex are thought to contribute to the ca/ts phenotype. However, several virus studies have elegantly demonstrated that the thermodynamic stability of certain RNA structures is critical for optimal virus replication (Berkhout et al., 1997; Mirmomeni et al., 1997; Rowe et al., 2000). Hence, this article presents an interesting new idea: cold-adapted, temperature-sensitive (ca/ts) phenotypes of influenza may depend on the temperature dependence of the RNA secondary structure. To test this hypothesis, we developed a new approach to predicting RNA sequences that may be responsible for this effect.

In this work we addressed the question of the contribution of the viral mRNA structures to the cold-adapted, temperature-sensitive phenotype. We developed a new bioinformatics tool to analyze the variations of the RNA structures when the temperature is shifted from 32 to 39°C, and to compare these variations in the ca/ts viral strains and in the parental viruses from which they are derived. We used RNAfold to predict the probability of each nucleotide to be involved in base pairing (via partition functions), rather than predicting secondary structures. We found that the nucleotides, whose probability to be base-paired was most affected by temperature, tend to cluster (organize into regions with several such changes). Moreover, these clusters were different for wild type and ca/ts mutant viruses. By comparing these viruses with a pool of artificial viral sequences in which synonymous mutations were randomly introduced (wt sequences with same number of mutations as in ca/ts sequences), we concluded that the existence of nine clusters in the ca/ts mutant viruses, but not in the wild type viruses, statistically correlated with the ca/ts phenotype. Additionally, the analysis revealed the existence of one cluster, present in the wt strain but absent in the ca/ts mutant, which could be attributed to introducing specific mutations causing the ca/ts phenotype.

It is worth noting that the length of the overlap between clusters (when we compare the location of clusters from two different strains) affects the strength of the statistical conclusion concerning appearance/disappearance of clusters resulting from randomly introduced mutations. The longer the overlap allowed, the easier it is to conclude that two clusters are different. Thus, we have restricted ourselves to short overlaps only. Any conclusion that two clusters are different, which was made based on short overlaps, would remain to be true if one allows longer overlaps. By contrast, a conclusion made on longer overlaps may not sustain a test with shorter overlaps. Thus, our results present the lower limit.

We also tested if the statistically significant clusters we have observed co-locate with those few RNA structures, which are already known for the influenza A virus. None such overlaps were observed. However, a comprehensive analysis of influenza RNA structures was never conducted. Several structures were discovered because of their biological importance and/or viral impairment in case these structures are disturbed. Still, relatively little is known about influenza virus RNA folding and its influence on influenza virus replication (Gultyaev et al., 2010). Thus, our results provide a rationale for testing experimentally whether RNA structures are indeed present at the locations of the clusters of highly temperature-sensitive positions at 32 and 39°C.

The research was designed by Alexander Shneider and me. The programming and all the computations were performed by me and Sebastian J. Kopetzky. The resulting data were analyzed by all the authors. The paper was written by myself, Sebastian J. Kopetzky, Dmitrij Frishman and Alexander Shneider.

RNA Biology 9:10, 1266–1274; October 2012; © 2012 Landes Bioscience

Specific temperature-induced perturbations of secondary mRNA structures are associated with the cold-adapted temperature-sensitive phenotype of influenza A virus

Andrey Chursov,[1,†] Sebastian J. Kopetzky,[1,†] Ignaty Leshchiner,[2,3] Ivan Kondofersky,[4] Fabian J. Theis,[4] Dmitrij Frishman[1,4,‡,*] and Alexander Shneider[3,‡,*]

[1]Department of Genome Oriented Bioinformatics; Technische Universität München; Wissenschaftzentrum Weihenstephan; Freising, Germany; [2]Genetics Division; Brigham and Women's Hospital; Harvard Medical School; Boston, MA USA; [3]Cure Lab, Inc.; Needham, MA USA; [4]Helmholtz Center Munich - German Research Center for Environmental Health (GmbH); Institute of Bioinformatics and Systems Biology; Neuherberg, Germany

[†]These authors contributed equally to this work and should be regarded as joint first authors. [‡]These authors contributed equally to this work and should be regarded as joint last authors.

Keywords: influenza, RNA, structure, temperature, vaccine

For decades, cold-adapted, temperature-sensitive (ca/ts) strains of influenza A virus have been used as live attenuated vaccines. Due to their great public health importance it is crucial to understand the molecular mechanism(s) of cold adaptation and temperature sensitivity that are currently unknown. For instance, secondary RNA structures play important roles in influenza biology. Thus, we hypothesized that a relatively minor change in temperature (32–39°C) can lead to perturbations in influenza RNA structures and, that these structural perturbations may be different for mRNAs of the wild type (wt) and ca/ts strains. To test this hypothesis, we developed a novel in silico method that enables assessing whether two related RNA molecules would undergo (dis)similar structural perturbations upon temperature change. The proposed method allows identifying those areas within an RNA chain where dissimilarities of RNA secondary structures at two different temperatures are particularly pronounced, without knowing particular RNA shapes at either temperature. We identified such areas in the NS2, PA, PB2 and NP mRNAs. However, these areas are not identical for the wt and ca/ts mutants. Differences in temperature-induced structural changes of wt and ca/ts mRNA structures may constitute a yet unappreciated molecular mechanism of the cold adaptation/temperature sensitivity phenomena.

Introduction

Influenza vaccines have been a great public health priority[1] and their future is man-made constructs created using molecular biology tools. Compared with other types of influenza vaccines, live attenuated influenza vaccines (LAIV) possess major advantages because of administration convenience and potency of the immune response.[2] There are alternative approaches which can lead to viral attenuation and be utilized for LAIV design.[3]

Since the late 1960s, cold-adapted temperature-sensitive (ca/ts) LAIVs have become an important vaccination instrument in the USSR. The ca/ts phenotype leads to impaired growth at an elevated temperature of approximately 39°C[4-9] while permitting viral growth at lower temperatures. Molecular mechanism(s) causing the ca/ts phenotype in influenza A viruses remain unclear. Significant effort was devoted to explaining temperature sensitivity through mutations in the coding regions and amino acid changes. Jin et al. found that certain non-silent mutations in PB1, PB2 and NP might lead to temperature-sensitivity when induced in A/Ann Arbor/6/60.[6] According to Song et al., three non-silent mutations in PB1 and one non-silent mutation in PB2 might lead to the ts phenotype.[4] Youil et al. investigated several A/Leningrad/134/17/57 subclones and found that the most temperature-sensitive one had amino acid changes in the PB1, PA and NS1 genes.[10] Furthermore, Snyder et al. found that it can be sufficient to induce the temperature-sensitive phenotype by replacing the two segments of coding for PA and M1/M2 of a wild type virus with those of A/Ann Arbor/6/60.[11] Interestingly, in all these cases at least one subunit of the viral polymerase (PA, PB1 and PB2) is affected.

In addition to the attempts to explain the ca/ts phenotype through mutations in viral proteins, there were also reports implicating RNAs in temperature sensitivity. A promising finding was made by Dalton et al.,[12] suggesting that, at an elevated

*Correspondence to: Dmitrij Frishman and Alexander Shneider; Email: d.frishman@wzw.tum.de and ashneider@curelab.com
Submitted: 07/17/12; Revised: 08/29/12; Accepted: 09/04/12
http://dx.doi.org/10.4161/rna.22081

temperature, viral polymerase tends to dissociate from the cRNA-promoter, thereby leading to a decreased vRNA synthesis while the synthesis of cRNA and mRNA remains approximately constant. A decrease in the synthesis of vRNA related to temperature sensitivity, which also maintained mRNA synthesis, was described by Chan et al.[9] In a more general vein, RNAs can serve as intracellular thermometers.[13] For example, a thermosensitive RNA switch was implicated in the propagation of tick-borne encephalitis virus.[14] Recent publications suggest that, apart from RNA abundance, RNA structures may play a comparably important role. The importance of mRNA secondary structures for expression of influenza virus genes was recently demonstrated by Ilyinskii et al.[15] Therefore, identification of previously unknown influenza RNA structures[16] and the analysis of their functional roles are areas of increasing interest.[17-19]

We hypothesized that changing temperature causes perturbations in mRNA secondary structures, which contributes to the cold-adapted, temperature-sensitive phenotype. To test this hypothesis, we have developed a new in silico method of analysis to reveal if the structures of two closely related RNA molecules would react differently to temperature elevation. Unfortunately, it is not possible to reliably calculate exact structures of each RNA molecule at two temperatures, compare the differences between the two structures, and then evaluate whether or not these differences are identical for two RNAs. First of all, at each particular temperature an RNA molecule may have different co-existing structures. Furthermore, since the number of possible structures increases rapidly with the length of the input sequence, the precision of RNA structure predictions suffers. Another limitation of RNA secondary structure predictions is that taking pseudoknots into account makes the task non-deterministic polynomial-time hard (NP-hard).[20] In this particular case NP-hard means that growth of RNA length elevates time necessary for computation to a restrictive duration. However, in support of our hypothesis, one does not need to know the exact structures before and after perturbation to conclude that the two structures have reacted differently. For example, if two windows are broken into a different number of pieces by soccer balls, we need to know neither the shapes of the windows and nor the exact forms of the pieces to conclude that the perturbations of the two glasses are not identical.

An ensemble of RNA structures can be represented via a partition function,[21,22] which is a sum of Boltzmann factors over every possible secondary structure. In using partition functions, one can calculate the probability for each nucleotide to be coupled within a double-stranded conformation.[23,24] An advantage of partition functions is that they take into account not just the minimum free energy structure, but rather an ensemble of energetically favorable structures. Thus, if one adenine would be bound to a particular uracil within a single highly likely structure, while another adenine would couple with ten uracils within ten less likely structures, parameters for these two adenines may be the same. Although partition functions are not precisely accurate, they are much more accurate than in silico predictions of the actual RNA structures. Partition functions were used instead of actual structures, for example, by Witwer et al.[25] and Thurner et al.[26] to investigate secondary structure conservation in Picornaviridae and Flaviviridae, respectively, and by Chursov et al.[27] for elucidating sequence-structure relationships in yeast mRNAs. However, so far, partition functions have not been used to assess and compare structural RNA perturbations caused by temperature elevation.

Based on partition functions, we have developed a technique to identify RNA sequence regions where probabilities of nucleotide coupling change the most with temperature elevation. We demonstrate that dense areas of altered nucleotide coupling are not identical for closely related wt and ca/ts RNAs. Thus, although, we cannot predict the exact RNA structures, we know that these structures are changing differently with temperature elevation.

Results

The propensity of nucleotides to appear in double-stranded conformations depends on temperature. As seen in **Figure 3**, all nucleotides change their base-pairing probabilities upon temperature elevation from 32°C to 39°C, with transitions from a double-stranded to a single-stranded conformation being expectedly more frequent (see **Table 2**). Between 62.8% and 75.2% of positions in each mRNA change their probability to be coupled to a lower value. Furthermore, between 3.9% and 10.9% of nucleotides in each mRNA change their base-pairing probabilities significantly (more than three standard deviations below or above the mean over all seven temperature increments between 33–39°C and 32°C (see the Materials and Methods section and **Table 3**). In all but one mRNAs, the majority of significantly changing positions (between 52% and 88.6%) shows a decrease in their base-pairing probability, whereas this percentage is somewhat lower (42.1%) for NS2 Arb/ca.

For each mRNA, we computed a density plot of significantly changing positions along the sequence as described in the Materials and Methods section (**Fig. 2**; **Figs. S1–19**). From these plots, it becomes immediately apparent that strongly temperature-sensitive positions are not evenly or randomly distributed along the sequence but rather aggregate in clusters. The numbers of clusters defined by the density-based algorithm for each mRNA are presented in **Table 4**. The only mRNA where no clusters were detected is NS2 Len/wt. The average length of clusters varies between 15.9 and 61.0 positions (**Table 5**) and the average density of significantly changing positions in the clusters is in the range of 22% to 53% (**Table 6**). Overall, very short clusters are required by the DBSCAN algorithm to have a very high density while the density in longer clusters can be as low as 21% (**Fig. 4**).

Furthermore, we found that patterns of cluster occurrence exhibit substantial differences between the wild type strains and their cold-adapted, temperature-sensitive mutants, as exemplified in **Figure 1** for a subsequence of the PA mRNA. In this case, a cluster of significantly changing positions is observed in Len/17/ca but not in Len/wt. This figure demonstrates that a perturbation of mRNA structure begins at a temperature of approximately 37°C. Out of 218 clusters of temperature-sensitive positions, 126 clusters are present in both wt and ca/ts strains, 38 clusters are

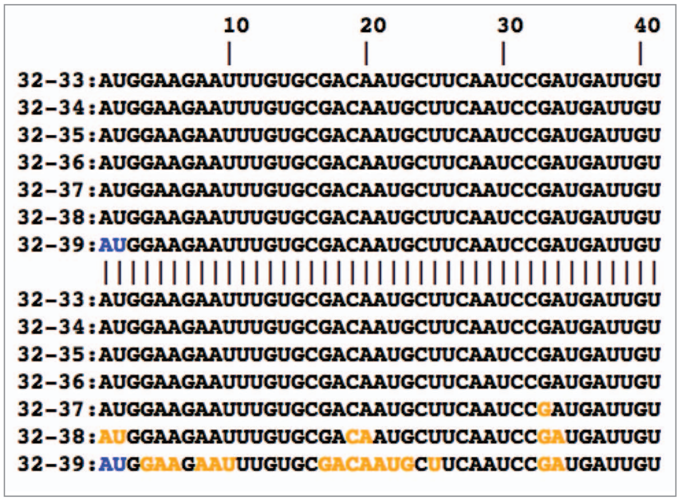

10
20
30
40
32-33:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-34:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-35:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-36:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-37:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-38:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-39:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-33:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-34:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-35:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-36:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-37:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-38:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU
32-39:AUGGAAGAAUUUGUGCGACAAUGCUUCAAUCCGAUGAUUGU

Figure 1. Comparison of significantly changing positions between the PA mRNA of Len/wt (upper 7 rows) and Len/17/ca (lower 7 rows). Each row corresponds to a difference vector v_{32-33}, …, v_{32-39} containing changes of base pairing probabilities between 32°C and a particular higher temperature. Positions in which base paring probabilities significantly change with temperature elevation in both sequences and those where these changes only affect one of the phenotypes are marked blue and orange, respectively. Only the first 40 bases of each sequence are shown; position numbers of the coding sequence are indicated at the top of the alignment.

present in wt strains but absent in ca/ts mutants, and 54 clusters are present in ca/ts mutants but absent in the wt counterpart (**Fig. 4 and Supplemental Data**).

The existence of clusters unique for ca/ts strains raises the question whether such clusters are associated with the mutations inducing the ca/ts phenotype or whether random mutations unrelated to the ca/ts phenotype would be as likely to induce these clusters. Likewise, one can ask whether the disappearance of some clusters present in wt strains from ca/ts mutants may be caused by particular ca/ts associated mutations. The best way to approach this problem would be to test whether or not the same pattern of cluster occurrence would be observed while comparing the wt strains investigated here with a high number of naturally occurring influenza virus strains as similar to the wt strains as their ca/ts mutants. However, there are currently not enough naturally occurring strains with the same extent of similarity to the wt as possessed by the ca/ts mutants.

We therefore compared wt sequences with computer-generated mutants possessing random synonymous mutations unrelated to the phenotype of interest. This analysis revealed existence of only one cluster that is present in wt strain (Len/wt) but absent in ca/ts mutant and could be attributed to introducing specific mutations causing the ca/ts phenotype (**Table 7**). The length of this cluster is 140 nucleotides and the density of significantly changing positions in it equals 38%. At the same time, there are nine clusters (one in Arb/ca, three in Len/17/ca, and five in Len/47/ca) present in ca/ts mutants, and not present in wt, that cannot be observed in the pool of in silico generated random mutants with statistically significant P-values. The length of these ca/ts associated clusters is in the range of 8 to 19 positions and the density of significantly changing positions in them varies between 32% and 80% (**Table 7**). All the clusters that can be associated with ca/ts phenotype are indicated in **Figure 4D**. The existence of such clusters suggests that the ca/ts phenotype may be associated with specific perturbations in mRNA secondary structures. Importantly, in all three ca/ts mutants, there are clusters located in the polymerase genes (PA or PB2), in line with previous reports where polymerase genes were consistently associated with the temperature-sensitive phenotype.[4,6,10,11]

Discussion

Temperature-sensitive mutants were reported for a variety of viruses.[39-42] Several studies have demonstrated that thermodynamic stability of certain RNA structures is critical for virus replication.[43-45] Temperature-sensitive, anti-viral and anti-bacterial vaccines remain to be promising public health instruments.[46-48] So far, cold-adapted temperature-sensitive anti-influenza vaccines have arguably made the largest contribution to the prevention of this infection around the world. Still, molecular mechanism(s) underlining the ca/ts influenza phenotype is poorly understood. Here, we have explored the hypothesis that ca/ts properties of known influenza strains can be (at least partially) explained by temperature-induced perturbations of mRNA structure.

It was, therefore, our intention to compare mRNAs at each of the temperatures of interest. However, despite the fact that significant attempts have been made toward theoretical predictions of RNA structure based on energy calculations[24,34,49] and co-variation analysis,[50,51] it is still not possible to calculate secondary structures of mRNAs accurately using currently available algorithms. At the same time, experimental technologies to determine RNA structures are only beginning to emerge[52,53] and are barely available for a broad spectrum of research projects. Thus, we had to develop an indirect computational method aimed to assess if two RNA molecules change their shapes differently in response to temperature elevation.

At each temperature, we calculate probability vectors that contain, for each nucleotide position, the probability to be coupled with another nucleotide within the same RNA, forming a double-helix structure. Apparently, this coupling is temperature-sensitive, with increasing temperature generally leading to a reduced likelihood of "weak" structures. Thus, (1) different structures may constitute an ensemble for the same RNA at different temperatures, and/or (2) at different temperatures the same structures may be present with different abundance. Both of these options are valid and may coexist because, in each given cell, multiple copies of the same RNA molecules may be distributed between alternative shapes.

The fact that the base-paring probability at each position within a probability vector changes with temperature elevation does not necessarily indicate that structural perturbations (or redistribution of alternative RNA structures) equally involve each nucleotide. Thus, we selected only those nucleotide positions that exhibited the most significant changes of their coupling

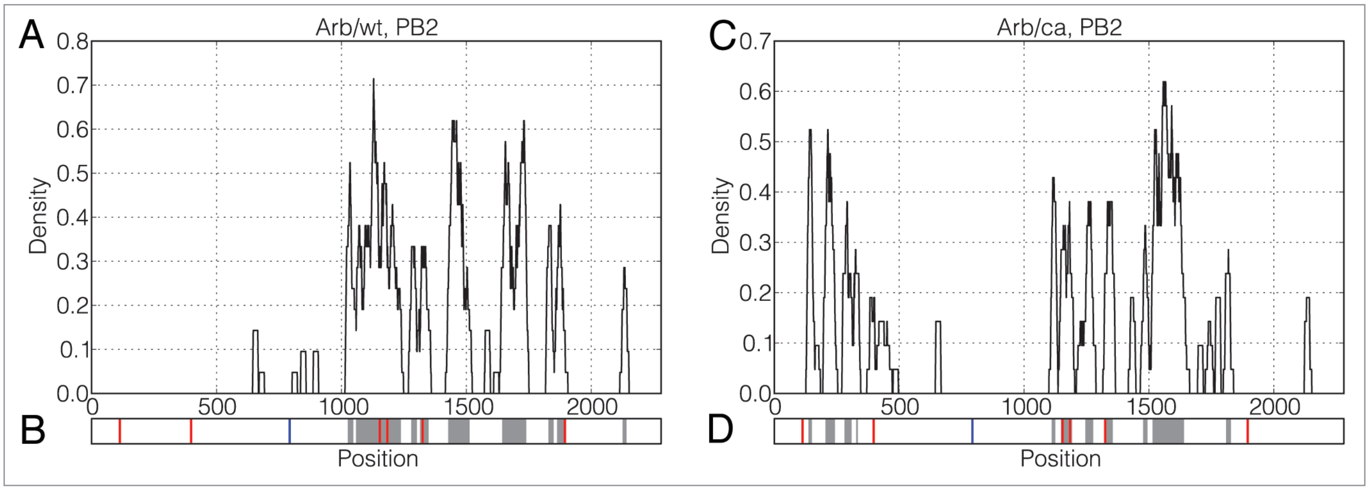


Figure 2. Distributions of significantly changing positions along the PB2 mRNAs of Arb/wt and Arb/ca. A sliding window of size 20 was moved in steps of 1 position over the vector v_{32-39} and the percentage of significantly changing positions in the window was calculated for each possible starting position. The resulting density plots are depicted in Figure **2A** and Figure **2C**. Location of clusters of significantly changing positions identified by the DBSCAN algorithm are depicted in Figure **2B** and Figure **2D** with gray color. Synonymous and non-synonymous mutations are depicted in Figure **2B** and Figure **2D** with red and blue vertical lines, respectively.

probabilities. We do not assert that if in two closely related RNA molecules the most temperature-sensitive positions coincide; these two RNA molecules undergo identical temperature-induced structural RNA perturbations. However, it is probably safe to assume that if two RNA variants manifest different nucleotide positions as the most temperature-sensitive ones within the probability vector, temperature elevation influences the structures of these RNA molecules in a different way. Thus, we have proposed here a new technique aimed at identifying mutations that influence temperature-dependent RNA behavior. The central finding upon which our approach is based is that temperature-sensitive positions are not randomly distributed along the length of RNA but rather form distinct clusters. We speculate that such clusters of temperature-sensitive positions may be located within RNA domains that change their shapes particularly strongly with temperature elevation. Although developed for a particular purpose, our method can be applied for studying the role of RNA structure perturbations in a wide range of temperature-related biological phenomena, such as the evolution of warm-bloodedness, thermophilic adaptation of prokaryptic organisms, or susceptibility of parasites and pathogens to increases in host temperature.

Differences in clusters of temperature-sensitive positions are a potential indicator that RNA structures of mutants react differently to temperature change. This raises the question whether these differences can be a causative factor for (or, at least, associated with) the unique ca/ts behavior of the particular influenza virus strains under study. We identified three types of clusters of temperature-sensitive positions that are (1) present in both wt and ca/ts mutants, (2) present in wt, but absent in ca/ts mutants, and (3) absent in wt, but appear in the ca/ts mutants. We, therefore, first tested whether the disappearance of some clusters in the mutants can indicate that they are causative for a rare phenotype, ca/ts. If these clusters would disappear in ca/ts mutants but remain in non-ca/ts RNA variants possessing the same number of mutations, one could conclude that the cluster disappearance and ca/ts behavior are associated. For all such clusters except one, a high number of computer-generated mutants, which are extremely unlikely to be ca/ts, also demonstrate disappearance of the same clusters. Thus, these clusters may simply correspond to temperature-sensitive regions within particular influenza mRNAs unrelated to ca/ts phenotype. Nevertheless, we did observe one cluster, which is associated with the ca/ts phenotype with statistically significant P-value. This cluster is present in the wt strain. It disappears specifically in the ca/ts mutant, but remains in the computer-generated mutants possessing the same number of mutations as the ca/ts one.

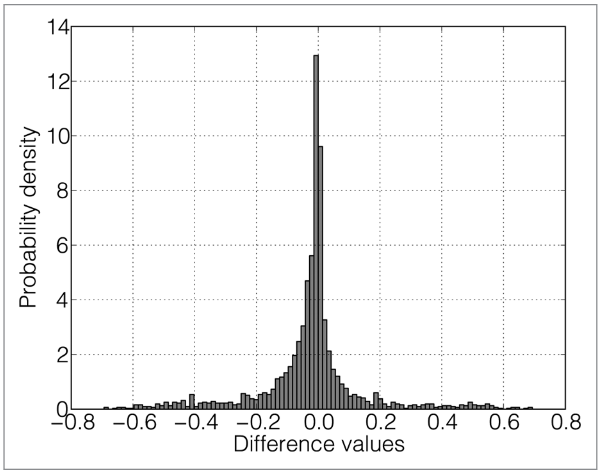


Figure 3. The histogram of differences of the probability values of nucleotides to be in a double-stranded conformation for PB1 Arb/wt upon temperature change between 32°C and 39°C. The vector of probabilities for 32°C was subtracted from the vector for 39°C.

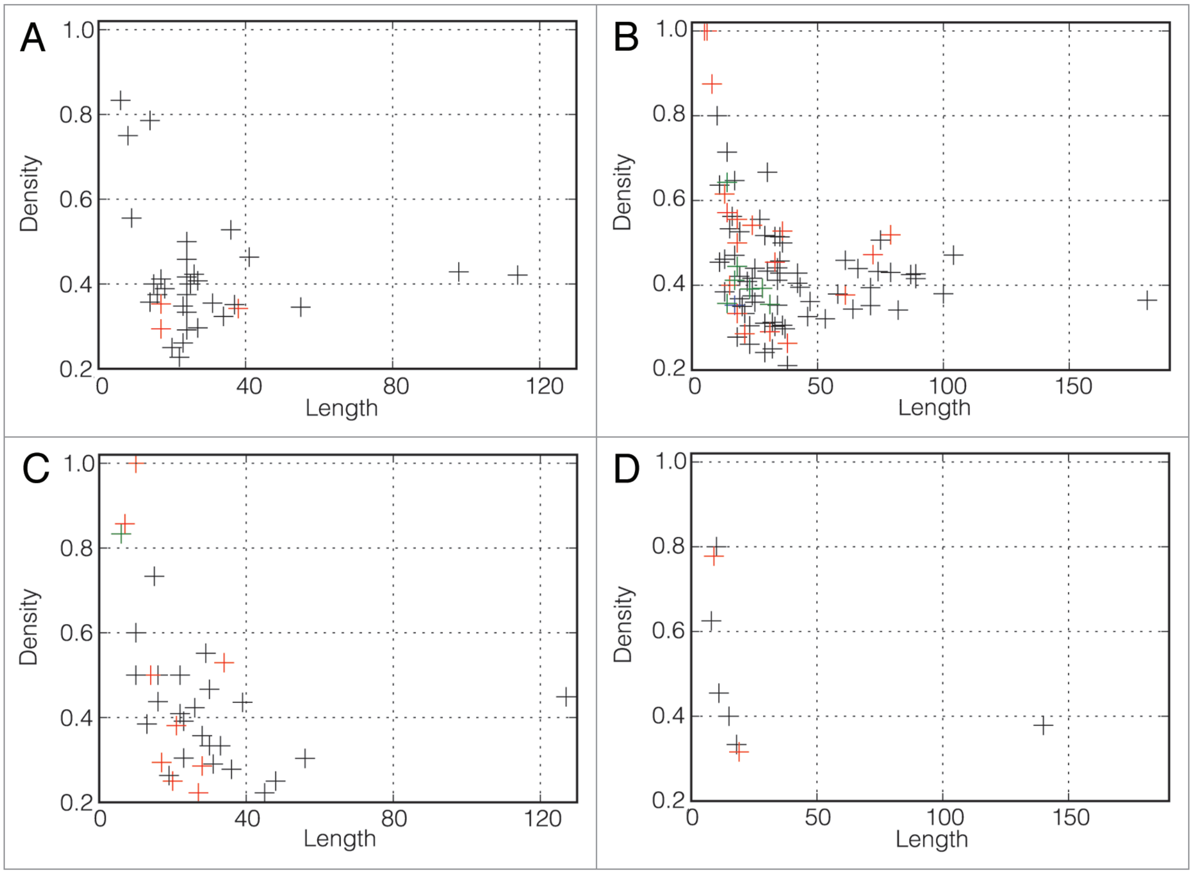


Figure 4. Density of significantly changing positions in determined clusters vs. length. (**A**) Clusters that occur only in wt mRNAs but are not statistically significant (37). (**B**) Clusters occurring in both wt and ca/ts mRNAs (126). (**C**) Clusters that occur only in ca/ts mutants but are not statistically significant (45). (**D**) Statistically significant clusters (9 of them occur only in ca/ts mutants and 1 of them occurs only in wt mRNA). Different colors show different numbers of clusters that have identical values of length and density. Black, one cluster; red, two clusters; green, three clusters; blue, four clusters.

Applying the same computational approach, we then tested if appearance of clusters of temperature-sensitive positions in ca/ts mutants, which are lacking in wt, is a phenotype-specific phenomenon. Based on comparisons with computer-generated mutants, we have demonstrated that nine particular clusters are unlikely to appear in mutants other than ca/ts. Thus, we hypothesize that changes in RNA structure caused by raising temperature could be a potential factor contributing to the molecular mechanisms of the temperature-sensitive and/or cold-adapted phenotype in influenza A.

Direct experimental evidence both on secondary structures of mRNAs and their interactions partners will be required to elucidate the exact role of temperature-induced structural changes in the acquisition of the ca/ts phenotype. For example, it is conceivable that conformational changes of influenza mRNA may play a role through altering the RNA ability to associate/dissociate with proteins and other molecules. Also, it cannot be ruled out that temperature-induced structural changes in the untranslated regions, which we have not considered in our analysis, contribute to the ca/ts phenotype. The current scarcity of sequence data for temperature-sensitive strains and their wild type counterparts notwithstanding, we here propose the hypothesis that temperature-induced structural RNA perturbations may be an underlying mechanism of the ca/ts behavior of influenza virus. Further research in this direction might contribute to the rational design of live-attenuated influenza vaccines.

Materials and Methods

Sequences. In our analysis, we have used the cold-adapted, temperature-sensitive mutants A/Ann Arbor/6/60 (Arb/ca) stemming from the wild type (Arb/wt) with the same name and the two mutants A/Leningrad/134/17/57 (Len/17/ca) and A/Leningrad/134/47/57 (Len/47/ca) stemming from the wild type (wt) A/Leningrad/134/57 (Len/wt). Since information on the location of UTRs was not available, only coding regions were used for the analysis. Information on the locations and sequences of coding regions was retrieved from EMBL-ENA (European Nucleotide Archive).[28] However, these sequences were adapted according to the publications where they originally were reported[29,30] since the mutations annotated in the database were not in agreement with those papers, and no further references were given. The files containing final sequences, used in the current analysis, are presented in the **Supplementary Data.**

The influenza A genome is composed of eight segments encoding 12 proteins: three polymerase subunits (PB1, PB2, and PA),

Table 1. The number of SNPs in the coding sequences of the ca/ts mutants compared with their wild type counterparts.

Strain	M1	M2	NP	NS1	NS2	PA	PB1	PB2
Arb/ca	0	1	2	1	1	3	7	7
Len/17/ca	1	2	0	0	1	3	3	1
Len/47/ca	1	2	1	0	1	3	4	3

The sequences of the ca/ts mutants of the M1, M2, NS2 and PA genes in Len/17/ca and Len/47/ca are identical.

a small proapoptotic mitochondrial protein (PB1-F2), hemagglutinin (HA), neuraminidase (NA), the nucleoprotein (NP), the matrix protein M1, an integral membrane protein M2, and the two nonstructural proteins NS1 and NS2.[31] Recently, Wise et al. showed that PB1 gene segment also encodes a twelfth gene product, N-terminally truncated version on the polypeptide, N40.[32] Sequences for NA and HA were not taken into consideration since these segments do not stem from attenuated viruses in the reassortant live vaccines, and thus cannot be associated with the temperature-sensitive phenotype. For all other genes, the numbers of single nucleotide polymorphisms (SNPs) in the coding sequences of the ca/ts mutants compared with their wild type counterparts are presented in **Table 1**.

Identification of significantly changing positions. For the first step, we wanted to identify those nucleotides within each mRNA that are the most prone to changing their coupling pattern with temperature elevation. These nucleotides would correspond to the most temperature labile positions within RNA chains. To achieve this goal, we proposed and implemented a new technique as discussed here.

At each particular temperature, an RNA sequence consisting of N nucleotides can be presented by a vector of probabilities (hereinafter referred to as "probability vector") for each nucleotide to be in a double-stranded conformation at this temperature. Thus, we substitute a sequence of N ribonucleotides with a sequence of N real numbers between 0.0 and 1.0. Then, we calculate the probability vectors for each of the influenza mRNAs for the temperatures 32°C up to 39°C (in increments of 1°C) using the RNAfold tool from the Vienna RNA package (v.1.8.5)[23,24,33-35] with the command line option –noLP that disallows base pairs that can only occur as helices of length 1. Performing the above described procedure, eight probability vectors were generated for each mRNA. Seven difference vectors v_{32-33}, …, v_{32-39} were calculated from the probability vectors for 33°C to 39°C for the same RNA and the vector at 32°C, containing the set of differences between the value for each position of the probability vector at higher temperature and the value for the same position at lower temperature. These positions in difference vectors of each mRNA that possess values more than three standard deviations apart from the mean calculated over all values of the seven difference vectors were considered temperature-sensitive. Such "significantly changing" positions are presumed to result from perturbations in secondary RNA structures due to the temperature elevation. Furthermore, to filter out possible calculation artifacts, we considered a position temperature-sensitive only if it appeared at some temperature and remained to be such at all higher temperatures.

Comparison of significantly changing positions between wt and ca/ts strains. To test whether significant temperature-induced structural changes in secondary RNA structures are the same or different for wild type strains and their cold-adapted, temperature-sensitive counterparts, we designed a visualization method allowing simultaneous comparison of temperature-induced changes for two RNAs. For example, **Figure 1** depicts a comparison of significantly changing positions between Len/wt and Len/17/ca for a subsequence of the PA mRNA.

Visualization of significantly changing positions demonstrated that such positions are not evenly distributed along the sequences but rather have a tendency to aggregate into clusters, *i.e.* regions with a high density of significantly changing positions. As a tool to analyze such clusters, we employed density plots obtained by sliding a 20-base long window over the vector v_{32-39} and calculating the percentage of significantly changing positions in the window for each possible starting position. For example, **Figures 2A and 2C** depict density plots for the PB2 mRNAs of Arb/wt and Arb/ca, respectively.

Identification of clusters of temperature-sensitive positions. We further sought to provide a definition of clusters of changing positions for each RNA, focusing on the difference vectors v_{32-39}. To these difference vectors, we applied the density-based spatial clustering of applications with noise (DBSCAN) algorithm.[36,37] This algorithm needs two parameters as input, a distance threshold r and a density threshold $MinPts$. For a given set of points D (in our case the set of significantly changing positions in mRNA according to the difference vector v_{32-39}), the density of every point p_i from D is calculated as the number of points q_i that are within a radius r around p_i. If $q_i > MinPts$, then the point p_i is classified as a *core point*. If the distance between two points is less than r, then they are said to be *directly-connected*. Two points are considered *density-connected* if they are connected to core points and these core points are, in turn, density-connected. A cluster is constructed as a maximally connected component of the set of points that have a distance of smaller than r to some core point. We used the implementation of DBSCAN from the scikit-learn Python module[38] with a distance threshold r equal to 11 and a density threshold $MinPts$ equals 4.

Generation of randomly mutated mRNAs. In order to assess whether the appearance of clusters of temperature-sensitive positions is specific for mutations inducing the ca/ts phenotype or whether random mutations unrelated to ca/ts phenotype would be as likely to induce these clusters, we adopted an approach similar to that employed in our previous paper.[27] For each wt mRNA, a data set consisting of 1000 mutant sequences was generated in silico. Each in silico generated variant contained the same number of mutations as the respective ca/ts mutant. All computer-generated mutations were synonymous ones and introduced into the sequences randomly. It is safe to assume that none (or extremely few) of the randomly generated in silico mutants would possess the ca/ts phenotype if tested in vitro and/or in vivo. Significantly changing positions in the sequences

Table 2. The number of positions in each mRNA where the probability of nucleotides to be in a double-stranded conformation decreases (increases) upon temperature elevation from 32°C to 39°C

Strain	M1	M2	NP	NS1	NS2	PA	PB1	PB2
Arb/wt	–/–	207/87	1045/452	445/209	231/135	1452/699	1433/841	1537/743
Arb/ca	–/–	211/83	1007/490	440/214	258/108	1437/714	1510/764	1531/749
Len/wt	534/225	221/73	900/507	–/–	233/133	1467/684	1428/846	1552/728
Len/17/ca	525/234	219/75	–/–	–/–	248/118	1462/689	1525/749	1578/702
Len/47/ca	525/234	219/75	899/508	–/–	248/118	1462/689	1510/764	1622/658

There are no positions at which the probability to be paired upon temperature change between 32°C and 39°C remains unchanged. Here, and in all subsequent tables for those mRNAs that were not considered in the analysis due to the absence of mutations, values are not shown.

Table 3. The number of nucleotides in each mRNA where the base pairing probability decreases (increases) significantly (more than three standard deviations from the mean over all temperature differences between 33°C to 39°C compared with 32°C) upon temperature change between 32°C and 39°C compared with other nucleotides in the same mRNA

Strain	M1	M2	NP	NS1	NS2	PA	PB1	PB2
Arb/wt	–/–	19/8	88/36	42/25	31/4	58/26	130/95	133/83
Arb/ca	–/–	13/12	78/34	29/18	8/11	102/50	100/57	132/68
Len/wt	46/17	17/8	49/34	–/–	15/12	84/53	129/84	130/72
Len/17/ca	50/16	19/13	–/–	–/–	23/6	114/39	125/60	131/62
Len/47/ca	50/16	19/13	48/39	–/–	23/6	114/39	123/60	137/52

Table 4. The number of clusters in each mRNA as determined by the DBSCAN algorithm

Strain	M1	M2	NP	NS1	NS2	PA	PB1	PB2
Arb/wt	-	1	7	5	3	5	12	9
Arb/ca	-	1	8	4	1	10	9	11
Len/wt	4	1	6	-	0	10	15	11
Len/17/ca	5	1	-	-	2	11	9	10
Len/47/ca	5	1	8	-	2	11	10	10

Table 5. Average cluster length in each mRNA

Strain	M1	M2	NP	NS1	NS2	PA	PB1	PB2
Arb/wt	-	42.0	24.9	27.0	18.3	19.4	37.8	58.1
Arb/ca	-	58.0	15.9	21.5	45.0	28.6	34.9	34.9
Len/wt	22.5	42.0	26.5	-	-	25.0	34.9	39.9
Len/17/ca	19.8	61.0	-	-	23.0	21.5	38.9	33.3
Len/47/ca	19.8	61.0	19.5	-	23.0	21.5	35.0	29.0

Table 6. Average density of significantly changing positions inside clusters in each mRNA

Strain	M1	M2	NP	NS1	NS2	PA	PB1	PB2
Arb/wt	-	0.43	0.39	0.38	0.45	0.37	0.44	0.39
Arb/ca	-	0.38	0.53	0.36	0.22	0.34	0.43	0.42
Len/wt	0.41	0.40	0.36	-	-	0.38	0.37	0.41
Len/17/ca	0.46	0.38	-	-	0.26	0.48	0.44	0.41
Len/47/ca	0.46	0.38	0.46	-	0.26	0.48	0.44	0.44

from the artificial data sets were determined as described above and used to calculate clusters of changing positions by applying the DBSCAN algorithm. Clusters from computer-generated sequences were compared with the clusters from naturally occurring wt and ca/ts mutants.

Statistical tests. For each particular cluster of interest identified in wt and/or the ca/ts mutants, the frequency of its occurrence in the in silico generated mutants was calculated. Using these frequencies we conducted a statistical analysis to test if occurrence/disappearance of a particular cluster is associated with the ca/ts phenotype. For each cluster, which we observed in a ca/ts mutant but not in the wt, we tested the null hypothesis (H_0) that the probability to observe this cluster among the computer generated sequences was 5% or higher. Conversely, for each cluster, which was observed in the wt but not in ca/ts strain, the null hypothesis (H_0) was that the probability to observe this cluster was less than 95%. In other words, a low frequency means that a cluster, which we observe in naturally occurring ca/ts strain although it is absent in the wt, is unlikely to occur by chance. Thus, the appearance of this cluster is likely to be associated with the ca/ts phenotype. Similarly, the fact that a cluster was present in the wt but disappeared in the ca/ts mutant can only be explained by the ca/ts phenotype if the probability to observe this cluster in the random mutants is 95% or higher.

To that end, we used one-sided binomial tests. The significance level for the test was Bonferroni-corrected by dividing the significance level of 5% by the total number of clusters in that sequence. H_0 was rejected for P-values lower than the adjusted significance level. For these calculations, a cluster was considered to be 'present' in an artificial sequence if that sequence contained

Table 7. Unique clusters potentially associated with the ca/ts phenotype. The P-values of all clusters in one sequence were checked against Bonferroni-corrected significance levels. For each Bonferroni correction, the total number of clusters located in the corresponding sequence was used (11 clusters in Arb/ca PB2, two clusters in Len/17/ca and Len/47/ca NS2, 11 clusters in Len/17/ca and Len/47/ca PA, ten clusters in Len/17/ca and Len/47/ca PB2, eight clusters in Len/47/ca NP, 11 clusters in Len/wt PB2).

Strain	Sequence	Position	Occurrence in the random data set[a]	95% confidence interval[b]	P-value
Arb/ca	PB2	329–336	15	[0.0, 0.023]	3.34E-09
Len/17/ca	NS2	290–308	16	[0.0, 0.024]	1.11E-08
Len/17/ca	PA	93–101	32	[0.0, 0.043]	0.0037
Len/17/ca	PB2	1293–1310	9	[0.0, 0.016]	5.24E-13
Len/47/ca	NP	1017–1026	29	[0.0, 0.039]	0.0007
Len/47/ca	NP	1178–1192	29	[0.0, 0.039]	0.0007
Len/47/ca	NS2	290–308	16	[0.0, 0.024]	1.11E-08
Len/47/ca	PA	93–101	32	[0.0, 0.043]	0.0037
Len/47/ca	PB2	808–818	3	[0.0, 0.008]	1.36E-18
Len/wt	PB2	1490–1629	982	[0.973, 1.0]	1.03E-07

[a]The number of times a particular cluster was found in a data set of 1000 sequences with randomly introduced mutations. [b]Estimated range of values which is likely to include the probability to find a particular cluster with the probability of 95%.

a cluster overlapping, by at least one position, with the cluster from the real sequence.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Funding

The DFG International Research Training Group 'Regulation and Evolution of Cellular Systems' (GRK 1563) and the Russian Foundation for Basic Research (RFBR 09–04–92742).

Acknowledgments

The authors gratefully acknowledge the support of the TUM Graduate School's Thematic Graduate Center Regulation and Evolution of Cellular Systems (RECESS) at the Technische Universität München.

Supplemental Materials

Supplemental materials may be found here: www.landesbioscience.com/journals/cc/article/22081

References

1. Stöhr K, Kieny MP, Wood D. Influenza pandemic vaccines: how to ensure a low-cost, low-dose option. Nat Rev Microbiol 2006; 4:565-6; PMID:16888876; http://dx.doi.org/10.1038/nrmicro1482.
2. Belsey MJ, de Lima B, Pavlou AK, Savopoulos JW. Influenza vaccines. Nat Rev Drug Discov 2006; 5:183-4; PMID:16557657; http://dx.doi.org/10.1038/nrd1988.
3. Ilyinskii PO, Thoidis G, Shneider AM. Development of a vaccine against pandemic influenza viruses: current status and perspectives. Int Rev Immunol 2008; 27:392-426; PMID:19065349; http://dx.doi.org/10.1080/08830180802295765.
4. Song H, Nieto GR, Perez DR. A new generation of modified live-attenuated avian influenza viruses using a two-strategy combination as potential vaccine candidates. J Virol 2007; 81:9238-48; PMID:17596317; http://dx.doi.org/10.1128/JVI.00893-07.
5. Falcón AM, Marión RM, Zürcher T, Gómez P, Portela A, Nieto A, et al. Defective RNA replication and late gene expression in temperature-sensitive influenza viruses expressing deleted forms of the NS1 protein. J Virol 2004; 78:3880-8; PMID:15047804; http://dx.doi.org/10.1128/JVI.78.8.3880-3888.2004.
6. Jin H, Lu B, Zhou H, Ma C, Zhao J, Yang CF, et al. Multiple amino acid residues confer temperature sensitivity to human influenza virus vaccine strains (FluMist) derived from cold-adapted A/Ann Arbor/6/60. Virology 2003; 306:18-24; PMID:12620793; http://dx.doi.org/10.1016/S0042-6822(02)00035-1.
7. Jin H, Zhou H, Lu B, Kemble G. Imparting temperature sensitivity and attenuation in ferrets to A/Puerto Rico/8/34 influenza virus by transferring the genetic signature for temperature sensitivity from cold-adapted A/Ann Arbor/6/60. J Virol 2004; 78:995-8; PMID:14694130; http://dx.doi.org/10.1128/JVI.78.2.995-998.2004.
8. Tsfasman TM, Markushin SG, Akopova II, Ghendon YZ. Molecular mechanisms of reversion to the ts+ (non-temperature-sensitive) phenotype of influenza A cold-adapted (ca) virus strains. J Gen Virol 2007; 88:2724-9; PMID:17872525; http://dx.doi.org/10.1099/vir.0.83014-0.
9. Chan W, Zhou H, Kemble G, Jin H. The cold adapted and temperature sensitive influenza A/Ann Arbor/6/60 virus, the master donor virus for live attenuated influenza vaccines, has multiple defects in replication at the restrictive temperature. Virology 2008; 380:304-11; PMID:18768193; http://dx.doi.org/10.1016/j.virol.2008.07.027.
10. Youil R, Kiseleva I, Kwan WS, Szymkowiak C, Toner TJ, Su Q, et al. Phenotypic and genetic analyses of the heterogeneous population present in the cold-adapted master donor strain: A/Leningrad/134/17/57 (H2N2). Virus Res 2004; 102:165-76; PMID:15084398; http://dx.doi.org/10.1016/j.virusres.2004.01.026.
11. Snyder MH, Clements ML, De Borde D, Maassab HF, Murphy BR. Attenuation of wild-type human influenza A virus by acquisition of the PA polymerase and matrix protein genes of influenza A/Ann Arbor/6/60 cold-adapted donor virus. J Clin Microbiol 1985; 22:719-25; PMID:4056002.
12. Dalton RM, Mullin AE, Amorim MJ, Medcalf E, Tiley LS, Digard P. Temperature sensitive influenza A virus genome replication results from low thermal stability of polymerase-cRNA complexes. Virol J 2006; 3:58; PMID:16934156; http://dx.doi.org/10.1186/1743-422X-3-58.
13. Shamovsky I, Ivannikov M, Kandel ES, Gershon D, Nudler E. RNA-mediated response to heat shock in mammalian cells. Nature 2006; 440:556-60; PMID:16554823; http://dx.doi.org/10.1038/nature04518.
14. Elväng A, Melik W, Bertrand Y, Lönn M, Johansson M. Sequencing of a tick-borne encephalitis virus from Ixodes ricinus reveals a thermosensitive RNA switch significant for virus propagation in ectothermic arthropods. Vector Borne Zoonotic Dis 2011; 11:649-58; PMID:21254926; http://dx.doi.org/10.1089/vbz.2010.0105.
15. Ilyinskii PO, Schmidt T, Lukashev D, Meriin AB, Thoidis G, Frishman D, et al. Importance of mRNA secondary structural elements for the expression of influenza virus genes. OMICS 2009; 13:421-30; PMID:19594376; http://dx.doi.org/10.1089/omi.2009.0036.
16. Moss WN, Priore SF, Turner DH. Identification of potential conserved RNA secondary structure throughout influenza A coding regions. RNA 2011; 17:991-1011; PMID:21536710; http://dx.doi.org/10.1261/rna.2619511.
17. Gultyaev AP, Fouchier RA, Olsthoorn RC. Influenza virus RNA structure: unique and common features. Int Rev Immunol 2010; 29:533-56; PMID:20923332; http://dx.doi.org/10.3109/08830185.2010.507828.

18. Priore SF, Moss WN, Turner DH. Influenza A virus coding regions exhibit host-specific global ordered RNA structure. PLoS ONE 2012; 7:e35989; PMID:22558296; http://dx.doi.org/10.1371/journal.pone.0035989.
19. Motard J, Rouxel R, Paun A, von Messling V, Bisaillon M, Perreault JP. A novel ribozyme-based prophylaxis inhibits influenza A virus replication and protects from severe disease. PLoS ONE 2011; 6:e27327; PMID:22110627; http://dx.doi.org/10.1371/journal.pone.0027327.
20. Lyngso RB. Complexity of pseudoknot prediction in simple models. Lect Notes Comput Sci 2004; 3142:919-31; http://dx.doi.org/10.1007/978-3-540-27836-8_77.
21. Pipas JM, McMahon JE. Method for predicting RNA secondary structure. Proc Natl Acad Sci USA 1975; 72:2017-21; PMID:1056009; http://dx.doi.org/10.1073/pnas.72.6.2017.
22. Zuker M, Sankoff D. Rna Secondary Structures and Their Prediction. Bull Math Biol 1984; 46:591-621.
23. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990; 29:1105-19; PMID:1695107; http://dx.doi.org/10.1002/bip.360290621.
24. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 2004; 10:1178-90; PMID:15272118; http://dx.doi.org/10.1261/rna.7650904.
25. Witwer C, Rauscher S, Hofacker IL, Stadler PF. Conserved RNA secondary structures in Picornaviridae genomes. Nucleic Acids Res 2001; 29:5079-89; PMID:11812840; http://dx.doi.org/10.1093/nar/29.24.5079.
26. Thurner C, Witwer C, Hofacker IL, Stadler PF. Conserved RNA secondary structures in Flaviviridae genomes. J Gen Virol 2004; 85:1113-24; PMID:15105528; http://dx.doi.org/10.1099/vir.0.19462-0.
27. Chursov A, Walter MC, Schmidt T, Mironov A, Shneider A, Frishman D. Sequence-structure relationships in yeast mRNAs. Nucleic Acids Res 2012; 40:956-62; PMID:21954438; http://dx.doi.org/10.1093/nar/gkr790.
28. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res 2005; 33(Database issue):D29-33; PMID:15608199; http://dx.doi.org/10.1093/nar/gki098.
29. Cox NJ, Kitame F, Kendal AP, Maassab HF, Naeve C. Identification of sequence changes in the cold-adapted, live attenuated influenza vaccine strain, A/Ann Arbor/6/60 (H2N2). Virology 1988; 167:554-67; PMID:2974219.
30. Klimov AI, Cox NJ, Yotov WV, Rocha E, Alexandrova GI, Kendal AP. Sequence changes in the live attenuated, cold-adapted variants of influenza A/Leningrad/134/57 (H2N2) virus. Virology 1992; 186:795-7; PMID:1733114; http://dx.doi.org/10.1016/0042-6822(92)90050-Y.
31. Nelson MI, Holmes EC. The evolution of epidemic influenza. Nat Rev Genet 2007; 8:196-205; PMID:17262054; http://dx.doi.org/10.1038/nrg2053.
32. Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, et al. A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. J Virol 2009; 83:8021-31; PMID:19494001; http://dx.doi.org/10.1128/JVI.00826-09.
33. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 1981; 9:133-48; PMID:6163133; http://dx.doi.org/10.1093/nar/9.1.133.
34. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast Folding and Comparison of Rna Secondary Structures. Monatsh Chem 1994; 125:167-88; http://dx.doi.org/10.1007/BF00818163.
35. Bompfünewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, et al. Variations on RNA folding and alignment: lessons from Benasque. J Math Biol 2008; 56:129-44; PMID:17611759; http://dx.doi.org/10.1007/s00285-007-0107-5.
36. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data MIning 1996:226-31.
37. Kriegel H-P, Kröger P, Sander J, Zimek A. Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2011; 1:231-40; http://dx.doi.org/10.1002/widm.30.
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011; 12:2825-30.
39. Kung YH, Huang SW, Kuo PH, Kiang D, Ho MS, Liu CC, et al. Introduction of a strong temperature-sensitive phenotype into enterovirus 71 by altering an amino acid of virus 3D polymerase. Virology 2010; 396:1-9; PMID:19906393; http://dx.doi.org/10.1016/j.virol.2009.10.017.
40. Lulla V, Sawicki DL, Sawicki SG, Lulla A, Merits A, Ahola T. Molecular defects caused by temperature-sensitive mutations in Semliki Forest virus nsP1. J Virol 2008; 82:9236-44; PMID:18596091; http://dx.doi.org/10.1128/JVI.00711-08.
41. Fan Y, Zhao Q, Zhao Y, Wang Q, Ning Y, Zhang Z. Complete genome sequence of attenuated low-temperature Thiverval strain of classical swine fever virus. Virus Genes 2008; 36:531-8; PMID:18401695; http://dx.doi.org/10.1007/s11262-008-0229-x.
42. Sparks JS, Donaldson EF, Lu X, Baric RS, Denison MR. A novel mutation in murine hepatitis virus nsp5, the viral 3C-like proteinase, causes temperature-sensitive defects in viral growth and protein processing. J Virol 2008; 82:5999-6008; PMID:18385240; http://dx.doi.org/10.1128/JVI.00203-08.
43. Berkhout B, Klaver B, Das AT. Forced evolution of a regulatory RNA helix in the HIV-1 genome. Nucleic Acids Res 1997; 25:940-7; PMID:9023102; http://dx.doi.org/10.1093/nar/25.5.940.
44. Mirmomeni MH, Hughes PJ, Stanway G. An RNA tertiary structure in the 3' untranslated region of enteroviruses is necessary for efficient replication. J Virol 1997; 71:2363-70; PMID:9032373.
45. Rowe A, Ferguson GL, Minor PD, Macadam AJ. Coding changes in the poliovirus protease 2A compensate for 5'NCR domain V disruptions in a cell-specific manner. Virology 2000; 269:284-93; PMID:10753707; http://dx.doi.org/10.1006/viro.2000.0244.
46. Sugimoto M, Yamanouchi K. Characteristics of an attenuated vaccinia virus strain, LC16m0, and its recombinant virus vaccines. Vaccine 1994; 12:675-81; PMID:8091843; http://dx.doi.org/10.1016/0264-410X(94)90215-1.
47. White MD, Bosio CM, Duplantis BN, Nano FE. Human body temperature and new approaches to constructing temperature-sensitive bacterial vaccines. Cell Mol Life Sci 2011; 68:3019-31; PMID:21626408; http://dx.doi.org/10.1007/s00018-011-0734-2.
48. Collins PL, Murphy BR. New generation live vaccines against human respiratory syncytial virus designed by reverse genetics. Proc Am Thorac Soc 2005; 2:166-73; PMID:16113487; http://dx.doi.org/10.1513/pats.200501-011AW.
49. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 2010; 11:129; PMID:20230624; http://dx.doi.org/10.1186/1471-2105-11-129.
50. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 2003; 31:3423-8; PMID:12824339; http://dx.doi.org/10.1093/nar/gkg614.
51. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 2008; 9:474; PMID:19014431; http://dx.doi.org/10.1186/1471-2105-9-474.
52. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature 2010; 467:103-7; PMID:20811459; http://dx.doi.org/10.1038/nature09322.
53. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr., Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 2009; 460:711-6; PMID:19661910; http://dx.doi.org/10.1038/nature08237.

2.3 RNAtips: Analysis of Temperature-induced Changes of RNA Secondary Structure

Andrey Chursov, Sebastian J. Kopetzky, Gennady Bocharov, Dmitrij Frishman and Alexander Shneider
Nucleic Acids Res., 41(W1):W486-W491, 2013

This article describes a web server for predicting temperature-induced changes in RNA secondary structure. A wide range of biological phenomena (e.g. elevation of body temperature due to an illness, adaptation to environmental temperature conditions, biology of cold-blooded vs. warm-blooded organisms) is related to the temperature effect on the organism. Although, thousands of papers have been published attempting to explain temperature effects through perturbations of protein structures, it has been almost impossible to assess if changes in RNA structures may contribute to these effects. The web server we have developed closes this gap and enables researchers to study temperature-induced perturbation of RNA structures.

Perturbations of secondary RNA structures may play an important role in an organism's reaction to temperature change. The RNAtips (**t**emperature-**i**nduced **p**erturbation of **s**tructure) web server can be used to predict regions of RNA secondary structures that are likely to undergo structural changes when prompted by a change in temperature. For a single RNA sequence, RNAtips identifies those nucleotides that have the greatest reaction to temperature change and the temperature-sensitive clusters of such nucleotides. When the research goal is to compare two RNA sequences and identify if they react to a temperature change differently, the locations of temperature-sensitive clusters within the two RNAs are compared. If the two sequences are homologs with a limited number of codon substitutions, an analysis can be performed to demonstrate if the difference in the location of temperature-sensitive clusters between the two sequences is specific to these particular nucleotide substitutions, or if it could be achieved with the same number of random mutations (synonymous or non-synonymous). The type of random computer-introduced mutations depends on whether the input sequences are coding or non-coding.

For input into RNAtips, one or two RNA sequences of the same length must be provided in FASTA-format (the header can be omitted). These sequences can either be uploaded as a text file (each file may only contain one sequence) or directly pasted into an input field. Sequence length is restricted to 9999 nucleotides. Additionally, a user has to select a temperature range for which the RNA structural perturbation should be calculated (the default is 32°C – 39°C). The minimal and the maximal temperatures allowed are 0°C and 99°C, respectively. Furthermore, the maximal temperature difference (i.e. $t_{max} - t_{min} + 1$) is restricted to 20°C (e.g. 30°C – 49°C). The advanced options allow the setting of the following: (i) a significance threshold (see the following paragraph on processing), (ii) the DBSCAN parameters ε and MinPts (the defaults are 11 for ε and 5 for MinPts), and (iii) parameters for RNAfold in order to decide if GU-pairs should be allowed.

In addition, to avoid incorrect attribution of coding or non-coding input sequences, we implemented a checkbox where the user can directly indicate that the input sequence(s) is (are) non-coding. If a user does not set this checkbox, then by default the server assumes that the input sequence(s) is (are) coding ones. Also, if a dataset of random mutants is generated, we indicate directly what type of mutations was introduced at the link for downloading this dataset. In the case of coding sequences, it says: “only synonymous mutations were introduced”; otherwise, it says: “random mutations were introduced.”

Processing is based on the method described in the previous paper. For each temperature in the given range, the probabilities of each nucleotide to be coupled within a double-stranded conformation are calculated with the RNAfold tool (Hofacker et al., 1994). For a temperature range $[t_{min} : t_{max}]$, the probability-differences for each nucleotide position for $(t_{min} + 1) - t_{min}, ..., t_{max} - t_{min}$ are then calculated. A threshold is applied (the default is three standard deviations) to the resulting distribution in order to identify the most temperature-sensitive positions. By using the DBSCAN algorithm (Ester et al., 1996) on the vector $t_{max} - t_{min}$, clusters of significantly changing positions are identified. P-values are obtained by applying one-sided binomial tests to evaluate the occurrence of identified clusters in the mutant dataset.

The HTML output provides colored visualization of temperature-sensitive positions and clusters. In addition, a table displays information on identified significant positions and clusters, including their total number as well as their average length and density. A second table shows the exact location of the clusters (if any). Results from the statistical evaluation (if conducted) are presented in a third table demonstrating whether the difference between two RNAs in their reaction to the temperature change is due to specific nucleotide substitutions or not. The output also contains histograms that were used to identify significantly changing positions (three sigmas away from the mean). In addition, there is a figure that shows the relationship between the length and density of the clusters. Finally, there is another figure that demonstrates the density of the most temperature-sensitive positions over the whole RNA sequence, together with the location of clusters and the location of nucleotide substitutions.

All figures and additional information, including sequences of the mutants generated for the statistical analysis, can be downloaded by the RNAtips user.

The project was designed by Alexander Shneider and me. The web server has been developed by me with an assistance of Sebastian J. Kopetzky and Gennady Bocharov. The paper was written by myself, Sebastian J. Kopetzky, Dmitrij Frishman and Alexander Shneider.

W486–W491 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue*
doi:10.1093/nar/gkt486

Published online 12 June 2013

RNAtips: analysis of temperature-induced changes of RNA secondary structure

Andrey Chursov[1], Sebastian J. Kopetzky[1], Gennady Bocharov[2], Dmitrij Frishman[1,3,]* and Alexander Shneider[4,]*

[1]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, Maximus-von-Imhof-Forum 3, D-85354 Freising, Germany, [2]Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina str. 8, 119333 Moscow, Russia, [3]Helmholtz Center Munich-German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany and [4]Cure Lab, Inc., 43 Rybury Hillway, Needham, MA 02492, USA

Received March 6, 2013; Revised May 1, 2013; Accepted May 11, 2013

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

ABSTRACT

Although multiple biological phenomena are related to temperature (e.g. elevation of body temperature due to an illness, adaptation to environmental temperature conditions, biology of coldblooded versus warm-blooded organisms), the molecular mechanisms of these processes remain to be understood. Perturbations of secondary RNA structures may play an important role in an organism's reaction to temperature change—in all organisms from viruses and bacteria to humans. Here, we present RNAtips (temperature-induced perturbation of structure) web server, which can be used to predict regions of RNA secondary structures that are likely to undergo structural alterations prompted by temperature change. The server can also be used to: (i) detect those regions in two homologous RNA sequences that undergo different structural perturbations due to temperature change and (ii) test whether these differences are specific to the particular nucleotide substitutions distinguishing the sequences. The RNAtips web server is freely accessible without any login requirement at http://rnatips.org.

INTRODUCTION

Structural perturbations in RNA molecules induced by temperature change may have important biological implications. For instance, the stability of mRNA structural elements in 5′-untranslated regions correlates with the translation rate in *Saccharomyces cerevisiae* (1). Another example is the temperature-sensitivity of cold-adapted influenza vaccine strains. For decades, it was a conundrum why wild-type influenza strains react differently to elevated temperature than their cold-adapted temperature-sensitive counterparts. Recently, it has been demonstrated that this difference in temperature sensitivity may be due to the difference in temperature-induced perturbations in mRNA secondary structures (2). Perhaps, the most widely known example is RNA thermometers, which at a particular temperature alter their structure, and regulate translation of heat-shock, cold-shock and virulence genes (3–8). Usually, RNA thermometers are located in 5′-untranslated regions, and their structures melt at an elevated temperature thereby permitting ribosomes to initiate the translation process.

There are several experimental approaches to measuring the melting temperature of an RNA structure (9), including ultraviolet absorbance (10,11), fluorescence-based techniques (12,13) and thermal gradient electrophoresis (14–16). Recently, temperature stability of RNA structural elements was assessed on a genome-wide basis (17). The Parallel Analysis of RNA Structures with Temperature Elevation technique was applied to the yeast transcriptome, and relative melting temperatures for RNA structures were obtained by probing RNA structures at different temperatures from 23 to 75°C. As a result of this assessment, thousands of potential RNA thermometers and highly temperature-stable structures were identified.

Temperature-induced perturbations of RNA structures may play crucial, and yet unknown, biological role(s) in a

---

*To whom correspondence should be addressed. Tel: +1 609 841 1201; Fax: +1 310 446 9814; Email: ashneider@curelab.com
Correspondence may also be addressed to Dmitrij Frishman. Tel: +49 8161 712134; Fax: +49 8161 71 2186; Email: d.frishman@wzw.tum.de

The RNAtips web server can be accessed at http://rnatips.org. This website is free and open to all users, and there is no login requirement.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

variety of processes. Elevation of body temperature is the most common symptom of many illnesses. The effects of elevated body temperature on RNA structures both in pathogens and their hosts are still unknown, although they may constitute a defense mechanism. Additionally, it would be interesting to assess whether RNA temperature sensitivity plays an evolutionary role in organism adaptation to different climate zones, as well as to seasonal and day–night temperature change. The latter question is especially important owing to global climate change. Is temperature sensitivity of RNA structures in bacteria living in geysers different from that of bacteria living at negative temperatures? Do RNA structures from warm-blooded organisms react to the temperature change similarly to their counterparts in cold-blooded animals? These and many other questions could not be systematically addressed, however, as (to the best of our knowledge) there is no convenient instrument to identify and compare temperature-sensitive regions of RNA molecules.

To close this gap, RNAtips (temperature-induced perturbation of structure) web server has been developed. For a single RNA sequence, RNAtips identifies (i) those nucleotides for which temperature change causes appreciable alteration of the probability to form Watson–Crick (W–C) pairs and (ii) clusters of such temperature-sensitive nucleotides. If the research goal is to compare two RNA sequences and identify whether they react differently to a temperature change, the locations of temperature-sensitive clusters within the two RNAs are compared. If the two sequences are homologs with a limited number of base substitutions, an analysis can be performed to demonstrate whether the difference in location of the temperature-sensitive clusters between the two sequences is specific to these particular nucleotide substitutions, or if it could be achieved with the same number of random mutations (synonymous and/or non-synonymous).

METHOD SUMMARY

The methodology implemented in RNAtips web server for assessing such impacts of temperature change was previously described and published by Chursov *et al.* (2). In short, each nucleotide within an RNA sequence has a probability of being paired via W–C bonds. This probability is temperature dependent; therefore, temperature changes influence the probability of forming W–C pairs for each and every nucleotide. However, some nucleotides change their pairing probabilities to a much greater extent than others. Moreover, these highly temperature-sensitive nucleotides may not be evenly distributed along the RNA sequence but rather form distinct clusters (2). Thus, the first task performed by the RNAtips web server is identification of those positions, which are prone through temperature elevation to significantly change their probability of being paired. This task is performed through the following steps. Step 1: Probabilities of nucleotides to be coupled within a double-stranded conformation are assessed at each temperature within the given range by using the RNAfold tool of the ViennaRNA package (18). Step 2: For each nucleotide, RNAtips calculates the difference in probability for it to be in a paired state at the lower temperature and at the higher one. These differences are calculated for the entire temperature range ($t_1 : t_2$) [i.e. for $(t_1 + 1) - t_1, \dots, t_2 - t_1$] and then combined into one data set. For example, if the temperature range is set to 32–39°C and the length of the sequence is 1000, then the changes of probabilities are considered for 33°C compared with 32°C, 34°C compared with 32°C,..., 39°C compared with 32°C, and the final data set would contain 7000 values. Step 3: The server identifies the most temperature-sensitive positions. For this purpose, the server selects those values (and their corresponding nucleotides) from the data set generated in Step 2, which are distant from the mean by more than three standard deviations (the default value can be changed by the user). The server then considers these positions to be the most temperature-sensitive, and they are then mapped on the original sequence. Furthermore, clusters of significantly changing positions are then identified by applying the density-based spatial clustering of applications with noise (DBSCAN) algorithm to the locations of such positions. The server default action is to apply the cluster analysis algorithm only to the highest temperature differences t_2-t_1, (32–39°C in the previous example) (19,20).

It may be important to assess whether structures of RNA molecules sharing sequence similarity react (dis)similarly to temperature change. For simplicity of explanation, assume that one RNA sequence was derived from another sequence via some mutations. Then, the second task, which can be performed by RNAtips server, is to identify whether structures of two homologous RNA sequences react differently to the temperature change and, if they do, whether this difference can be attributed to the specific mutations separating the two homologous sequences. Thus, if a user inputs two sequences, RNAtips identifies clusters of temperature-sensitive positions, which could be either common for both sequences or uniquely present in only one of the two RNA molecules. If the clusters of temperature-sensitive positions are not identical for the two sequences, the server offers statistical analysis identifying whether the difference in temperature sensitivity is specific to the particular nucleotide substitutions naturally differentiating the sequences or whether any set of mutations comparable in size could lead to the same difference.

Therefore, assume that N nucleotide substitutions differentiate sequence A from sequence B. The server generates a data set of derivative sequences for A introducing N substitutions into each derivative sequence. There are two different methods of introducing random substitutions into a sequence(s) depending on whether the sequence(s) is(are) non-coding or coding. If A is a coding sequence (default), mutants will be generated by introducing synonymous mutations only. If A is a non-coding sequence, the user should mark a checkbox: 'The input sequence(s) is(are) non-coding'. In this case, *in silico* mutations will be introduced at random positions mimicking frequencies of nucleotide substitutions naturally occurring between A and B (e.g. if 25% of nucleotide substitutions between A and B are T->C, then T->C substitutions will be introduced in 25% of random *in silico* mutations). For

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

each computer-generated sequence, the server will calculate its clusters of the temperature-sensitive positions as described earlier in the text. If sequence B has a sequence-specific cluster of temperature-sensitive positions not present in A, some of *in silico* derivatives of A may possess clusters overlapping with the sequence-specific cluster observed in B. Let us assume that 1% or less of computer-generated sequences possess such clusters overlapping with the sequence-specific cluster in B. This means that 99% of random mutation sets did not lead to the appearance of this sequence-specific cluster of temperature sensitivity specific for sequence B, but not for A. Thus, one can conclude that the RNA structure of sequence B reacts to the temperature change differently than the structure of sequence A because it possesses a specific set of mutations as opposed to just N non-specific mutations. The RNAtips server performs a statistical analysis calculating a *P*-value for every sequence-specific cluster by performing a one-sided binomial test. For sequence-specific clusters occurring in the first sequence but not in the second one, the null hypothesis (H_0) is that the probability to observe this cluster is <95%. Consequently, a small *P*-value shows that the cluster is unlikely to disappear in the second sequence by chance. For sequence-specific clusters occurring in the second sequence but not in the first one, the null hypothesis (H_0) is that the probability to observe this cluster amongst the mutants generated *in silico* is $\geq 5\%$. Therefore, a small *P*-value shows that the cluster is unlikely to appear in the second sequence by chance.

WEB SERVER

Input data

The input for RNAtips consists of either one or two RNA sequences of the same length that should be provided in FASTA-format (the header can be omitted). The sequences can either be uploaded as text files (each file may contain only one sequence), or the sequences may be directly pasted into an input field. The sequences may contain the characters A, C, G, U and T (for further computations, all Thymidines will be replaced with Uracils automatically). The maximal length of the sequences is limited to 9999 nt. To see an example of possible input sequences, the user can click on the 'sample' link on the Start page. Influenza strains A/Leningrad/134/57 and its cold-adapted temperature-sensitive mutant A/Leningrad/134/47/57 are used as sample sequences.

Additionally, a user has to specify two temperatures t_1 and t_2 (in °C) to define the temperature range ($t_1 : t_2$) for which the RNA structural perturbation should be calculated (the default range is 32–39°C). t_2 is the temperature for which the actual cluster identification will be performed. The minimal allowed temperature is 0°C, and the maximal allowed temperature is 99°C. Furthermore, the maximal allowed temperature difference (i.e. $t_2 - t_1 + 1$) is restricted to 20°C (e.g. 30–49°C).

If a user inputs two sequences, two options are available. The default option is to perform a statistical analysis of the sequence-specific clusters identified in each of the two sequences and to test whether these sequence-specific clusters result from the particular set of mutations distinguishing the sequences. However, this analysis takes some time and may be unnecessary for the particular user. In this case, the user can choose the checkbox for the 'Don't Create a Mutant Dataset' option (this option is only relevant for two input sequences).

An advanced user can deviate from this default setup and input his parameters of choice. It was described in the 'Method Summary' section that the statistical threshold for identifying significantly changing positions is 3.0 standard deviations. However, in a custom calculation, a user may also choose a threshold level other than 3.0. The next two advanced parameters, ε and MinPts, are both parameters for the clustering algorithm called DBSCAN. As the first step, the algorithm randomly selects one significantly changing position. MinPts specifies the minimal number of significantly changing positions in a cluster. ε specifies the distance from the chosen nucleotide. If the number of significant positions specified by MinPts is located within distance ε, the sequence segment is then considered part of a cluster. The default values for ε and MinPts are 11 and 5, respectively.

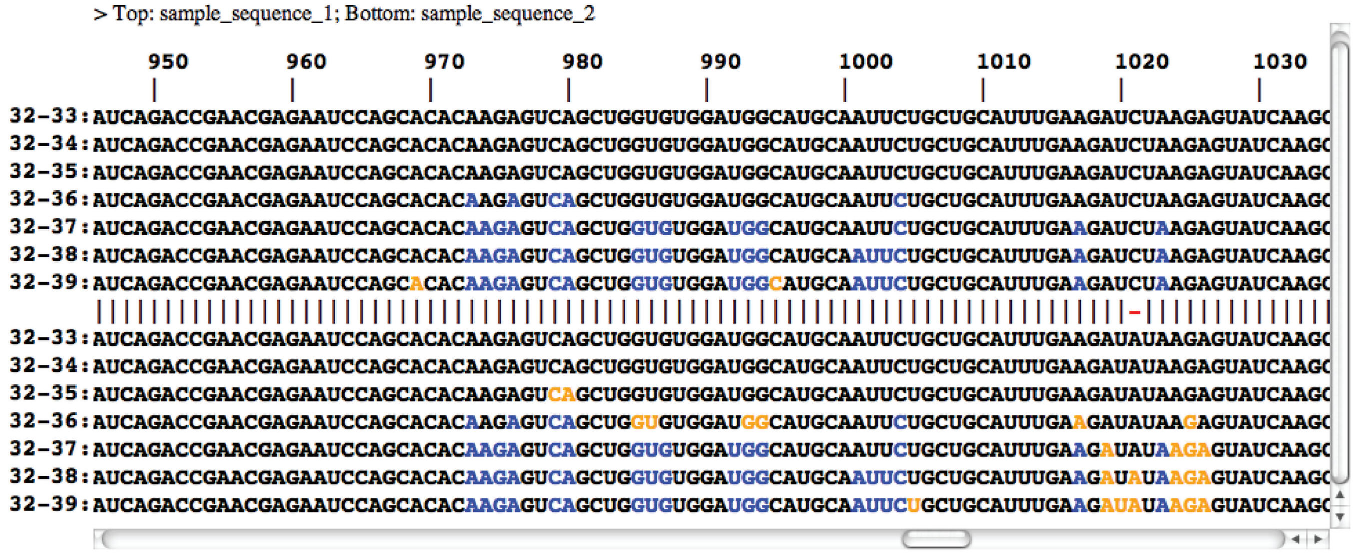
Finally, a user can then select checkbox options: 'Don't allow GU pairs at the end of helices' and/or 'Don't allow GU pairs'. These selections instruct the server whether GU pairs should be considered in calculations when the probabilities of nucleotides to be in a double-stranded confirmation at any given temperature are calculated. These two checkboxes are converted into the –noCloseGU and –noGU parameters of RNAfold during the calculations of probabilities of nucleotides to be coupled.

Server output

At the top of the results page from RNAtips, the HTML output provides colored visual representation of identified temperature-sensitive positions (Figure 1). The left column contains values for each temperature within the temperature range ($t_1 : t_2$). The right column presents the input sequence with those nucleotides—which are the most temperature-sensitive at this temperature—marked in either blue or orange color. The header line presents the FASTA header of the sequence(s). Position numbers are indicated under the header line. In the case of two input sequences, a line between the results for both sequences indicated matching positions with '|' and mismatches (mutations) with '-'. Additionally, significant positions that are sequence specific to one of the two sequences only are displayed in orange color. Positions that change their probabilities to be paired significantly in both sequences are displayed in blue color. If only one sequence is used as an input, then all positions demonstrating the highest potency to change their likelihood of forming W–C bonds are displayed in orange color. In addition, the HTML output demonstrates the temperature initiating a perturbation of the RNA structure.

All tables and figures presenting more detailed results are shown in the lower part of the page and described in the following paragraphs. The first table displays general information on identified significant positions and

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013



```
> Top: sample_sequence_1; Bottom: sample_sequence_2
         950       960       970       980       990      1000      1010      1020      1030
          |         |         |         |         |         |         |         |         |
32-33:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
32-34:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
32-35:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
32-36:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
32-37:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
32-38:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
32-39:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUCUAAGAGUAUCAAGC
      |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||-||||||||||||||
32-33:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
32-34:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
32-35:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
32-36:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
32-37:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
32-38:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
32-39:AUCAGACCGAACGAGAAUCCAGCACACAAGAGUCAGCUGGUGUGGAUGGCAUGCAAUUCUGCUGCAUUUGAAGAUAUAAGAGUAUCAAGC
```

Figure 1. Comparison of significantly changing positions between two RNAs. Top and bottom halves of the figure demonstrate influenza strains A/Leningrad/134/57 and its cold-adapted temperature-sensitive mutant A/Leningrad/134/47/57, respectively. Each row corresponds to a particular temperature difference. Positions in which base pairing probabilities significantly change with temperature elevation in both sequences and those where these changes only affect one of the sequences are marked blue and orange, respectively. Position numbers are indicated at the top of the alignment.

clusters. For every input sequence, it has the following fields: 'Sequence' (shows the ID of the input sequence); '#significant positions/total length' (the number of significantly changing positions and the total length of the input sequence); 'signif. pos. < 0/signif. pos. > 0' [the numbers of significantly changing positions that decrease (or increase) their probability to be paired with temperature elevation]; 'Number of clusters' (the total number of identified clusters of significantly changing positions); 'Avg. cluster density' (the average density of significantly changing positions in the identified clusters); and 'Avg. cluster length' (the average length of the identified clusters). The probability difference values are calculated by subtracting the value at the highest temperature from the value at the lowest temperature ($p_{39^\circ C}$ - $p_{32^\circ C}$ for the previous example). Cluster density is calculated as the number of significantly changing positions in a cluster divided by the total length of the cluster.

If a sequence contains 1000 nt and the temperature changes from 32 to 39°C, there are 7000 values reflecting how much each nucleotide would change its probability to form W–C couples when the temperature increases from 32 to 33°C, from 32 to 34°C, … , from 32 to 39°C. The output for this example would contain a histogram over all these 7000 data points (Figure 2). These histograms are used to identify the most temperature-sensitive positions (by default, further than 3 standard deviations away from the mean value). Overall, a histogram contains (t_2 - t_1)*(sequence length) values. For every input sequence, one histogram is presented at the output page. Thus, if two closely related sequences were used to compare their temperature sensitivity, the output would possess two histograms.

The exact location (start and end positions) of the identified clusters (if any) is shown in the following table.

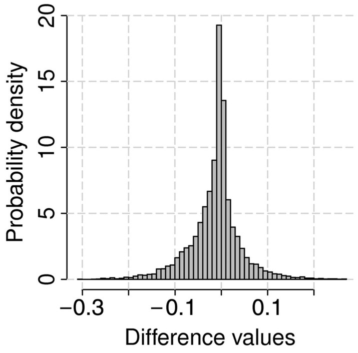


Figure 2. The histogram of differences in probability values of nucleotides to be in a double-stranded conformation for mRNA of nucleoprotein (NP) of influenza strain A/Leningrad/134/57 on temperature change between 32 and 39°C. The probability values of nucleotides to be paired for 32°C were subtracted from the probability values for every temperature from 33 to 39°C. All the differences were combined into one data set.

The accompanying output figure shows the relationship between the length and density of the clusters (Figure 3). In this figure, each point represents one cluster. The cluster density is plotted versus the cluster length. Several clusters can have the same properties, and in such a case, the corresponding points will overlap. Therefore, the total number of apparent points can be different from the total number of clusters. Such tables and figures are presented for every sequence in which clusters of the most temperature-sensitive positions were identified. Otherwise, the web server directly indicates that no clusters were identified for a particular sequence.

For every input sequence, the following figure demonstrates density of the most temperature-sensitive positions

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

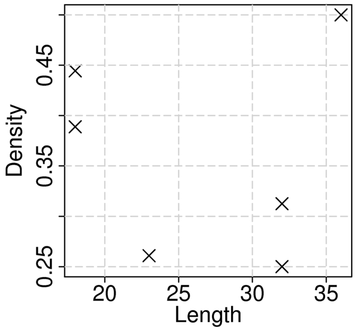


Figure 3. Density of significantly changing positions in determined clusters versus length of those clusters. Clusters were identified for mRNA of nucleoprotein (NP) of influenza strain A/Leningrad/134/57 by applying default parameters of the web server.

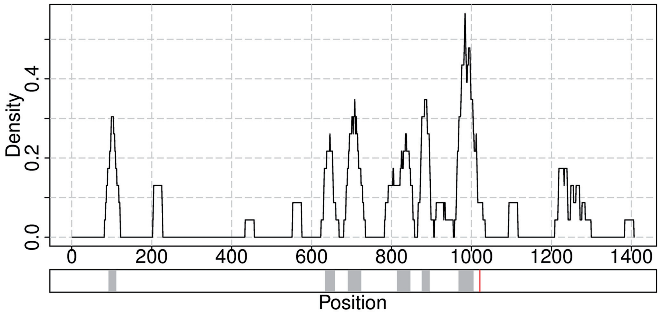


Figure 4. The upper panel demonstrates a density plot of significantly changing positions along the input sequence. A sliding window of size $2^*\varepsilon+1$ is moved in steps of 1 position over the sequence with the highest temperature difference t_2-t_1. The percentage of significantly changing positions in the window is calculated for each possible starting position. The bottom panel shows location of clusters of significantly changing positions identified by the DBSCAN algorithm depicted with gray color, and mutation is depicted with red vertical line. The depicted mutation corresponds to the nucleotide difference between influenza strains A/Leningrad/134/57 and its cold-adapted temperature-sensitive mutant A/Leningrad/134/47/57.

over the whole RNA sequence together with localization of clusters and localization of nucleotide substitutions (if any) (Figure 4). The upper part of this figure is created by moving a sliding window of size $2^*\varepsilon+1$ over the corresponding sequence and determining the density of significantly changing positions within it. The lower part shows the localization of clusters and mutation sites on the sequence.

As described earlier in the text, if two homologous RNA sequences constitute an input, one of the sequences may possess clusters of temperature-sensitive nucleotides, which are not present in the other RNA molecule (i.e. clusters that can be found for the given DBSCAN parameters in one RNA, and they do not overlap with any clusters from the other RNA). Appearance of these sequence-specific clusters may be a specific consequence of the particular nucleotide substitutions differentiating the RNAs. Alternatively, the clusters could result from a high number of non-specific mutations. Results of the statistical analysis presented in the last table (if conducted) demonstrate whether a sequence-specific temperature-sensitive cluster observed in one RNA but not in another is due to specific nucleotide substitutions taking place in the sequences. In other words, these data demonstrate whether such a specific difference between the two RNAs can be achieved by introducing the same number of random mutations. The server generates a data set of *in silico* mutants for the first RNA as described in the 'Method Summary' section. Some of these *in silico* mutants may possess temperature-sensitive clusters, which are not present in the original RNA sequence. The table shows positions of sequence-specific clusters observed in the RNA sequence, the frequency for each sequence-specific cluster to be overlapping with a cluster in the computer-generated mutants (at least, by one position), the *P*-value and 95% confidence interval calculated from the binomial test for each sequence-specific cluster to be a result of a random mutation set introduced into the original RNA.

All figures and additional information can be downloaded by RNAtips users. The results page enables a user to download a zip-file of all sequences of the *in silico* mutants (if generated). Results of every job will be stored on the server for at least 3 days. Every submitted job receives a unique URL and a user can browse the results during this period.

Implementation

RNAtips web server has a user-friendly interface and runs under the Linux operating system. The server's back-end, including the core part of computations as well as implementation of the DBSCAN algorithm, is written in Python. Statistical tests and generation of plots are implemented in R programming language. Calculation of probabilities of nucleotides to be paired in a double-stranded conformation is performed by using the RNAfold tool of the ViennaRNA package. The front-end part of the web server is implemented in HTML markup language with dynamic parts written in JavaScript programming language. A MySQL database is used to store the input parameters and results of the computations. The server contains a help page with detailed explanation of its functionality.

DISCUSSION

Before this presentation of RNAtips web server, researchers did not have a simple and feasible way to evaluate the affect of temperature change on secondary RNA structure. RNAtips is based on the analysis proposed and described by Chursov *et al.* (2). The name RNAtips stands for 'temperature-induced perturbation of structure'. This server can be used to analyze localization of temperature-induced changes in the secondary structures of RNA and to compare such changes between two sequences of the same length. There are at least three advantages of using RNAtips web server instead of simply calculating the probabilities of nucleotides to be paired at two different temperatures and then comparing those probabilities. First, RNAtips deciphers those

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

nucleotides within the RNA sequence, which change the most in their probability to form W–C bonds in response to a given temperature change. The web server demonstrates clusters of these positions within a sequence, which constitute the most temperature-sensitive structural regions. The second major benefit of RNAtips is the tool it provides to compare whether RNA structures of two closely related sequences would react (dis)similarly to a temperature change. If two RNA molecules possess different clusters of temperature-sensitive positions, their RNA structures react to the temperature change differently. Furthermore, if two RNA sequences are distinct in some nucleotide substitutions, RNAtips can be used to analyze whether either the difference in temperature sensitive clusters is specific to these particular nucleotide substitutions or whether it was likely to be caused by a similar number of non-specific nucleotide substitutions. Finally, the top RNAtips' results page is an HTML output that presents the temperature initiating a perturbation of secondary structure in a particular temperature-sensitive region. To the best of our knowledge, no other server provides these options. RNAtips web server can be applied to a broad spectrum of research topics such as drug development, molecular diagnostic and disease prognosis, evolutionary mechanisms, ecology, investigation of climate change effects and many more. In addition, currently, we are preparing a downloadable version of the source code for local usage.

FUNDING

DFG International Research Training Group 'Regulation and Evolution of Cellular Systems' [GRK 1563]; Russian Foundation for Basic Research [RFBR 09-04-92742]. Funding for open access charge: DFG International Research Training Group 'Regulation and Evolution of Cellular Systems' [GRK 1563].

Conflict of interest statement. None declared.

REFERENCES

1. Ringnér,M. and Krogh,M. (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, e72.
2. Chursov,A., Kopetzky,S.J., Leshchiner,I., Kondofersky,I., Theis,F.J., Frishman,D. and Shneider,A. (2012) Specific temperature-induced perturbations of secondary mRNA structures are associated with the cold-adapted temperature-sensitive phenotype of influenza A virus. *RNA Biol.*, **9**, 1266–1274.
3. Shamovsky,I., Ivannikov,M., Kandel,E.S., Gershon,D. and Nudler,E. (2006) RNA-mediated response to heat shock in mammalian cells. *Nature*, **440**, 556–560.
4. Shamovsky,I. and Nudler,E. (2008) New insights into the mechanism of heat shock response activation. *Cell. Mol. Life Sci.*, **65**, 855–861.
5. Chowdhury,S., Maris,C., Allain,F.H. and Narberhaus,F. (2006) Molecular basis for temperature sensing by an RNA thermometer. *EMBO J.*, **25**, 2487–2497.
6. Storz,G. (1999) An RNA thermometer. *Genes Dev.*, **13**, 633–636.
7. Narberhaus,F., Waldminghaus,T. and Chowdhury,S. (2006) RNA thermometers. *FEMS Microbiol. Rev.*, **30**, 3–16.
8. Chowdhury,S., Ragaz,C., Kreuger,E. and Narberhaus,F. (2003) Temperature-controlled structural alterations of an RNA thermometer. *J. Biol. Chem.*, **278**, 47915–47921.
9. Mergny,J.L. and Lacroix,L. (2003) Analysis of thermal melting curves. *Oligonucleotides*, **13**, 515–537.
10. Marmur,J. and Doty,P. (1959) Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids. *Nature*, **183**, 1427–1429.
11. Marmur,J. and Doty,P. (1962) Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *J. Mol. Biol.*, **5**, 109–118.
12. Reed,G.H., Kent,J.O. and Wittwer,C.T. (2007) High-resolution DNA melting analysis for simple and efficient molecular diagnostics. *Pharmacogenomics*, **8**, 597–608.
13. Baaske,P., Duhr,S. and Braun,D. (2007) Melting curve analysis in a snapshot. *Appl. Phys. Lett.*, **91**, 133901.
14. Thatcher,D.R. and Hodson,B. (1981) Denaturation of proteins and nucleic acids by thermal-gradient electrophoresis. *Biochem. J.*, **197**, 105–109.
15. Rosenbaum,V. and Riesner,D. (1987) Temperature-gradient gel electrophoresis. Thermodynamic analysis of nucleic acids and proteins in purified form and in cellular extracts. *Biophys. Chem.*, **26**, 235–246.
16. Wienken,C.J., Baaske,P., Duhr,S. and Braun,D. (2011) Thermophoretic melting curves quantify the conformation and stability of RNA and DNA. *Nucleic Acids Res.*, **39**, e52.
17. Wan,Y., Qu,K., Ouyang,Z., Kertesz,M., Li,J., Tibshirani,R., Makino,D.L., Nutter,R.C., Segal,E. and Chang,H.Y. (2012) Genome-wide measurement of RNA folding energies. *Mol. Cell*, **48**, 169–181.
18. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
19. Ester,M., Kriegel,H.-P., Sander,J. and Xu,X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data MIning*. Association for the Advancement of Artificial Intelligence, Portland, Oregon, pp. 226–231.
20. Kriegel,H.-P., Kröger,P., Sander,J. and Zimek,A. (2011) Density-based clustering. *Wiley Interdiscip. Rev. Data. Min. Knowl. Discov.*, **1**, 231–240.

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

2.4 Conservation of mRNA Secondary Structures May Filter Out Mutations in *Escherichia coli* Evolution

Andrey Chursov, Dmitrij Frishman and Alexander Shneider
Nucleic Acids Res., 41(16):7854-7860, 2013

In this paper, we stipulate and address the hypothesis stating that conservation of mRNA structures and mRNA minimum folding energy serves as a previously unknown factor of bacterial evolution. Our main finding is that purifying selection tends to eliminate those mutations in essential genes that lead to greater changes of MFE values and, therefore, may be more disruptive for the corresponding mRNA secondary structures. This effect implies that mutations disrupting mRNA secondary structures may directly affect the fitness of the organism. Thus, our results support the hypothesis of the paper and imply conservation of mRNA structures as a previously unknown factor of bacterial evolution.

The present research was devoted to testing one single hypothesis: “Does disruption of mRNA structure in bacteria serve as a functional gene knockout?” This question is a direct continuation of a previous work by Ilyinski et al. (Ilyinskii et al., 2009) who have demonstrated that perturbation of mRNA structure serves as a functional gene knockout in a non-bacterial system. Thus, we have hypothesized that the same effect may take place in bacteria as well. If this hypothesis is correct, then mutations that are disruptive of mRNA structures would be “filtered out” only if they took place in essential genes. By definition, knockout of an essential gene renders bacteria nonviable. Alternatively, if a mutation perturbs mRNA structure of a nonessential gene making the gene non-functional, this mutation would not be eliminated. Following this logic and conducting a hypothesis-driven research, we have focused on essential vs. nonessential genes only. Our enthusiasm in testing this hypothesis was further encouraged by discussions with Richard Lenski and other bacteriologists considering this particular question of high biological relevance.

Additionally, it is important to note that we originally selected the essentiality data produced by Gerdes et al. (Gerdes et al., 2003) instead of the Keio dataset (Baba et al., 2006; Yamamoto et al., 2009) based on a biological reason. These two datasets utilize alternative definitions of essentiality both of which make biological sense. To produce the Keio collection, Mori's group have knocked out genes one by one and observed if the particular clone can grow by itself without assistance and/or interaction with other clones. In contrast, Gerdes and colleagues have analyzed the ability to grow bacteria lacking a gene, while the bacteria was surrounded with other bacterial cells possessing this particular gene. Obviously, the system used by Gerdes et al. is more similar to the situation of Lenski's experiment. Additionally, it is known, that some bacteria lacking a gene could not grow by itself, but can maintain growth in the presence of a helper strain, for example the report by D'Onofrio et al. (D'Onofrio et al., 2010). Thus, utilizing the Keio collection as a proxy for Lenski's experimental design, we would be at a higher risk to identify some genes as false-positive and/or false-negative than with Gerdes dataset.

The research was designed by Alexander Shneider and me. The research was performed by myself. All the authors participated in analyzing the resulting data and in writing the paper.

Nucleic Acids Research Advance Access published June 19, 2013

Nucleic Acids Research, 2013, 1–7
doi:10.1093/nar/gkt507

Conservation of mRNA secondary structures may filter out mutations in *Escherichia coli* evolution

Andrey Chursov[1], Dmitrij Frishman[1,2,*] and Alexander Shneider[3,*]

[1]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, Maximus-von-Imhof-Forum 3, D-85354, Freising, Germany, [2]Helmholtz Center Munich—German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany and [3]Cure Lab, Inc., 43 Rybury Hillway, Needham, MA 02492, USA

Received January 11, 2013; Revised April 29, 2013; Accepted May 1, 2013

ABSTRACT

Recent reports indicate that mutations in viral genomes tend to preserve RNA secondary structure, and those mutations that disrupt secondary structural elements may reduce gene expression levels, thereby serving as a functional knockout. In this article, we explore the conservation of secondary structures of mRNA coding regions, a previously unknown factor in bacterial evolution, by comparing the structural consequences of mutations in essential and nonessential *Escherichia coli* genes accumulated over 40 000 generations in the course of the 'long-term evolution experiment'. We monitored the extent to which mutations influence minimum free energy (MFE) values, assuming that a substantial change in MFE is indicative of structural perturbation. Our principal finding is that purifying selection tends to eliminate those mutations in essential genes that lead to greater changes of MFE values and, therefore, may be more disruptive for the corresponding mRNA secondary structures. This effect implies that synonymous mutations disrupting mRNA secondary structures may directly affect the fitness of the organism. These results demonstrate that the need to maintain intact mRNA structures imposes additional evolutionary constraints on bacterial genomes, which go beyond preservation of structure and function of the encoded proteins.

INTRODUCTION

Increasing experimental (1) and computational (2,3) evidence points to the existence of extensive RNA structures in the coding regions of mRNA molecules. RNA secondary structures have been implicated in regulation of translation initiation, elongation and termination in both prokaryotes and eukaryotes (4,5). In particular, the anti-correlation between translation efficiency and the thermodynamic stability of local secondary structure in the vicinity of the translation initiation site has been thoroughly documented (6). RNA hairpins are thought to be involved in controlling mRNA decay (7), localization (8–10) and interaction with other molecules (11). Overall, the mRNA coding regions appear to be more structured than the untranslated regions (1) and have lower minimum folding free energies. Hence, the mRNA coding regions appear to have more stable structures than codon-randomized sequences (12). Owing to the need to simultaneously preserve both the function and structure of the encoded protein, as well as the structural elements of the RNA molecule itself, mRNA coding regions are subject to dual selection pressure.

Using a mammalian system, we have recently shown that mutations altering secondary structures of influenza mRNAs may serve as a functional knockout of the corresponding genes (13). More recently, Moss *et al.* (14) established a direct connection between mutation patterns in the influenza virus genome and the hydrogen-bonding patterns shaping RNA structures. Thus, preservation of viral RNA structures and elimination of mutations disruptive for RNA structures may be a previously unknown mechanism of viral evolution. In the present article, we put forward the hypothesis that conservation of RNA structures may also play a role in bacterial evolution. To examine this hypothesis, we compared the genomes of parental and progeny *Escherichia coli* clones standing 40 000 generations apart. The 'long-term evolution experiment' (15–18) tracking genetic changes in 12 populations of *E. coli* was started by Richard Lenski in February 1988. All 12 replicate populations have originated from a single cell of the baseline strain, which was an *E. coli* B clone, and have been propagated at 37°C in liquid culture. Every 500 generations, samples for each population were frozen

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

---

*To whom correspondence should be addressed. Tel: +49 179 538 2799; Fax: +49 8161 712 186; Email: d.frishman@wzw.tum.de
Correspondence may also be addressed to Alexander Shneider. Tel: +1 609 841 1201; Fax: +1 260 572 3405; Email: ashneider@curelab.com

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

away at −80°C and retained for sequencing and comparison with their predecessors.

If our hypothesis is correct, mutations in essential genes that disrupt mRNA secondary structures would lead to insufficient gene expression and, due to the essentiality of those genes, such mutations would be filtered out as lethal. By contrast, selection against mutations disrupting mRNA secondary structures of nonessential genes would be expected to be less pronounced because an altered expression level of nonessential genes would not influence bacterial propagation. Supplementary Figure S1 exemplifies predicted structural perturbations induced by mutations altering minimal free energy of an *E. coli* mRNA.

To demonstrate this effect, one would ideally need to calculate exact secondary structures for both the original and the mutated mRNAs, compare them and make an inference about the changes in the RNA structure caused by mutations. However, a single RNA molecule may fold into more than one conformation (19,20). With increasing sequence length, the number of possible structures that an RNA molecule can adopt with similar (in many cases even the same) values of folding energy increases as well (21), thereby resulting in diminished prediction accuracy. Another well-known complication is that predicting secondary structures with pseudoknots is an NP-hard problem (22), which necessitates using approximations in structure prediction algorithms. Therefore, instead of calculating explicit secondary structure shapes for mRNAs, we pursued an indirect method of assessing whether mutation(s) affect secondary structures by quantifying minimum free energy (MFE) change. While different RNA structures may have exactly the same MFE, different MFE values are guaranteed to correspond to different structures. Despite the fact that a mutation did not change MFE does not mean that the RNA structure remained the same, yet, an opposite situation is reliably conclusive. The more mutations change the MFE, the greater affect on a secondary RNA structure they have.

Using this approach, we investigated how mutations observed in essential and nonessential genes influence the MFE values of mRNA structures. This article presents evidence that mutations in essential genes of *E. coli* that occurred during the 'long-term evolution experiment' changed the MFE of mRNA secondary structures to a lesser extent than mutations in nonessential genes. We emphasize that we focus exclusively on the conservation of secondary structures of mRNA coding regions and do not consider noncoding RNAs. This finding supports our hypothesis that mutations disrupting the mRNA structure of essential genes are filtered out during the course of bacterial evolution.

MATERIALS AND METHODS

Experimental data on evolutionary mutations in *E. coli*

In our analysis, we used data on genetic polymorphisms in *E. coli* accumulated in the course of the 'long-term evolution experiment' (16–18). Specifically, mutations in the 40 000th generation of one of the populations (Ara-1), with the ancestral strain REL606 (GenBank accession number NC_012967.1), were investigated. In this 40 000th clone, 627 single-nucleotide polymorphisms (SNPs) and 26 deletions, insertions and other polymorphisms were detected. Hereinafter, we take into account only SNPs. Ninety-two mutations occurring in intergenic regions as well as six mutations in pseudogenes and one mutation in an insertion sequence element were excluded from consideration. We also ignored one SNP owing to an inconsistency between the mutated nucleic acid, as reported in (18) and the nucleic acid occurring at this position in the complete genome sequence. Two genes with available SNP data were not considered: one with an inconsistency between its nucleotide and amino acid sequences, and another that had one of the reported mutations in its start codon. Our final data set contained 523 mutations involving 485 genes.

Data on essential and nonessential genes of *E. coli*

There is no essentiality data available for the B strain of *E. coli*, but it is closely related to the well-studied *E. coli* K-12 MG1655 (23,24). For this latter strain (GenBank accession number U00096.2) Gerdes *et al.* (25) experimentally identified 620 genes as essential and 3126 genes as dispensable using a genetic footprinting technique. Because of the numerous discrepancies between the gene names, we conducted similarity-based transfer of essentiality data from the MG1655 strain to the REL606 strain, using the bidirectional best hit strategy to identify orthologous genes. Using blastp (26), we aligned all mutated genes from the REL606 genome against all genes from the MG1655 genome and *vice versa*. Genes from the two genomes were considered orthologous if they were the best hits for each other, with amino acid sequence identity >75% and e-value $<10^{-25}$. This procedure enabled us to map 456 out of the 485 mutated genes in REL606 to the MG1655 strain, of which 48 were essential, 348 dispensable and 60 had undefined essentiality according to the MG1655 annotation.

MFE values of RNA secondary structures

For each of the 48 essential and 348 nonessential mutated genes, we calculated MFE values of secondary RNA structures for both the original ancestral mRNAs and their 40 000th generation counterparts. We used the RNAfold tool from the Vienna RNA Package with the command line option noLP, which disallowed base pairs that can only occur as helices of length 1 (27).

Generation of randomly mutated mRNAs

For each gene reported in (18) as possessing mutation(s) in the 40 000th generation, we produced an *in silico* family of random counterparts. Synonymous random mutations were introduced into ancestral mRNAs. When compared with the ancestral strain, each computer-generated RNA sequence had the same number of point mutations as the respective 40 000th generation mutant. There are six types of possible nucleotide substitutions: C:G → A:U; A:U → C:G; A:U → U:A; C:G → G:C; C:G → U:A; A:U → G:C. We introduced random mutations in such a

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

way that the frequency for any given nucleotide substitution type was similar for essential and nonessential gene groups (Supplementary Table S1) and free of transition to transversion bias. For each gene, the ratio of transitions to transversions was calculated, and distributions of these ratios were compared for essential and nonessential genes. According to the Mann–Whitney U test, these distributions do not differ ($P = 0.38$). For the purpose of this work, we did not have to simulate *in silico* relative frequencies of nucleotide substitution observed reported by Wielgoss *et al.* (28). The MFE of secondary RNA structure was calculated for each computer-generated sequence by the RNAfold tool as described above.

The number of computer-generated sequences varied from gene to gene dependent on gene length (Figure 1). If a short gene possesses only one nucleotide substitution *in vitro*, the number of conceivable *in silico* generated sequences having only one nucleotide changed is limited to an exhaustive set of synonymous point mutations (e.g. 516 variants for the gene *yciT* of length 750). For sufficiently long genes (typically >1300 bases), the subset of 1000 sequences with randomly introduced SNPs was used for further analysis.

Statistical test

To find out whether *in vitro* mutations in essential and nonessential genes differ in their affect on MFE and mRNA secondary structures, we applied the following analysis. First, for each gene, we determined the absolute value of the difference between the MFE of the ancestor RNA and the MFE of the *in vitro* mutant, as well as that of each of the computer-generated mutants. Then, we calculated the fraction of computer-generated mutants whose absolute values of MFE differences were lower than the corresponding *in vitro* mutant. Each gene in the data set of essential genes and in the data set of dispensable genes was thus characterized by a percentile value. The Mann–Whitney U test was then used with the null hypothesis (H_0) that the percentile values for essential and nonessential genes are from the same distribution.

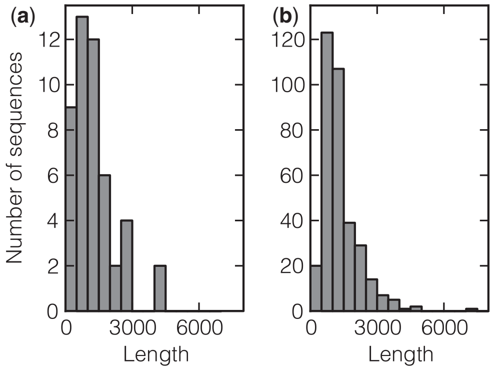


Figure 1. Histograms of gene lengths for essential (**a**) and nonessential (**b**) genes.

Data availability

The lists of defined essential and nonessential genes with the corresponding MFE values are presented in the Supplementary Tables S2 and S3, respectively.

RESULTS

The main scientific questions we addressed in this study are whether purifying selection tends to eliminate mutations that are disruptive for mRNA structures, and whether this effect is more pronounced in essential genes compared with dispensable ones. Our methodology involved a comparison of actual mutations observed in the 40 000th generation of the 'long-term evolution experiment' (18) with a pool of random computer-generated mutations.

As an example, for a gene harboring two point mutations, we generated a thousand *in silico* mutants with two mutations each. MFE was calculated for the ancestor mRNA as well as for the mRNA of the 40 000th generation mutant experimentally observed in a Petri dish and for those of the *in silico* mutants. Owing to slight sequence changes, the RNA folding energies of both experimentally recorded and computer-generated mutants will be somewhat different from the MFE of the ancestor's mRNA. We calculated the fraction of *in silico* mutants with a lesser extent of MFE change than the mutant observed *in vitro*. Suppose, for example, that the MFE of the ancestor mRNA was −5 kcal/mol and that the MFE of the *in vitro* mutant differs from it by 2 kcal/mol (it does not matter whether the MFE went down to −7 or went up to −3 kcal/mol). If 700 out of 1000 computer-generated mutants have their MFEs either >−3 or <−7 kcal/mol, it means that for this particular gene a mutant recorded in the *in vitro* experiment changes its MFE to a greater extent than 30% of the randomly mutated sequences.

Suppose that experimentally observed mutations in essential genes lead to bigger MFE changes than only 10% of random mutations, while in the data set of mutations in nonessential genes, MFE changes bigger than those of random mutants are observed in 50% of the cases. This would indicate that the evolutionary constraints acting on mRNA structure in essential genes are stronger that those acting on dispensable genes.

In the *E. coli* genome sampled at 40 000th generation, 523 nucleotide substitutions occurred in 485 genes (11.5% of all *E. coli* genes), of which 48 genes were essential, 348 nonessential and 89 genes either could not be successfully mapped from the REL606 to the MG1655 strain or had unknown essentiality status (Table 1). A great majority of the mutated genes (92.2%) have only one SNP mutation. For the mutated genes, the ratio between the number of essential and nonessential genes is 0.138; while for nonmutated genes, this ratio is 0.206. The latter finding is in agreement with the report by Jordan *et al.* (29) showing that essential genes in bacteria accumulate mutations less frequently than nonessential genes do. The majority of mutations are nonsynonymous (Table 2), with the ratio of synonymous to nonsynonymous

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

Table 1. The number of all, essential and nonessential genes in which a particular number of SNPs occurred during the 'long-term evolution experiment' between the first and the 40 000th generations

Genes type	Number of mutations per gene			Total
	1	2	3	
All genes[a]	447	36	2	485
Essential genes	43	5	0	48
Nonessential genes	318	29	1	348

[a]Including those with unknown essentiality status.

Table 2. The divisions of synonymous and nonsynonymous mutations

Mutation type	All genes	Essential genes	Nonessential genes
Synonymous	83	4	70
Nonsynonymous	442	49	309

mutations in essential genes (0.082) being somewhat lower than in nonessential genes (0.227). The *P*-value calculated using a binomial test (4 synonymous SNPs out of 53 in essential genes *vs* 70 synonymous SNPs out of 379 in nonessential genes) equals 0.048.

SNPs cause changes in the MFE and structural perturbations in many, though not all, mRNAs (Table 3). Specifically, 27.1% of essential genes do not change the MFE value, while only 14.7% of nonessential genes demonstrate the same MFE values for both original and mutated mRNAs. In general, SNPs in essential genes change the MFE values (median = 0.69 kcal/mol) to a smaller extent than do mutations in nonessential genes (median = 1.10 kcal/mol) (Figure 2). We compared the properties of essential and nonessential genes that could influence MFE calculations, but found no confounding factors (data not shown). Both groups of genes have the same average GC content. While essential genes tend to be somewhat shorter than nonessential ones (Figure 1), neither in essential nor in nonessential genes do the differences in MFE between the native and mutated sequences depend on mRNA length. Additionally, different types of mutations (e.g. $C \rightarrow G$) occur in these two data sets equally often. At the same time, mutations observed at the 40 000th generation *in vitro* are more likely to reduce MFE of nonessential genes than the essential ones. MFE value decreased in 56.0% of the nonessential mutants, while only 45.8% of the essential ones demonstrated MFE reduction. A possible interpretation could be that ancestral essential genes were folded in structures that caused the values of their folding energies to be close to the minimum (robust); in contrast, nonessential genes had MFEs more distant from the minimal values. Thus, mutations were less likely to reduce energies of essential genes.

We subsequently compared the absolute values of MFE changes caused in each mRNA by naturally occurring and an equal number of randomly introduced synonymous mutations, thus avoiding those mutations in the sequences, generated *in silico*, that could be eliminated by purifying selection due to their effect on the encoded protein. For each mRNA, we determined the fraction of *in silico* derivatives, which change their MFEs less than the mutant observed *in vitro*. These percentages are much lower in essential *E. coli* genes than in nonessential genes (Table 4), implying that mutations accumulated in essential *E. coli* genes affect MFEs (and hence secondary structure) to a lesser extent than mutations in nonessential genes. This effect is further demonstrated by the fact that the cumulative distribution function corresponding to essential genes elevates considerably faster at the beginning (Figure 3). The difference between values for essential and nonessential genes is statistically significant according to the Mann–Whitney U test ($P = 0.044$). Our results suggest that mRNA secondary structure imposes substantially smaller selective pressure at the mutations taking place in nonessential genes because the median of their effect on MFE is 50.2% of what the random set of mutations would cause. By contrast, the median value of how mutations occurring in essential genes influence MFEs is only 32.6% of what random mutations would do.

Table 3. The number of essential and nonessential genes that decrease, increase or do not change their MFE value on mutation

Gene type	$MFE_{\text{mutant}} - MFE_{\text{original}}$			Total
	<0	=0	>0	
Essential	22 (45.8%)	13 (27.1%)	13 (27.1%)	48
Nonessential	195 (56.0%)	51 (14.7%)	102 (29.3%)	348

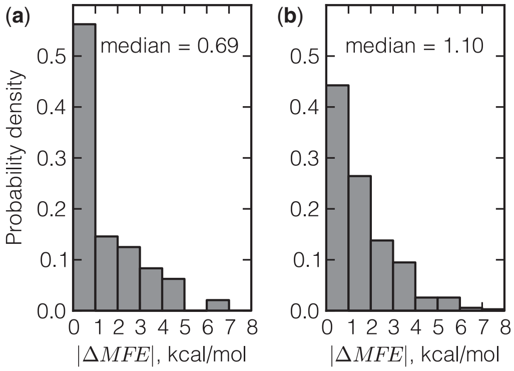


Figure 2. Histograms of absolute changes in MFE values for essential (**a**) and nonessential (**b**) genes.

DISCUSSION

The importance of mRNA secondary structure for gene expression was demonstrated for many organisms including bacteria (30–33), human (34,35) and *Drosophila* (36). These studies showed that synonymous SNPs altering mRNA folding may result in decreased mRNA stability and may also change expression efficiency. In a recent *in vitro* study, we introduced mutations

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

Table 4. Summary of MFE changes in mRNA secondary structures of essential and nonessential genes

Gene type	Number of genes	Lower quartile %	Median %	Upper quartile %
Essential	48	9.4	32.6	71.1
Nonessential	348	25.5	50.2	75.0

Lower quartile, median and upper quartile values are presented for the distributions of percentages of computer-generated mutants with randomly introduced mutations that change the MFE less than the naturally occurring mutations.

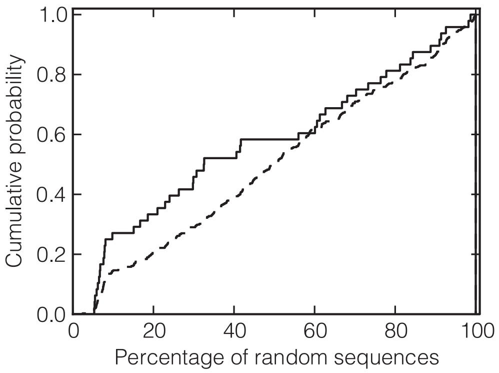


Figure 3. Cumulative distribution functions of the percentages of randomly introduced *in silico* mutations that change the MFE values less than the mutations occurring *in vitro* for essential genes (solid line) and nonessential genes (dashed line). Each curve gives the probability that the MFE change in a particular gene due to an actual mutation will be higher than MFE changes observed in a given percentage of genes with randomly introduced mutations.

into an influenza gene, particularly into a region of the gene encoding for a functionally important protein domain (13). As a result of the perturbations in the RNA structure caused by these mutations, gene expression was significantly reduced. Mutations altering RNA structures thus had a functional knockout effect. We also demonstrated that mutations disruptive to RNA structure may impair transcription without facilitating mRNA degradation. Thus, a new mechanism of viral evolution was proposed (13). We hypothesized that mutations disruptive to RNA structures would likely be eliminated to preserve the gene regions encoding for functionally important sites of viral proteins. Following this line of thought, the goal of the present study was to examine whether preservation of mRNA structures is implicated in the evolution of bacteria.

An important evolutionary characteristic of bacterial genes is their essentiality for organism survival, which can be experimentally assessed based on absence of growth on knockout. We hypothesized that if some of the mutations causing perturbations in mRNA structures also result in reduction in expression levels of bacterial genes, these mutations are more likely to be eliminated by purifying selection if they take place in essential rather than nonessential genes. Indeed, we found that mRNA secondary structures of essential genes are more conserved than those of nonessential genes in bacteria. Previous work revealed that essential bacterial genes are more evolutionarily conserved than nonessential ones (29,37). It was shown that in the *E. coli* genome paired DNA bases have lower propensities to mutate than unpaired bases (38,39). Based on the comparison of the *E. coli* and *Salmonella typhi* genomes, it was concluded that homologous RNAs of polycistronic genes in both organisms have significantly higher folding potential than randomized sequences, which is a sign that natural selection is acting to preserve RNA secondary structure in the coding regions of polycistronic genes (7). However, to the best of our knowledge, preservation of intact mRNA structures of individual genes has not yet been assessed as a potential constraint on the evolution of bacterial genes.

As the best available proxy for *E. coli* B essential genes we used experimentally determined the essentiality status of genes in the closely related *E. coli* K-12 genome (25). Such homology mapping may not always be accurate, even between similar organisms, owing to possible differences in gene regulation, posttranslational modifications and other cellular processes. An additional factor potentially masking the true magnitude of the effect is that we used changes in MFE as an indirect indication that the secondary structure of the mRNA has changed. However, changes in RNA sequence and the resulting perturbations of its structure may in fact take place without causing MFE changes (Table 3). Owing to these obvious limitations, we believe that our results represent a conservative estimate of the role played by mRNA structure in constraining mutations.

Our results point to the preservation of coding mRNA structures as a previously unappreciated factor influencing bacterial evolution. Until now, selective pressure in coding regions was thought to primarily act against mutations that either impair protein function and stability or affect robustness against mistranslations (40). In particular, selection against mistranslation-induced protein misfolding is currently considered to be the major factor determining the strong dependence of protein evolutionary rate on the level of expression (41). The bulk of this research has thus been devoted to the 'protein half of the equation'—translation, folding and function. In the past few years, attention is being increasingly focused on the noncoding selective pressure in coding regions, which is manifested by the presence of synonymous constraint (42–44). Such noncoding selective pressure may be caused, on one hand, by the presence of various functional elements, such as microRNA binding sites, transcription factor binding sites and splicing enhancers in eukaryotic mRNAs, and, on the other hand, by the formation of RNA structural elements playing a role in mRNA localization, degradation and interactions with other molecules. This article presents the first statistical evidence linking mRNA folding to bacterial evolution. Our principal finding is that purifying selection tends to eliminate those mutations in essential genes that lead to greater changes of MFE values and, therefore, may be more disruptive for the corresponding mRNA secondary structures. This effect is implying that synonymous mutations disrupting mRNA secondary structures may directly affect the fitness of the organism.

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Richard Lenski for productive and stimulating discussions, Dr Olivier Tenaillon for illuminating comments on the article and Prof. Andrei Osterman for helpful comments on the essentiality data. The authors gratefully acknowledge the support of the TUM Graduate School's Thematic Graduate Center Regulation and Evolution of Cellular Systems (RECESS) at the Technische Universität München.

FUNDING

DFG International Research Training Group 'Regulation and Evolution of Cellular Systems' [GRK 1563]; Russian Foundation for Basic Research [RFBR 09-04-92742]. Funding for open access charge: DFG International Research Training Group 'Regulation and Evolution of Cellular Systems' [GRK 1563].

Conflict of interest statement. None declared.

REFERENCES

1. Wan,Y., Kertesz,M., Spitale,R.C., Segal,E. and Chang,H.Y. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, **12**, 641–655.
2. Meyer,I.M. and Miklós,I. (2005) Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.*, **33**, 6338–6348.
3. Findeiß,S., Engelhardt,J., Prohaska,S.J. and Stadler,P.F. (2011) Protein-coding structured RNAs: a computational survey of conserved RNA secondary structures overlapping coding regions in drosophilids. *Biochimie*, **93**, 2019–2023.
4. Gray,N.K. and Hentze,M.W. (1994) Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.*, **19**, 195–200.
5. Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.
6. Gu,W., Zhou,T. and Wilke,C.O. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **6**, e1000664.
7. Katz,L. and Burge,C.B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.*, **13**, 2042–2051.
8. Gonzalez,I., Buonomo,S.B., Nasmyth,K. and von Ahsen,U. (1999) ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation. *Curr. Biol.*, **9**, 337–340.
9. Chartrand,P., Meng,X.H., Singer,R.H. and Long,R.M. (1999) Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle *in vivo*. *Curr. Biol.*, **9**, 333–336.
10. Olivier,C., Poirier,G., Gendron,P., Boisgontier,A., Major,F. and Chartrand,P. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell Biol.*, **25**, 4752–4766.
11. Ankö,M.L. and Neugebauer,K.M. (2012) RNA-protein interactions *in vivo*: global gets specific. *Trends Biochem. Sci.*, **37**, 255–262.
12. Seffens,W. and Digby,D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, **27**, 1578–1584.
13. Ilyinskii,P.O., Schmidt,T., Lukashev,D., Meriin,A.B., Thoidis,G., Frishman,D. and Shneider,A.M. (2009) Importance of mRNA secondary structural elements for the expression of influenza virus genes. *Omics*, **13**, 421–430.
14. Moss,W.N., Priore,S.F. and Turner,D.H. (2011) Identification of potential conserved RNA secondary structure throughout influenza A coding regions. *RNA*, **17**, 991–1011.
15. Lenski,R.E., Rose,M.R., Simpson,S.C. and Tadler,S.C. (1991) long-term experimental evolution in *Escherichia coli* .I. Adaptation and divergence during 2,000 generations. *Am. Nat.*, **138**, 1315–1341.
16. Cooper,T.F., Rozen,D.E. and Lenski,R.E. (2003) Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **100**, 1072–1077.
17. Blount,Z.D., Borland,C.Z. and Lenski,R.E. (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **105**, 7899–7906.
18. Barrick,J.E., Yu,D.S., Yoon,S.H., Jeong,H., Oh,T.K., Schneider,D., Lenski,R.E. and Kim,J.F. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, **461**, 1243–1247.
19. Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
20. Höbartner,C. and Micura,R. (2003) Bistable secondary structures of small RNAs and their structural probing by comparative imino proton NMR spectroscopy. *J. Mol. Biol.*, **325**, 421–431.
21. Tinoco,I. Jr and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
22. Lyngsø,R.B. (2004) Complexity of pseudoknot prediction in simple models. *Lect. Notes Comput. Sc.*, **3142**, 919–931.
23. Jeong,H., Barbe,V., Lee,C.H., Vallenet,D., Yu,D.S., Choi,S.H., Couloux,A., Lee,S.W., Yoon,S.H., Cattolico,L. *et al.* (2009) Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.*, **394**, 644–652.
24. Studier,F.W., Daegelen,P., Lenski,R.E., Maslov,S. and Kim,J.F. (2009) Understanding the differences between genome sequences of *Escherichia coli* B Strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. *J. Mol. Biol.*, **394**, 653–680.
25. Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balázsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S. *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, **185**, 5673–5684.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
27. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
28. Wielgoss,S., Barrick,J.E., Tenaillon,O., Wiser,M.J., Dittmar,W.J., Cruveiller,S., Chane-Woon-Ming,B., Médigue,C., Lenski,R.E. and Schneider,D. (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl Acad. Sci. USA*, **110**, 222–227.
29. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.
30. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–258.
31. Hall,M.N., Gabay,J., Débarbouillé,M. and Schwartz,M. (1982) A role for mRNA secondary structure in the control of translation initiation. *Nature*, **295**, 616–618.
32. Qing,G., Xia,B. and Inouye,M. (2003) Enhancement of translation initiation by A/T-rich sequences downstream of the initiation codon in *Escherichia coli*. *J. Mol. Microbiol. Biotechnol.*, **6**, 133–144.

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

33. Griswold,K.E., Mahmood,N.A., Iverson,B.L. and Georgiou,G. (2003) Effects of codon usage versus putative 5′-mRNA structure on the expression of *Fusarium solani* cutinase in the *Escherichia coli* cytoplasm. *Protein Expr. Purif.*, **27**, 134–142.
34. Nackley,A.G., Shabalina,S.A., Tchivileva,I.E., Satterfield,K., Korchynskyi,O., Makarov,S.S., Maixner,W. and Diatchenko,L. (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, **314**, 1930–1933.
35. Duan,J., Wainwright,M.S., Comeron,J.M., Saitou,N., Sanders,A.R., Gelernter,J. and Gejman,P.V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.*, **12**, 205–216.
36. Carlini,D.B., Chen,Y. and Stephan,W. (2001) The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics*, **159**, 623–633.
37. Wilson,A.C., Carlson,S.S. and White,T.J. (1977) Biochemical evolution. *Annu. Rev. Biochem.*, **46**, 573–639.
38. Wright,B.E., Reschke,D.K., Schmidt,K.H., Reimers,J.M. and Knight,W. (2003) Predicting mutation frequencies in stem-loop structures of derepressed genes: implications for evolution. *Mol. Microbiol.*, **48**, 429–441.
39. Hoede,C., Denamur,E. and Tenaillon,O. (2006) Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. *PLoS Genet.*, **2**, e176.
40. Pál,C., Papp,B. and Lercher,M.J. (2006) An integrated view of protein evolution. *Nat. Rev. Genet.*, **7**, 337–348.
41. Drummond,D.A. and Wilke,C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**, 341–352.
42. Chen,H. and Blanchette,M. (2007) Detecting non-coding selective pressure in coding regions. *BMC Evol. Biol.*, **7(Suppl.1)**, S9.
43. Lin,M.F., Kheradpour,P., Washietl,S., Parker,B.J., Pedersen,J.S. and Kellis,M. (2011) Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res.*, **21**, 1916–1928.
44. Chamary,J.V. and Hurst,L.D. (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.*, **6**, R75.

Downloaded from http://nar.oxfordjournals.org/ at UB der TU Muenchen on June 22, 2013

Chapter 3

Discussion

A large number of papers, both theoretical and experimental, have demonstrated the importance of the different conformations in which non-coding RNAs have to fold for the regulation of different cellular processes (Eddy, 2001; Kung et al., 2013). Furthermore, secondary structures of unstranslated regions of messenger RNAs have been shown to play a role in translation, degradation, stability and others (Mignone et al., 2002). Recently, functionally important secondary structural elements in coding regions of mRNAs have also been reported for different living organisms and viruses. Nonetheless, a comprehensive analysis of the diversity and functional roles of the structures of mRNAs is still needed.

One of the main obstacles in the analysis of the functional roles of RNA structures is the lack of well developed experimental methods for probing RNA structures. As a result, most of the publications on RNA structures published so far have been based on analysis of theoretical predictions, which are subjected to various limitations and are thus characterized by low accuracy. One standard approach for predicting RNA secondary structure is based on free energy minimization, but the free energy nearest-neighbor model used in it is incomplete (Lu et al., 2006). However, it is worth noting that the model is continuously improved by integration of new measurements. Another limitation of the theoretical approaches is caused by the fact that the number of possible conformations for an RNA sequence rises very fast with the length of the sequence. Moreover, taking pseudoknots into consideration even makes the problem of predicting RNA structure NP-hard. That is why all prediction algorithms have to use some simplifications and approximations. Furthermore, several experimental works have confirmed that an RNA may have different co-existing conformations that will be important for fulfilling different functions (Höbartner and Micura, 2003; Schultes and Bartel, 2000). Thus, predicting a single optimal structure for an RNA sequence may not accurately reflect the reality.

To alleviate some of these problems, alternative approaches have been developed, one of which for example is based on the prediction not of a structure but of probabilities of nucleotides to be in a double-stranded conformation. This can be done by computing a partition function, which is the sum of Boltzmann factors over all possible conformations of an RNA molecule. Then the probability of every conformation can be assessed as a ratio of the Boltzmann factor for this particular conformation over the partition function. Finally, the probability of every nucleotide to be paired can be assessed as a sum of the probabilities corresponding to those structures in which that nucleotide is in a double-stranded conformation. This approach is widely used because instead of trying to predict one optimal structure, it considers the entire ensemble of possible conformations, and it was actively applied in the course of the present thesis.

The main goal of this thesis was to increase the understanding of the mechanisms and role of secondary structure of messenger RNAs. One of the underlying basis for this thesis was the paper by Ilyinsky and co-authors (Ilyinskii et al., 2009), which showed that disruption of a particular secondary structural element in coding regions of influenza mRNAs leads to a significant decrease in gene expression level. Mutations altering RNA structures thus had a functional knockout effect.

In the first publication of this thesis the first ever transcriptome-wide measurement of mRNA secondary structures was analyzed. Based on that experimental data and on theoretical predictions we wanted to examine the hypothesis that structure (and potentially, as a consequence, function) is more conserved than sequence in mRNAs. We compared probabilities of nucleotides to be in a double-stranded conformation between mRNAs of homologous genes in yeasts. Our results showed that the probabilities of nucleotides to be paired are unrelated for those sequences, the sequence identity level of which is below 85-90%. When the sequence identity level was in the range of 90100%, the distance between the structures (evaluated as root mean square deviation between vectors of probabilities) and the sequence similarity followed linear correlation. Our findings demonstrate that the structures of coding regions of mRNAs are less evolutionary conserved than those of non-coding RNAs. Thus, from the evolutionary point of view, it is more important to eliminate misfoldings of proteins than to maintain global structures of messenger RNAs. In the second publication of the thesis, we suggested a computational method of finding those regions in RNA molecule, the secondary structure of which is most likely to alter with changes in temperature. With this technique, cold-adapted (ca) temperature-sensitive (ts) influenza strains, which are widely used as live attenuated influenza vaccines, were analyzed and compared to wild type (wt) influenza strains. We showed that there are temperature-sensitive regions in influenza mRNAs, in which nucleotides are the most prone to changing their probability to be paired with temperature elevation. These regions differed between the wt strains and their ca/ts counterparts. To assess the statistical significance of such differences, mutants were generated in silico by introducing the same number of single nucleotide polymorphisms as there is between the wt and the real ca/ts strains. The conducted statistical analysis revealed the existence of ten regions, the difference in which is likely to be associated with the ca/ts phenotype. Nine of those temperature-sensitive areas were not observed in the wt strains, but were detected in the ca/ts mutants. The tenth region had the opposite behavior, it was present in the wt strain, but was absent in the ca/ts mutant. Thus, based on the developed computational method, we demonstrated that changes in mRNA secondary structures caused by temperature elevation may potentially determine temperature-sensitivity of cold-adapted influenza strains.

Next, based on the suggested indirect methodology for identification of temperature-sensitive regions in RNA secondary structures, the publicly available web service RNAtips (temperature-induced perturbations of structure) was implemented. Prior to our development, researchers studying broad range of temperature-related biological phenomena did not have a simple way of analyzing the consequences of changes in temperature on RNA conformations. The service provides every scientist with an instrument for evaluation of temperature-induced perturbations in the secondary structure of RNAs. RNAtips has a user-friendly graphical interface, allows customization of different parameters of interest, and generates high-resolution output plots. Additionally, if a researcher is interested in comparing the areas of temperature sensitivity between two homologous sequences, RNAtips is capable of conducting statistical analysis of the difference in the location of the temperature-sensitive regions to determine if that difference is specific to a particular set of nucleotide substitutions. For this, a dataset of random mutants will be automatically generated by introducing into one of the input sequences the same number of mutations as there is between the original sequences. Also, the original algorithm was upgraded as initially we worked with coding regions of mRNAs and hence for building datasets of random mutants only synonymous mutations were inserted. However, for those scientists, who are interested in non-coding RNAs, RNAtips would generate datasets of random mutants by introducing any random mutations. The service is freely accessible to all users and may be applied in studying multiple biological phenomena related to temperature.

Last but not least, we demonstrated that RNA structures may play a role in bacterial evolution. Mutations in E. coli occurring in the long-term evolution experiment being conducted by Richard Lenski and their influence on the minimum free energy values of messenger RNAs were analyzed. According to the lows of thermodynamics, every RNA molecule tends to fold into a structure with the lowest possible free energy. Therefore, we calculated minimal free energy (MFE) values for the original mRNAs and for the mutated mRNAs from the 40,000$^{\text{th}}$ generation and assessed how it has changed. It is important to note that if ancestral and progeny RNA sequences have the same MFE value, it does not mean that their structures are identical. However, different MFE values would indicate that the structure has changed in the course of evolution. Comparing changes in the minimum free energy values of messenger RNAs of essential and nonessential genes in E. coli, we determined that in general mutations in essential genes tend to change the MFE value to a lesser extent compared to mutations in nonessential genes. Next, we compared the changes in MFE values attributed to the mutations that were observed between the first and the 40,000$^{\text{th}}$ generations to the MFE changes that could be caused by the mutations that were filtered out or could have potentially taken place. To do that, in silico mutants were generated by introducing the same number of synonymous mutations as observed in the real experiment mutant to the ancestral mRNAs of each gene. We calculated the difference in the MFE values between the original E. coli mRNAs and the computer-generated mutants and computed the percentage of those in silico mutants which changed the MFE value to a lesser extent than the real mutant did. Cumulative distribution functions for those percentage values for essential and nonessential genes clearly demonstrated that such percentages are much lower in essential genes compared to nonessential ones. Our results clearly show that mutations accumulated in essential E. coli genes affect the minimum free energy values (and, hence, the secondary structure of mRNAs) to a lesser extent than those in nonessential genes. Thus, we can conclude that mutations, which disrupt the secondary structures of mRNAs of essential genes may be filtered out during the evolution. This in turn suggests that mRNA secondary structure imposes selective pressure on single nucleotide polymorphisms taking place in essential genes, whereas its effect is much weaker in nonessential genes.

So far, studying RNA structures has been largely hampered by the absence of experimentally measured data. However, genome-scale methods for measuring RNA structures are becoming increasingly available. Without any doubts, further technical improvements will allow for measuring RNA conformations more accurately and faster, and hence, will greatly contribute to increasing our notion of the plethora of functions associated with the different RNA structures. The structural RNA field is still marked by a large number of unanswered questions. For instance, investigating the dynamics of RNA conformations will be crucial for understanding how structure alters and what functions the same molecule may perform at different moments in time. Another unexplored area that would be beneficial for both experimental measurements and theoretical predictions is concerned with measuring or taking into consideration potential interactions between the thousands of different molecules within the cell.

The work done in this thesis contributes to understanding the great potential that the field of structural RNA holds and in particular demonstrates the power of computational tools for unraveling important functional aspects of RNA structures.

Chapter 4

Conclusions and outlook

Until recently it had been considered that proteins accomplish most of the regulatory functions in living organisms. Results from the ENCODE (ENCyclopedia Of DNA Elements) project (Feingold et al., 2004; Thomas et al., 2007; ENCODE Project Consortium, 2011) aimed to identify all functional elements encoded in the human genome revealed that more than 80% of the genome have some biochemical functions. There are over 20,000 protein-coding genes in the human genome, but they cover only a tiny portion of the genome. Much of this 80% is transcribed to non-protein-coding RNAs performing some regulatory functions (Dunham et al., 2012; Djebali et al., 2012; Bánfai et al., 2012). Therefore, as we can see there are many more very important functions that RNA molecules perform within modern cells and there is a lot to learn about RNA.

Experimental techniques of measuring RNA structures have been developed to the point that large RNAs can be probed. Hence, it is likely that more RNA genomes will be probed and, due to the obvious reasons, viral genomes are of the greatest interest. Information about RNA conformations of different viruses will help to understand viral biology or pathogenicity (Wan et al., 2011). At the same time, as the amount of data generated by next-generation technologies for probing the structure of RNA molecules increases, the importance of applying bioinformatics approaches to analyzing the data properly and to accurately interpreting the results will increase as well. In addition, further improvements of experimental techniques likely will allow accurate determining how different RNAs interact with each other and with RBPs. Such data will also contribute significantly to our understanding of the different biological processes and will be very important for drugs design.

I also believe that accurate bioinformatics analysis may indicate specific things that should be tested experimentally. Our research has demonstrated that further studies aimed to identify exact mRNA sequences (including UTRs) of influenza strains are needed. 5´- as well as 3´-untranslated regions might have a major impact on structural elements in the coding region. However, at the present time, the only available information we possess is cRNA/vRNAs and CDSs. Obviously, 5´- as well as 3´-untranslated regions have important RNA structures. These structures have to be preserved due to their crucial biological functions.

It would be interesting to test experimentally if the clusters, which we predicted in our work, are indeed the regions of the viral mRNAs in which the RNA structure is most sensitive to temperature. This would be the first ever experimental work of its kind (after all, the very concept of temperature-sensitive clusters is proposed in our paper). The experimental analysis, demonstrating that mRNA structures are the most temperature-sensitive in the regions corresponding to the clusters we have discovered, would require a comprehensive experimental analysis of influenza mRNAs at two temperatures and a comparison of structural RNA perturbations both within and outside of the clusters.

Further research of secondary structures of mRNAs in *Escherichia coli* is needed as well. Experiments aimed to measure expression levels of essential and nonessential genes, which can support our hypothesis, are of potential interest. Difference in the expression levels between essential and nonessential genes, if it is observed, can serve as an experimental evidence of selective pressure on the essential genes in *Escherichia coli*.

Currently, we are observing only the tip of the iceberg. It is very likely that many more classes of RNAs, as well as new unexpected functions of RNAs, will be discovered. This in turn will require development of new bioinformatics tools and algorithms for data analysis and predictions. The present work is a small step toward our understanding of biological roles of RNAs, yet it can serve as a basis for the further hypotheses.

Bibliography

Abrahams, J. P., van den Berg, M., van Batenburg, E., and Pleij, C. (1990). Prediction of rna secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res*, 18(10):3035–3044.

Akutsu, T. (2000). Dynamic programming algorithms for rna secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1):45–62.

Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., and Murphy, K. P. (2007). Efficient parameter estimation for rna secondary structure prediction. *Bioinformatics*, 23(13):i19–i28.

Atkins, J. F. and Gesteland, R. F. (1996). A case for trans translation. *Nature*, 379(6568):769–771.

Auffinger, P. and Westhof, E. (1998). Simulations of the molecular dynamics of nucleic acids. *Curr Opin Struct Biol*, 8(2):227–236.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006). Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Molecular systems biology*, 2(1).

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, Jr, W. E., Kundaje, A., Gunawardena, H. P., Yu, Y., Xie, L., Krajewski, K., Strahl, B. D., Chen, X., Bickel, P., Giddings, M. C., Brown, J. B., and Lipovich, L. (2012). Long noncoding rnas are rarely translated in two human cell lines. *Genome Res*, 22(9):1646–1657.

Baudin, F., Ehresmann, C., Romby, P., Mougel, M., Colin, J., Lempereur, L., Bachellerie, J.-P., Ebel, J.-P., and Ehresmann, B. (1987). Higher-order structure of domain iii in escherichia coli 16s ribosomal rna, 30s subunit and 70s ribosome. *Biochimie*, 69(10):1081–1096.

Belasco, J. G. and Chen, C. Y. (1988). Mechanism of puf mrna degradation: the role of an intercistronic stem-loop structure. *Gene*, 72(1-2):109–117.

Benedetti, G. and Morosetti, S. (1995). A genetic algorithm to search for optimal and suboptimal rna secondary structures. *Biophys Chem*, 55(3):253–259.

Berkhout, B., Klaver, B., and Das, A. T. (1997). Forced evolution of a regulatory rna helix in the hiv-1 genome. *Nucleic acids research*, 25(5):940–947.

Bernhart, S. H. and Hofacker, I. L. (2009). From consensus structure prediction to rna gene finding. *Brief Funct Genomic Proteomic*, 8(6):461–471.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008). Rnaalifold: improved consensus structure prediction for rna alignments. *BMC Bioinformatics*, 9:474.

Bevilacqua, P. C. and Blose, J. M. (2008). Structures, kinetics, thermodynamics, and biological functions of rna hairpins. *Annu Rev Phys Chem*, 59:79–103.

Bolton, P. H. and Kearns, D. R. (1975). Nmr evidence for common tertiary structure base pairs in yeast and e. coli trna. *Nature*.

Bredenbeek, P. J., Pachuk, C. J., Noten, A. F., Charité, J., Luytjes, W., Weiss, S. R., and Spaan, W. J. (1990). The primary structure and expression of the second open reading frame of the polymerase gene of the coronavirus mhv-a59; a highly conserved polymerase is expressed by an efficient ribosomal frameshifting mechanism. *Nucleic Acids Res*, 18(7):1825–1832.

Brierley, I., Pennell, S., and Gilbert, R. J. C. (2007). Viral rna pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol*, 5(8):598–610.

Brion, P. and Westhof, E. (1997). Hierarchy and dynamics of rna folding. *Annu Rev Biophys Biomol Struct*, 26:113–137.

Brunel, C. and Romby, P. (2000). Probing rna structure and rna-ligand complexes with chemical probes. *Methods Enzymol*, 318:3–21.

Cai, L., Malmberg, R. L., and Wu, Y. (2003). Stochastic modeling of rna pseudoknotted structures: a grammatical approach. *Bioinformatics*, 19 Suppl 1:i66–i73.

Candeias, M. M., Malbert-Colas, L., Powell, D. J., Daskalogianni, C., Maslon, M. M., Naski, N., Bourougaa, K., Calvo, F., and Fähraeus, R. (2008). P53 mrna controls p53 activity by managing mdm2 functions. *Nat Cell Biol*, 10(9):1098–1105.

Cashmore, A. R., Brown, D. M., and Smith, J. D. (1971). Selective reaction of methoxyamine with cytosine bases in tyrosine transfer ribonucleic acid. *J Mol Biol*, 59(2):359–373.

Cech, T. R. (1986). The generality of self-splicing rna: relationship to nuclear mrna splicing. *Cell*, 44(2):207–210.

Chamorro, M., Parkin, N., and Varmus, H. E. (1992). An rna pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger rna. *Proc Natl Acad Sci U S A*, 89(2):713–717.

Chang, D.-J., Kimmer, C., and Ouyang, M. (2010). Accelerating the nussinov rna folding algorithm with cuda/gpu. In *Signal Processing and Information Technology (ISSPIT), 2010 IEEE International Symposium on*, pages 120–125. IEEE.

Chang, S. E. (1973). Selective modification of cytidine and uridine residues in escherichia coli formylmethionine transfer ribonucleic acid. *J Mol Biol*, 75(3):533–547.

Chen, S. J. and Dill, K. A. (2000). Rna folding energy landscapes. *Proc Natl Acad Sci U S A*, 97(2):646–651.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823.

Christiansen, J., Brown, R. S., Sproat, B. S., and Garrett, R. A. (1987). Xenopus transcription factor iiia binds primarily at junctions between double helical stems and internal loops in oocyte 5s rna. *EMBO J*, 6(2):453–460.

Crick, F. (1962). Towards the genetic code.

Crick, F. et al. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crick, F. H. (1958). On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138.

Crick, F. H. (1968). The origin of the genetic code. *J Mol Biol*, 38(3):367–379.

Daniel, Jr, W. and Cohn, M. (1975). Proton nuclear magnetic resonance of spin-labeled escherichia coli trnaf1met. *Proc Natl Acad Sci U S A*, 72(7):2582–2586.

Daou-Chabo, R. and Condon, C. (2009). Rnase j1 endonuclease activity as a probe of rna secondary structure. *RNA*, 15(7):1417–1425.

Darty, K., Denise, A., and Ponty, Y. (2009). Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, 25(15):1974–1975.

de Bruijn, M. H. and Klug, A. (1983). A model for the tertiary structure of mammalian mitochondrial transfer rnas lacking the entire 'dihydrouridine' loop and stem. *EMBO J*, 2(8):1309–1321.

Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009). Accurate shape-directed rna structure determination. *Proc Natl Acad Sci U S A*, 106(1):97–102.

Ding, Y. (2006). Statistical and bayesian approaches to rna secondary structure prediction. *RNA*, 12(3):323–331.

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*, 32(Web Server issue):W135–W141.

Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for rna secondary structure prediction. *Nucleic Acids Res*, 31(24):7280–7301.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

Do, C. B., Woods, D. A., and Batzoglou, S. (2006). Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98.

D'Onofrio, A., Crawford, J. M., Stewart, E. J., Witt, K., Gavrish, E., Epstein, S., Clardy, J., and Lewis, K. (2010). Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chemistry & biology*, 17(3):254–264.

Doshi, K. J., Cannone, J. J., Cobaugh, C. W., and Gutell, R. R. (2004). Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction. *BMC Bioinformatics*, 5:105.

Dowell, R. D. and Eddy, S. R. (2004). Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC Bioinformatics*, 5:71.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kelllis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elinitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Starnatoyannopoulos, J. A., Tennebaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E. C., Trout, D., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Antoshechkin, I., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Gunter, C., Newberry, J. S., Levy, S. E., Absher, D. M., Mortazavi, A., Wong, W. H., Wold, B., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Parker, S. C. J., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J., Davidson, C., Derrien, T., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Howald, C., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Kokocinski, F., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tanzer, A., Tapanari, E., Tress, M. L., van Baren, M. J., Walters, N., Washieti, S., Wilming, L., Zadissa, A., Zhengdong, Z., Brent, M., Haussler, D., Kellis, M., Valencia, A., Gerstein, M., Raymond, A., Guigó, R., Harrow, J., Hubbard, T. J., Landt, S. G., Frietze, S., Abyzov, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Cheng, C., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyenger, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Larnarre-Vincent, N., Leng, J., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Pei, B., Raha, D., Ramirez, L., Reed, B., Rozowsky, J., Sboner, A., Shi, M., Sisu, C., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.-K., Yang, X., Yip, K. Y., Zhang, Z., Struhl, K., Weissman, S. M., Gerstein, M., Farnham, P. J., Snyder, M., Tenebaum, S. A., Penalva, L. O., Doyle, F., Karmakar, S., Landt, S. G., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Patacsil, D., Slifer, T., Victorsen, A., Yang, X., Snyder, M., White, K. P., Auer, T., Centarin, L., Eichenlaub, M., Gruhl, F., Heerman, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Weng, Z., Whitfield, T. W., Wang, J., Collins, P. J., Aldred, S. F., Trinklein, N. D., Partridge, E. C., Myers, R. M., Dekker, J., Jain, G., Lajoie, B. R., Sanyal, A., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Jain, G., Johnson, A. K., Johnson, E. M., Kutyavin, T. M., Lajoie, B. R., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sabo, P., Sanchez, M. E., Sandstrom, R. S., Sanyal, A., Shafer, A. O., Stergachis, A. B., Thomas, S., Thurman, R. E., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. A., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Kaul, R., Dekker, J., Stamatoyannopoulos, J. A., Dunham, I., Beal, K., Brazma, A., Flicek, P., Herrero, J., Johnson, N., Keefe, D., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Wilder, S. P., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Kundaje, A., Hardison, R. C., Miller, W., Giardine, B., Harris, R. S., Wu, W., Bickel, P. J., Banfai, B., Boley, N. P., Brown, J. B., Huang, H., Li, Q., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Hoffman, M. M., Sahu, A. O., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., Weng, Z., Iyer, S., Dong, X., Greven, M., Lin, X., Wang, J., Xi, H. S., Zhuang, J., Gerstein, M., Alexander, R. P., Balasubramanian, S., Cheng, C., Harmanci, A., Lochovsky, L., Min, R., Mu, X. J., Rozowsky, J., Yan, K.-K., Yip, K. Y., and Birney, E. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Durbin, R. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Eddy, S. R. (2001). Non-coding rna genes and the modern rna world. *Nat Rev Genet*, 2(12):919–929.

Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.-P., and Ehresmann, B. (1987). Probing the structure of rnas in solution. *Nucleic Acids Research*, 15(22):9109–9128.

Emory, S. A., Bouvet, P., and Belasco, J. G. (1992). A 5'-terminal stem-loop structure can stabilize mrna in escherichia coli. *Genes Dev*, 6(1):135–148.

ENCODE Project Consortium (2011). A user's guide to the encyclopedia of dna elements (encode). *PLoS Biol*, 9(4):e1001046.

Esteller, M. (2011). Non-coding rnas in human disease. *Nat Rev Genet*, 12(12):861–874.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data MIning*. Kdd.

Farnebo, M., Bykov, V. J. N., and Wiman, K. G. (2010). The p53 tumor suppressor: a master regulator of diverse cellular processes and therapeutic target in cancer. *Biochem Biophys Res Commun*, 396(1):85–89.

Favorova, O. O., Fasiolo, F., Keith, G., Vassilenko, S. K., and Ebel, J. P. (1981). Partial digestion of trna–aminoacyl-trna synthetase complexes with cobra venom ribonuclease. *Biochemistry*, 20(4):1006–1011.

Feingold, E., Good, P., Guyer, M., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F., Gingeras, T., Kampa, D., Sekinger, E., et al. (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640.

Felden, B., Himeno, H., Muto, A., McCutcheon, J. P., Atkins, J. F., and Gesteland, R. F. (1997). Probing the structure of the escherichia coli 10sa rna (tmrna). *RNA*, 3(1):89–103.

Fernández, A. (1992). A parallel computation revealing the role of the in vivo environment in shaping the catalytic structure of a mitochondrial rna transcript. *Journal of theoretical biology*, 157(4):487–503.

Fields, D. S. and Gutell, R. R. (1996). An analysis of large rrna sequences folded by a thermodynamic method. *Fold Des*, 1(6):419–430.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–811.

Florentz, C., Briand, J. P., Romby, P., Hirth, L., Ebel, J. P., and Glegé, R. (1982). The trna-like structure of turnip yellow mosaic virus rna: structural organization of the last 159 nucleotides from the 3' oh terminus. *EMBO J*, 1(2):269–276.

Flower, A. M. and McHenry, C. S. (1990). The gamma subunit of dna polymerase iii holoenzyme of escherichia coli is produced by ribosomal frameshifting. *Proc Natl Acad Sci U S A*, 87(10):3713–3717.

Garret, M., Labouesse, B., Litvak, S., Romby, P., Ebel, J. P., and Giegé, R. (1984a). Tertiary structure of animal trnatrp in solution and interaction of trnatrp with tryptophanyl-trna synthetase. *Eur J Biochem*, 138(1):67–75.

Garret, M., Romby, P., Giegé, R., and Litvak, S. (1984b). Interactions between avian myeloblastosis reverse transcriptase and trnatrp. mapping of complexed trna with chemicals and nucleases. *Nucleic Acids Res*, 12(5):2259–2271.

Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., Somera, A., Kyrpides, N., Anderson, I., Gelfand, M., et al. (2003). Experimental determination and system level analysis of essential genes in escherichia coli mg1655. *Journal of bacteriology*, 185(19):5673–5684.

Giedroc, D. P., Theimer, C. A., and Nixon, P. L. (2000). Structure, stability and function of rna pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol*, 298(2):167–185.

Giege, R., Moras, D., and Thierry, J. (1977). Yeast transfer rna¡ sup¿ asp¡/sup¿: A new high-resolution x-ray diffracting crystal form of a transfer rna. *Journal of molecular biology*, 115(1):91–96.

Giegerich, R., Voss, B., and Rehmsmeier, M. (2004). Abstract shapes of rna. *Nucleic Acids Res*, 32(16):4843–4851.

Gilbert, W. (1986). Origin of life: The rna world. *Nature*, 319(6055).

Gillet, R. and Felden, B. (2001). Emerging views on tmrna-mediated protein tagging and ribosome rescue. *Mol Microbiol*, 42(4):879–885.

Gluick, T. C. and Draper, D. E. (1994). Thermodynamics of folding a pseudoknotted mrna fragment. *J Mol Biol*, 241(2):246–262.

Gonzalez Jr, R. L. and Tinoco Jr, I. (1999). Solution structure and thermodynamics of a divalent metal ion binding site in an rna pseudoknot. *Journal of molecular biology*, 289(5):1267–1282.

Gorodkin, J., Heyer, L. J., Brunak, S., and Stormo, G. D. (1997a). Displaying the information contents of structural rna alignments: the structure logos. *Comput Appl Biosci*, 13(6):583–586.

Gorodkin, J., Heyer, L. J., and Stormo, G. D. (1997b). Finding the most significant common sequence and structure motifs in a set of rna sequences. *Nucleic Acids Research*, 25(18):3724–3732.

Gottesman, S. and Storz, G. (2011). Bacterial small rna regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*, 3(12).

Greenleaf, W. J., Frieda, K. L., Foster, D. A. N., Woodside, M. T., and Block, S. M. (2008). Direct observation of hierarchical folding in single riboswitch aptamers. *Science*, 319(5863):630–633.

Gregorian, Jr, R. and Crothers, D. M. (1995). Determinants of rna hairpin loop-loop complex stability. *J Mol Biol*, 248(5):968–984.

Greider, C. W. and Blackburn, E. H. (1985). Identification of a specific telomere terminal transferase activity in tetrahymena extracts. *Cell*, 43(2):405–413.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubck, R., and Hofacker, I. L. (2008). The vienna rna websuite. *Nucleic Acids Res*, 36(Web Server issue):W70–W74.

Guerrier-Takada, C. and Altman, S. (1984). Catalytic activity of an rna molecule prepared by transcription in vitro. *Science*, 223(4633):285–286.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The rna moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–857.

Gultyaev, A. P., Fouchier, R. A., and Olsthoorn, R. C. (2010). Influenza virus rna structure: unique and common features. *International reviews of immunology*, 29(6):533–556.

Gultyaev, A. P., van Batenburg, F. H., and Pleij, C. W. (1999). An approximation of loop free energy values of rna h-pseudoknots. *RNA*, 5(5):609–617.

Gutschner, T. and Diederichs, S. (2012). The hallmarks of cancer: a long non-coding rna point of view. *RNA Biol*, 9(6):703–719.

Hendrix, D. K., Brenner, S. E., Holbrook, S. R., et al. (2005). Rna structural motifs: building blocks of a modular biomolecule. *Quarterly reviews of biophysics*, 38(3):221–244.

Higgs, P. G. (1993). Rna secondary structure: a comparison of real and random sequences. *Journal de Physique I*, 3(1):43–59.

Higgs, P. G. (1995). Thermodynamic properties of transfer rna: a computational study. *J. Chem. Soc., Faraday Trans.*, 91(16):2531–2540.

Higgs, P. G. (2000). Rna secondary structure: physical and computational aspects. *Quarterly reviews of Biophysics*, 33(03):199–253.

Hingerty, B., Brown, R., and Jack, A. (1978). Further refinement of the structure of yeast trna¡ sup¿ phe¡/sup¿. *Journal of molecular biology*, 124(3):523–534.

Höbartner, C. and Micura, R. (2003). Bistable secondary structures of small rnas and their structural probing by comparative imino proton nmr spectroscopy. *J Mol Biol*, 325(3):421–431.

Hofacker, I. L., Fekete, M., Flamm, C., Huynen, M. A., Rauscher, S., Stolorz, P. E., and Stadler, P. F. (1998). Automatic detection of conserved rna structure elements in complete rna virus genomes. *Nucleic Acids Res*, 26(16):3825–3836.

Hofacker, I. L., Fekete, M., and Stadler, P. F. (2002). Secondary structure prediction for aligned rna sequences. *J Mol Biol*, 319(5):1059–1066.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188.

Hofacker, I. L. and Stadler, P. F. (1999). Automatic detection of conserved base pairing patterns in rna virus genomes. *Comput Chem*, 23(3-4):401–414.

Holbrook, S. R. and Kim, S.-H. (1997). Rna crystallography. *Biopolymers*, 44(1):3–21.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465.

Huarte, M. and Rinn, J. L. (2010). Large non-coding rnas: missing links in cancer? *Hum Mol Genet*, 19(R2):R152–R161.

Hutchinson, E. C., von Kirchbach, J. C., Gog, J. R., and Digard, P. (2010). Genome packaging in influenza a virus. *J Gen Virol*, 91(Pt 2):313–328.

Igo-Kemenes, T. and Zachau, H. G. (1969). On the specificity of the reduction of transfer ribonucleic acids with sodium borohydride. *Eur J Biochem*, 10(3):549–556.

Igo-Kemenes, T. and Zachau, H. G. (1971). Involvement of 1-methyladenosine and 7-methylguanosine in the three-dimensional structure of (yeast)trnaphe. *Eur J Biochem*, 18(2):292–298.

Ilyinskii, P. O., Schmidt, T., Lukashev, D., Meriin, A. B., Thoidis, G., Frishman, D., and Shneider, A. M. (2009). Importance of mrna secondary structural elements for the expression of influenza virus genes. *OMICS A Journal of Integrative Biology*, 13(5):421–430.

Jacks, T., Madhani, H. D., Masiarz, F. R., and Varmus, H. E. (1988). Signals for ribosomal frameshifting in the rous sarcoma virus gag-pol region. *Cell*, 55(3):447–458.

Jacks, T. and Varmus, H. E. (1985). Expression of the rous sarcoma virus pol gene by ribosomal frameshifting. *Science*, 230(4731):1237–1242.

Jackson, R. J. and Kaminski, A. (1995). Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. *RNA*, 1(10):985–1000.

James, B. D., Olsen, G. J., and Pace, N. R. (1989). Phylogenetic comparative analysis of rna secondary structure. *Methods Enzymol*, 180:227–239.

Jentsch, S. (1996). When proteins receive deadly messages at birth. *Science*, 271(5251):955–956.

Jossinet, F., Ludwig, T. E., and Westhof, E. (2007). Rna structure: bioinformatic analysis. *Curr Opin Microbiol*, 10(3):279–285.

Kean, J. M. and Draper, D. E. (1985). Secondary structure of a 345-base rna fragment covering the s8/s15 protein binding domain of escherichia coli 16 s ribosomal rna. *Biochemistry*, 24(19):5052–5061.

Kearns, D. R. and Shulman, R. G. (1974). High-resolution nuclear magnetic resonance studies of the structure of transfer ribonucleic acid and other polynucleotides in solution. *Accounts of Chemical Research*, 7(2):33–39.

Keiler, K. C. (2008). Biology of trans-translation. *Annu Rev Microbiol*, 62:133–151.

Keiler, K. C., Waller, P. R., and Sauer, R. T. (1996). Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger rna. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 990–993.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of rna secondary structure in yeast. *Nature*, 467(7311):103–107.

Kikuchi, Y., Ando, Y., and Shiba, T. (1986). Unusual priming mechanism of rna-directed dna synthesis in copia retrovirus-like particles of drosophila. *Nature*, 323(6091):824–826.

Kim, S. H., Quigley, G., Suddath, F. L., McPherson, A., Sneden, D., Kim, J. J., Weinzierl, J., Blattmann, P., and Rich, A. (1972). The three-dimensional structure of yeast phenylalanine transfer rna: shape of the molecule at 5.5-a resolution. *Proc Natl Acad Sci U S A*, 69(12):3746–3750.

Kim, S. H., Quigley, G., Suddath, F. L., and Rich, A. (1971). High-resolution x-ray diffraction patterns of crystalline transfer rna that show helical regions. *Proc Natl Acad Sci U S A*, 68(4):841–845.

Kim, S. H., Quigley, G. J., Suddath, F. L., McPherson, A., Sneden, D., Kim, J. J., Weinzierl, J., and Rich, A. (1973). Three-dimensional structure of yeast phenylalanine transfer rna: folding of the polynucleotide chain. *Science*, 179(4070):285–288.

Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H., Seeman, N. C., and Rich, A. (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer rna. *Science*, 185(4149):435–440.

Kloc, M., Foreman, V., and Reddy, S. A. (2011). Binary function of mrna. *Biochimie*, 93(11):1955–1961.

Konings, D. A. and Gutell, R. R. (1995). A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rrnas. *RNA*, 1(6):559–574.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-splicing rna: autoexcision and autocyclization of the ribosomal rna intervening sequence of tetrahymena. *Cell*, 31(1):147–157.

Kung, J. T. Y., Colognori, D., and Lee, J. T. (2013). Long noncoding rnas: past, present, and future. *Genetics*, 193(3):651–669.

Ladner, J. E., Jack, A., Robertus, J. D., Brown, R. S., Rhodes, D., Clark, B. F., and Klug, A. (1975). Atomic co-ordinates for yeast phenylalanine trna. *Nucleic Acids Res*, 2(9):1629–1637.

Larson, S. B. and McPherson, A. (2001). Satellite tobacco mosaic virus rna: structure and implications for assembly. *Curr Opin Struct Biol*, 11(1):59–65.

Layton, D. M. and Bundschuh, R. (2005). A statistical analysis of rna folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res*, 33(2):519–524.

Levitt, M. (1969). Detailed molecular model for transfer ribonucleic acid. *Nature*, 224(5221):759–763.

Li, F., Zheng, Q., Ryvkin, P., Dragomir, I., Desai, Y., Aiyer, S., Valladares, O., Yang, J., Bambina, S., Sabin, L. R., Murray, J. I., Lamitina, T., Raj, A., Cherry, S., Wang, L.-S., and Gregory, B. D. (2012). Global analysis of rna secondary structure in two metazoans. *Cell Rep*, 1(1):69–82.

Litt, M. (1969). Structural studies on transfer ribonucleic acid. i. labeling of exposed guanine sites in yeast phenylalanine transfer ribonucleic acid with kethoxal. *Biochemistry*, 8(8):3249–3253.

Lockard, R. E. and Kumar, A. (1981). Mapping trna structure in solution using double-strand-specific ribonuclease v1 from cobra venom. *Nucleic Acids Res*, 9(19):5125–5140.

Low, J. T. and Weeks, K. M. (2010). Shape-directed rna secondary structure prediction. *Methods*, 52(2):150–158.

Lowman, H. B. and Draper, D. E. (1986). On the recognition of helical rna by cobra venom v1 nuclease. *J Biol Chem*, 261(12):5396–5403.

Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A., and Summers, M. F. (2011). Nmr detection of structures in the hiv-1 5'-leader rna that regulate genome packaging. *Science*, 334(6053):242–245.

Lu, Z. J., Turner, D. H., and Mathews, D. H. (2006). A set of nearest neighbor parameters for predicting the enthalpy change of rna secondary structure formation. *Nucleic Acids Res*, 34(17):4912–4924.

Lück, R., Steger, G., and Riesner, D. (1996). Thermodynamic prediction of conserved secondary structure: application to the rre element of hiv, the trna-like element of cmv and the mrna of prion protein. *J Mol Biol*, 258(5):813–826.

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011). Multiplexed rna structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Proc Natl Acad Sci U S A*, 108(27):11063–11068.

Lyngsø, R. B. (2004). Complexity of pseudoknot prediction in simple models. In *Automata, Languages and Programming*, pages 919–931. Springer.

Lyngsø, R. B. and Pedersen, C. N. (2000). Rna pseudoknot prediction in energy-based models. *Journal of computational biology*, 7(3-4):409–427.

Mackie, G. A. and Genereaux, J. L. (1993). The role of rna structure in determining rnase e-dependent cleavage sites in the mrna for ribosomal protein s20 in vitro. *J Mol Biol*, 234(4):998–1012.

Madison, J., Everett, G., and Kung, H. (1966). Nucleotide sequence of a yeast tyrosine transfer rna. *Science*, 153(3735):531–534.

Marraffini, L. A. and Sontheimer, E. J. (2010). Crispr interference: Rna-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*, 11(3):181–190.

Mathews, D. H. (2006). Revolutions in rna secondary structure prediction. *J Mol Biol*, 359(3):526–532.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proc Natl Acad Sci U S A*, 101(19):7287–7292.

Mathews, D. H., Moss, W. N., and Turner, D. H. (2010). Folding and finding rna secondary structure. *Cold Spring Harb Perspect Biol*, 2(12):a003665.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *J Mol Biol*, 288(5):911–940.

Mathews, D. H. and Turner, D. H. (2002a). Dynalign: an algorithm for finding the secondary structure common to two rna sequences. *J Mol Biol*, 317(2):191–203.

Mathews, D. H. and Turner, D. H. (2002b). Experimentally derived nearest-neighbor parameters for the stability of rna three- and four-way multibranch loops. *Biochemistry*, 41(3):869–880.

Mathews, D. H. and Turner, D. H. (2006). Prediction of rna secondary structure by free energy minimization. *Curr Opin Struct Biol*, 16(3):270–278.

Matzke, M. A. and Birchler, J. A. (2005). Rnai-mediated pathways in the nucleus. *Nat Rev Genet*, 6(1):24–35.

Mauger, D. M. and Weeks, K. M. (2010). Toward global rna structure analysis. *Nature biotechnology*, 28(11):1178–1179.

McCarthy, J. E. and Gualerzi, C. (1990). Translational control of prokaryotic gene expression. *Trends Genet*, 6(3):78–85.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119.

Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005). Rna structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *Journal of the American Chemical Society*, 127(12):4223–4231.

Mignone, F., Gissi, C., Liuni, S., and Pesole, G. (2002). Untranslated regions of mrnas. *Genome Biol*, 3(3):REVIEWS0004.

Mirmomeni, M. H., Hughes, P. J., and Stanway, G. (1997). An rna tertiary structure in the 3'untranslated region of enteroviruses is necessary for efficient replication. *Journal of virology*, 71(3):2363–2370.

Miura, K., Iwano, T., Tsuda, S., Ueda, T., Harada, F., and Kato, N. (1982). Chemical modification of cytosine residues of trnava1 with hydrogen sulfide (nucleosides and nucleotides. xl). *Chem Pharm Bull (Tokyo)*, 30(11):4126–4133.

Miura, K., Tsuda, S., Iwano, T., Ueda, T., Harada, F., and Kato, N. (1983a). Chemical modification of cytosine residues of mouse 5 s ribosomal rna with hydrogen sulfide. (nucleosides and nucleotides 43). *Biochim Biophys Acta*, 739(2):181–189.

Miura, K., Tsuda, S., Ueda, T., Harada, F., and Kato, N. (1983b). Chemical modification of guanine residues of mouse 5 s ribosomal rna with kethoxal. (nucleosides and nucleotides 46). *Biochim Biophys Acta*, 739(3):281–285.

Montange, R. K. and Batey, R. T. (2008). Riboswitches: emerging themes in rna structure and function. *Annu Rev Biophys*, 37:117–133.

Mortimer, S. A. and Weeks, K. M. (2007). A fast-acting reagent for accurate analysis of rna secondary and tertiary structure by shape chemistry. *J Am Chem Soc*, 129(14):4144–4145.

Mougel, M., Ehresmann, B., and Ehresmann, C. (1986). Binding of escherichia coli ribosomal protein s8 to 16s rrna: kinetic and thermodynamic characterization. *Biochemistry*, 25(10):2756–2765.

Nudler, E. and Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *Trends Biochem Sci*, 29(1):11–17.

Nussinov, R. and Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded rna. *Proc Natl Acad Sci U S A*, 77(11):6309–6313.

Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. (1978). Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82.

Orgel, L. E. (1968). Evolution of the genetic apparatus. *J Mol Biol*, 38(3):381–393.

Pace, N. R., Thomas, B. C., and Woese, C. R. (1999). Probing rna structure, function, and history by comparative analysis. *COLD SPRING HARBOR MONOGRAPH SERIES*, 37:113–142.

Palmenberg, A. C. and Sgro, J.-Y. (1997). Topological organization of picornaviral genomes: statistical prediction of rna structural signals. In *Seminars in VIROLOGY*, volume 8, pages 231–241. Elsevier.

Peattie, D. A. and Gilbert, W. (1980). Chemical probes for higher-order structure in rna. *Proc Natl Acad Sci U S A*, 77(8):4679–4682.

Perret, V., Garcia, A., Puglisi, J., Grosjean, H., Ebel, J. P., Florentz, C., and Giegé, R. (1990). Conformation in solution of yeast trna(asp) transcripts deprived of modified nucleotides. *Biochimie*, 72(10):735–743.

Pieler, T., Digweed, M., Bartsch, M., and Erdmann, V. A. (1983). Comparative structural analysis of cytoplasmic and chloroplastic 5s rrna from spinach. *Nucleic Acids Res*, 11(3):591–604.

Pley, H. W., Flaherty, K. M., and McKay, D. B. (1994). Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372(6501):68–74.

Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N., Blazewicz, J., and Adamiak, R. W. (2012). Automated 3d structure composition for large rnas. *Nucleic Acids Res*, 40(14):e112.

Pyle, A. M. and Green, J. B. (1995). Rna folding. *Curr Opin Struct Biol*, 5(3):303–310.

Qiu, H., Kaluarachchi, K., Du, Z., Hoffman, D. W., and Giedroc, D. P. (1996). Thermodynamics of folding of the rna pseudoknot of the t4 gene 32 autoregulatory messenger rna. *Biochemistry*, 35(13):4176–4186.

Quigley, G. J., Wang, A. H., Seeman, N. C., Suddath, F. L., Rich, A., Sussman, J. L., and Kim, S. H. (1975). Hydrogen bonding in yeast phenylalanine transfer rna. *Proc Natl Acad Sci U S A*, 72(12):4866–4870.

Reeder, J. and Giegerich, R. (2005). Consensus shapes: an alternative to the sankoff algorithm for rna consensus structure prediction. *Bioinformatics*, 21(17):3516–3523.

Reid, B. R., Ribeiro, N. S., Gould, G., Robillard, G., Hilbers, C. W., and Shulman, R. G. (1975). Tertiary hydrogen bonds in the solution structure of transfer rna. *Proc Natl Acad Sci U S A*, 72(6):2049–2053.

Reid, B. R. and Robillard, G. T. (1975). Demonstration and origin of six tertiary base pair resonances in the nmr spectrum of e. coli trna1val. *Nature*, 257(5524):287–291.

Ren, J., Rastegari, B., Condon, A., and Hoos, H. H. (2005). Hotknots: heuristic prediction of rna secondary structures including pseudoknots. *RNA*, 11(10):1494–1504.

Rich, A. and RajBhandary, U. (1976). Transfer rna: molecular structure, sequence, and properties. *Annual review of biochemistry*, 45(1):805–860.

Rivas, E. and Eddy, S. R. (1999). A dynamic programming algorithm for rna structure prediction including pseudoknots. *J Mol Biol*, 285(5):2053–2068.

Robertson, M. P. and Joyce, G. F. (2012). The origins of the rna world. *Cold Spring Harb Perspect Biol*, 4(5).

Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F., and Klug, A. (1974a). Correlation between three-dimensional structure and chemical reactivity of transfer rna. *Nucleic Acids Res*, 1(7):927–932.

Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F., and Klug, A. (1974b). Structure of yeast phenylalanine trna at 3 a resolution. *Nature*, 250(467):546–551.

Rocca-Serra, P., Bellaousov, S., Birmingham, A., Chen, C., Cordero, P., Das, R., Davis-Neulander, L., Duncan, C. D. S., Halvorsen, M., Knight, R., Leontis, N. B., Mathews, D. H., Ritz, J., Stombaugh, J., Weeks, K. M., Zirbel, C. L., and Laederach, A. (2011). Sharing and archiving nucleic acid structure mapping data. *RNA*, 17(7):1204–1212.

Romaniuk, P. J. (1985). Characterization of the rna binding properties of transcription factor iiia of xenopus laevis oocytes. *Nucleic acids research*, 13(14):5369–5387.

Romby, P., Moras, D., Bergdoll, M., Dumas, P., Vlassov, V. V., Westhof, E., Ebel, J. P., and Giegé, R. (1985). Yeast trnaasp tertiary structure in solution and areas of interaction of the trna with aspartyl-trna synthetase. a comparative study of the yeast phenylalanine system by phosphate alkylation experiments with ethylnitrosourea. *J Mol Biol*, 184(3):455–471.

Romby, P., Moras, D., Dumas, P., Ebel, J. P., and Giegé, R. (1987). Comparison of the tertiary structure of yeast trna¡ sup¿ asp¡/sup¿ and trna¡ sup¿ phe¡/sup¿ in solution: Chemical modification study of the bases. *Journal of molecular biology*, 195(1):193–204.

Rousset, F., Pelandakis, M., and Solignac, M. (1991). Evolution of compensatory substitutions through gu intermediate state in drosophila rrna. *Proceedings of the National Academy of Sciences*, 88(22):10032–10036.

Rowe, A., Ferguson, G. L., Minor, P. D., and Macadam, A. J. (2000). Coding changes in the poliovirus protease 2a compensate for 5 ncr domain v disruptions in a cell-specific manner. *Virology*, 269(2):284–293.

Ruan, J., Stormo, G. D., and Zhang, W. (2004). An iterated loop matching approach to the prediction of rna secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66.

Sankoff, D. (1985). Simultaneous solution of the rna folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825.

SantaLucia, Jr, J. and Turner, D. H. (1997). Measuring the thermodynamics of rna secondary structure formation. *Biopolymers*, 44(3):309–319.

Schmitz, M. and Steger, G. (1996). Description of rna folding by ”simulated annealing”. *J Mol Biol*, 255(1):254–266.

Schneemann, A. (2006). The structural and functional role of rna in icosahedral virus assembly. *Annu Rev Microbiol*, 60:51–67.

Schultes, E. A. and Bartel, D. P. (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, 289(5478):448–452.

Schultes, E. A., Hraber, P. T., and LaBean, T. H. (1999). Estimating the contributions of selection and self-organization in rna secondary structure. *J Mol Evol*, 49(1):76–83.

Seetin, M. G. and Mathews, D. H. (2012). Rna structure prediction: an overview of methods. *Methods Mol Biol*, 905:99–122.

Seffens, W. and Digby, D. (1999). mrnas have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res*, 27(7):1578–1584.

Serganov, A. and Patel, D. J. (2007). Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat Rev Genet*, 8(10):776–790.

Shapiro, B. A. and Wu, J. C. (1996). An annealing mutation operator in the genetic algorithms for rna folding. *Comput Appl Biosci*, 12(3):171–180.

Shapiro, B. A., Wu, J. C., Bengali, D., and Potts, M. J. (2001). The massively parallel genetic algorithm for rna folding: Mimd implementation and population variation. *Bioinformatics*, 17(2):137–148.

Shapiro, B. A., Yingling, Y. G., Kasprzak, W., and Bindewald, E. (2007). Bridging the gap in rna structure prediction. *Curr Opin Struct Biol*, 17(2):157–165.

Shepard, P. J. and Hertel, K. J. (2008). Conserved rna secondary structures promote alternative splicing. *RNA*, 14(8):1463–1469.

Sigler, P. B. (1975). An analysis of the structure of trna. *Annual review of biophysics and bioengineering*, 4(1):477–527.

Soussi, T. and Wiman, K. G. (2007). Shaping genetic alterations in human cancer: the p53 mutation paradigm. *Cancer Cell*, 12(4):303–312.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006). Rnashapes: an integrated rna analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503.

Stern, S., Wilson, R. C., and Noller, H. F. (1986). Localization of the binding site for protein s4 on 16 s ribosomal rna by chemical and enzymatic probing and primer extension. *J Mol Biol*, 192(1):101–110.

Stojanovski, M. Z., Gjorgjevikj, D., and Madjarov, G. (2012). Parallelization of dynamic programming in nussinov rna folding algorithm on the cuda gpu. In *ICT Innovations 2011*, pages 279–289. Springer.

Su, Q., Jiang, J., and Fu, Y. (2013). A hardware implementation of nussinov rna folding algorithm. In *Computer Engineering and Technology*, pages 84–91. Springer.

Suddath, F. L., Quigley, G. J., McPherson, A., Sneden, D., Kim, J. J., Kim, S. H., and Rich, A. (1974). Three-dimensional structure of yeast phenylalanine transfer rna at 3.0angstroms resolution. *Nature*, 248(443):20–24.

Sussman, J. L., Holbrook, S. R., Warrant, R. W., Church, G. M., and Kim, S. H. (1978). Crystal structure of yeast phenylalanine transfer rna. i. crystallographic refinement. *J Mol Biol*, 123(4):607–630.

Tabaska, J. E., Cary, R. B., Gabow, H. N., and Stormo, G. D. (1998). An rna folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699.

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding rnas: regulators of disease. *J Pathol*, 220(2):126–139.

Tarasow, T. M. and Eaton, B. E. (1998). Dressed for success: realizing the catalytic potential of rna. *Biopolymers*, 48(1):29–37.

Taufer, M., Licon, A., Araiza, R., Mireles, D., van Batenburg, F. H. D., Gultyaev, A. P., and Leung, M.-Y. (2009). Pseudobase++: an extension of pseudobase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Res*, 37(Database issue):D127–D135.

Theimer, C. A. and Giedroc, D. P. (1999). Equilibrium unfolding pathway of an h-type rna pseudoknot which promotes programmed -1 ribosomal frameshifting. *J Mol Biol*, 289(5):1283–1299.

Theimer, C. A., Wang, Y., Hoffman, D. W., Krisch, H. M., and Giedroc, D. P. (1998). Non-nearest neighbor effects on the thermodynamics of unfolding of a model mrna pseudoknot. *Journal of molecular biology*, 279(3):545–564.

Thomas, D. J., Rosenbloom, K. R., Clawson, H., Hinrichs, A. S., Trumbower, H., Raney, B. J., Karolchik, D., Barber, G. P., Harte, R. A., Hillman-Jackson, J., et al. (2007). The encode project at uc santa cruz. *Nucleic acids research*, 35(suppl 1):D663–D667.

Tinoco, Jr, I., Borer, P. N., Dengler, B., Levin, M. D., Uhlenbeck, O. C., Crothers, D. M., and Bralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol*, 246(150):40–41.

Tinoco, Jr, I. and Bustamante, C. (1999). How rna folds. *J Mol Biol*, 293(2):271–281.

Tinoco, Jr, I., Uhlenbeck, O. C., and Levine, M. D. (1971). Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367.

Troutt, A., Savin, T. J., Curtiss, W. C., Celentano, J., and Vournakis, J. N. (1982). Secondary structure of bombyx mori and dictyostelium discoideum 5s rrna from s1 nuclease and cobra venom ribonuclease susceptibility, and computer assisted analysis. *Nucleic Acids Res*, 10(2):653–664.

Tsuchihashi, Z. and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of dna polymerase iii holoenzyme. *Proc Natl Acad Sci U S A*, 87(7):2516–2520.

Turner, D. H. and Mathews, D. H. (2010). Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38(Database issue):D280–D282.

Turner, D. H., Sugimoto, N., and Freier, S. M. (1988). Rna structure prediction. *Annu Rev Biophys Biophys Chem*, 17:167–192.

Ulveling, D., Francastel, C., and Hubé, F. (2011). When one is better than two: Rna with dual functions. *Biochimie*, 93(4):633–644.

Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., and Haussler, D. (2010). Fragseq: transcriptome-wide rna structure probing using high-throughput sequencing. *Nat Methods*, 7(12):995–1001.

van Batenburg, F. H., Gultyaev, A. P., and Pleij, C. W. (1995). An apl-programmed genetic algorithm for the prediction of rna secondary structure. *J Theor Biol*, 174(3):269–280.

Van Stolk, B. J. and Noller, H. F. (1984). Chemical probing of conformation in large rna molecules. analysis of 16 s ribosomal rna using diethylpyrocarbonate. *J Mol Biol*, 180(1):151–177.

Vary, C. P. and Vournakis, J. N. (1984a). Rna structure analysis using methidiumpropyl-edta.fe(ii): a base-pair-specific rna structure probe. *Proc Natl Acad Sci U S A*, 81(22):6978–6982.

Vary, C. P. and Vournakis, J. N. (1984b). Rna structure analysis using t2 ribonuclease: detection of ph and metal ion induced conformational changes in yeast trnaphe. *Nucleic acids research*, 12(17):6763–6778.

Vlassov, V. V., Giegé, R., and Ebel, J. P. (1981). Tertiary structure of trnas in solution monitored by phosphodiester modification with ethylnitrosourea. *Eur J Biochem*, 119(1):51–59.

Vlassov, V. V., Kern, D., Romby, P., Giegé, R., and Ebel, J. P. (1983). Interaction of trnaphe and trnaval with aminoacyl-trna synthetases. a chemical modification study. *Eur J Biochem*, 132(3):537–544.

Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the transcriptome through rna structure. *Nat Rev Genet*, 12(9):641–655.

Wan, Y., Qu, K., Ouyang, Z., and Chang, H. Y. (2013). Genome-wide mapping of rna structure using nuclease digestion and high-throughput sequencing. *Nat Protoc*, 8(5):849–869.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D. L., Nutter, R. C., Segal, E., and Chang, H. Y. (2012). Genome-wide measurement of rna folding energies. *Mol Cell*, 48(2):169–181.

Wang, Z., Treder, K., and Miller, W. A. (2009). Structure of a viral cap-independent translation element that functions via high affinity binding to the eif4e subunit of eif4f. *Journal of Biological Chemistry*, 284(21):14189–14202.

Warf, M. B. and Berglund, J. A. (2010). Role of rna structure in regulating pre-mrna splicing. *Trends Biochem Sci*, 35(3):169–178.

Watson, J. D. and Crick, F. H. (1953a). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967.

Watson, J. D. and Crick, F. H. (1953b). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, Jr, J. W., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009). Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–716.

Weeks, K. M. (2010). Advances in rna structure analysis by chemical probing. *Curr Opin Struct Biol*, 20(3):295–304.

Westhof, E., Dumas, P., and Moras, D. (1985). Crystallographic refinement of yeast aspartic acid transfer rna. *J Mol Biol*, 184(1):119–145.

Westhof, E., Dumas, P., and Moras, D. (1988a). Hydration of transfer rna molecules: a crystallographic study. *Biochimie*, 70(2):145–165.

Westhof, E., Dumas, P., and Moras, D. (1988b). Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer rna crystals. *Acta Crystallogr A*, 44 (Pt 2):112–123.

Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C., and Weeks, K. M. (2008). High-throughput shape analysis reveals structures in hiv-1 genomic rna strongly conserved across distinct biological states. *PLoS Biol*, 6(4):e96.

Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2005). Rna shape chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in trna(asp) transcripts. *J Am Chem Soc*, 127(13):4659–4667.

Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (shape): quantitative rna structure analysis at single nucleotide resolution. *Nat Protoc*, 1(3):1610–1616.

Wimberly, B. T., Brodersen, D. E., Clemons, Jr, W., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. (2000). Structure of the 30s ribosomal subunit. *Nature*, 407(6802):327–339.

Woese, C. R. (1967). *The genetic code: the molecular basis for genetic expression*. Harper & Row New York.

Wong, K., Bolton, P., and Kearns, D. (1975). Tertiary structure in e. coli trna arg and trna val. *Biochimica et biophysica acta*, 383(4):446.

Wong, T. W. and Clayton, D. A. (1986). Dna primase of human mitochondria is associated with structural rna that is essential for enzymatic activity. *Cell*, 45(6):817–825.

Workman, C. and Krogh, A. (1999). No evidence that mrnas have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, 27(24):4816–4822.

Wrede, P. and Rich, A. (1979). Stability of the unique anticodon loop conformation of e.coli trnafmet. *Nucleic Acids Res*, 7(6):1457–1467.

Wrede, P., Woo, N. H., and Rich, A. (1979a). Initiator trnas have a unique anticodon loop conformation. *Proc Natl Acad Sci U S A*, 76(7):3289–3293.

Wrede, P., Wurst, R., Vournakis, J., and Rich, A. (1979b). Conformational changes of yeast trnaphe and e. coli trna2glu as indicated by different nuclease digestion patterns. *Journal of Biological Chemistry*, 254(19):9608–9616.

Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. (1999). Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers*, 49(2):145–165.

WuJu, L. and JiaJin, W. (1998). Prediction of rna secondary structure based on helical regions distribution. *Bioinformatics*, 14(8):700–706.

Wurst, R. M., Vournakis, J. N., and Maxam, A. M. (1978). Structure mapping of 5'-32p-labeled rna with s1 nuclease. *Biochemistry*, 17(21):4493–4499.

Wyatt, J. R., Puglisi, J. D., and Tinoco, Jr, I. (1990). Rna pseudoknots. stability and loop size requirements. *J Mol Biol*, 214(2):455–470.

Xayaphoummine, A., Bucher, T., Thalmann, F., and Isambert, H. (2003). Prediction and statistics of pseudoknots in rna structures using exactly clustered stochastic simulations. *Proc Natl Acad Sci U S A*, 100(26):15310–15315.

Xia, T., SantaLucia, Jr, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of rna duplexes with watson-crick base pairs. *Biochemistry*, 37(42):14719–14735.

Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., et al. (2009). Update on the keio collection of escherichia coli single-gene deletion mutants. *Molecular systems biology*, 5(1).

Zaug, A. J., Grabowski, P. J., and Cech, T. R. (1983). Autocatalytic cyclization of an excised intervening sequence rna is a cleavage-ligation reaction. *Nature*, 301(5901):578–583.

Zaug, A. J., Kent, J. R., and Cech, T. R. (1984). A labile phosphodiester bond at the ligation junction in a circular intervening sequence rna. *Science*, 224(4649):574–578.

Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.-S., and Gregory, B. D. (2010). Genome-wide double-stranded rna sequencing reveals the functional significance of base-paired rnas in arabidopsis. *PLoS Genet*, 6(9):e1001141.

Zuker, M. (1989). On finding all suboptimal foldings of an rna molecule. *Science*, 244(4900):48–52.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415.

Zuker, M. and Jacobson, A. B. (1995). "well-determined" regions in rna secondary structure prediction: analysis of small subunit ribosomal rna. *Nucleic Acids Res*, 23(14):2791–2798.

Zuker, M. and Sankoff, D. (1984). Rna secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148.

Acknowledgements

The present thesis is a result of over three years of research and cooperation with a lot of people. I owe all of them a debt of gratitude for their help, stimulating discussions and great advice.

First of all, I would like to thank Prof. Dmitrij Frishman for giving me the opportunity to work in his group. I appreciate very much the freedom that I have been given, the productive discussions, the exchange of ideas with him and especially his open-mindedness.

Equally, I want to thank Dr. Alexander Shneider for all the lessons he has taught me about the different aspects of biology and management, for his wonderful support and for being a great source of inspiration to me. In addition, I am thankful to Prof. Ilya Muchnik who, in fact, played one of the most crucial roles in order for this PhD to happen.

I am also very grateful to the Dean of the Wissenschaftszentrum Weihenstephan Prof. Alfons Gierl who kindly agreed to be the Chair of my examining committee.

I would like to express my gratitude to my colleagues, who have also become very close friends of mine. Thanks to Dr. Dmitry Ivankov and Dr. Natalia Bogatireva who have always supported me in many different ways since the moment I came to Germany. Thanks to Florian Goebels for many interesting conversations during coffee breaks. And special thanks go to Stefka Tyanova for constantly providing me with opportunities to practice my acumen.

I would also like to thank Claudia Luksch and Leonie Corry for all their excellent administrative help and for always being positive during my years in the lab. Also, thanks go to Sebastian Toepel and Jonathan Hoser for the constant maintaining of our servers and the IT infrastructure, and to Elizabeth Hamzi-Schmidt for the great impact on my personal development.

Thanks to Sebastian Kopetzky, and Yanping Zhang, who nicely agreed to collaborate with me. It was a great pleasure to have the opportunity to work with them.

Hopefully, this collaboration was as useful and helpful to them as it was to me. Also many thanks go to the other RECESS members from Germany and from Russia for all the nice times and the great fun we had during our RECESS retreats.

I want to thank Prof. Andrey Mironov, Prof. Oleg Kiselev, and Prof. Alexander Gultyaev, who enthusiastically participated in discussing the projects I have been working on, for their delightful comments and scientific input.

And last but not least, I would like to thank my family and numerous friends who have supported me in many different ways along this route to successful completion of my PhD thesis.

Andrey Chursov

Studium

seit 2010	**Technische Universität München, Freising, Deutschland**. Promotion in Bioinformatik, Spezialisierung auf bioinformatischer Analyse von RNA Sekundärstruktur
2008–2009	**Moskauer Institut für Physik und Technologie, Moskau, Russland**. Masterabschluss in Management mit Spezialisierung auf Unternehmensführung
2007–2009	**Yandex School of Data Analysis, Moskau, Russland**. Zweijähriges Masterstudium im Bereich der Datenanalyse, Maschinelles Lernen, Algorithmus, etc.
2007–2009	**Moskauer Institut für Physik und Technologie, Moskau, Russland**. Masterabschluss in angewandte Mathematik und Physik mit Spezialisierung auf Informatik und Datenanalyse
2003–2007	**Moskauer Institut für Physik und Technologie, Moskau, Russland**. Bachelorstudium in angewandte Mathematik und Physik mit Spezialisierung auf Informatik

Berufliche Erfahrung

2007–2010	**Software Ingenieur**, *Yandex*, Moskau, Russland. Entwicklung von Algorithmen und Überwachungssystemen für die Entdeckung von fast identischen Dokumenten in Suchergebnissen. Entwicklung eines Systems zur Analyse von Benutzer-Reaktionen auf Suchergebnisse.
2006–2007	**Software Ingenieur**, *Moscow Center of SPARC Technologies*, Moskau, Russland. Entwicklung von Gerätetreibern für Linux- und Solaris-Betriebssysteme sowie deren Leistungstest

Kenntnisse

Computerkenntnisse	Windows, Linux, Mac OS, Microsoft Office, LaTeX	
Programmiersprachen	C, C++, Java, Python, mySQL, MATLAB, R, HTML, Javascript	
Sprachkenntnisse	Englisch: verhandlungssicher	Deutsch: Grundkenntnisse
	Russisch: Muttersprache	

Publikationen

- **Andrey Chursov**, Mathias C. Walter, Thorsten Schmidt, Andrei Mironov, Alexander Shneider and Dmitrij Frishman. Sequence-structure relationships in yeast mRNAs. *Nucleic Acids Res.,* 40(3):956-962, 2012.
- **Andrey Chursov**, Sebastian J. Kopetzky, Ignaty Leshchiner, Ivan Kondofersky, Fabian J. Theis, Dmitrij Frishman and Alexander Shneider. Specific temperature-induced perturbations of secondary mRNA structures are associated with the cold-adapted temperature-sensitive phenotype of influenza A virus. *RNA biology,* 9(10):1266-1274, 2012.

- **Andrey Chursov**, Sebastian J. Kopetzky, Gennady Bocharov, Dmitrij Frishman and Alexander Shneider. RNAtips: analysis of temperature-induced perturbations of RNA secondary structures. *Nucleic Acids Res.,* 41(W1):W486-W491, 2013.
- **Andrey Chursov**, Dmitrij Frishman and Alexander Shneider. Conservation of mRNA secondary structures may filter out mutations in *Escherichia coli* evolution. *Nucleic Acids Res.,* 41(16):7854-7860, 2013.

Glossary/Abbreviations

cDNA – complementary deoxyribonucleic acid
CDS – coding DNA sequence
CMCT – 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate
cRNA – complementary ribonucleic acid
DEPC – diethyl pyrocarbonate
DMS – dimethyl sulfate
DNA – deoxyribonucleic acid
ENU – ethylnitrosourea
lncRNA – long non-coding ribonucleic acid
MFE – minimum free energy
miRNA – micro ribonucleic acid
mRNA – messenger ribonucleic acid
ncRNA – non-coding ribonucleic acid
NMR – nuclear magnetic resonance
RBP – RNA binding protein
RNA – ribonucleic acid
rRNA – ribosomal ribonucleic acid
siRNA – short interfering ribonucleic acid
snoRNA – small nucleolar ribonucleic acid
SNP – single nucleotide polymorphism
tRNA – transfer ribonucleic acid
UTR – untranslated region
vRNA – viral ribonucleic acid