Optimizing RNA Secondary Structure Over All Possible Encodings of a Given Protein

Barry Cohen bcohen@cs.sunysb.edu

Steven Skiena skiena@cs.sunysb.edu

Department of Computer Science, SUNY Stony Brook, Stony Brook, NY 11794-4400, USA

Keywords: RNA folding, RNA secondary structure, dynamic programming, codon selection

1 Introduction

RNA sequences act as templates for building proteins, according to the scheme in which 61 different codons (sequences of three consecutive nucleotide bases) in RNA code for 20 different residues (amino acids) in proteins. Each residue has as few as 1 or as many as 6 corresponding codons. This freedom implies that the number of possible RNA sequences coding for a given protein grows exponentially in the length of the protein. For example, a typical modest-size plant hemoglobin of 147 residues (accession number AB010831) may be coded for by 1.7×10^{75} distinct RNA sequences.

Why is one particular sequence chosen from this huge space of possibilities? We note that each RNA sequence folds to a characteristic shape. The shape or structure of a protein plays a major role in determining its interaction with other molecules, and hence its function. The same is true for certain RNAs which have known functions, for example tRNAs. It is therefore natural to ask whether any functional significance has evolved in the structure of other RNA sequences which code for proteins.

We are developing a set of computational tools which answer certain questions about the space of possible RNA sequences which code for a target protein, to answer such questions as:

- Among the RNA sequences coding for a given protein P, which has the minimum energy (maximum structure)?
- What is the maximum energy (least folded) RNA sequence corresponding to a target protein P?
- Among the RNA sequences coding for a given protein P, do any fold into a specific target structure T?
- What is the topology of the RNA sequence space corresponding to a given protein? An RNA sequence space defines a corresponding RNA conformation space. Which parts of this RNA conformation space are densely populated by sequences, and which are sparse?

We present efficient algorithms for finding the minimum energy RNA structure associated with a given protein, along with experimental results on real proteins. We expect that such tools will be useful in investigating whether there functional significance to the structure of an RNA coding sequence, by placing actual coding sequences in context of the space of possibilities.

2 Background

The correlation of sequence to structure is a central problem of computational biology. The problem occurs in two flavors. In the folding problem, one is given a sequence and asked what shape it folds into; that is, what is its structure? In the inverse problem, one is given a structure and asked to design a sequence which produces it.

Sequence-structure problems occur both on the level of proteins, where the sequence is a series of residues (amino acids), and on the level of nucleic acids, where the sequence is a series of bases (A,C,G,T for DNA, A,C,G,U for RNA). The ultimate goal in the folding problem is the prediction of the three dimensional structure of an organic macromolecule from its sequence.

For RNA, secondary structure prediction is a stepping stone toward this goal. RNA has a well-defined secondary structure. Since it is single-stranded, its component bases tend to bond with other bases in the strand in analogous combinations to the Watson-Crick pairs formed in double-stranded DNA (AU, UA, CG, GC) or 'wobble' pairs (UG, GU). The set of such pairs constitutes the secondary structure of an RNA strand.

In the currently most widely used model, such pairs demarcate the RNA strand into nested loops – hairpin turns, helices, bulges, two-sided internal loops and multisided internal loops.

Zuker *et al.* [2] has developed algorithms which achieve substantial success (on average, 73% of known base pairs on domains of fewer than 700 nucleotides [3]) in RNA secondary structure prediction. Measurements in Turner's laboratory [1] of energy involved in various nucleotide interactions form the basis of the energy function employed. Dynamic programming is used to optimize the secondary structure under the observed energy functions. The Zuker-Turner model forms the framework for the problems we are working on.

3 Algorithmic and Computational Results

We have developed an efficient dynamic programming algorithm for the *minimum energy RNA sequence problem*, where we are given a target protein P, and seek to find the most stable (lowest energy) RNA sequence which codes for P, and the structure into which this sequence folds.

Asymptotically, our algorithm runs in $O(n^3)$ time, the same time as state-of-the-art programs [2] which only fold a given RNA sequence. Our algorithm exploits the fact that each residue can be coded in only a constant number of ways. Its performance can be substantially enhanced by exploiting regularities in the RNA energy function to shortcut the calculation of the values of the cells in the energy table. These regularities include:

- That an RNA sequence (by assumption) folds to the minimum energy conformation among the set of possible conformations. Therefore, given an RNA sequence S, the entropic energy of any supersequence $S+$ of S will be at most the energy of S.
- That the destabilizing energies attributed to certain structures (loops, bulges and hairpins) are monotonically nondecreasing above a given size.

Our implementation computes the optimal RNA sequence and structure among the potential 10^{75} RNA sequences for the 147-residue hemoglobin protein in 30 minutes on a desktop computer. The solution to the minimum energy problem is not in general unique, and our algorithm identifies which codons/positions in an optimal RNA sequence are undetermined or partially determined.

References

[1] Jaeger, J.A., Turner, D.H. and Zuker, M., Improved predictions of secondary structures for RNA, *Proc. Natl. Acad. Sci. USA*, 86(20):7706–7710, 1989.

[2] Lyngso, R., Zucker, M., and Pedersen, C., Internal loops in RNA secondary structure prediction, *Proc. RECOMB '99*, 260–267, 1999.

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