EVALUATING CONTEMPORARY MINIMUM FREE ENERGY BASED METHODS FOR RNA FOLDING

Andreas Smedemark Hammer Student id: 201808573

Supervisor: Christian Storm Pedersen

Master Thesis
Bioinformatics Research Centre
June 2023

Evaluating contemporary minimum free energy based methods for RNA folding

Andreas Smedemark Hammer

June 14, 2023

Abstract

Only 2% of the human genome is known to be protein coding, a large portion of the non coding is dedicated to RNAs, which we now know to fill a large variety of functions in the cell. In order to better our ability to find RNAs, knowing their secondary structure or how to recognize it is of great importance, as RNA is often more evolutionarily conserved in structure than they are by sequence. The purpose of this project is to provide an overview of common methods of predicting folding / secondary structure formation of RNAs using Minimum Free Energy based models, by reviewing contemporary solutions and comparing the most prevalent against each other. For the purpose of testing the select programs a dataset was curated from the bpRNA database, with emphasis on low sequence similarity, and varied types of RNA, to best represent RNA as broadly as possible. The experiments done indicate that strictly accuracy wise, traditional methods are unlikely to compete broadly with machine learning based approaches, however they were able to compete in some cases (for specific RNA types) albeit not as consistently. The resource use of the traditional algorithms was markedly lower than the machine learning methods, with ViennaRNA proving the most resource efficient. Furthermore results highlight the differences in prediction accuracy of different RNA families, as it affects both traditional MFE based algorithmic solutions as well as ML based, and suggests that we may not be able to solely rely on free energy minimization for prediction.

Contents