The focus of this thesis is on RNA structure and evolution. We performed computational analyses to improve our understanding of the evolution of RNA molecules, and especially of what the effect of RNA structure is on the evolution. In addition we made several practical contributions in the form of methods and algorithms. In this chapter the main results of the thesis are summarized and discussed. We also provide some future perspectives for research on RNA.

7.1 Evolutionary patterns in nucleotide composition

We have extensively studied the nucleotide composition in ribosomal RNA molecules (Chapter 2). We investigated the evolutionary changes in the composition in both rRNA subunits (SSU and LSU) and all three domains of life (archaea, bacteria, and eukaryotes). The first interesting observation was that the nucleotide composition of each unpaired structural element (loops, bulges, and junctions) was significantly different. In the data set we studied, all of the regions were rich in purines, which was to be expected given the overall purine bias in rRNA sequences (Schultes *et al.*, 1997). Surprisingly though, these regions did not contribute equally to the bias. In addition, the observed patterns of variation were highly similar between the two subunits and three domains of life despite billions of years of divergence in which functionally unconstrained parts of the molecule could presumably vary independently.

A possible explanation for these trends in composition was that the structural elements are under different selective constraints, just like the different codon positions in mRNA molecules exhibit different compositional patterns because of the effect that mutations in each frame have (Muto and Osawa, 1987). To test this hypothesis we examined the patterns of nucleotide composition in randomized sequences and their predicted structures. Since these sequences are not homologous and have not undergone any form of selection, but only have the same overall composition, we had expected these compositional differences between the unpaired regions to disappear. However, contrary to our expectation, the distinction between loops, bulges, and junctions was also present in this data set. We even performed several sensitivity tests trying to explain this phenomenon, but the results proved robust. This implied that the compositional biases can not be explained by natural selection, even though the rRNA sequences themselves are of course under strong selection and highly conserved. The compositional biases seemed to be related to the overall base composition of the sequence and to appear as a result of the RNA folding process. These findings suggested that these compositional patterns, if they would generalize to all RNAs, could be used to improve secondary structure prediction methods by evaluating the predictions on their ability to reproduce the expected compositional biases.

7.2 Using nucleotide composition for structure prediction

As a follow-up on the rRNA study described above, we decided to examine the compositional patterns of variation in other RNA families to see if we could use the information for structure prediction (Chapter 5). This turned out not to be as easy as we thought by the time of the rRNA study, because the exact biases are family-specific. Eventually though, after the realization that we could use the patterns observed under the real biological structure *relative* with respect to those observed in other (incorrect) structures, we were able to identify several generic features shared across all RNAs. This means for example that the variation in UC content in the paired regions will be tiny in one family and larger in another family, but in both families the variation is very small compared to other structures each family could have folded into. Hence, the key idea of this study was to use the compositional patterns as observed in biological structures to distinguish between realistic and unrealistic foldings.

We implemented this idea by designing a score function that describes the generic features. These shared properties include the characteristic axis-like distribution of the stems (much variation in GC content, little variation in UC and UG content) and the distinct tight distributions of the unpaired regions (as observed in the ribosomal RNAs). The score function is used to assess the plausibility of a structure for a set of sequences that presumably fold into the same structure. Specifically, given an alignment of homologous RNA sequences, an ensemble of candidate structures is generated, and subsequently the most realistic structures are selected by the score function.

We demonstrated that accurate structure predictions can be made using the patterns of nucleotide information as the only form of evidence. However, we expect that the accuracy can be improved further when nucleotide composition is used in combination with other information sources (such as covariation and thermodynamics), since the most successful methods in structure prediction adopt this strategy (Hofacker *et al.*, 2002; Knight *et al.*, 2004; Mathews *et al.*, 2004). Thus, future work could be aimed at integrating our method in a package such as BayesFold (Knight *et al.*, 2004) in order to combine multiple sources of information. Applications other than structure prediction could also be explored. For example, since the patterns result from a matching alignment and consensus structure, they might be used to find the optimal alignment given the (experimentally determined) structure, contrasting with the opposite approach we just described.

7.3 Rates of change are structure and lineage specific

Besides the direction of change in ribosomal RNA molecules, which is just one aspect of evolution, we also investigated the rate of change in the rRNA molecules (Chapter 3). The generally accepted model of RNA evolution is that the bulk of the changes are compensatory mutations in stems and that unpaired regions are rather conserved and homogeneous. As outlined in Chapter 3 we had reasons to believe this view could be refined by a detailed analysis of the rates of change in all unpaired categories using the huge amount of sequence data available today. First of all, we found that loops, bulges, and junctions change at different rates: in general loops evolve fastest, and junctions slowest. Hence, we conclude that these structural elements should be treated separately, because they do not only differ in composition (as explained in section 7.1), but also in rate of change. Moreover, the detailed look at the separate unpaired elements revealed that the traditional view of fast-evolving stems was violated in eukaryotes. In this phylogenetic domain changes in hairpin loops dominated the evolution, and eukaryotes thus do not fit the traditional model which was observed in archaea and bacteria. The eukaryotic pattern was not only caused by the insertions and deletions in this domain, because the patterns persisted when these regions were excluded from the analysis. These results suggest that evolutionary models specific for structural elements and phylogenetic groups should improve the construction of phylogenetic trees. Future work could try to assess the effects of accounting for the observed rate differences on the phylogeny.

These observations raised another question: could these differences be explained by the distribution of the structural elements in the three-dimensional molecule? It is known that fast-evolving residues are mostly found on the outside of the molecule, and highly conserved residues are mostly located on the inside, close to the catalytic core (Wuyts *et al.*, 2001a). Thus if the fast-evolving stems would be found mostly on the outside, this could explain the observed rates. We could test this hypothesis in bacteria only, because only a crystal structure from a bacterial ribosome is available. First, the distance from functional features clearly affected the conservation in all structural elements. Second, the elements were not distributed equally throughout the 3D structure. Combining these observations, we predicted the rates of change in each structural element based on their distribution in the ribosome. However, we found only a very weak correlation between the predicted and actual rate distributions. This suggests that the distance from functional elements in the ribosome can not fully explain the rate distributions in all structural elements, but that the structural element itself influences the rates. Since the rates of evolution in the structural elements differ in each phylogenetic domain, the observed rates are likely to be a poor predictor for the secondary structure. It would be interesting to repeat this analysis for eukaryotic species, thus we eagerly await the publication of a crystal structure for an eukaryotic ribosome.

7.4 Messenger RNAs are not robust against errors

Messenger RNAs (mRNAs) are under very different functional constraints than structural RNAs such as those found in the ribosome. These messengers are templates for a protein molecule, and changes at the nucleotide level might affect the protein that is encoded. The genetic code itself is highly optimized to minimize the effect of changes in the DNA on the protein that is produced, but whether the genetic messages (the mRNA molecules) are also robust against translation errors is studied far less. In Chapter 4 we compared the level of error resistance in real biological mRNA molecules to that in various randomized “messages”. To our surprise, biological messages were less resistant to translation errors than their randomized counterparts. This means the data did not support the hypothesis that species choose their codons in a way that minimizes errors. But intriguingly the range in error values (expressing the error resistance) observed in real mRNAs was much smaller than in the randomized messages, suggesting they might be under selection for a specific level of errors, which would give them an evolutionary advantage.

A second property we investigated was the relationship between the nucleotide composition at each codon position and the error minimization. If the differential compositions at the three codon positions would be caused by selection for error-resistant messages, we would expect to see a correlation between these two properties. However, the data did not show such a correlation. Thus, natural selection for error-resistant codons can not explain the universal compositional patterns in mRNAs. Future work could be targeted at testing other selective forces that could explain the differential composition. One idea we might test is whether mRNA molecules are optimized to reduce secondary structure formation. Stable stem-loop structures in mRNA molecules could make the ribosome stall during translation, which is of course unfavorable for the message. We might thus expect the structural potential in real biological messages to be smaller than in randomized messages.

7.5 The definition of pseudoknots

One of the practical contributions in this thesis is the development of several methods for pseudoknot removal from structural models. This is a common task that needs to be performed because many computational tools can not handle pseudoknots. Often researchers work with ad hoc definitions, which are poorly documented. This hinders the reproducibility of studies, and causes duplicated efforts in developing pseudoknot-removal methods. In Chapter 6 we described that many different criteria can be used to define (and remove) pseudoknots. Moreover, we demonstrate that different methods applied to the same nested structure can result in different pseudoknot-free solutions. This could have an important effect on downstream structure-based analyses, and the consequences of choosing one method over another should be considered. Assessing the effect of different methods was outside the scope of this work and was left for future analyses. For example, in case of searching for RNA family members with a pseudoknot-free structural model, choosing less conserved helices over more conserved helices might result in finding more hits.

7.6 Future perspectives

The functional and structural characterization of novel RNA molecules—ribozymes, riboswitches, small RNAs, etc.—will stay an appealing subject for research in the next few years. The growing amount of RNA data, including sequences, structures, and functional annotations, will facilitate more complex analyses. In addition to studying the RNA molecules in isolation, studying them in complex with other molecules will be interesting. How does RNA interact with other molecules? Can we predict these interactions? Much of the work will be driven by disease-related questions, but also biotechnology will benefit from our increasing knowledge of RNA. The models of RNA evolution will continue to be refined, hopefully bringing us closer to understanding the origin of life.

Bioinformatics already is and will be a vital component of RNA research. Its value has been demonstrated in for example RNA structure prediction, tertiary structure analysis, detecting novel RNAs (such as riboswitches) in the genome, and miRNA and siRNA target prediction. The rapidly growing data resources related to RNA create a demand for more and better analysis methods, but also efficient storage and visualization in biologically meaningful ways are increasingly important. The existence of decent software packages will support the development of new analyses and allow the construction of larger projects, such as integrating multiple data sources and tools. A package such as PyCogent (Knight *et al.*, 2007) is an excellent example: it provides many core objects in bioinformatics (sequences, alignments, structures, profiles, etc.), data retrieval methods, and application controllers to run third-party programs in an integrative way. It also supports many standard bioinformatics analyses and several visualization tools. It will take continuous effort to keep the package up to speed with the rapid developments in the field. An increase in users and developers of the package will be necessary, and one important step will be the creation of tutorial-style documentation to make the package suitable for students to use. This state-of-the-art software package has great potential and should be used in educational settings to train the next generation of bioinformaticians.

In conclusion, more and more exciting discoveries about RNA are being made, and it is clear by now that RNA has a prominent role in the cell and is involved in many fundamental processes. Altogether, RNA research has a bright future ahead of itself.