Elucidation of RNA Secondary Structure Using Genetic Algorithms

Ian O'Dwyer

October 15, 2001

RNA (ribonucleic acid) is a critical molecule in many biologically interesting systems. RNA is implicated in viral infections such as AIDS and the common cold as well as protein production and numerous other biological functions. The collection of base pairs which occur in the three dimensional structure of RNA, also known as the secondary structure, is of crucial importance for the functionality of the RNA. For example, both ribosomal (rRNA) and transfer RNA (tRNA) derive their function from this secondary structure and flaws in the structure are believed to render the RNA inactive. Other types of RNA, such as messenger RNA (mRNA), which codes amino acids to form the basis of proteins, rely more on their base-pair sequence to provide function and information. However, there is still a dependence on secondary structure to control the rate of coding the proteins by the mRNA. It is therefore of great interest to study and understand the secondary structure of RNA as a step to understanding RNA functionality. It is possible to determine RNA secondary structure from X-ray crystallography and nuclear magnetic resonance techniques, but such experiments are often complex, difficult and time consuming to perform, so alternative methods are sought. This paper focuses on the computational methods and algorithms used by molecular biologists to elucidate RNA secondary structure. In particular, we review genetic algorithms (GAs) and follow recent work by Chen et al, [1] supplemented with other sources as required. As an astronomy graduate student frequently using numerical models, I felt this would be a topic in molecular biology to which it might be possible for physicists to bring some new insights.

There are two main approaches to the prediction of RNA secondary structure: Comparative sequence analysis involves computing common foldings for a family of aligned, homologous RNAs. Underlying this method is the assumption that structure is conserved more than sequence. The difficulty with this method as a predictive tool is that prior knowledge of the alignment of the RNA sequences is required, implying some prior knowledge of primary and secondary RNA structure. Algorithms to perform this alignment without a priori knowledge of the RNA structure exist, but are currently limited to small sequences of RNA. The second approach is a thermodynamic one. Energies are assigned to features of RNA secondary structure such as loops and mismatched pair stacking and then RNA secondary structures which minimize, or almost minimise, free energy are sought. This method relies on the quality of the thermodynamic input data, although this is continuously being refined, and suffers from the fact that there are a number of almost degenerate free energy minima from which an optimal solution must be chosen. The thermodynamic optimization method tends to break down when pseudoknots are encountered, since it is not known how to assign energies to these structures, and it is often assumed that pseudoknots are a part of RNA tertiary structure to avoid this issue. Although the results from energy optimization are sub-optimal without *a priori* knowledge of the RNA structure, thermodynamic optimization techniques can be persuaded to deliver reasonable results, especially when the domain of possible RNA structures is poorly defined.

GAs are a stochastic optimization technique used in studying a variety of biological and non-biological systems, which can be applied to the problem of secondary RNA structure prediction. They were originally characterized by Holland [2] in 1974 and have subsequently gained popularity in many applications. An attractive feature of GAs is that their behaviour mimics natural genetic evolution, including concepts like mutation, reproduction and survival of the fittest. GAs perform operations on a population of possible solutions, each encoded with a representation of the genetic material of an individual in nature (e.g. a chromosome). The GA randomly changes some solutions, called GA mutation, and combines features of optimal parental solutions, called GA crossover, in order to find a new generation of solutions which comes closer to satisfying some fitness criteria (typically minimizing the free energy of the RNA structure). The GA then continues to iterate until it can no longer improve the solution. Unfortunately, free energy minimization alone does not lead to optimal RNA structures and some form of comparative sequencing is used to supplement the GA results in order to select the optimal structure, thus limiting the usefulness of GAs in RNA secondary structure prediction. Chen et al have developed an approach which can predict the secondary structure of the RNA, *without knowing or finding the alignment of sequences*, hence producing a GA which finds optimal structures with much greater frequency.

The basic approach is to use not only the free energy, but also the structural similarity and stability among sequences as fitness criteria. It is this structural similarity which underpins phylogenetic comparisons, hence adding extra predictive power to the GA. Initially, a GA is applied to a population of randomly generated structures with the free energy as the only fitness criterion. The GA iterates crossover, mutation and selection until all the structures in the population reach some predefined level of stability. In the second stage, each structure is assigned a measure which reflects the degree to which structural features have been conserved among sequences. The GA is applied again, but now the degree of structure conservation across generations is used as the fitness criteria to produce possible common structures for the sequences. The selected structures are now ranked on a scale closely related to the structure conservation measure and these are examined to find a convincing structure.

Chen et al applied their GA to 20 tRNA sequences, 25 5S rRNAs, seven *rev* response elements (RREs) in HIV-2 and 10 RREs in SIV. The RNA structures from the GA were compared to known structures for the 20 tRNA, which contained 432 base pairs. The most favourable structure from the GA correctly predicted 87.7% of known base pairs and one of the 10 ranked ordered structures contained 98.8% of known base pairs. For the 25 5S rRNAs, the most favourable model predicted 95.3% of the 910 known base pairs on average, and one of the top ten structures contained 98.6% of known base pairs on average. These results seem to indicate that the GA performs well in predicting secondary structure and it may be especially useful in cases where there is little knowledge of the domain of possible solutions. Conversely, an accuracy of 87.7% is very poor in applications which require exact knowledge of RNA secondary structure and, in this context, GAs can currently only be viewed as a first order approximation to RNA secondary structure. A major issue of concern for this approach is the amount of computational time and power required to find optimal structures. The number of secondary structures grows exponentially with sequence length and whilst structural similarity and stability criteria can be used to limit the number of structures, the same criteria may not be applicable to all sequences. For example, S.acidoc 5S rRNA produced 64 possible structures whilst B.brevis 5S rRNA produced 2386 structures starting from the same initial criteria. Clearly, the optimal values for the fitness criteria are different in both cases and it is difficult to determine in advance what values to use. This leads to the adoption of fairly loose criteria to account for the wide variation in number of structures, which results in a large number of possible structures being collected in the first stage of the algorithm. This causes a heavy computational burden in the final stage of the algorithm when the fitness criteria is the degree of structural conservation. In order to reduce this computational burden, Chen et al suggest the use of a fairly restricted GA which will be satisfied by most of the sequences, then using looser criteria for the remaining sequences. Whilst this seems like a reasonable approach, it is not clear that this would be practical for a large number of sequences. Overall this GA requires $O(n^2 m^2 N^2)$ computation time, where m is the maximum number of structures among N sequences. For typical runs in the paper, this amounted to a few hours of cpu time (<40 hours).

Dynamic programming algorithms (DPAs) are a more popular and widely used alternative to GAs. A DPA is essenitally a method for aligning the RNA sequences by recursion, but with caching of intermediate results, typically reducing the computational complexity of the problem from $O(2^n)$ to $\mathrm{O(n^2)}$. Sub-optimal folding algorithms based on dynamic programming will therefore take considerably less time to run than a GA. Proponents of DPAs deride GAs as 'black box' methods which lack theoretical underpinning [6] . DPAs work at the resolution of the individual base pairs, whereas the basic unit of GAs is the *stem*. The stem is simply a chain of base pairs whose size is based on several considerations, available computing power being one. The resolution of the GA is therefore limited by the minimum stem size in the model and, since secondary structure is possible at the level of the single base pair, a DPA should always predict structure at least as accurately as a GA. The weakness of the DPA is that, like the previous available GAs, it works only to minimise free energy and so requires auxilliary information to determine the optimal structure from the set of optimal solutions it produces. The great strength of the new approach illustrated by Chen et al, is that this structural information is already included in the evolution of the algorithm.

By way of conclusion, and to examine one possible future direction for the use of GAs in RNA secondary structure prediction, we look at the problem of pseudoknots. As mentioned previously, basic GAs cannot cope with pseudoknots in the RNA structure, although it has recently been shown that DPAs can deal with this complication [3] (again with the caveat that they require auxilliary structural information to determine the optimal structure). However, the massively parallel GA of Shapiro and Navetta [4, 5] is able to cope with pseudoknots in a reasonable amount of computing time. The GA was run on a parallel processor MasPar MP-2 machine (16,384 processors) and predicted about 85% of the base pairs predicted by a DPA. With current improvements in the availability and performance of multi-processor machines, it would seem worthwhile to attempt to combine the parallel algorithm with the structural similarity and stability fitness criteria of Chen et al to further improve secondary structure prediction from GAs.

References

[1] Chen, J.H., Le, S.-Y. and Maizel,J.V. (2000) Prediction of common secondary structures of RNAs: a genetic algorithm approach. Nucleic Acids Res., 28, 991-999

[2] Holland, J.H. (1975) Adaptation in Natural and Artificial Systems. Univ. of Mich. Press, Ann Arbor, Mich.

[3] Rivas, E. and Eddy, S.R. (1999)A Dynamic Programming Algorithm for RNA structure Prediction Including Pseudoknots. J. Mol. Biol. 285, 2053-2068

[4] Shapiro, B.A. and Navetta, J. (1994) A massively Parallel Genetic Algorithm for RNA Secondary Structure Prediction. The Journal of Supercomputing, Vol. 8, 195-207

[5] Shapiro, B.A. and Wu, J.C. (1997) Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm Comput. Appl. Biosci. 13, 459-471.

[6] Zuker, M. (2000) Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol. 10, 303-310

[7] A basic introduction to GAs - http://cs.felk.cvut.cz/ xobitko/ga/