# Simulating the RNA-world and Computational Ribonomics

A thesis presented
for the degree of

Doctor of Philosophy
in
BioMathematics

at Massey University, Palmerston North,
New Zealand.

Paul Phillip Gardner
2003

# Abstract

**Project 1:** Experiments by Piccirilli *et al.* (*Nature, Lond.* **343,** 33-37 (1990)) have shown that the canonical RNA genetic alphabet, `AUCG` (or `ATCG` in DNA), is not the only possible nucleotide alphabet. In this work we address the question "Is the canonical alphabet optimal?" Computational tools are used to infer RNA secondary structures (shapes) from RNA sequences of various possible alphabets, and measures of RNA shape are gathered with respect to alphabet size. Then, simulations based upon replication and selection of fixed sized RNA populations are used to investigate the effect of alternative alphabets upon RNAs ability to evolve through a fitness landscape. These results imply that for low copy fidelity the canonical alphabet is fitter than two, six and eight letter alphabets. Under high copy fidelity conditions, a six letter alphabet out-performed the four letter alphabets, which suggests that the canonical alphabet is indeed a relic of the RNA-world.

**Project 2:** Non-coding RNA genes produce functional RNA molecules rather than proteins. One such family is the H/ACA snoRNAs. Unlike the related C/D snoRNAs, these have resisted automated detection until recently.

We develop an algorithm for screening the *Saccharomyces cerevisiae* genome for novel H/ACA snoRNAs. To achieve this, we introduce some new methods to facilitate the search for non-coding RNAs in genomic sequences which are based on properties of predicted minimum free energy (MFE) secondary structures. The algorithm has been implemented and can be generalised to enable screening of other eukaryote genomes. We find that use of primary sequence data alone is insufficient for identifying novel H/ACA snoRNAs. The use of secondary structure filters reduces the number of candidates to a manageable size. On the basis of genomic location data, we identify three strong H/ACA snoRNA candidates. These together with a further 47 candidates obtained by our analysis are being screened experimentally and investigated (along with known H/ACA snoRNAs) using comparative genomic analysis.

# Acknowledgements

First and foremost I would like to say a large thank you to my supervisors Mike Hendy and David Penny for convincing me that doing a PhD was going to be "good for me", I certainly hope having to supervise me was "good for them".

Vincent Moulton has been a great motivator and supplier of travel funds to exotic Sundsvall and Uppsala in Sweden with the help of the STINT grant.

A big thanks to Sverker Edvardsson, I've benefited considerably from your hospitality and wisdom during the kiwi/swedish summers, particularly of scientific programming.

Barbara Holland has had the occasional fruitful idea (specifically `Revolver` in chapter 2) and has mercilessly edited my desire to almost never be quite definite about anything (and changing tense in mid-sentence). She has also been an excellent person to drink coffee, beer and limoncello with.

I would like to commiserate with the unfortunate biologists Ant Poole, Alicia Gore and Anu Idicula who were willing to don lab-coats, pick up pipettes and seek out snoRNA candidates.

Thanks to:

- The `Sisters` and `Helix` administrators particularly Lutz Grosz and Andre Barczak for allowing me to monopolise their souper-computer and providing excellent manuals for newbie parallel programmers.

- The numerous groups who have allowed me to drop in and give talks and share ideas with on my way to or from Sweden: Ivo Hofacker, Peter Schuster, Peter Stadler *et. al.* with the Theoretical Biochemistry Institute in Vienna, Britt-Marie Sjöberg and Marie Öhman at Stockholm University, Skip Fournier and Wayne Decatur at the University of Massachusetts, and Robert Giegerich at Bielefeld University in Germany.

- The nascent Allan Wilson Centre group for keeping life interesting.

- Patrick Rynhart (a.k.a. Sunshine), Tim White and Brett Ryland for your helpful assistance with a variety of computational problems, and also the occasional entertaining cup of coffee, glass of beer or shoot-em-up.

- The Pagans (Netball team) for getting me away from a computer once a week.

Finally I'd like to say thanks to my whanau for tolerating and supporting a "professional student" in their midst. To my father, Ross, who has been great for financial and emotional support over the years and my mother, Kim, who has always

been there for me. To my siblings Robin, Rick and Christy for never ever letting me get a big head. To my ever supportive grandparents, Vena and sadly departed Doug and, Pip and Jenny. Lastly, I'd like to express my gratitude to *mein Liebling* Erna, for your care and support throughout this degree.

Paul P. Gardner April 1, 2003.

# Preface

**Motivation:** RNA is a fascinating biopolymer, which is fundamental to all known cellular life-forms. It has both a coding role (like DNA) and a functional role (like protein) in modern organisms. This means genotype and phenotype are encoded in the same molecule in contrast to the usual situation as laid out by the "central dogma of molecular biology" where genotype is encoded by DNA and phenotype is expressed in the form of protein. This has led several evolutionary biologists to hypothesise an ancient RNA-world stage in the evolution of modern life. In an RNA-world RNA preceded protein and DNA, by performing both a catalytic and carrier of genetic information role for these ancient life-forms. This circumvents the "which came first, the chicken or the egg?" problem with the role of chicken replaced by protein and egg replaced with DNA (Gesteland & Atkins, 1993; Gesteland *et al.*, 1999).

Another RNA related field is the study of "Ribonomics" which entails determining the genomic locations and sequences of functional RNA coding genes. This problem has proved difficult to solve, due to the fact that functional RNAs don't utilise start-stop codons or conserve sequence information to the same degree as proteins. In RNA, the only usable signals are generally short protein recognition (sequence) motifs and/or a conserved secondary structure. The degree of cellular life's reliance upon functional RNA is still largely unknown. Whilst estimates of numbers of protein coding genes for many organisms are frequently cited, it is not known how many functional RNAs exist, or even the order of magnitude this is likely to be. Except for a few specific examples, such as the DNA-protein translators tRNA and rRNA, few functional RNA groups have been categorised. However progress is being made in this direction, particularly now that comparative genomics techniques, are being applied to this problem (Mattick & Gagen, 2001; Rivas & Eddy, 2001; Dennis, 2002).

**Thesis Outline:** This thesis is comprised of three chapters. Chapter one is a brief review of the current computational RNA literature and provides essential background material to the rest of the thesis.

Chapter two consists of an investigation into optimal genetic alphabet sizes. It begins with a "Context, Overview and Preliminary Results" section, followed by a manuscript which is currently in press, entitled "Optimal Alphabets for an RNA-world".

Chapter three discusses attempts to computationally locate H/ACA box (a.k.a.

pseudouridylation guide) snoRNA coding genes in *Saccharomyces cerevisiae*. The chapter is primarily comprised of manuscripts. The first manuscript, which is currently in press, is entitled "A search for H/ACA snoRNAs using predicted MFE secondary structures" and the second (unpublished) manuscript, entitled "Locating H/ACA snoRNAs using a combination of comparative genomics and MFE structure prediction", consists of a preliminary investigation of using comparative genomic techniques to locate H/ACA snoRNA coding genes.

The appendix contains a published account, entitled "RNA Folding Argues Against a Hot-Start Origin of Life", this is comprised of: (1) experimental work I carried out using equipment in the lab of Laurie Creamer at the Dairy Research Institute and, (2) a computer-based investigation I carried out into the properties of random RNA sequences with respect to temperature and base-composition. This work is included to provide background material and is not to be examined.

# Contents

# List of Figures

# List of Tables