# Modeling RNA Evolution: In-Silico and In-Vivo

Part 1: RNA Evolution in the RNA World.

Part 2: The Evolution of a RNA Virus: RS Virus in New Zealand

A thesis presented

for the degree of

Master of Science

in

BioMathematics

at Massey University, Palmerston North

New Zealand

James William Matheson

2004

## Abstract

We look at two aspects of the evolution of RNA.

First we look at RNA replication dynamics in an early RNA world context. Experimental evidence (Spiegelman *et al.* 1965, Biebricher *et al.* 1981) shows that under some conditions RNA evolves towards small quickly replicating molecules. We investigate what conditions are sufficient for a population of RNA molecules to evolve towards a balanced population of molecules. This is a population not completely dominated by a single length of molecule. We consider two models: A linear model in which indel rate is inversely proportional to length and a game theory model in which reproductive efficiency depends on the distribution of molecule lengths within a population (this is linked to catalytic efficiency). Models are investigated using analytic, numerical and simulation methods. The linear model is not sufficient to support a population with balanced length distribution. Simulation methods show that the game theory model may support such a population.

We next look at RNA evolution in the context of RNA virus evolution. Using virus samples taken over a thirty year period we investigate the evolution of Respiratory Syncytial Virus (RSV) in New Zealand. RSV most strongly affects infants and the elderly, causing cold like symptoms in mild cases and bronchiolitis or occasionally death in severe cases. New Zealand has a higher incidence of RSV bronchiolitis per head of population than many other developed countries. We compare New Zealand strains of the virus to those isolated overseas to investigate if New Zealand may have significantly different strains. We look at the evolution of the virus within New Zealand looking for evidence of antigenic drift, as well as analysing substitution rates and selection at individual codon sites. No evidence is found to suggest that New Zealand has significantly different strains of RSV from other countries. We conclude the higher rate of severe RSV in New Zealand must be caused by factors other than virus strain. The portion of the virus analysed shows strong evidence of being under positive selective pressure. This and other analyses suggest that RSV may be undergoing antigenic drift.

# Acknowledgements

I would like to say a special thank you to my supervisors, Mike Hendy, David Penny and Barbara Holland for their support, and for the many drafts of this document they have read and corrected. All mistakes that remain are my own.

I would like to thank the Allan Wilson Centre and IFS for their financial support. Thank you to my friends and colleagues there: Klaus, Michael, Tim, Bhalchandra and Barbara for keeping me sane. Thank you to Tim for putting up with so many computer questions and last but not least to Susan and Joy for helping with all the day to day things.

Thank you to Joanna Kirman, Fenella Rich and Catherine Cohet at the Malaghan Institute, without whom I would not have been able to do the RSV section of this thesis. Thank you for putting up with my ignorance of immunology and for sequencing all those snot samples.

Thank you also to Dr. Keith Grimwood of Wellington Hospital and Sue Huang of ESR for your part in the RSV project.

Thanks go to:

- Wim Hordijk for help with the initial simulations.
- Igor Boglaev and Andreas Dress for your help with the differential equations description of the game theory.
- Allen Rodrigo for taking the time to tell me about your work.
- Brett Ryland for suggesting a bilinear interpolation.

- James Chai, for putting up with my many enquiries about Helix.

Finally thanks go to my family for being so supportive, especially in my transition from PhD to Masters.

James W. Matheson

5th of December 2004.

# Preface

This project initially started as a PhD thesis and was transmuted into a Masters thesis after one year for personal reasons. In its initial stages the project focused on RNA replication dynamics in an RNA world situation. Simulations for this study were done on the Helix supercomputing cluster.
I also wanted to do some more practical work so when data became available at the Malaghan Institute on Respiratory Syncytial Virus (RSV) I agreed to analyse it as part of a project investigating RSV in New Zealand lead by Dr. Joanna Kirman of the Malaghan Institute.

The majority of the analyses in both parts of this project were done using R (R Development Core Team 2004). High speed simulation code was written in ANSI C. Parallel code used the MPI Parallel library.

**Motivation:** RNA, with its ability to encode both genotype (sequence) and phenotype (folding) in the same molecule is thought to have preceded DNA and protein as a carrier of genetic information by some scientists. An RNA world, in which RNA is the primary information carrying and catalytic molecule, is postulated to have been the first stage of evolution. The plausibility and structure of such a world rests on how RNA behaves. This motivates the study of early RNA evolution.

The study of RSV is motivated primarily by its medical significance. RSV has it largest effect in infants where it causes cold like symptoms in mild cases and bronchiolitis and occasionally death in severe cases. RSV data over a long time period (30 years) is not often available, and there is little data available on New Zealand RSV, so the discovery of New Zealand RSV samples from 1967 to 1997 in an ESR freezer offers a useful opportunity to study this virus.

# Contents

# List of Figures

# List of Tables

# Glossary of terms and abbreviations.

**Asymmetric Indel rate** ..... 6

The rate at which insertions occur does not equal the rate at which deletions occur.

**Asymmetry Parameter** ..... 7

Determines the relative probability of insertions and deletions.

**Bilinear Interpolation** ..... 35

Estimation of internal points of a surface by linear interpolation from two boundaries.

**Cluster Measure** ..... 67

Measures the degree of homogeneity of a sequence.

**Copy Error** ..... 5

The rate at which copy errors occur during replication given as a probability per replication of an error occurring.

**Cytodomain** ..... 49

The part of a protein inside the viral capsid.

**Degree of Congruence** ..... 65

Extent to which split partitions coincide with partitions defined by identical property values.

**Degree of Homogeneity** ..... 66

Measure of the degree to which terms of a sequence are identical.

**DNA** ..... 2

Deoxyribonucleic Acid. Molecule that encodes genetic information of all cellular life forms.

# 1 Introduction

This thesis investigates several aspects of the evolution of biological molecules. RNA is the molecule we concentrate on. We take a brief look at RNA in section 1.1.

In Section 2 we look at how RNA might have been involved in the early development of life. Eigen (1971) considers constant length self replicating biomolecules (specifically RNA). Eigen's model goes from early chemical self organization in evolution to the formation of hypercycles, which are self-replicating autocatalytic cycles. There is some argument over the veracity of his claims for hypercycles as the path to more complex biomolecules (Boerlijst and Hogeweg 1991, Zintzaras *et al.* 2002), however our area of interest lies before the formation of hypercycles in the area of RNA evolution. We look at variable length self replicating biomolecules and ask the question: Can we find a set of conditions sufficient to create a stable population of molecules with a balanced length distribution? This question arises from experimental observations (Spiegelman *et al.* 1965, Biebricher *et al.* 1981) showing that, under some conditions, RNA will evolve towards a highly biased (short) length distribution. This unlike what is seen in nature today. We attempt to answer this question by using mathematical models and simulation methods to investigate RNA replication dynamics in different model systems.

In section 3 we move from the theoretical to the practical implications of RNA evolution. We look at the evolution of Respiratory Syncytial Virus (RSV) in New Zealand. RSV is common in infants. Mild infection causes symptoms similar to a common cold, severe infection can cause bronchiolitis and death. New Zealand has a higher incidence of hospital admissions from RSV bronchiolitis than many other developed countries (Vogel *et al.* 2003). We investigate if this is due to New Zealand having different strains of RSV to other countries, as well as looking at the characteristics of the virus's evolution in New Zealand.

## 1.1 What is RNA?

RNA stands for **R**ibo**N**ucleic **A**cid which is a single stranded cousin to DNA. RNA like DNA is constructed from a linear chain of nucleotides attached to a sugar phosphate backbone. Unlike DNA, RNA is usually single stranded and the nucleotides used are Adenine, Cytosine, Guanine and Uracil (not Thymine) often abbreviated to A, C, G and U.

Due to their single stranded nature RNA molecules are free to fold back on themselves. This means that nucleotides that fold to be adjacent can form hydrogen bonds (base-pair) creating 2D structures such as loops and hairpins. Different nucleotides form different strengths of bond. The strongest bond is $C \equiv G$ followed by $A = U$. There is also weak binding affinity in the bond $C - A$. Bases with strong base pairing affinities are said to complement each other, C and G are complementary base pairs as are A and U. The 2D structures formed by RNA can in turn fold in 3D space to form complex structures. Some of these 3D structures will provide the chemical binding sites which allow RNA to catalyse chemical reactions.

Modern theories of the origin of life assume an RNA-world stage (Yarus 1999). This is a stage of evolution that is dominated by RNA. In these theories founding populations of RNA molecules are produced by natural RNA synthesis from nucleotides on ancient earth. RNA replication is aided by catalytic RNA called ribozymes. The founding population gradually evolves, by mechanisms such as that discussed in Eigen (1971), towards the production of protein and eventually DNA. The situation in the RNA-world differs from the modern situation; in the RNA-world RNA was both catalyst and information carrying molecule. In the modern situation the functions of catalyst and information carrier are separated between protein and DNA respectively. RNA catalysts for RNA processing (ribozymes) are essential to the RNA-world hypothesis as they form the basis for theories of self-replicating systems of RNA molecules. Though the existence of efficient ribozymes is yet to be experimentally proven, there is reason to believe it will be, with groups such as that of David Bartel (Lawrence and Bartel 2003) finding molecules that have good RNA polymerase activity but limited processivity (their ability to catalyse other molecules is

not long lived).

In a hypothetical RNA world environment containing free nucleotides and ribozymes some RNA sequences can undergo replication (Spiegelman *et al.* 1965). This process involves a complementary copy of the molecule being created from the original by pairing each base in the original unwound strand with its complement. This complement is complemented in turn to create a replica of the original sequence. There are no known error correcting mechanisms in this process so RNA replication is prone to errors. Errors can take the form of miscoded bases (called 'substitutions') or the addition or deletion of bases from the molecule (these are called insertions and deletions respectively and are collectively referred to as indels).

In this thesis we look at a formal model of aspects of the early RNA world (section 2) as well as how RNA, in the form of RNA viruses, evolves in the world today (section 3).