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

Chapter 1

Introductory Material

1.1 RNA Chemistry

1.1.1 Chemical Structure of RNA

The biological polymers RNA (ribonucleic acid) and DNA (deoxyribonucleic acid) are long strings of monomer units known as nucleotides. Each nucleotide base is composed of a base, a sugar and a phosphate group. The sugar and phosphate groups form a contiguous unit known as the sugar-phosphate backbone, which performs a structural role. In contrast, the information content of the molecule is encoded by the order of the different bases attached to the sugar group.

The sugar-phosphate backbone is linked via a phosphate group joining the 5' carbon of one ribose (deoxyribose in DNA) unit and the 3' carbon of the adjacent ribose sugar. This imposes a directionality upon the molecule, the free ends are referred to as the 5' and 3' ends as one end has an un-bonded 5' carbon and the other an un-bonded 3' carbon. Sequences are generally written in the 5' to 3' direction (see Figure 1.1).

DNA molecules are usually encountered as two complementary strands; RNA, on the other hand, is often single stranded and can form *intra-molecular* interactions driven by hydrogen bonding and stacking of paired bases. The most common nucleotide bases are cytosine (`C`), guanine (`G`), adenine (`A`) and uracil (`U`) (replaced by thymine (`T`) in DNA), although other bases do exist such as pseudo-uracil (Ψ) and the synthetically produced complementary bases κ and χ (Piccirilli *et al.*, 1990). Bases are often classified as either one or two ringed nitrogenous units called pyrimidines (`C`, `U` and `T`) or purines (`A` and `G`) respectively.

Base-pairs within RNA shapes are usually of the canonical Watson-Crick type, that are formed by three hydrogen bonds between `C` and `G` and two hydrogen bonds between `A` and `U`. "Wobble" pairing can also occur, this is a non-canonical pairing

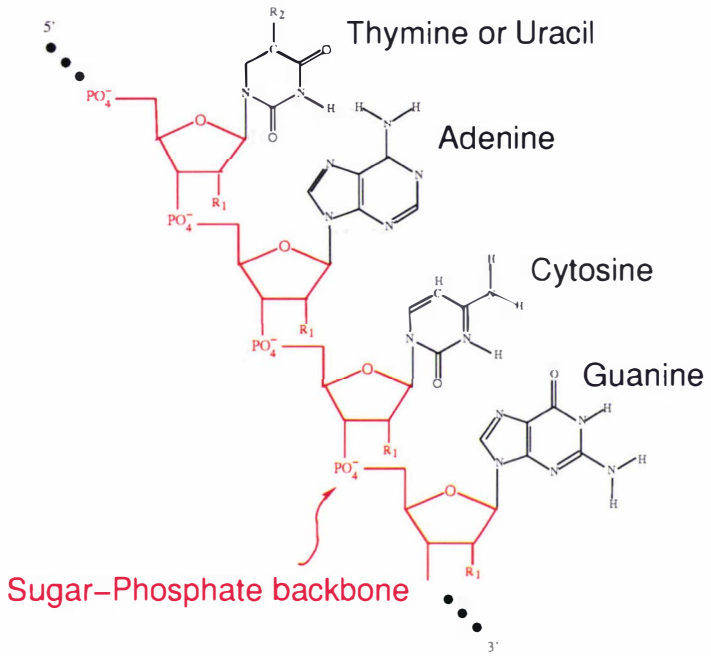


Figure 1.1: A polynucleotide sequence consisting of a series of $5' - 3'$ sugar-phosphate links, forming a backbone from which nucleotide bases can protrude. In DNA, R_1 and R_2 are -H and $-CH_3$ respectively, whereas in RNA they are -OH and -H respectively. Hence the R_1 group is the reason DNA is called **deoxy**ribonucleic acid compared to ribonucleic acid and DNA uses the base thymine whereas RNA uses uracil.

between G and U that is often found in RNA secondary structure. Probes of RNA structure using X-ray diffraction, NMR and thermodynamic studies have revealed other non-canonical (rare) pairings such as G with A, and U with C.

1.1.2 Primary, Secondary and Tertiary Structure of RNA

RNA structure can be classified at 3 different levels of information content. The *primary* structure of an RNA refers to the nucleotide sequence for that RNA. *Secondary* structure refers to regions of self complementarity that an RNA can form with itself. The definition of secondary structure is often restricted to nested base-pairs (non-nested interactions are known as "pseudo-knots" and are discussed further in section 1.2.1) and canonical plus wobble base-pairs and can be represented as a planar tree-like structure (figure 1.8B). Finally, *tertiary* structure of an RNA refers to the spatial arrangements of all elements of the RNA in three dimensions (see figure 1.2 for a graphical representations of primary, secondary and tertiary RNA structure).

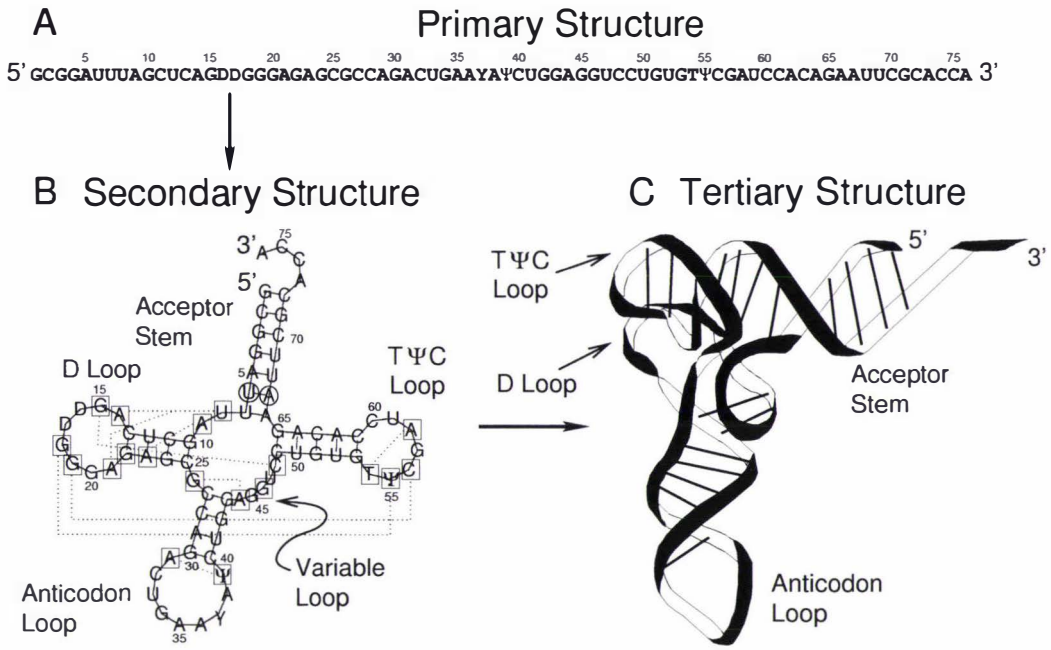


Figure 1.2: Primary, secondary and tertiary structures of yeast phenylalanine tRNA. **A:** The sequence was obtained from "The Genomic tRNA Database" (Lowe & Eddy, 1997*b*; Lowe, 2002). **B:** The secondary structure was inferred from an alignment of yeast tRNA-PHE sequences by `RNAalifold` (Hofacker *et al.*, 2002), circled bases indicate neutral mutations with respect to the displayed secondary structure. Pseudo-knots and non-canonical base-pairs are indicated with a dashed line connecting squared bases (Sundaralingham & Rao, 1975). **C:** A cartoon representation of tRNA tertiary structure, based upon tertiary structures obtained from the Protein Databank Bank (ID 6TNA,1EHZ) (Kim *et al.*, 1974; Shi & Moore, 2000).

1.1.3 Functional RNAs

Preliminaries

For the following discussion we will define a *functional RNA* (fRNA) as **any RNA** performing a function other than encoding a protein or viral genome, i.e. it is functional in its own right. Most fRNAs can be classified as either ribozymes and/or *non-coding RNAs* (ncRNAs). ncRNAs **are cellular RNAs** that perform a function other than encoding a protein. Classic examples of ncRNAs are the cellular translation apparatus formed in part by ribosomal RNA (rRNA) and transfer RNA (tRNA). This nomenclature is not universally agreed upon, ncRNAs have also been referred to as non-messenger RNA (nmRNA), non-protein-coding RNA (npcRNA) and possibly others.

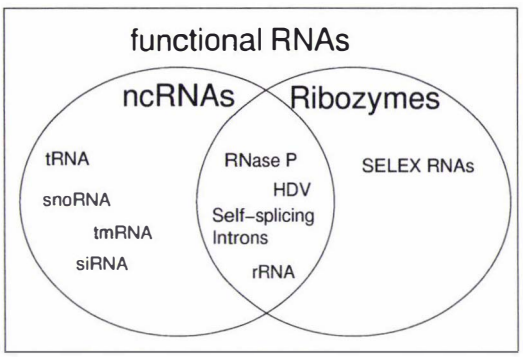


Figure 1.3: We will define a *functional RNA* (fRNAs) as **any RNA** performing a function other than encoding a protein or viral genome. Most fRNAs can be classified as at least one of the 2 groups known as ribozymes and/or non-coding RNAs (ncRNAs). A few examples of each group are shown.

Non-coding RNAs

Perhaps the most commonly known and best characterised ncRNA is **transfer RNA** (tRNA), which, along with ribosomal RNA, is integral to protein synthesis. tRNA serves as an adapter molecule that translates protein coding genes from the 4 letter nucleotide alphabet to the 20 letter protein alphabet. It has a characteristic clover-leaf secondary structure and an L-shaped 3D structure, as pictured in figure 1.2B & C.

Ribozymes: enzymatic RNA

RNA was once thought to be a passive carrier of genetic information for viruses and as an intermediate during the translation of protein from DNA, however it is now known that some RNA molecules (ribozymes) are catalytically active. Examples of ribozymes which catalyse key cellular reactions are: the hammerhead ribozyme, ribonuclease P, self-splicing introns, hepatitis delta virus and the hairpin ribozyme (see Doudna & Cech, 2002 for an excellent discussion of these). In particular the ribosome itself is basically an RNA enzymatic machine (Nissen *et al.*, 2000).

In addition to "naturally" evolved ribozymes, laboratory experiments have used a procedure known as SELEX (Systematic Evolution of Ligands by Exponential enrichment), to produce artificially evolved ribozymes by using successive rounds of affinity chromatography (selection) followed by polymerase chain reaction (PCR) (amplification) (Tuerk & Gold, 1990; Gesteland & Atkins, 1993; Szostak, 1993).

1.1.4 The Central Dogma and functional RNAs

The historical "Central Dogma of Molecular Biology", first posed in the early 1950s (Gesteland & Atkins, 1993), states that a DNA encoded gene is **transcribed** into messenger RNA (mRNA), which is subsequently **translated** into protein (See Figure 1.4). In general this concept holds, but is not the complete picture. There are, as yet, an unknown number of RNAs which do not serve a coding role, they are functional transcripts in themselves (Eddy, 2002). In addition, the dogma implies that each each gene has a single corresponding gene product, the discovery of "alternative splicing" whereby introns (non-coding regions of a gene) are spliced in variable patterns also contradicts the dogma (Modrek *et al.*, 2001).

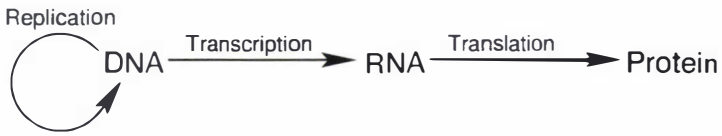


Figure 1.4: The Central Dogma of Molecular Biology: genes are perpetuated as sequences of nucleic acid, but usually function by being expressed in the form of proteins (Gesteland & Atkins, 1993).

Perhaps an adjustment to the Central Dogma which distinguishes between messenger RNA (mRNA) and functional RNA (fRNA) would be useful (Dennis, 2002; Mattick & Gagen, 2001). By adding an extra branch to the diagram we illustrate that gene products may also be functional RNAs (See Figure 1.5). RNAs often need maturing in the form of covalent modification of certain key nucleotides. For example mature ribosomal RNAs and transfer RNAs (rRNAs and tRNAs are an integral part of the cells machinery for synthesising protein) have a number of methylated ribose sugars and uridine isomers known as Pseudo-uridine (this is discussed further in Chapter 3).

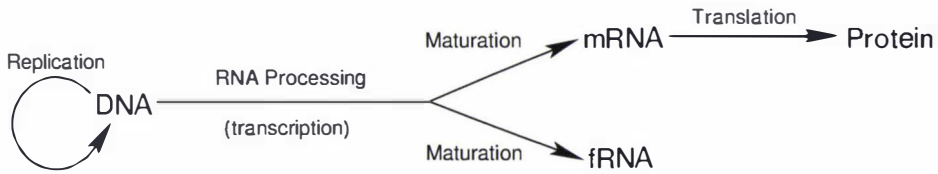


Figure 1.5: The modern central dogma of Molecular Biology: genes are perpetuated as sequences of nucleic acid, which function by being expressed in the form of proteins or mature functional RNAs (Dennis, 2002; Mattick & Gagen, 2001).

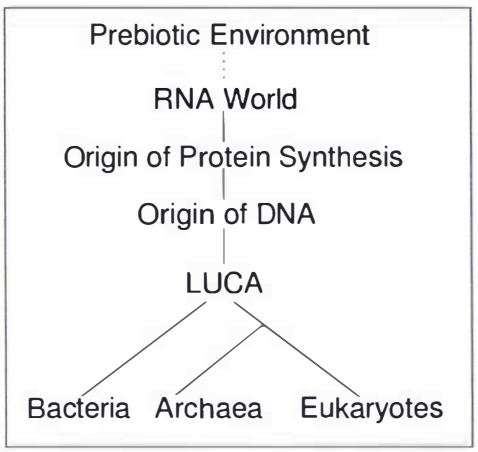


Figure 1.6: A hypothetical series in the origin of life that includes an RNA-World stage, adapted from (Poole *et al.*, 1999).

1.1.5 The RNA-world

The original hypothesis that RNA based evolution may have preceded DNA/protein synthesis, upon which current biota rely, was first proposed in the late 1960s by a number of different groups (Woese, 1967; Crick, 1968; Orgel, 1968). The subsequent, startling discovery that RNA has catalytic abilities (Kruger *et al.*, 1982; Cech, 1986; Cech, 1987) has added significant weight to the theory and led to the coining of the term "the RNA world" (Gilbert, 1986).

All RNA-world hypotheses rely upon two basic assumptions; firstly that RNA preceded protein as the main macro-molecular catalyst, and secondly that RNA preceded DNA as the main informational molecule. Hypotheses differ over aspects, such as, what forms pre-RNA-world life may have taken (Joyce, 2002), the metabolic complexity of the RNA-world (Spirin, 2002), and the roles of various RNA cofactors (White, 1976; Jadhav & Yarus, 2002).

An interesting corollary of this theory is that many of the metabolically active RNAs of today's organisms may be relics or molecular fossils from the RNA-World (Poole *et al.*, 1998; Jeffares *et al.*, 1998). Additionally, it appears that the RNA-world hypothesis is not compatable with the currently held belief that life originated in under-water volcanic vents. This is due to the fact that RNA cannot maintain a stable, active tertiary structure at high temperature, this is discussed in more detail in Appendix I (Moulton *et al.*, 2000*a*).

1.2 RNA Informatics

1.2.1 RNA Secondary Structures and Structural Elements

Definitions.

The primary structure of single stranded ribonucleic acid (RNA) is represented by a linear sequence of the symbols $\mathbf{S} = s_1 s_2 \cdots s_n$, where each s_i is one of the nucleotides `A`, `U`, `C` or `G` (although other bases do occur they are rare and usually ignored). The sequence **S** is the *primary structure* of the RNA molecule. Recall that the bases can also form base-pairs, the Watson-Crick pairs are `A` with `U`, and `C` with `G`. Additionally wobble `G` with `U` pairs are frequently allowed, other pairings exist but are infrequent and are ignored in preliminary investigations.

RNA secondary structure is frequently represented as a planar graph (see Figure 1.8A, for an example), these satisfy a few basic criteria.

DEFINITION 1 (RNA SECONDARY STRUCTURE GRAPH) *An RNA secondary structure graph consists of an ordered set of vertices* $(a_1 a_2 \cdots a_n)$ *which represent bases, and edges* $(a_i a_j)$ *which represent adjacencies within the nucleotide sequence and hydrogen bonds between complementary bases. The adjacency matrix* A *of the secondary structure graph must satisfy the following constraints (Waterman, 1995; Hofacker* et al., *1998):*

1. $a_{i,i+1} = 1$ *for* $1 \leq i \leq n-1$ *(This represents the sugar-phosphate backbone of RNA).*
2. $\forall\ i$ *there is at most one* $k \neq i \pm 1$ *satisfying* $a_{ik} = 1$ *(Nucleotide* s_i *and* s_k *are said to be paired).*
3. $\forall\ a_{ik} = 1$ $(k \neq i \pm 1)$ *then* $|k - i| > 3$ *(Paired bases must be at least 3 nucleotides apart due to torsional constraints on the backbone).*
4. *If* $a_{ij} = a_{kl} = 1$ $(i < j,\ k < l)$ *and* $i < k < j$ *then* $k < l < j$ *(This is the "no pseudo-knot" criterion that forces all base-pairs to be nested (see figure 1.7 for some example pseudo-knots)).*

   A vertex k *is* **interior** *to the base-pair* (i, j) *if* $i < k < j$*, furthermore if there exists no base-pair* (m, n) *such that* $i < m < k < n < j$ *then* k *is said to be* **immediately interior** *to* (i, j).

DEFINITION 2 (THE COMPONENTS OF RNA SECONDARY STRUCTURES) *RNA secondary structures can be decomposed into* **stacks**, **loops** *and* **external elements** *(Hofacker* et al., *1998).*

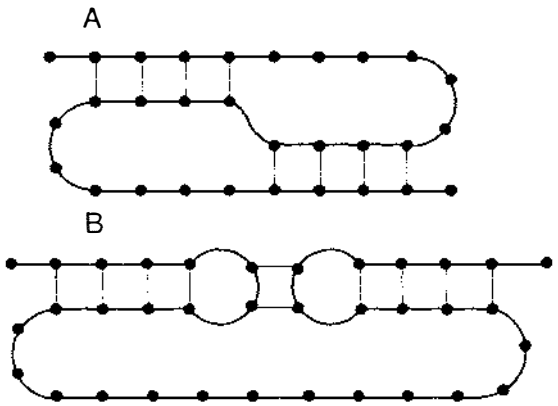


Figure 1.7: Two examples of non-nested base-pairing, usually called pseudo-knots. **A** is an example of a "simple pseudo-knot", **B** is called the "kissing hair-pins". These are examples of structures that are not represented in a standard secondary structure graph.

---

1. *A* **stack** *consists of nested, consecutive base-pairs. With indices satisfying,* $i < j$ *and* $(i,j),(i+1,j-1),\cdots,(i+k,j-k)$ *and* $k \leq \frac{1}{2}(j-i-3)$. *The length of the stack is* $k+1$ *and the edge* (i,j) *is the* **terminal base-pair** *of the stack.*

2. *A* **loop** *is defined as a sequence of consecutive, unpaired vertices immediately interior to a base-pair.*

3. *An* **external vertex** *is any unpaired vertex not contained in a loop. Consecutive external vertices are called* **external elements**.

LEMMA 1 (UNIQUE SECONDARY STRUCTURE DECOMPOSITION) *Any secondary structure* **S** *can be uniquely decomposed into external elements, loops and stacks (Hofacker* et al., *1998).*

Proof. Each vertex involved in a base-pair belongs to a unique stack. All unpaired vertices are either interior or exterior to a base-pair. Interior unpaired vertices must belong to a loop, likewise exterior vertices must belong to an external element. Since each stack, loop and external element is unique, the decomposition is also unique.

DEFINITION 3 (CHARACTERISATION OF LOOPS) *The* **degree** *of a loop is defined as* $1+k$, *where* k *is the number of terminal base-pairs that are internal to the closing base-pair (Hofacker* et al., *1998).*

1. *Loops of degree 1 are called* **hairpin loops** *(H).*

2. *Loops of degree 2 are divided into two groups. Consider an external base-pair* (i, j) *and an internal base-pair* $(i + p, j - q)$, $(i < p < q < j)$, *where* p *and* q *are non-negative integers and vertices* $i+1, \cdots, i+p-1$ *and* $j-1, \cdots, j-q+1$ *are unpaired.*

   (a) *A* **bulge** *(B) occurs when exactly one of* p *or* q *is equal to zero.*

   (b) *An* **internal loop** *(I) occurs when both* p *and* q *are greater than zero.*

3. *Loops of degree 3 or more are known as* **multi-loops** *(M).*

See Figure 1.8A for a graphical representations of characteristic loops.

Representations of Secondary Structure.

DEFINITION 4 (CLASSICAL REPRESENTATIONS) *The nodes represent nucleotides, the edges* (a_i, a_{i+1}) *represent the sugar-phosphate backbone, all other edges represnt a base-pair. (see Figures 1.8A, C & D). The representation in Figure 1.8A is the more commonly used representation by biologists, representations in figures 1.8C & D have been used by theoreticians in the field for showing how a base-pair divides an RNA (for example, Waterman (1995)).*

DEFINITION 5 (TREE REPRESENTATION(S)) *There are a variety of ways to represent RNA secondary structures as trees (Hofacker* et al., *1998; Moulton* et al., *2000b). Some compress substructures into single labeled vertices, which is equivalent to the coarse-grained representations discussed on page 11. In figure 1.8B the secondary structure is translated into a rooted tree embedded in the plane. An additional node (the diamond) is introduced in order to root the tree and a base-pair* (i, j) *is represented by a vertex* x *such that the children* $y_1, \cdots, y_k$ *of* x *correspond to the base-pairs* $(i_1, j_1) \cdots, (i_k, j_k)$ *immediately interior to* (i, j). *Unpaired vertices are leaves added to the vertex representing the closing pair of the loop containing the unpaired vertex. The tree representation is often used by people wishing to employ known tree results for secondary structure analyses. For example, tree metrics can be used as a measure of distance between secondary structures (Moulton* et al., *2000b)*

DEFINITION 6 (DOT-BRACKET REPRESENTATION) *This is a string of n parentheses and periods. A base-pair between nucleotides* i *and* j *is indicated by an open bracket '(' at position* i *and a close bracket ')' at position* j. *Unpaired nucleotides are indicated with a period '.' (Hofacker* et al., *1994) (See Figure 1.8E). This representation is particularly suited for storage on a computer.*

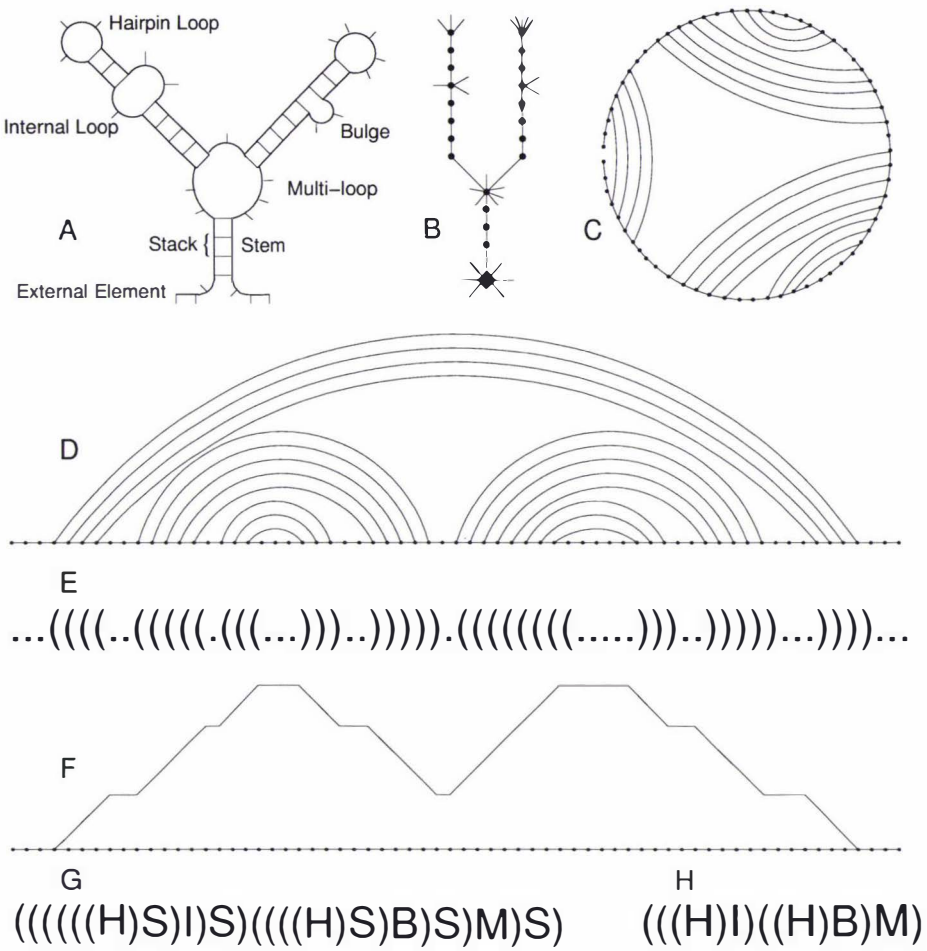


Figure 1.8: Equivalent representations of an RNA secondary structure. **A:** is generally known as the "biological representation" and is the most commonly used representation. **C** and **D:** are circle and arch representations respectively of this structure, note that representations **A**, **C** and **D** are isomorphic to each other. **B:** is a tree representation, internal nodes indicate a base-pair and leaves are unpaired positions. The diamond-shaped node is added to maintain a rooted tree for structures with external elements. **E:** is the dot-bracket notation, in which base-pairs are indicated by matching parentheses in corresponding positions in the string, and unpaired positions are indicated by periods. **F:** is the mountain plot representation. The mountain plot corresponds well with the dot-bracket notation. The plot is incremented by one for an opening bracket '(', and decremented by one for a closing bracket ')', otherwise the plot remains at the same height. **G** and **H:** are coarse-grained representations of the secondary structure. Loops are condensed to their type (thus ignoring size information): hairpin (H), internal (I), bulge (B), multi-loop (M) and stack (S). Since each loop is enclosed by a stack this information is often considered redundant and not indicated, which produces representation **H**.

DEFINITION 7 (MOUNTAIN PLOT REPRESENTATION) *The bracket notation leads naturally to the mountain representation (see equation 1.1 and figure 1.8E & F). This is a two-dimensional graph consisting of the points with x-coordinate k corresponding to the kth nucleotide and y-coordinate y_k a count of the number of base-pairs enclosing this nucleotide (Hogeweg & Hesper, 1984) (See figure 1.8F). The mountain representation is useful for visualising and comparing structures (see figure 1.2.4 for an example), it has been used to derive a metric (Moulton* et al., *2000*b*) and also used to search genomic sequences for related secondary structures (section 3.2).*

$$\begin{aligned} y_0 &= 0 \\ y_{k+1} &= \begin{cases} y_k + 1 & \text{if } S_k = \text{'('} \\ y_k - 1 & \text{if } S_k = \text{')'} \\ y_k & \text{otherwise.} \end{cases} \end{aligned} \tag{1.1}$$

DEFINITION 8 (COARSE-GRAINED REPRESENTATION(S)) *For many applications it is useful to ignore some of the data displayed in standard secondary structure representations. Coarse-grained representations, which, do not retain the full information of the secondary structure are displayed in figure 1.8G. This condensed form represents loops and stacks as a single character: 'H' for hairpin a loop, 'I' for an interior loop, 'B' for a bulge, 'M' for a multi-loop, and 'S' for a stack (Shapiro, 1988). A more compressed representation (shown in figure 1.8H), is possible. This is obtained by considering only the loops, as each loop is always closed by a stack the 'S' is a redundant character for this form (Hofacker* et al., *1994).*

1.2.2 RNA Shape-Space

Here we discuss a few generic properties of the mapping from RNA sequences to secondary structure. First, some new terminology needs to be introduced: The set of all sequences of a fixed length N, generated from a given alphabet (`AUCG` for example) is called *sequence-space*; The corresponding set of all secondary structures of fixed length (N) is refered to as *shape-space*. Most of the following discoveries were found via a brute-force/exhaustive enumeration approach to explore the nature of the sequence to structure mapping (Grüner *et al.*, 1996).

1. **One shape, many sequences:** RNA sequence-space is significantly larger than RNA shape-space. The cardinality (size) of sequence-space is 4^N (where N is the sequence length). By considering the dot-bracket alphabet, the cardinality (size) of shape-space has an upper bound of 3^N. However, each left

parenthesis must have a corresponding right parenthesis, and if pseudo-knots are ignored, the brackets must also be nested, these considerations lead to a tighter upper bound, C_N, the Nth term of the series called the *Catalan Numbers* (see equation 1.2). These bounds do not incorporate limitations on minimal stack and loop size. Recursions and generating functions which incorporate these have been used to show that the cardinality of RNA shape-space is given by S_N (see equation 1.3) (Schuster *et al.*, 1994; Waterman, 1995; Hofacker *et al.*, 1998). Analysis of results from exhaustive minimum free energy (MFE) folding of sequences of a fixed length show that the number of shapes actually used is considerably lower than this upper bound (Grüner *et al.*, 1996).

$$\begin{aligned} C_N &= \frac{(2N)!}{(N!)^2} \times \frac{1}{N+1} \\ &\sim \frac{1}{\sqrt{\pi}} \times N^{\frac{-3}{2}} \times 4^N \end{aligned} \tag{1.2}$$

$$S_N \sim 1.4848 \times N^{\frac{-3}{2}} \times (1.84892)^N \tag{1.3}$$

In conclusion, RNA sequence-space grows by a factor of 4 for each base added to the length, however, RNA shape-space only grows by ~ 2 for each added base. This means there are many more sequences than shapes and therefore many sequences are mapped to the same shape.

2. **Common shapes:** On average there are $\frac{4^N}{S_N}$ sequences for each unique secondary structure, but some secondary structures occur more frequent than others. A *common structure* is defined as any structure that occurs more frequently than the expected average. Analysis of results from an exhaustive enumeration search of a coarse-grained shape-space for all GC derived sequences of length 30 show that common structures form approximately 10% of the total shapes encountered, yet a total of 93% of the sequences fold into these structures (Grüner *et al.*, 1996; Fontana, 2002).

3. **Neutral Networks:** These consist of sequences of the same shape which differ by $1-2$ point mutations. Nearest neighbours can "percolate" throughout sequence-space. Paths of structure-neutral mutations traverse sequence-space (see figure 1.9), thus a primary sequence can be completely altered but the same shape maintained (Schuster, 1993).

4. **Shape-space covering:** Any random sequence has neighbours within a small distance (a Hamming distance for example, see section 1.2.3) of itself which

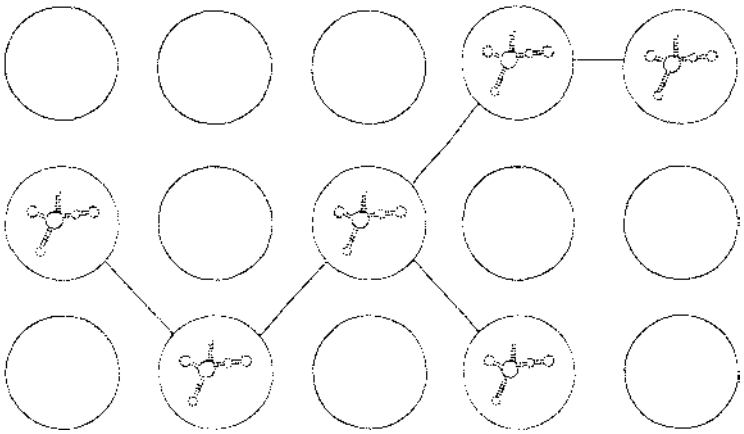


Figure 1.9: A neutral network linking common shapes. Sequences differing by a Hamming distance of 1 (or 2) from each other frequently map to the same common shape. Consequently paths of shape neutral mutations traverse much of sequence-space (Schuster, 1993; Fontana, 2002).

fold into all the common shapes (see figure 1.10 for an illustration of this concept). For random sequences derived from the AUCG alphabet of length 100 an average of 15 mutations was sufficient to recover all the common shapes (Schuster *et al.*, 1994; Grüner *et al.*, 1996). This result is supported by an interesting laboratory demonstration: Two structurally and enzymatically unrelated ribozymes (a class III self-ligating ribozyme and an HDV self-cleaving ribozyme) had, within approximately 40 mutations of each other a novel ribozyme that performed both catalytic tasks (Schultes & Bartel, 2000). So not only are there "short" paths through sequence-space between different shapes there are also paths through sequence-space connecting functional elements.

1.2.3 Metrics on RNA structures

Measuring the degree of difference between different RNA structures is a useful tool for exploring properties of RNA shape-space. In the special case of RNA primary sequences of the same length a natural measure is the Hamming metric. This is a count of the number of positions in which two sequences of equal length differ (Hamming, 1950). In other cases an alignment cost is often computed (Waterman, 1995; Durbin *et al.*, 1998). For tertiary structures a minimised 'root mean square' (RMS) distance is generally used (Lesk, 1991).

Ways to compute metric distances between RNA secondary structures are less obvious. If the structures are inferred from sequences of the same length and are in dot-bracket notation then the Hamming metric can be used.

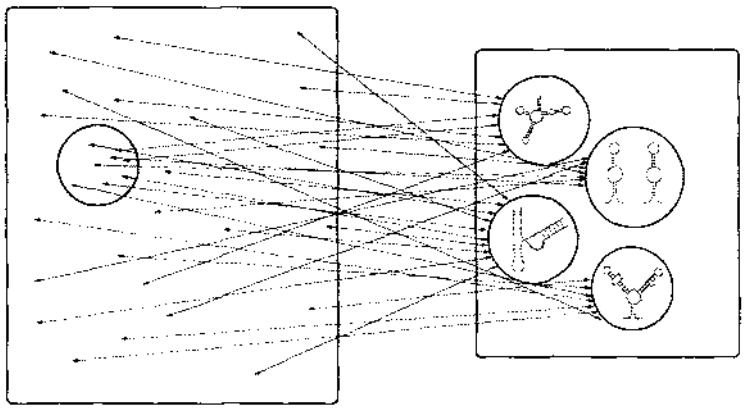


Figure 1.10: An outline of the map from RNA sequence to shape, any random sequence has a ball of neighbours which map to (almost) all common shapes (Schuster, 1993; Fontana, 2002).

A useful, but coarse metric, is the *base-pair metric*, this is defined as the cardinality of the symmetric difference of the sets of base-pairs in 2 secondary structures, i.e. the number of base-pairs the structures **do not** have in common.

Other popular methods to use are *tree metrics*, a tree representation of a secondary structure (T_1) is transformed into another tree (T_2) via a series of edit operations such as node insertion/deletion which have a predefined (often arbitrary) cost. A dynamic programming algorithm has been developed to compute the minimum cost of an edit-path between any two trees (Shapiro, 1988; Shapiro & Zhang, 1990). The advantage of tree metrics is that structures of different lengths can be compared. Additionally, coarse-grained structures can be used if global rather than local properties of shape-space are more relevant. A disadvantage of using tree metrics is that they can be time consuming to compute for long sequences.

Another metric that has proved useful is the *mountain metric* (Moulton *et al.*, 2000*b*; Edvardsson *et al.*, 2003), these are essentially RMS values computed between the mountain plot representations of different secondary structures. For example, the mountain metric $d_M(A,B)$ between two structures A and B with corresponding mountain plot values y_i^A and y_i^B, $(1 \leq i \leq N)$ is given by:

$$d_M(A,B) := \sqrt{\sum_{i=1}^{N}(y_i^A - y_i^B)^2} \tag{1.4}$$

A flaw in this method is that additional external base-pairs are penalised significantly more than extra internal base-pairs. A "fix" which is supposed to buffer

the effect of this problem is proposed in Moulton *et al.*, (2000), where a modified mountain plot is used to calculate the metric (compare equations 1.1&1.5). Unfortunately, this seems to make little to no difference in many cases. For example, if A =(((((...))))) and A_{in} =((((.....)))) is the same structure as A except with the innermost base-pair deleted and A_{out} =.((((...)))). is A except the outermost base-pair is deleted, then the percent difference between $d_M(A, A_{in})$ and $d_M(A, A_{out})$ in both the fixed and unfixed versions is 53.6%. Ideally, a secondary structure metric (such as the Hamming and base-pair metrics) would return a percent difference between $d(A, A_{in})$ and $d(A, A_{out})$ of or near zero.

Advantages of the mountain metric are that it is fast and easy to compute, easy to implement, and rescaling can be used to compare structures of different lengths (Edvardsson *et al.*, 2003).

$$\begin{aligned} y_0 &= 0 \\ y_{k+1} &= \begin{cases} y_k + \frac{1}{|k-l|} & \text{if } k \cdot l \text{ is a base-pair and } k < l. \\ y_k - \frac{1}{|k-l|} & \text{if } l \cdot k \text{ is a base-pair and } k > l. \\ y_k & \text{otherwise.} \end{cases} \end{aligned} \tag{1.5}$$

1.2.4 Prediction of RNA Secondary Structure

Preliminaries

A current problem in bioinformatics is to computationally determine the secondary and tertiary structure of any RNA sequence. Secondary and the resultant tertiary structures determine the function of the molecule, as has been shown by crystallographic studies of $tRNA^{Phe}$ (Kim *et al.*, 1974).

The folding of the one-dimensional primary structure into the three-dimensional tertiary structure can decomposed into two steps:

1. Formation of the RNA secondary structure by the Watson-Crick base-pairings, G≡C, A=U, and the weaker G-U pairs.

2. Folding of the planar secondary structure into a three-dimensional tertiary structure in the presence of divalent metal ions such as Mg^{2+}.

The driving-force behind secondary structure formation is the stacking of base-pairs. The formation of an energetically favourable stack base-pair (helical) region, however, also implies the formation of an energetically unfavourable loop region. This "frustrated" energetics leads to a vast number of helix and loop arrangements (Wuchty *et al.*, 1999).

Secondary structure conservation is sufficient for maintaining an active tertiary structure. This is supported by well documented conservation of secondary structure in evolution (Grüner *et al.*, 1996; Sankoff *et al.*, 1978; Zuker & Le, 1990). Thus a good point to begin the study of RNA structure is at the level of secondary structure.

There are two dominant methods for inferring the secondary structure of RNA: *Comparative sequence analysis* (Woese & Pace, 1993) and the *minimisation of free energy* (Zuker, 2000). Hybrids of these two methods have also been successful for some examples (Witwer *et al.*, 2001; Hofacker *et al.*, 2002).

Comparative rather than energy methods, are more robust for large RNA molecules (Akmaev *et al.*, 2000; Le & Zuker, 1991). However, using phylogenetic information to predict the secondary structure of RNA relies upon a sequence alignment and knowledge of the consensus secondary structure of homologous RNAs. This alignment step often requires labour intensive manual intervention (Woese *et al.*, 1983).

Often homologous RNA sequences (or structures) are unavailable, thus rendering comparative techniques impotent. Hence the need for energy minimisation, which currently relies upon thermodynamic parameters and dynamic programming (or stochastic context free grammars), to determine minimum and near minimum free energy secondary structures.

Secondary structure prediction for a single RNA sequence

Maximum Base-pairs: There are many ways to infer a secondary structure from a given RNA sequence, one of the simplest methods is to maximise the number of base-pairs (Waterman, 1995). Unfortunately the solution is not usually unique. For example, the phylogenetically inferred secondary structure of the 77 nucleotide histidine tRNA ($tRNA^{his}$) has 22 base-pairs, but Waterman's algorithm finds 149126 different secondary structures with 26 base-pairs (Wuchty *et al.*, 1999; Fontana, 2002).

Minimum Free Energy (MFE): Since RNA molecules comply to the laws of thermodynamics, it is theoretically possible to deduce the structure of an RNA molecule from its sequence by locating the conformation with the lowest free energy. The advantage of this approach is that it does not require a multiple sequence alignment.

Experimentally derived energy parameters are available for the contribution of an individual loop as a function of its size, its delimiting base-pairs, and the sequence of the unpaired bases. These are usually measured for $T = 37^oC$ and $1M$ sodium chloride solutions. For the base-pair stacking the enthalpic (ΔH) and

entropic (ΔS) contributions are known separately. Contributions from all other loop types are assumed to be purely entropic. This allows one to compute the temperature dependence of the free energy (ΔG) contributions:

$$\Delta G_{stack} = \Delta H_{37,stack} - T\Delta S_{37,stack}$$

$$\Delta G_{loop} = -T\Delta S_{37,loop}$$

Parameter Estimation: From studies of oligoribonucleotides Tinoco *et al* (1971) were among the first to experimentally estimate thermodynamic parameters for RNA secondary structure prediction . These values have recently been updated (Antao & Tinoco, 1992; Mathews *et al.*, 1999). An alternative to using laboratory techniques to determine thermodynamic folding parameters, is to use known biological structures, and vary algorithm parameters until the correct secondary structure is predicted (Papanicolaou *et al.*, 1984).

It is not feasible to study the thermodynamics of every possible sequence and compatible structure, so a simplified model is necessary to estimate the folding properties of all sequences from data obtained using a limited number of sequences. The most popular thermodynamic model is the nearest neighbour model. Since hydrogen bonding and stacking are both short range interactions, the stability is assumed to depend only on the identity of the adjacent pairs (Borer *et al.*, 1974). This model has been experimentally validated as a reasonable approximation for many cases (Kierzek *et al.*, 1986). The free energy contribution of loops and bulges are more difficult to estimate. Originally these regions were assumed to be solely dependent upon the number of unpaired nucleotides the loop contained. However the discovery of unusually stable tetra-loops (4-base hairpins) `GNRA` and `UNCG` have shown that this model needs expanding (Antao & Tinoco, 1992).

In 1981 M. Zuker and P. Stiegler published a recursive algorithm for obtaining an optimal folding for "large" RNA sequences, the algorithm utilises thermodynamic parameters . It starts a systematic search in all sub-fragments for the lowest free energy structure containing at least one base-pair. The first pass will calculate the MFE structures of all possible subsequences of length 5, the second pass will use these previously calculated values to find the MFE of all subsequences of length 6. The algorithm is incremented until the MFE structure of the entire sequence has been calculated. The lowest free energy structures are calculated, for each fragment, with and without the constraint that the terminal nucleotides are paired, and stored in the matrices V and W.

$$W(l,j) = \begin{cases} 0, & \text{if } j-l \leq 3. \\ \min(W(l,j-1), \min_{l \leq i \leq j}(V(i,j) + W(l,j-1))) & \text{if } 3 < j-l \leq n. \end{cases} \quad (1.6)$$

$$V(i,j) = \begin{cases} \min(H(i,j), S(i,j) + V(i+1,j-1), B(i,j), M(i,j)), & \forall\ i < j. \\ \infty, & \forall\ i \geq j. \end{cases} \quad (1.7)$$

Where $H(i,j)$, $S(i,j)$, $B(i,j)$, and $M(i,j)$ are functions that return free energy values dependent upon whether the base-pair between nucleotides i and j closes a hairpin loop, a stacked pair, a bulge/internal loop or a multi-loop respectively. This algorithm has been implemented in `MFOLD` (Mathews & Turner, 1999; Zuker *et al.*, 1999) and `RNAfold` (Hofacker *et al.*, 1994).

The advantage of this algorithm over other methods is speed, but to keep the algorithm computationally feasible several simplifications had to be made. For example, tertiary interactions such as pseudo-knot formation are ignored and only "nearest-neighbour" interactions are evaluated. A recent algorithm utilising a recursive stochastic context free grammar (SCFG) developed by Elena Rivas and Sean Eddy is now available which predicts pseudo-knots for short ($\leq$ 100 nucleotides) RNA sequences (Rivas & Eddy, 2000*b*).

There are some problems associated with using MFE methods to predict RNA secondary structure. For example, the energy parameters which the folding algorithm relies upon are inevitably imprecise. Hence, the true MFE structure might be one that is suboptimal with respect to the parameters used. In addition the biological structure may not be the MFE structure due to unknown constraints that may change relative energies, turning an otherwise suboptimal structure into the favourable one. Also, under physiological conditions RNA sequences may exist in alternative states whose energy difference is small (Giegerich *et al.*, 1999).

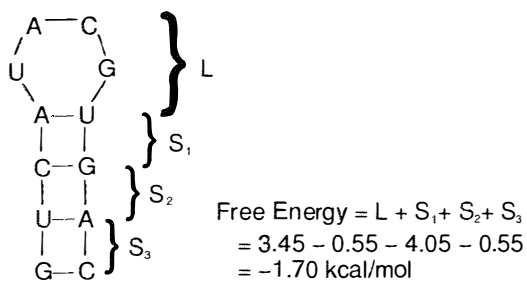


Figure 1.11: An example of how the free energy of an RNA sequence and structure is computed.

Secondary structure prediction using an alignment of RNA sequences

Background: Comparative sequence analysis has been immensely successful in predicting secondary structure of functional RNA molecules. It was used to predict the structure of tRNAs, rRNAs, and a number of ribozymes (Dirheimer *et al.*, 1995; Pace *et al.*, 1989; Woese & Pace, 1993). This form of analysis can be used to predict secondary structure pairings, and even some tertiary interactions. Many results are later verified when the structure of the molecule is solved using X-ray diffraction or nuclear magnetic resonance (NMR). For example the crystal structure of yeast Phe-tRNA when solved by X-ray diffraction verified all predicted secondary structure interactions (Kim *et al.*, 1974). More recently the structure of P RNA has been predicted using this approach (Parsch *et al.*, 2000). Several new structural elements such as pseudo-knots, non-canonical pairings and tetra-loops were proposed on the basis of comparative analysis, and have been substantiated by high-resolution experimental methods (Antao & Tinoco, 1992).

Constraints upon secondary structure can be revealed by compensatory base changes. When a point mutation occurs in a double stranded region (which will generally be disadvantageous to the organism) then the process of natural selection ensures that those organisms preserving the original (better) secondary structure either through a reversal or compensatory mutation will have a better chance of surviving to the next generation (Parsch *et al.*, 2000).

Mutual Information Content: In order to detect compensatory mutations the *Mutual Information Content* ($H(m,n)$) of two columns m and n in an alignment is calculated. The frequencies of each base ($B_i \in \{\texttt{A},\texttt{U},\texttt{C},\texttt{G}\}$) in column m, $f_m(B_i)$ and column n, $f_n(B_j)$ are counted, i.e. $\{f_m(\texttt{A}), f_m(\texttt{U}), f_m(\texttt{C}), f_m(\texttt{G})\}$. Additionally, the joint frequencies of each combination of two nucleotides in positions m and n (and row k) of the alignment are also counted, $f_{m,n}(B_i, B_j)$. If the base frequencies in any two columns are independent of one another, then h_{mn} is expected to equal 1, otherwise h_{mn} is expected to be greater than 1. Where, h is:

$$h_{mn} = \frac{f_{m,n}(B_i, B_j)}{f_m(B_i) \times f_n(B_j)}. \tag{1.8}$$

If columns m and n are perfectly covarying then $f_{m,n}(B_i, B_j) = f_m(B_i) = f_n(B_j)$ and if columns m and n are completely independent then $f_{m,n}(B_i, B_j) = f_m(B_i) \times f_n(B_j)$. The mutual information content between columns m and n in bits is thus defined as:

$$H(m,n) = \sum_{B_i,B_j} f_{m,n}(B_i, B_j) \times \log_2\{h_{mn}\}. \tag{1.9}$$

The terms of $H(m, n)$ are 0 when there is no correlation between columns m and n and increase (in absolute value) depending upon how well correlated m and n are and upon the number of sequences in the alignment (Durbin *et al.*, 1998; Mount, 2001; Eddy & Durbin, 1994; Gorodkin *et al.*, 1997).

RNAalifold, A Hybrid Method: Recently a group based in Vienna had some success with combining MFE and comparative approaches to secondary structure inference upon "sufficiently divergent" RNA sequences (Hofacker *et al.*, 2002) (see figure 1.2.4). The basic approach Hofacker *et al.*, (2002) have implemented is to condense an input alignment to a consensus sequence and use standard MFE folding upon the consensus sequence, the difference is in the evaluation of the cost function, bonus scores are given to those base-pairs where compensatory mutations have occurred. The method works well for RNA alignments which are sufficiently close for them to be aligned using standard algorithms (`CLUSTALW` for example (Thompson *et al.*, 1994)), but divergent enough that compensatory mutations can be observed in the alignment.

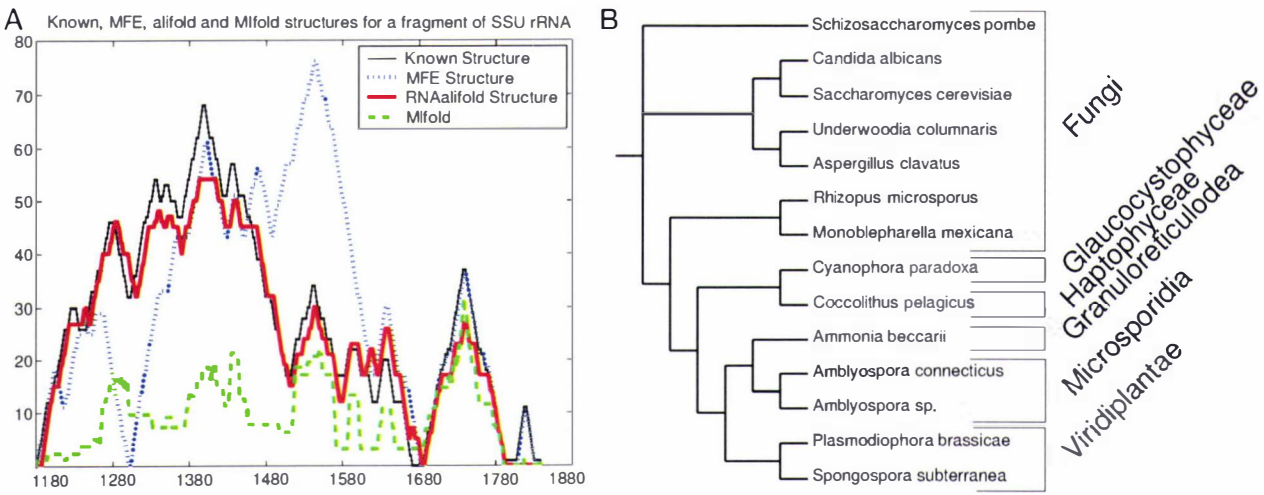


Figure 1.12: **A:** Mountain-plots of secondary structures that were inferred using a variety of techniques for a fragment of *Saccharomyces cerevisiae* SSU rRNA (nucleotides 1179 to 1806). The minimum free energy secondary structure was inferred using `RNAfold` (Hofacker *et al.*, 1994). An alignment of 13 SSU rRNA, downloaded from "The rRNA WWW Server" (`http://rrna.uia.ac.be/index.html`), was used to infer secondary structures using `RNAalifold` (Hofacker *et al.*, 2002) and the mutual information based method Mlfold (see pages 79 & 80). **B:** A phylogeny produced using neighbour-joining (Thompson *et al.*, 1994) of the taxa used to infer the secondary structures.

CHAPTER 2

Genetic Alphabet Size and the RNA-world

"We not only want to know how nature is (and how her transactions are carried through), but we also want to reach, if possible, a goal which may seem utopian and presumptuous, namely, to know why nature is such and not otherwise."

–A. Einstein. (translated by A. Eschenmoser)

2.1 Context, Overview and Preliminary Results

This project attempts to answer the question- Why is a 4-letter genetic alphabet (`ATCG` in DNA, `AUCG` in RNA) used by all modern life on Earth? There is no obvious optimum at 4 and alternative alphabets are certainly possible (Piccirilli *et al.*, 1990; Szathmáry, 1991; Szathmáry, 1992; Mac Dónaill, 2002) To pursue this question further we use the concept of an RNA-world, within which RNA performed the major coding and catalytic roles of life (see page 6), and assume the 4-letter alphabet became fixed at this developmental stage.

I was the principal author of the manuscript in this chapter "Optimal Alphabets for an RNA-world". It discusses simulations and results produced by myself (Paul Gardner) in collaboration with Barbara Holland, Vince Moulton, Mike Hendy and David Penny who have each operated largely in an advisory role. The manuscript has undergone the review process and been accepted for publication with minimal adjustments in "*Proceedings of the Royal Society of London, series B*". I begin this chapter with an extended discussion of the simulations used to investigate optimality of different nucleotides.

2.1.1 Simulation 1: Statistical Measures of RNA Secondary Structure

Motivation:

The work of Schultes *et al.*, (1999) and Seffens and Digby (1999), suggests that functional RNAs (fRNAs) have lower free energy than one would expect by chance. Encouraged by this we thought it would be interesting to investigate the free energies of randomly generated RNAs from different alphabets. In particular, Schultes *et al.*, (1999) studied statistics such as fraction pairing (P), the Shannon entropy (Q), and average stem length and observed a statistically significant difference between fRNAs and their shuffled counterparts for each of these statistics. These three statistics are included in the following study along with a few of our own measures of secondary structure.

Later research by Workman & Krogh (1999) and, Rivas & Eddy (2000) revealed that the earlier claims by Schultes *et al.*, (1999) and Seffens and Digby (1999) were flawed due to the shuffling techniques employed. They found that fRNAs that have been randomised using a shuffling routine that preserves the di-nucleotide frequencies of the original RNAs have no significant difference between the free-energy distributions of the fRNAs and randomised sequences. This is due to the fact that algorithms for inferring secondary structure by minimising the free-energy, evaluate energies in the di-nucleotide domain (see figure 1.11 for an illustration). Hence, shuffled sequences with different di-nucleotide frequencies to their un-shuffled counterparts show significantly different distributions of free-energy dependent statistics (Workman & Krogh, 1999). However, useful results can still be gleaned from such studies.

Algorithm description and implementation

A modified version of `RNAfold`[1] is used here; `RNAfold` is distributed with the `Vienna v1.4` package. A feature of this package that was found to be useful for this work is that it can infer secondary structures for sequences derived from artificial alphabets {`A,B,C,D,E,F,`$\cdots$} where `A` pairs `B`, `C` pairs `D`, `E` pairs `F`, etc. A global variable "`energy_set`" is used to determine the energy parameters used for each pair `energy_set=1` uses `G:C` energies for each base-pair, `energy_set=2` uses `A:U` energies and `energy_set=3` alternates between `G:C` and `A:U` energy assignments. The default is `energy_set=0` for sequences generated from the `AUCG`$\kappa\chi$ alphabet where `A` pairs `U`, `C` pairs `G`, the artificially synthesised nucleotides κ and χ base-pair

[1]See Appendix II for information regarding software

(Piccirilli *et al.*, 1990) and a wobble `G` with `U` pair is also allowed. 1000 sequences of length 100 were generated with a uniformly distributed nucleotide composition for canonical and non-canonical alphabets, each was folded using `RNAfold`, and a variety of secondary structure dependent measures were calculated from the output (see figure 2.1).

Statistical measures

The statistics considered here are: P, which is the fraction of paired bases within the optimal structure, S, the average length of the stems within a structure, and the free energy (MFE) of the inferred structures. P, S and MFE are all measures of the amount of pairing within the inferred structures. We have also used Q, the Shannon entropy (defined in equation 2.1) of the base-pairing probability matrix $[p_{ij}]$ which is calculated using the partition function (McCaskill, 1990). The value p_{ij} gives the probabilty that base i pairs with base j in an RNA sequence. The Shannon Entropy gives a measure of the number of alternative folds an RNA sequence can fold into. Also, F is used (known locally as the Gardner Uniqueness of Folding Function, $GUFF$) which is the Frobenius Norm of the base-pairing probability matrix (defined in equation 2.2). Q (and to a lesser extent F) are measures of how well defined a fold is. The p_{ij} values for a "well-defined" structure (in other words a structure with little conflict in its assignment) are near to either zero or one, hence the function $p_{ij} \log_2 p_{ij}$ is zero for well-defined base-pairs (one can use L'Hopital's rule to prove this). A large Q therefore indicates significant conflict within the structure assignment and a potentially erroneous minimum free energy structure assignment. These measures are discussed in more detail in the paper later in this chapter (Gardner *et al.*, 2003).

$$Q := \frac{-1}{Q_{max}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij} \log_2 p_{ij} \quad \text{where,} \quad Q_{max} = \frac{1}{2} N \log_2 N. \tag{2.1}$$

$$F := \sqrt{\frac{1}{N} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}^2 \right)}. \tag{2.2}$$

Results

The first thing we note when studying 'measures of secondary structure' versus 'alphabet size' plots is that there is no obvious optimum at 4 (see figure 2.1 & figure 1 in the attached chapter). In fact they each monotonically increase or decrease with respect to alphabet size. However there are features one should note. For example,

the amount of base-pairing decreases and the well definedness of the structures increases (see Q and F) with increasing alphabet size. Generally there is a significant difference between 2 and 4 letter alphabets, wheras the transition from 4 to 6 is considerably less significant.

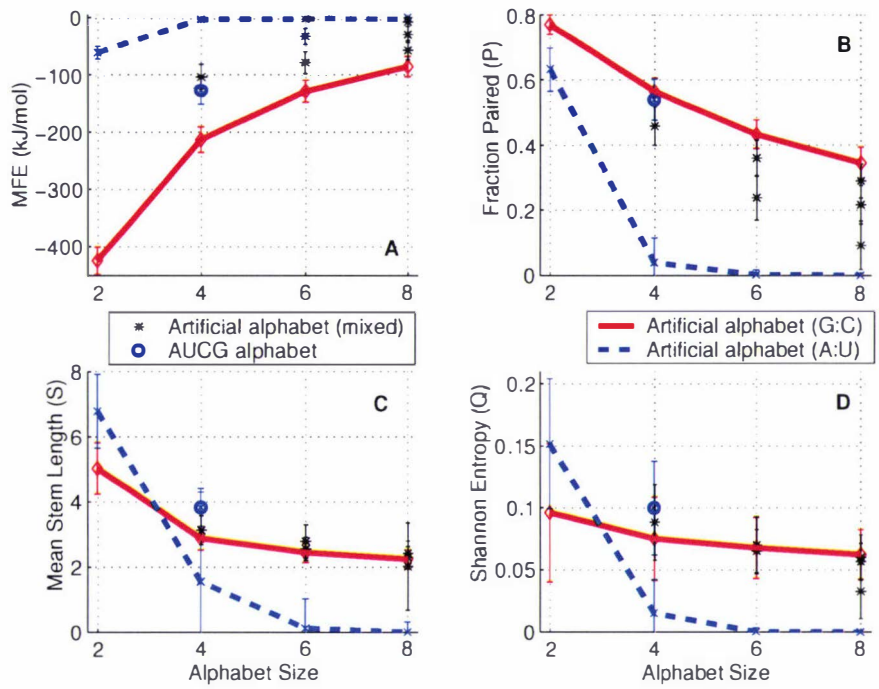


Figure 2.1:
Average minimum free energy (MFE) (A), fraction of paired bases (B), mean stem length (C) and Shannon entropy (D) values were calculated from 1000 random sequences generated by different alphabets and folded using a variety of different energy parameters.
Mean values for sequences folded using only G:C energy parameters are connected with a solid red line, those folded using only A:U energy parameters are joined by a dashed blue line, folds using a mix of A:U and G:C parameters are shown with a black "*", the canonical alphabet in each case is shown with a blue "o".

2.1.2 Simulation 2: `Revolver`

Motivation

The previous section explored statistical measures of RNA secondary structures, which could be thought of as exploring phenotype differences of alphabets in a putative RNA-world. But perhaps a more relevant question is: "What are the abilities of different RNA coding regimes to evolve in an RNA-world?" To explore this fur-

ther we construct a modified "flow-reactor" (Schuster, 2000; Gardner *et al.*, 2003). A slight difference between the flow-reactor of Schuster (2000) and our algorithm (dubbed `Revolver`) is in the expression of the fitness of individuals in the population of RNAs. Inside the Schuster (2000) flow-reactor, the fitness of an individual determines its replication rate, whereas in `Revolver` the fitness of the individual determines its likelihood of survival to the next generation. Also, the Schuster (2000) flow-reactor is initialised with just one randomly generated sequence, `Revolver` is initialised with a number of random sequences. Even so, the empirical behaviour of these two systems appear to be quite similar.

Algorithm description and implementation of `Revolver`

An initial population of RNA sequences is randomly generated from an alphabet of 2, 4, 6, or 8 letters. This is followed by successive rounds of amplification (with point mutations) and selection. The selection method is biased towards those RNAs folding into shapes "close" to a predefined target shape. This results in a population of sequences with MFE structures "close" to the target shape.

This system models SELEX (systematic evolution of ligands by exponential amplification) laboratory experiments, whereby RNA aptamers binding specifically to another molecule are artificially evolved. This is accomplished through successive rounds of the reverse transcriptase polymerase chain reaction (RT-PCR), **amplification**, and affinity chromatography, **selection**, upon a population of RNA sequences (Tuerk & Gold, 1990).

1. **Initialisation:** `Revolver` is initialised with a randomly generated population of RNA sequences. Each sequence is created in a character (nucleotide) by character fashion, each time a character is required it is randomly selected using a random number generator (`rand()` in `C`) from the alphabet string.

   Structures are inferred using `RNAfold` from each of the nascent sequences for later fitness evaluation. The resultant random population then undergoes an amplification-selection step to generate the population at generation 1.

2. **Amplification:** Prior to a selection step $\mathcal{AF}$ (amplification factor) copies of each sequence are made. The algorithm used here copies each sequence character by character, each time a character is copied there is a probability ($\mathcal{P}$) that a point mutation occurs. If a mutation does occur then a **different** character is randomly selected from the alphabet.

3. **Fitness evaluation:** Secondary structure metrics are employed to calculate $\mathcal{F}_i$ as detailed in equation 2.3. The function $d_{metric}(S_{target}, S_i)$ returns a met-

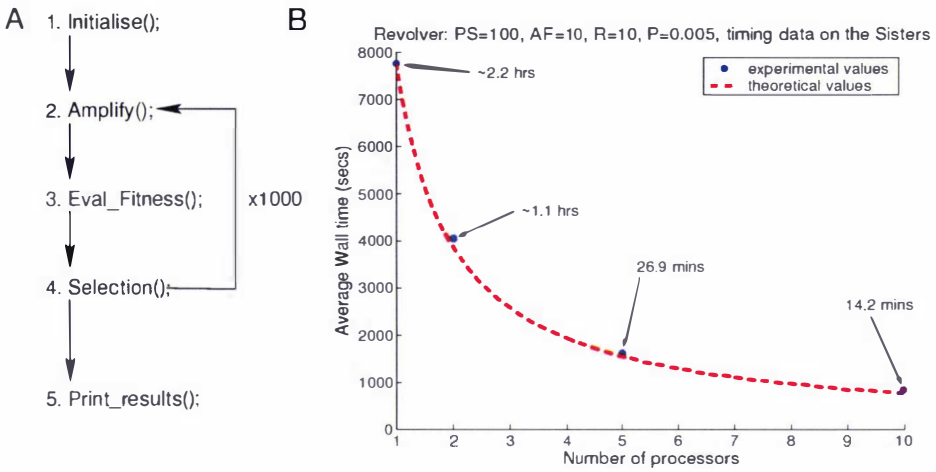


Figure 2.2: `Revolver`: **A:** A flowchart of the `Revolver` algorithm (see the text for a detailed discussion of steps 1-5).
B: A plot displaying averages from 4 timed trial runs of `Revolver`. Each run timed 10 repeats of `Revolver` on 10, 5, 2 and 1 nodes of a 16 node Beowulf computer[3]. The population size was 100, the amplification factor 10, the per site probability of mutation 0.005, and the number of generations 1000. Times in human readable format are shown within the plot. This result shows that `Revolver` scales well with the number of processors.

---

ric distance between secondary structures S_{target} and S_i. $d_{max}(S_{target}, S_i)$ is the maximum d_{metric} value between the structure S_i and fixed S_{target}. Hence if $S_i = S_{target}$ then $d_{metric} = 0$ and $\mathcal{F}_i = 1.0$ and if S_i is very different to S_{target} then $d_{metric} = d_{max}$ and $\mathcal{F}_i = 0.0$. Several metrics have been investigated and the base-pair metric was found to be the most discriminatory (see section 1.2.3) and therefore one we will use throughout this work.

$$\mathcal{F}_i = 1 - \frac{d_{metric}(S_{target}, S_i)}{d_{max}(S_{target}, S_i)} \tag{2.3}$$

Population fitness: At each generation the population fitness is measured by the "mode" of the individual fitnesses. The mode is used as the fitness distribution at each generation it is usually skewed to the left (see figures 2.4 & 2.5). The mean and also the median were too sensitive to outliers to be useful measures of centrality for this work.

4. **Selection:** The selection process used here is known in the genetic algorithm community as roulette wheel selection which is a "stochastic, proportionate method" that "samples with replacement" (Goldberg, 1989). Fitness values

$\mathcal{F}_i$ (see equation 2.3) are assigned to each sequence (i) in a population based upon how far the inferred structure was from the predefined optimum. Each sequence occupies a "slice of pie" (hence roulette wheel), of size $\mathcal{F}_i$ out of a total $\mathcal{F}_{tot} = \sum_{j=1}^{N} \mathcal{F}_j$. The probability of a sequence being selected is $P(i) = \frac{\mathcal{F}_i}{\sum_{j=1}^{N} \mathcal{F}_j}$. A uniformly distributed random number $R \in [0, \mathcal{F}_{tot}]$ is generated. If R lies between the cumulative fitnesses of the j^{th} and $(j+1)^{th}$ sequences, then i is selected. This is repeated until the required number of sequences for the next population-to-be-amplified has been filled (Holland, 1975; Goldberg, 1989; Man *et al.*, 1999).

Parallelisation strategy: The original single-node (1 CPU) version of `Revolver` which was written in the `C` programming language was time consuming to run (see figure 2.2), as the fitness evaluation phase requires up to 1000 sequences to be folded at each generation. This was highly dependent upon the mutation rate.

Therefore, a parallel version of `Revolver` was implemented upon a 16-node Beowulf cluster (see `http://sisters.massey.ac.nz`). Using the `C` programming language, the Message Passing Interface (`MPI`) and a "very robust queueing system" known as the Portable Batch System (`PBS`) to accomplish this (Kernighan & Ritchie, 1988; Message Passing Interface Forum, 1994; Henderson & Tweten, 1996).

A central dogma for saving computational time when computing in parallel is to minimise the amount of message passing between nodes (CPUs). The ideal situation is to construct an "embarrassingly parallel" implementation of code where each node in the system operates independently of the other nodes. An extreme variant of this approach is the "poor man's parallelism" approach, which entails multiple instances of a serial code running simultaneously on multiple nodes. This approach has no communication over-heads and will scale perfectly with the number of nodes.

When implementing parallel processes a master-slave arrangement is generally used as it comparatively simple to understand, debug and encode. The master-slave arrangement defines a single node (usually 0) as the master whilst all the other nodes (1-N) serve as slaves. The task of the master node is usually to divide a job up between the slaves and collate the resultant data, the slave nodes process their assigned jobs and send a signal to the master upon completion.

`Revolver` is amenable to a parallel implementation, there were several ways of doing this that were considered. At each generation the required number of fitness evaluation jobs could be farmed out to the slave nodes. The slaves would fold each received sequence and return the results back to the master node for the selection phase of the algorithm. However, since the average behaviour of the

different RNA coding regimes is of primary interest and `Revolver` is to be run several times for each parameter setting and results averaged over all the runs. Therefore it seems sensible for each node to calculate independent runs of `Revolver` and take an average over the runs from each node upon completion. This is then an "embarrassingly parallel" implementation of many `Revolver` runs. A further step could have been made by implementing `Revolver` using a "poor man's parallel" however this would have resulted in serious problems whilst attempting to collate data from hundreds of repeats of the experiment.

Times: The timing data of this implementation displayed in figure 2.2 shows that as expected the embarrassingly parallel implementation of `Revolver` scales very well with an increasing number processors. The communication overheads are minimal as the nodes only ever communicate with the master node at the end of an entire run when data is returned for the print step of the algorithm.

Results

It is wise to explore the parameter space of `Revolver` before comparing the "evolvabilities" of the different RNA alphabets. The parameters of interest are: the optimal shape, the per-site-probability-of-mutation ($\mathcal{P}$), and the effect of different base-pair energy parameters on the system. I will assume that the larger the population size and amplification factor are the better the final results will be. But first let's study a single run of `Revolver`.

A single run of `Revolver`: For this experimental run of `Revolver` we have used the base-pair metric to evaluate the fitness, the clover-leaf was the target structure, the alphabet used to generate sequences was the canonical alphabet (`AUCG`), sequences were folded using `RNAfold` with the default energy parameters (`energy_set=0`), the population size was fixed to 100, the amplification factor was 10 and the per site probability of mutation ($\mathcal{P}$) was 0.005. The results of this run are displayed in figures 2.3, 2.4 & 2.5. Observe in figure 2.3 that the fitness evolution progresses in leaps separated by periods of no apparent progression towards the optimal structure. During the periods of no apparent progress, neutral and near-neutral mutations are accrued by the population. These provide a platform from which fitter mutants may spring. Much of the behaviour of a quasi-species is displayed by this data, whereby a population of mutants forms a "cloud" about a master sequence (or sequences) (Eigen *et al.*, 1988; Eigen *et al.*, 1989; Schuster, 2000).

It is interesting to consider the fitness distributions at generations 1, 630, 680 and 1000. Generations 630 and 680 are either side of an adaptive leap and generations 1 and 1000 at opposing extremes with respect to population fitness and

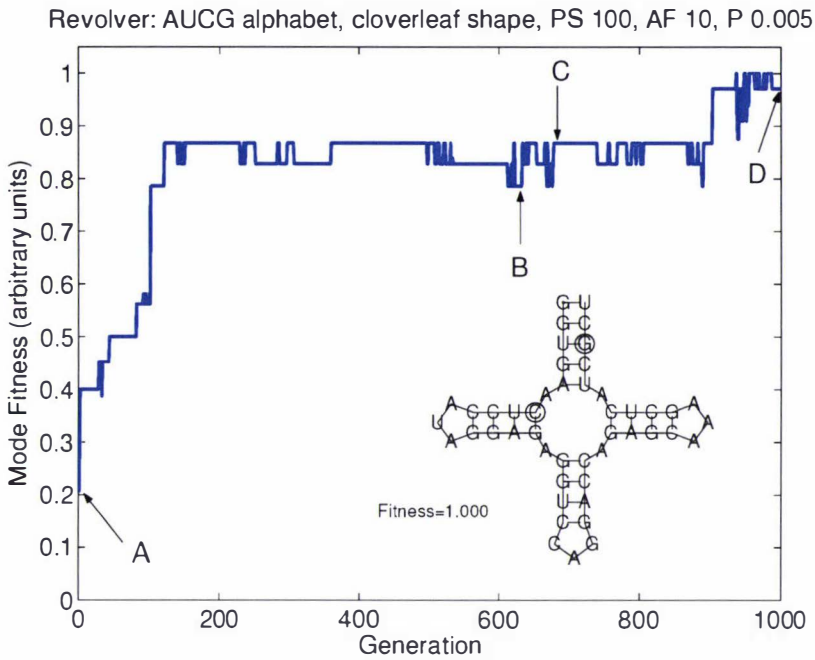


Figure 2.3: `Revolver`: The results of a single run. The target structure is the clover-leaf (an example "evolved" in this run is shown above), the population size was fixed to 100, the amplification factor was 10 for each generation, and the per site probability of mutation was 0.005. The plot shows how the mode fitness of the population of RNA sequences progresses as a function of generation in the flow-reactor.
The arrows labelled A, B, C, and D indicate where the frequency distributions of the fitness shown in figures 2.4 & 2.5 were obtained (generation 1, 630, 680 and 1000 respectively).

generation (see figure 2.3). These are shown in figures 2.4 & 2.5 and are discussed in detail below:

- *Generation 1:* The population of RNAs has only under-gone one selection step, consequently the resultant sequences are still near-random and the fitness distribution is near-normal, yet still slightly skewed to the left.

  On display in figure 2.4A are the three structures with fitness 0.207, which is the mode all the fitness values at this generation. Base-pairs in common with the target shape are indicated with arrows. Note that each structure has 3 base-pairs in common with the target and 9 extraneous base-pairs which are not-in-common with the target. As the target structure has 17 base-pairs the fitness of these structures is evaluated by the following:
  $\mathcal{F} = 1 - \frac{9+(17-3)}{(9+3)+17} \approx 0.207.$

- *Generation 630:* The distribution shown in figure 2.4B is approximately bi-

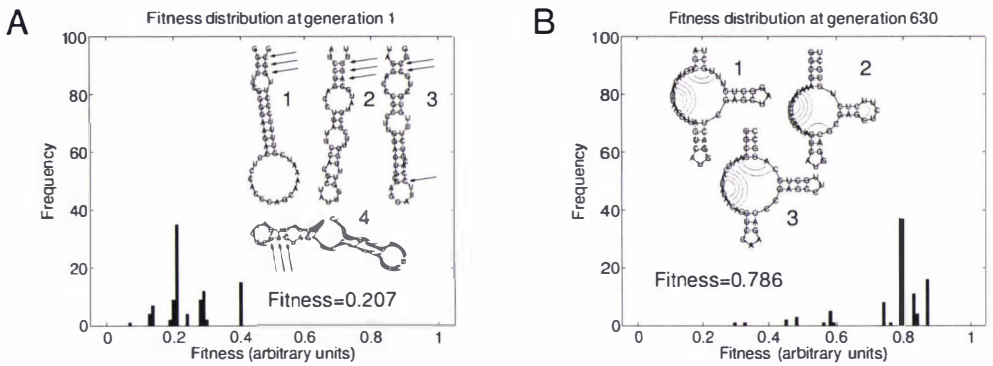


Figure 2.4: Plots **A&B** show the fitness distributions of a single run of `Revolver` at generations 1 and 630 respectively (see figure 2.3). Inset into plot **A** are the 4 structures (numbered 1, 2, 3 & 4) which have the mode fitness (0.207) at generation 1, the arrows indicate base-pairs in common with the target structure. Likewise the 3 structures (numbered 1, 2 & 3) inset into plot **B** are the structures with the mode fitness (0.786) at generation 680, the arcs indicate the base-pairs required before these sequences have an optimal fitness of 1.000.

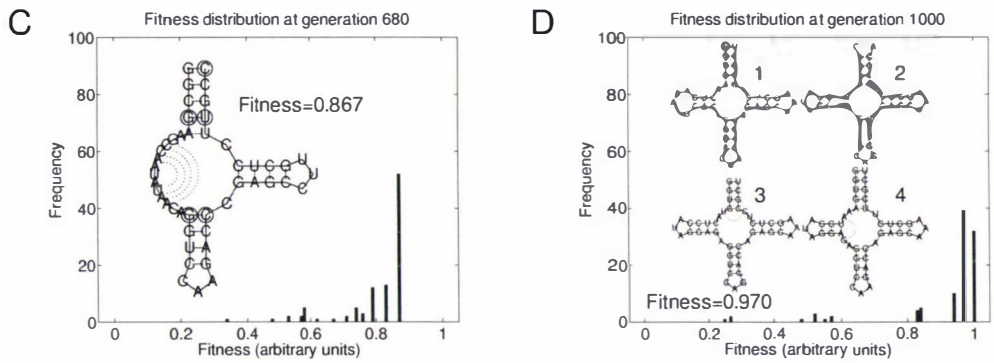


Figure 2.5: Plots **C&D** show the fitness distributions of a single run of `Revolver` at generations 680 and 1000 respectively (see figure 2.3). Inset into plot **C** is the structure which has the mode fitness (0.867) at generation 680. The circled bases indicate structure-neutral mutations within the population of sequences with this shape (care of `RNAalifold`). Likewise the structures inset into plot **D** are the structures with the mode fitness (0.970) at generation 1000. The arcs indicate the base-pairs required before these sequences have an optimal fitness of 1.000.

modal with the current dominant sequences with $\mathcal{F} = 0.786$ under pressure from sequences with $\mathcal{F} = 0.867$. There is a definite tail on the left of the distribution showing the "elderly" and recent deleterious mutations in the population, these will be eroded away with time.

Each of the three structures with $\mathcal{F} = 0.786$ at generation 630 are shown.

Each has 11 base-pairs in common with the optimal structure, but 6 base-pairs are still required (shown with dotted arcs) to map these structures to the target. The fitness of these structures was evaluated like so:
$\mathcal{F} = 1 - \frac{0+(17-11)}{11+17} \approx 0.786$.

- *Generation 680:* By generation 680 the competition between the sequences with $\mathcal{F} = 0.786$ and $\mathcal{F} = 0.867$ for space in the flow-reactor is inevitably being won by the latter sequences which have been more successful at being selected (see figure 2.5C). Although the former are still present in the population it is unlikely that they will return to their earlier levels.

  The unique structure with the mode fitness of 0.867 is shown. A total of 17 unique sequences fold into this structure, each of which was derived from the same common ancestor. As several sequences folding into the same structure were available, `RNAalifold` (Hofacker *et al.*, 2002) was used to infer the structure shown here. This program displays structure-neutral mutations (see the circled nucleotides in figure 2.5C). These mutations are indicated by the circled nucleotides. The structure has 13 base-pairs in common with the 17 base-pairs of the target structure, and has no extraneous base-pairs. The fitness is therefore:
  $\mathcal{F} = 1 - \frac{0+(17-13)}{13+17} \approx 0.867$.

- *Generation 1000:* The fitness distribution at generation 1000 (shown in figure 2.5D) is definitely skewed to the left. The mode fitness is 0.970 yet individuals folding into the target structure with maximal fitness of 1.000 are present.

  A total of four different structures mapped to the mode fitness of 0.970 at generation 1000. Each has 16 base-pairs in common with the clover-leaf target, thus requiring just one more base-pair before the full clover-leaf is obtained. The sequence folding into the clover-leaf structure is displayed in figure 2.3.

  It is of interest to note that the sequences in generations 630, 680 and 1000 share a common ancestor (structure 1 shown in figure 2.4A).

`Revolver` and various shapes: Figure 2.6 displays the results of several runs of `Revolver` with a variety of shapes as the target structure. We have used relatively short (length=47) sequences to keep the computation time low and avoided tetra-loops (hairpins of size four), many of which are given a bonus score by `RNAfold` with the canonical alphabet, which may confer a towards this this alphabet. We have used: a single-stemmed structure which maximises the number of base-pairs, a double stemmed structure which is reminiscent of the consensus H/ACA snoRNA

structure, a Y-shaped structure that is one of the simplest shapes with a multi-loop, and a clover-leaf shape which displays all the characteristics of the modern day tRNA (see figure 1.2).

The curve corresponding to the clover-leaf shape stands out considerably from the rest. The other three curves climb much more steeply over the first 100 generations with the single-stem being the steepest followed by the dual-stem then the Y-shape. Then each of these three curves reaches a plateau above which the mean population fitness does not rise. The height of each plateau for the Y-shape, dual-stem and single-stem is related to the number of base-pairs in the target (which is a factor in the evaluation of the base-pair metric). The clover-leaf curve in contrast has still not reached a plateau after 1000 generations, it is climbing steadily and will probably reach a plateau higher than the other structures after a 1000 more generations.

We selected the clover-leaf structure for further studies as it displays behaviour which is potentially useful for discriminating between different parameter settings (alphabet size and probability of mutation for example) and is related to the well known tRNA secondary structure.

`Revolver` and `RNAfold` energy parameters: Earlier results (see figure 2.1) show that direct measures of RNA secondary structure are very sensitive to varying the energy parameter selection and that these measures monotonically decrease (or increase) with alphabet size. Yet the `Revolver` mean population fitnesses for different energy parameters show quite different results (see figure 2.7). The fitness-values decrease rapidly with increasing alphabet size when no `CG` energy parameters are used (`energy_set=1`, see figure 2.1B). This implies that at least one "strong" (`CG` type base-pair) is required for any alphabet to evolve to complex phenotypes.

The alphabets that contain at least one base-pair with `CG` pairing strength (`energy_set={2,3}`, see figure 2.1C) show a dramatic increase in alphabet fitness for the transition from a two to four letter alphabet-size. Yet, the alphabet fitnesses do not alter considerably for the four to six alphabet-size transition. However, the alphabet fitnesses decrease in the six to eight transition. Therefore, an optimum exists for either a four or six letter alphabet size.

For `CG` energy-parameters the fitness peaks at the four letter alphabet size (best observed in figure 2.7C). Whereas, the peak shifts for the `AUCG`, `AUCGKX` and mixed (`AU+CG`) energy-parameters to the six letter alphabet size. These indicate that the focus of further investigations should be upon whether four or six letter alphabets were optimal in an RNA-world (see "Optimal Alphabets for an RNA-world", section 2.2 for further discussion and to observe the affects of varying mutation rates upon these results).

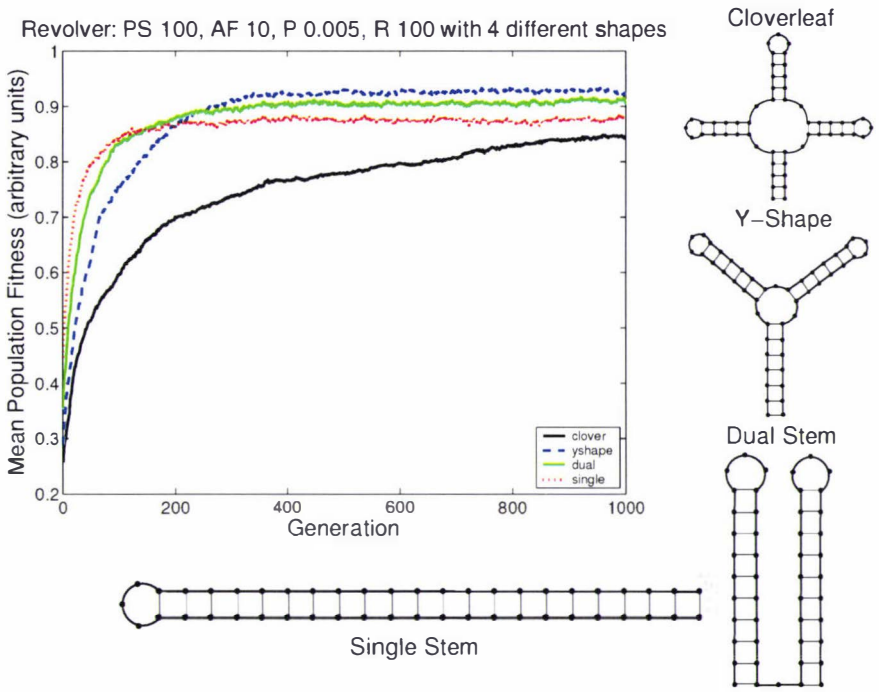


Figure 2.6: `Revolver`: the results of an average over 100 repeats of `Revolver` run for 1000 generations with 4 different target structures. The population size was fixed to 100, the amplification factor was 10 for each generation, and the per site probability of mutation was 0.005.

It seems that as a general rule-of-thumb the more "complex" structures are more difficult for an RNA population to evolve toward. As the clover-leaf was the most difficult shape (studied here) to achieve it was selected for further study. Hence we are assuming that active shapes in the RNA-world had a certain degree of "complexity", and that active shapes in the RNA-world were difficult to make. Additionally, the clover-leaf displays many of the characteristics of the modern day ncRNA known as tRNA.

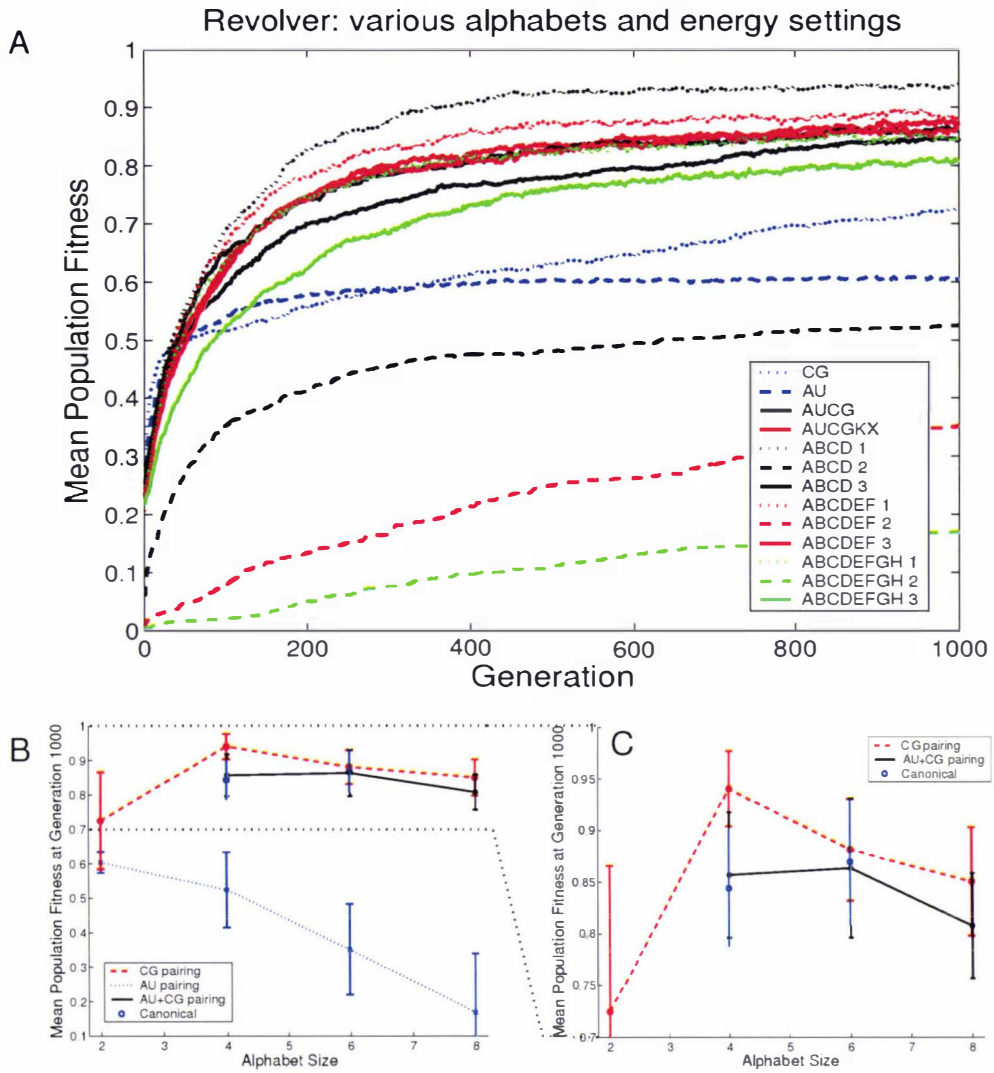


Figure 2.7: `Revolver`: the results of an average over 100 repeats of `Revolver` with several different energy parameters. The target structure was the clover-leaf, the population size was fixed to 100, the amplification factor was 10 for each generation, and the per site probability of mutation was 0.005.
Plot **A** displays the mean population fitness versus generation for several different alphabets and `RNAfold` energy settings.
Plots **B&C** show the mean population fitnesses as a function of alphabet size (with error bars indicating 1 standard deviation either side of the mean). A dashed red line connects the points using only `CG` energy parameters, the dotted blue line connects the points using only `AU` energy parameters. A solid black line connects the points using `AU+CG` energy parameters and a blue "o" shows values for the canonical alphabet `AUCG` and the `AUCGKX` alphabet. Plots **B&C** display the same data, the difference is that in plot **C** the `AU` energy parameter data is ignored. This displays the alphabet fitnesses at a better resolution.

2.1.3 Simulation 3: RiboRace

Algorithm description and implementation

Motivation: Interesting results may be gleaned by allowing two (or possibly more) alphabets compete for space in the flow-reactor. The advantages of this experiment are that an outcome can be determined in fewer generations than with Revolver, and it could also discriminate between the 4 and 6 letter alphabets at a better resolution than with Revolver.

The major differences between Revolver and RiboRace occur at the initialisation stage and in the parallel implementation strategy. The parallelisation is less CPU time efficient but provides a saving in human time for the data accumulation and analysis phase (compare figures 2.2A & 2.8A).

Initialisation: RiboRace, like Revolver, is initialised with a randomly generated population of RNA sequences. The resultant population consists of 50 sequences generated by each of the two input alphabets.

Parallelisation strategy: The strategy used to parallelise Revolver whilst providing excellent scaling and greatly increased run time for many repeats of Revolver was limited in a number of ways. Firstly, in order to keep the computational load balanced on each of the available nodes in the system, the number of times an experiment was run had to be a factor of the number of nodes. Secondly, collating data at the end of an experiment could be troublesome and time consuming. Either, a large amount of data could be passed back to the master node from the slaves, or the data could be dumped in a node specific file. Neither option was particularly appealing. Subsequently, I decided that trading algorithm speed for researcher time at the data analysis phase was worth the trouble of implementing a new algorithm. Now the required number of computationally intensive tasks (foldings) at each generation are farmed out by the master node to x slave nodes. From figure 2.8B note that the new strategy scales well with the number of nodes.

Results

For a broad range of mutation rates ($10^{-4} - 2 \times 10^{-2}$) the 4 letter AUCG defeats the AUCGKX in a race for space in the flow-reactor (see figure 2.9B). Then the mutational load becomes too great for the 4 letter alphabet at which time it no longer passes sufficient information to following generations. The 6 letter alphabet is more robust to this information melt-down. In fact, a higher mutation rate is advantageous to the 6 letter alphabet as paths to points in sequence space from a random start are longer, therefore "faster" movement through the space is advantagous (see figure 2.9A). At a mutation rate of ~ 0.4 the mutational load becomes so large that

no information is passed to the offspring. Consequently the population is essentially re-initialised at each generation. The 4 letter alphabet has a higher propensity for base-pairing than the 6 letter alphabet (see figure 2.1), consequently it is more likely to have base-pairs in common with the target structure than the 6 letter alphabet. Therefore the size of the 6 letter group is eroded over time.

These `RiboRace` results appear to contradict the earlier results gathered from `Revolver` (shown and discussed later in this chapter (Gardner *et al.*, 2003)). The `RiboRace` results suggest that a 6 letter alphabet is optimal when mutation rates are high, and that the 4 letter alphabet is optimal for low and extremely high mutation rates. Much of the mutation range shown in figure 2.9 were not explored with `Revolver`, however the results that were gathered suggest that the 6 letter alphabet should have been optimal in the mutation rate range 0.001-0.006 and from 0.006 upwards the 4 letter alphabet should dominate. It is likely that these results are an artifact of the `RiboRace` system (or perhaps buggy code, although this has been extensively studied), due to the relatively few generations (generally 15-30) before an alphabet dominates the system.

Future experiments for `RiboRace` are runs initialised with a sequences that fold into a fixed structure, no mixing of the alphabets for a fixed number of generations and a less stringent selection proceedure could be used. Once these results have been studied then a variety of different competing alphabets can be explored to determine optimal conditions for each.

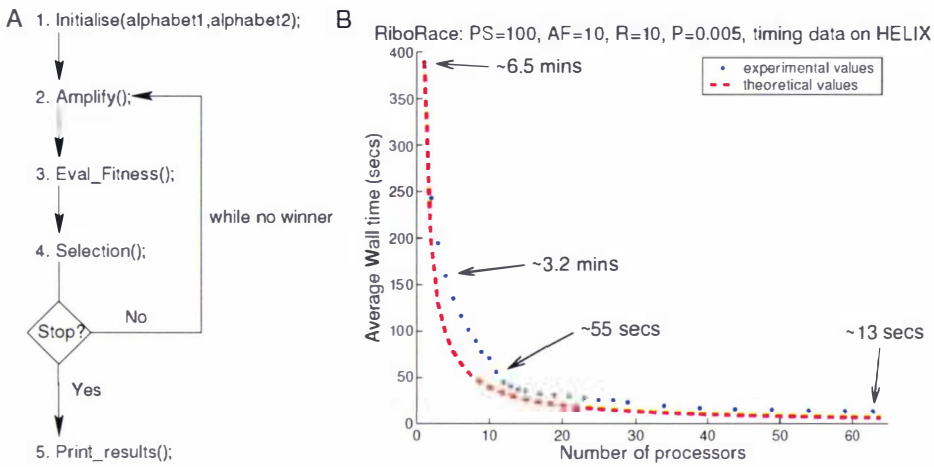


Figure 2.8: `RiboRace`: **A**: An outline of the `RiboRace` algorithm. The code is essentially the same as `Revolver`. The differences are that the population is initialised with two alphabets instead of one, and there is an extra **Stop** condition; **Stop** if one of the alphabets fills 100% of the flow reactor or if the generation limit of 1000 has been reached.

B: Timing data of a single run of `RiboRace` on the `HELIX` Beowulf computer (`http://helix.massey.ac.nz`), the clover-leaf was the target structure, the population size was 100 (50 strings generated from each of the `AUCG` and `AUCGKX` alphabets), the amplification factor was 10 and the mutation rate was 0.005, the experiment was repeated 10 times with different initial populations. Wall times were averaged over 3 runs. Times in human readable format are indicated within the plot.

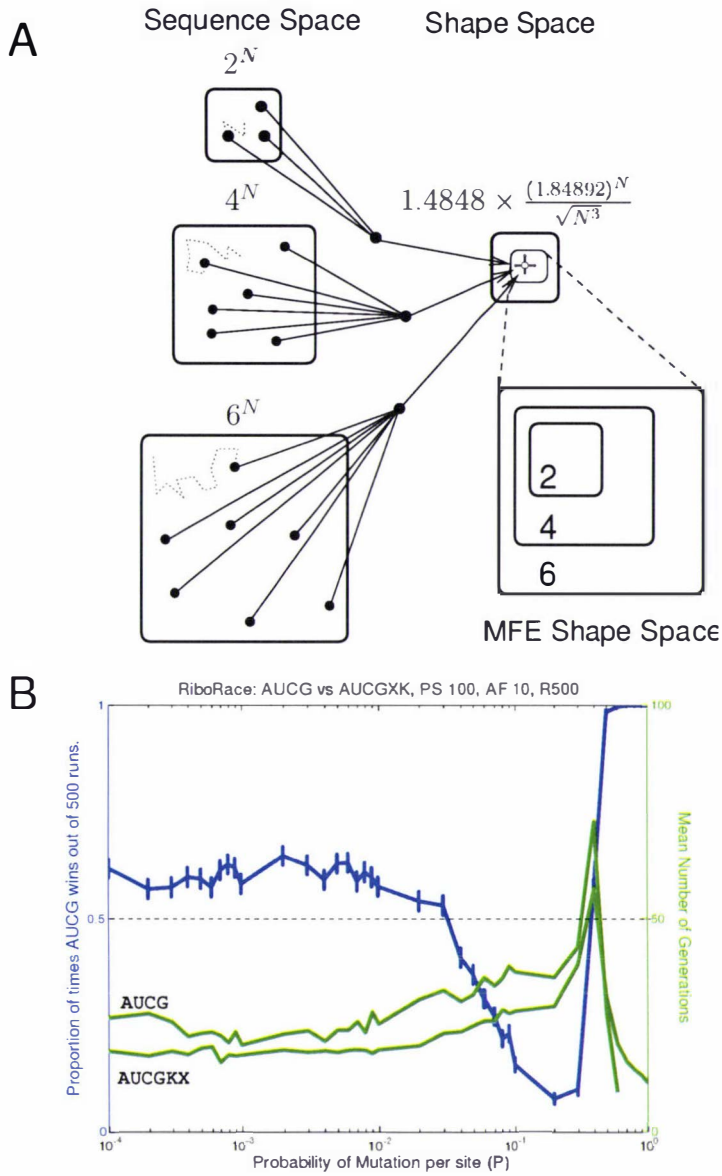


Figure 2.9: **A:** Sequences spaces of 2, 4, and 6 letter alphabets each map onto the same shape space. The cardinality of each space is shown as a function of sequence length (N). Exhaustive analysis of MFE structures show that little of shape space is actually realised (Grüner *et al.*, 1996). As the sequences generated from the 4 letter alphabet are a subset of sequences generated by the 6 letter alphabet, then all of the MFE structures for the 4 letter sequences can be realised by 6 letter sequences. Hence, MFE shape space for the 2, 4 and 6 letter alphabets can be viewed as a series of nested sets.

B: `RiboRace`: `AUCG` vs `AUCGKX`. The left axis and solid line shows the proportion of times the `AUCG` defeats the `AUCGKX` in a race to fill the flow-reactor from 500 repeats of `RiboRace`. The right axis and dashed lines shows the mean number of generations required for the 2 alphabets to defeat the other. The population size is 100, amplification factor 10, target shape the clover-leaf for and range of different per-site-probabilities-of-mutation are used.

2.2 *Paper 1,* Optimal Alphabets for an RNA-world

Author: Paul P. Gardner, Barbara R. Holland, Michael D. Hendy, Vincent Moulton, & David Penny

Year: 2003

Journal: *Proceedings of the Royal Society of London, Series B.*

Volume: 270

Number: 1520

Page: 1177-1182

THE ROYAL SOCIETY FirstCite® e-publishing

Received 11 October 2002
Accepted 27 January 2003
Published online

Optimal alphabets for an RNA world

Paul P. Gardner[1,3*], Barbara R. Holland[1,3], Vincent Moulton[4], Mike Hendy[1,3] and David Penny[2,3]

[1]*Institute of Fundamental Sciences, and* [2]*Institute of Molecular Biosciences, Massey University, PB 11 222, Palmerston North, New Zealand*
[3]*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Auckland, New Zealand*
[4]*Linnaeus Center for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden*

Experiments have shown that the canonical AUCG genetic alphabet is not the only possible nucleotide alphabet. In this work we address the question 'is the canonical alphabet optimal?' We make the assumption that the genetic alphabet was determined in the RNA world. Computational tools are used to infer the RNA secondary structure (shape) from a given RNA sequence, and statistics from RNA shapes are gathered with respect to alphabet size. Then, simulations based upon the replication and selection of fixed-sized RNA populations are used to investigate the effect of alternative alphabets upon RNA's ability to step through a fitness landscape. These results show that for a low copy fidelity the canonical alphabet is fitter than two-, six- and eight-letter alphabets. In higher copy-fidelity experiments, six-letter alphabets outperform the four-letter alphabets, suggesting that the canonical alphabet is indeed a relic of the RNA world.

Keywords: genetic systems; ribozymes; systematic evolution of ligands by exponential amplification (SELEX); RNA world

1. INTRODUCTION

For current models of the origin of modern life an obligatory step is an RNA world (Gesteland *et al.* 1999). This is a point in time when RNA was the predominant biomolecule and served both as the carrier of genetic information and as the primary catalyst for metabolism. These days coding is predominantly carried out by DNA, and metabolic processes by proteins. However, several key roles are still performed by RNAs, leading to the suggestion that some coding and metabolic aspects of current living systems are relicts of the RNA world (Szathmáry 1992; Poole *et al.* 1998; Jeffares *et al.* 1998).

In the RNA world the genotype and the phenotype are expressed in the same molecule. The genotype refers to the sequence of nucleotides within the molecule and the phenotype can be viewed as the specific three-dimensional conformation of the catalytically active functional RNA (fRNA). As a result, the 'fitness' of a ribo-organism can be inferred from the phenotype (Higgs 2000; Joyce 2000). In the case of RNA, the stabilizing forces conferred by the formation of a secondary structure are much greater than those conferred by the rearrangement of secondary structural elements in three-dimensional space (Grüner *et al.* 1996). This fact is supported by the well-documented secondary-structure conservation in fRNAs (Sankoff *et al.* 1978; Hofacker *et al.* 1998; Eddy 1999; Parsch *et al.* 2000). Thus, primary to secondary structure mappings are relevant for studies of evolution in the RNA world. However, an RNA can fold into many near-optimal secondary structures (Zuker 2000; Higgs 2000). In other words, the genotype does not specify a unique phenotype. Part of our study is to find the conditions under which RNA sequences are expected to fold into a small number of stable structures without the assistance of chaperones.

Based on our current knowledge, an RNA world would have been dominated by four coding nucleotides—the two pairs A : U and C : G (although other ribonucleotides may have also been involved as cofactors in catalysis; White 1976). A natural series of questions relevant to understanding the RNA world include the following.

(i) Are two pairs of nucleotides optimal?
(ii) Is there an advantage of a four-nucleotide system over two nucleotides (one pair, either A : U or C : G)?
(iii) If four is better than two, then is six better than four, and eight better than six?

Some researchers have used tools from organic chemistry to investigate the properties of non-canonical RNA systems (here we will refer to the AUCG alphabet as canonical and to the alternatives as non-canonical) such as those with differing nucleotide bases and alternative sugar groups (Rich 1962; Switzer *et al.* 1989; Piccirilli *et al.* 1990; Bain *et al.* 1992; Eschenmoser 1999). Computational experiments have been used to compare non-canonical RNA systems (Szathmáry 1991, 1992; Grüner *et al.* 1996). Szathmáry (1991, 1992) in particular has posed the question 'what is the optimal size for the genetic alphabet?' He concludes that the four-letter genetic alphabet is a 'frozen evolutionary optimum' that was determined in the RNA world. More recently Mac Dónaill (2002) investigated the optimality of the nucleotide alphabet in terms of error minimization using informatic techniques; he concluded that the canonical alphabet is one of the 'better' possibilities. We explore this question further in the context of maps from RNA primary

* Author for correspondence (P.P.Gardner@massey.ac.nz).

structure to RNA secondary structure and from RNA secondary structure to ribo-organism fitness.

In particular, we consider two new approaches to exploring the implications of different alphabet sizes. First, we investigate the average properties of secondary structures derived from both canonical and non-canonical RNAs, and, second, we study how random structures evolve towards some predefined structures for each of the alternative alphabet sizes. From the first investigation we discovered that, while different alphabets lead to secondary structures that have very different properties, beyond some obvious conclusions, it is difficult to determine exactly which of these properties would have been optimal for the RNA world. However, using a 'flow-reactor' simulation (Fontana & Schuster 1998) inspired by systematic evolution of ligands by exponential amplification (SELEX) experiments (Tuerk & Gold 1990), we show that in the RNA world the canonical four-letter (two base pair) alphabet outperforms the non-canonical alphabets under several sets of evolutionary conditions.

2. STATISTICAL MEASURES OF THE RNA SECONDARY STRUCTURE

We study the statistical properties of the 'molecular morphospace' (Schultes *et al.* 1999) of random RNA with respect to alphabet size. We use a modified distribution of Vienna v. 1.4, in particular the dynamic programming algorithm, RNAfold (available from www.tbi.univie.ac.at/~ivo/RNA/). This program uses empirically derived energy values to infer a minimum free-energy secondary structure from a single RNA sequence (Hofacker *et al.* 1994). Note that, while RNAfold may not always predict the 'exact' biological RNA secondary structure (Zuker 2000), we are interested in only the average behaviour of different RNA coding regimes, and, therefore, any inaccuracies inherent in this method are not expected to have a significant effect upon our results (Moulton *et al.* 2000*a*). Another important point is that using statistical measures to discriminate between different alphabet sizes is generally 'easier' than discriminating between evolved and random sequences, at least for the canonical alphabet (Rivas & Eddy 2000; Schattner 2002). Hence, for this work, fruitful results may be obtained by studying random sequences.

A feature of RNAfold that we exploit throughout this paper is that it allows the prediction of secondary structures for sequences generated from artificial alphabets; ABCD... (where A pairs B, C pairs D, and so on). We restrict our attention to alphabets with two, four, six and eight letters since, as discussed in Szathmáry (1992), there are only $2^3 = 8$ unique hydrogen-donor/acceptor configurations between any two complementary nucleotides and a nucleotide that can be either purine or pyrimidine, yielding 16 unique letters and eight nucleotides. Owing to time and space considerations, we will consider a maximum of only eight letters. Alphabets with an odd number of bases are not considered as this requires two different bases to compete for a complementary site during replication; the base with lower affinity would be lost after just a few generations. There are four energy-parameter options that RNAfold uses: the default (0) uses parameters for the canonical alphabet, otherwise it folds sequences generated by an artificial alphabet with (1) G : C, (2) A : U or (3) alternating G : C and A : U energy-parameter assignments for the base pairs AB, CD, ...

Statistics were gathered from 1000 randomly generated sequences of fixed length $N = 120$ for all possible RNAfold energy-parameter selections (see figure 1). The statistics shown are: P, the fraction of paired bases within the optimal structure; Q, the Shannon entropy of the base-pairing probability matrix (p_{ij} is calculated using the partition function; McCaskill 1990), defined by

$$Q := \frac{-1}{Q_{\max}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij} \log_2 p_{ij} \text{ where, } Q_{\max} = \frac{1}{2} N \log_2 N;$$

and F (known locally as the Gardner uniqueness of folding function), that is the Frobenius norm of the base-pairing probability matrix, defined by

$$F := \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}^2\right)}.$$

Q is intended to be a measure of how well defined the predicted secondary structure is—a high Q value indicates uncertainty in the structure assignment, with many alternative structures near to optimal, whereas a low Q value indicates a well-defined assignment. Note from the melting curves for Q that, as the temperature increases Q also initially increases until a threshold is reached when the number of valid structures starts to decrease until the only possible assignment is the completely unfolded structure (see Schultes *et al.* (1999) for further discussion). Unlike the Shannon entropy, F will distinguish between a 'well-folded' stable secondary structure and a completely unfolded molecule. From figure 1 we see that base pairing (P) decreases as the alphabet size increases. For example, the canonical alphabet has an average fraction of paired bases of 0.54 whereas the maximums for the six- and eight-letter alphabets were 0.43 and 0.34, respectively. This can be explained by the fact that there is a higher probability of any given base matching its complement with a smaller alphabet. Similarly, the measure of disorder, Q, is lower indicating a lower degree of structural uncertainty with increasing alphabet size.

Comparing the canonical alphabet and the ABCD alphabet with RNAfold energy parameter 3 we observe that the canonical alphabet has more base pairing, owing to the additional G : U pair that is allowed in the canonical system. As a consequence of this there is also more uncertainty (larger Q values) in the predicted structure.

Each statistic is sensitive to the energy-parameter selection to varying degrees, and is therefore also sensitive to the base composition, which we keep constant (on average). In particular note that a regime containing solely A : U pairing clearly has insufficient pairing potential to maintain RNA secondary structures (from random sequences) of any complexity. From this we note that any pairing regime must contain at least one pair of G : C type to be viable.

Earlier work in this area has shown that a four-letter alphabet would have been a significant improvement upon a two-letter alphabet (Fontana *et al.* 1993; Schuster 1993; Grüner *et al.* 1996). Although ribozymes generated from two-letter alphabets have proven to be adequate enzymes

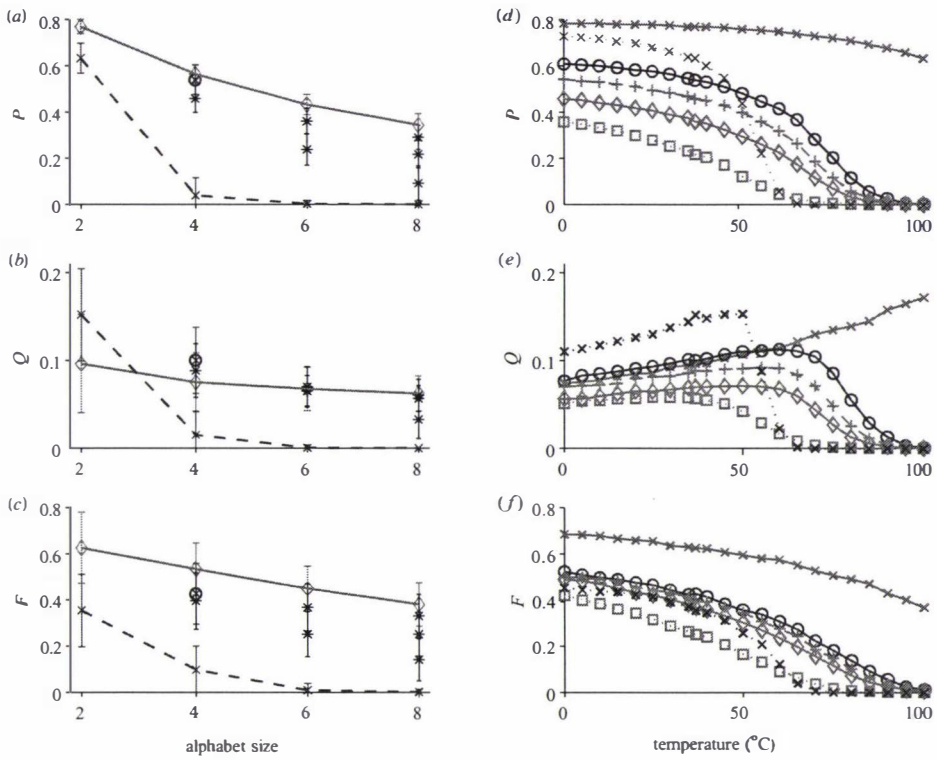


Figure 1. Average RNA secondary structure statistics from 1000 randomly generated sequences with respect to alphabet size (*a–c*) and melting curves (*d–f*). The sequence length is fixed to 120 nucleotides and the temperature is 37 °C (*a–c*). (*a–c*) The solid line connecting the diamonds corresponds to all base-pairs assigned an energy value equivalent to a G : C pair; the dashed line connecting the crosses used only A : U energy parameters. Asterisks, those artificial alphabets where a mixture of energy parameters was used; open circles, the canonical (AUCG) alphabet. (*d–f*) Crosses, points corresponding to the 2-letter alphabets AU (dotted line) and CG (solid line); open circles, the canonical alphabet; plus signs, the 4-letter alphabet with mixed energy parameters (no G : U base pairs); diamonds, the 6-letter alphabet with mixed energy parameters; squares, the 8-letter alphabet with mixed energy parameters.

(Reader & Joyce 2002), the results presented here show that the differences in the statistics between the two- and four-letter alphabets are marked, but the differences between the four- and six-letter alphabets are considerably less so. This suggests that a two-letter alphabet (if one ever existed) would have been rapidly outcompeted by a four-letter one. But does the same argument hold for the four- versus six-letter situation?

3. EVOLVING RNA *IN SILICO*

Having described some statistical properties of random sequences we now consider the impact of alphabet size on the ability of a population of sequences to evolve towards a predefined target structure. We compare the ways in which various non-canonical RNA systems evolve through a fitness landscape. To achieve this we constructed a modified 'flow reactor'. This is a stochastic discrete-time model, with capacity limited to a fixed number of sequences (Fontana & Schuster 1998). This system models SELEX laboratory experiments, where the goal is to artificially evolve RNA aptamers binding specifically to another molecule (Tuerk & Gold 1990). For SELEX experiments the final shape(s) is generally unknown; however, since it is difficult to infer computationally how well an RNA will bind to another molecule, a target structure is defined in advance. Then the probability of survival to the next generation is made a function of the distance to the target.

At generation zero the flow reactor is filled with a pool of randomly generated sequences; successive rounds of amplification with replication error, followed by selection, are used to generate 'evolved' sequences with a corresponding shape that is near some target structure (see figure 2*a*). The probability of selection is a function of the fitness (W), which is dependent upon the distance between the secondary structures of the individual (S_i) and the target (S_{target}). As a distance measure we use the base-pair metric ($d_{BP}(S_{target}, S_i)$), which is a count of the base pairs that two secondary structures (S_{target}, S_i) of equal length do not have in common (see Moulton *et al.* (2000*b*) for a technical description). When compared with

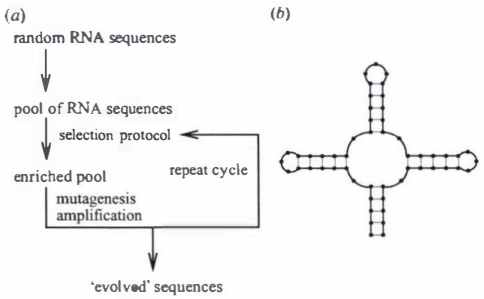


Figure 2. (*a*) An outline of the simulated flow reactor, adapted from Szostak (1993). (*b*) A biological representation of the target structure.

alternative metrics such as the hamming distance and mountain metric (Moulton *et al.* 2000*b*) the base-pair metric showed better discrimination between phenotypes within the evolving population. The metric is scaled to take values bounded by 0 and 1, with 1 being a perfect match to the target structure and 0 being a structure with no base pairs in common with the optimal structure:

$$W = 1 - \frac{d_{\mathrm{BP}}(S_{\mathrm{target}}, S_i)}{\max(d_{\mathrm{BP}}(S_{\mathrm{target}}, S_i))}.$$

The selected target, the clover leaf (see figure 2*b*), contained no tetra-loops (which may confer an advantage to the canonical alphabet of Antao & Tinoco (1992)), had a reasonable degree of 'structural complexity' (which would be important for function in the RNA world) and displayed many of the characteristics of a modern ribozyme. Our preliminary studies showed that the empirical behaviour of the flow reactor is largely independent of the energy parameter used (data not shown); this is in contrast to the data in § 2. For all the following experiments we use alternating A : U and C : G energy parameters for the artificial alphabets.

Fontana & Schuster (1998) note that the evolution of this system progresses in leaps towards the target structure, interspersed with periods of no apparent adaptive progress, during which neutral mutations are accrued before the next adaptive step. Here, we are more interested in how possible non-canonical RNA alphabets might perform against each other in an RNA world, so we average over many runs to eliminate the noisy effects of the adaptive leaps.

The function that infers a secondary structure is computationally intensive ($O(N^3)$, where N is the sequence length), and is called many times. To ensure that the computations were completed in a timely fashion we constructed an implementation of the flow reactor that ran on 10 nodes of an Intel-based Beowulf Cluster (code available upon request).

4. COMPUTATIONAL RESULTS

We define the population fitness ($\bar{W}(i)$) as the modal fitness of all the individuals in a population at generation i. Figure 3*a* was generated by taking the mean $\bar{W}(i)$ of 100 runs of the flow reactor for the two-, four-, six- and eight-letter alphabets. The flow reactor was run for 1000 generations and the probability of mutation at each site during replication was 0.01. Inset is a plot of the mean population fitness at generation 1000 as a function of alphabet size showing an optima for the canonical alphabet for this particular parameter set.

The data in figure 3*b* were generated by collecting the mean population fitnesses at generation 1000 for mutation rates ranging from 0 to 0.01. Observe that the four-letter alphabets outperformed the alternatives for high mutation rates. But, as the mutation rate decreases, first the four-letter (ABCD) alphabet then the six-letter alphabet outperform the canonical alphabet. In an RNA world it is unlikely that the mutation rate was low, therefore we can conclude that in this world the canonical RNA alphabet was indeed superior to the alternatives considered here. Only when the *copy fidelity* increased were the four-letter alphabets outperformed by the six-letter alphabets, which is in agreement with the results of Szathmáry (1992). But, by comparing the canonical and non-canonical alphabets where the only difference is that wobble (G : U) base pairs are allowed, we can conclude that allowing wobble pairing for the six-letter alphabet will reduce the advantage of a six-letter alphabet over a four-letter alphabet.

5. DISCUSSION

We have presented a novel approach to investigating the optimal alphabet size in the RNA world. We know that the six- and eight-letter alphabets can cover just as much (if not more) of shape space as the canonical (AUCG) and non-canonical (ABCD) four-letter alphabets, but it would seem from our simulations that the paths from random shapes to our target through sequence space are shorter for the four-letter regimes when the copy fidelity is relatively low. Otherwise the peak shifts to the six-letter alphabet. Although this effect may not be as significant as shown here, owing to the fact that we use the base-pair metric, which penalizes extraneous base pairs, from figure 1 we observe that the four-letter alphabet generally has more base pairs than the larger alphabet sizes. Thus, we conclude that the canonical alphabet was very likely to have been optimal in the RNA world but could indeed be outcompeted (as Szathmáry (1991, 1992) has already suggested) by an alternative six-letter system under a high copy fidelity regimen (although the effects of wobble pairing have not been taken into account here). In addition, copy fidelity decreases with increasing alphabet size (Szathmáry 1992; Mac Dónaill 2002) so it is more realistic to compare high-fidelity two- and four-letter alphabet fitnesses with low-fidelity six- and eight-letter alphabet fitnesses. This would have the effect of increasing the fidelity range where a four-letter alphabet outcompetes a six-letter alphabet.

Ribozymes are usually much larger and more complex than the structure that we use as a target in the flow reactor. Using a more complex structure as the target will improve discrimination between the alphabets. However, experiments to study larger structures will take longer to compute owing to the complexity of the RNA-folding algorithm.

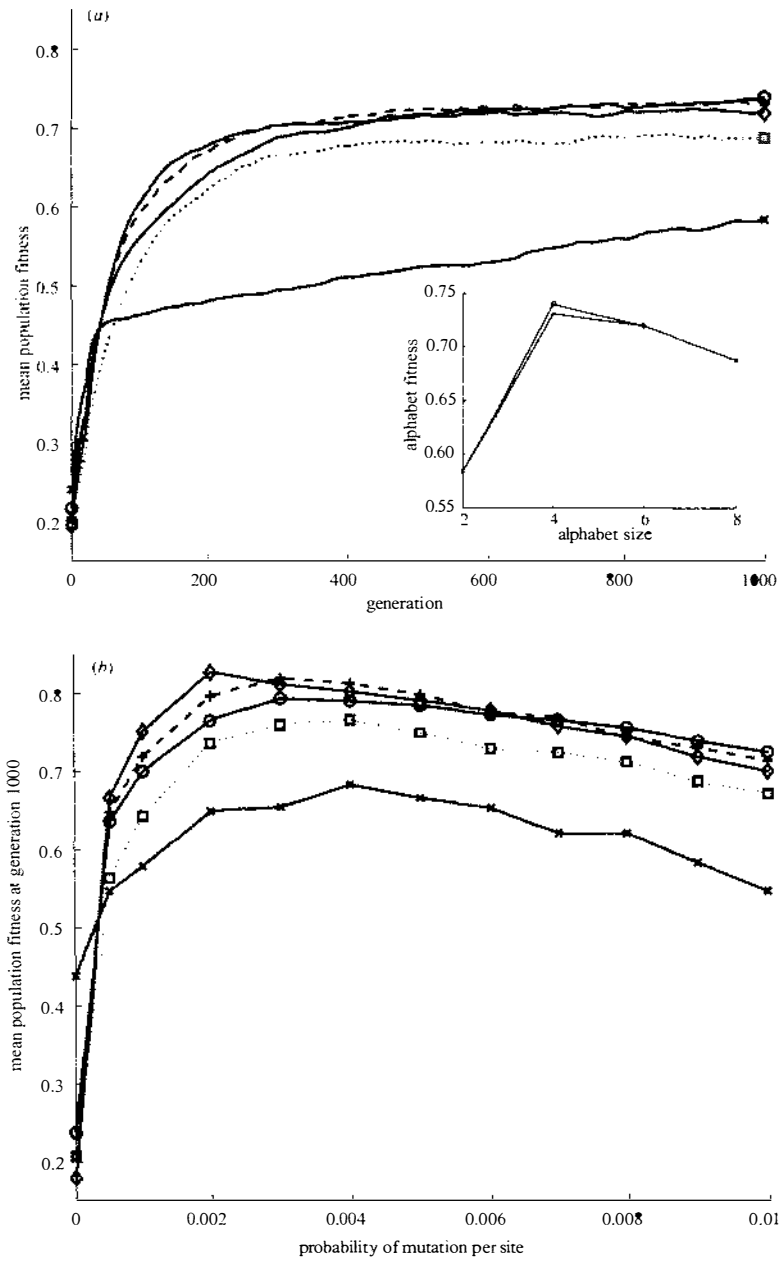


Figure 3. (*a*) Evolving towards a target structure. Results were gathered from the flow reactor (see § 4) for two-, four-, six- and eight-letter artificial genetic alphabets and the canonical AUCG alphabet. The probability of mutation is 0.009 and the population size is maintained at 100. The inset is a plot of the alphabet fitness, which is defined to be mean population fitness at generation 1000 as a function of alphabet size. (*b*) The mean population fitness at generation 1000 as a function of the probability of mutation per site. The results are for two-, four-, six- and eight-letter artificial alphabets (represented by crosses, addition signs, diamonds and squares, respectively) and the canonical AUCG alphabet (circles).

These experiments do not, however, take into account the additional metabolic cost of using longer alphabets. But if we observe four- and six-letter alphabets performing in an almost indistinguishable manner then one can argue that four is optimal owing to the added difficulty of synthesizing an additional base pair (although ribozymes may have been more specific). For the larger 10–16-letter alphabets not included in this study the general trend is an asymptotic approach to a low fitness (data not shown). Additionally, the amount of base pairing with larger alphabets decreases to negligible levels, although the effects of hydrophobic interactions between bases may alter these conclusions (Wu *et al.* 2000).

The authors thank I. Hofacker for implementing the changes in VIENNA necessary for this research, and B. Ryland (bug exterminator). We also thank the administrators of the parallel computing facility used for much of this work (http://sisters.massey.ac.nz). The authors were funded by the New Zealand Marsden Fund and the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), Swedish Research Council (VR).

REFERENCES

Antao, V. P. & Tinoco, I. 1992 Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.* **20**, 819–824.

Bain, J., Switzer, C., Chamberlin, A. & Benner, S. 1992 Ribosome-mediated incorporation of a nonstandard amino-acid into a peptide through expansion of the genetic code. *Nature* **356**, 537–539.

Eddy, S. R. 1999 Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9**, 695–699.

Eschenmoser, A. 1999 Chemical etiology of nucleic acid structure. *Science* **284**, 2118–2124.

Fontana, W. & Schuster, P. 1998 Continuity in evolution: on the nature of transitions. *Science* **280**, 1451–1455.

Fontana, W., Konings, D., Stadler, P. & Schuster, P. 1993 Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404.

Gesteland, R., Cech, T. & Atkins, J. (eds) 1999 *The RNA world*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Schuster, P. & Stadler, P. F. 1996 Analysis of RNA sequence structure maps by exhaustive enumeration. *Monatshefte Chem.* **127**, 355–374.

Higgs, P. G. 2000 RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.* **33**, 199–253.

Hofacker, I. L., Fontana, W., Bonhoeffer, S. & Stadler, P. F. 1994 Fast folding and comparison of RNA secondary structures. *Monatshefte Chem.* **125**, 167–188.

Hofacker, I., Fekete, M., Flamm, C., Huynen, M., Rauscher, S., Stolorz, P. & Stadler, P. 1998 Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26**, 3825–3836.

Jeffares, D. C., Poole, A. M. & Penny, D. 1998 Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36.

Joyce, G. F. 2000 Perspectives: RNA structure—ribozyme evolution at the crossroads. *Science* **289**, 401–402.

McCaskill, J. S. 1990 The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers* **29**, 1105–1119.

Mac Dónaill, D. 2002 A parity code interpretation of nucleotide alphabet composition. *Chem. Commun.* **18**, 2062–2063.

Moulton, V., Gardner, P., Pointon, R., Creamer, L., Jameson, G. & Penny, D. 2000*a* RNA folding argues against a hot-start origin of life. *J. Mol. Evol.* **51**, 416–421.

Moulton, V., Zuker, M., Steel, M., Pointon, R. & Penny, D. 2000*b* Metrics on RNA secondary structures. *J. Comput. Biol.* **7**, 277–292.

Parsch, J., Braverman, J. & Stephan, W. 2000 Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154**, 909–921.

Piccirilli, J., Krauch, T., Moroney, S. & Benner, S. 1990 Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343**, 33–37.

Poole, A. M., Jeffares, D. C. & Penny, D. 1998 The path from the RNA world. *J. Mol. Evol.* **46**, 1–17.

Reader, J. & Joyce, G. 2002 A ribozyme composed of only two different nucleotides. *Nature* **420**, 841–844.

Rich, A. 1962 *Horizons in biochemistry*, pp. 103–126. New York: Academic.

Rivas, E. & Eddy, S. R. 2000 Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**, 583–605.

Sankoff, D., Morin, A. & Cedergren, R. 1978 The evolution of 5S RNA secondary structures. *Can. J. Biochem.* **56**, 440–443.

Schattner, P. 2002 Searching for RNA genes using base composition statistics. *Nucleic Acids Res.* **30**, 2076–2082.

Schultes, E. A., Hraber, P. T. & LaBean, T. H. 1999 Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.* **49**, 76–83.

Schuster, P. 1993 RNA based evolutionary optimization. *Origins Life Evol. Biosphere* **23**, 373–391.

Switzer, C., Moroney, S. & Benner, S. 1989 Enzymatic incorporation of a new base-pair into DNA and RNA. *J. Am. Chem. Soc.* **111**, 8322–8323.

Szathmáry, E. 1991 Four letters in the genetic alphabet: a frozen evolutionary optimum? *Proc. R. Soc. Lond.* B **245**, 91–99.

Szathmáry, E. 1992 What is the optimal size for the genetic alphabet? *Proc. Natl Acad. Sci. USA* **89**, 2614–2618.

Szostak, J. W. 1993 Ribozymes—evolution *ex vivo*. *Nature* **361**, 119–120.

Tuerk, C. & Gold, L. 1990 Systematic evolution of ligands by exponential enrichment—RNA ligands to bacteriophage-T4 DNA-polymerase. *Science* **249**, 505–510.

White, H. B. 1976 Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101–104.

Wu, Y., Ogawa, A., Berger, M., McMinn, D., Schultz, P. & Romesburg, F. 2000 Efforts toward expansion of the genetic alphabet: optimization of interbase hydrophobic interactions. *J. Am. Chem. Soc.* **122**, 7621–7632.

Zuker, M. 2000 Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.* **10**, 303–310.

CHAPTER 3

Automated Identification of Pseudouridylation Guide snoRNA Genes

Little Bo-Peep has lost her sheep,
And doesn't know where to find them;
Leave them alone, and they'll come home,
Wagging their tails behind them.

–Mother Goose nursery rhyme about lost sheep.

3.1 Context and Overview

This project relates attempts to computationally identify "pseudouridylation guide small nucleolar RNAs" (H/ACA snoRNAs). To date, 44 pseudouridines have been identified on yeast rRNAs (see Table 3.1) and 17 H/ACA snoRNAs have been shown to guide modification of 21 of these (Ofengand & Fournier, 1998; Samarsky & Fournier, 1999). Based on this data we suspect that perhaps 10-20 yeast H/ACA snoRNAs remain to be identified. The H/ACA snoRNAs maintain 4-6 short primary motifs and a global secondary structure which may provide sufficient information for a genome wide search for candidate H/ACA snoRNA Genes (see Paper 2, figure 1).

The follwing chapter is composed of two papers, Paper 2 has been published in "Bioinformatics" and Paper 3 is a draft detailing a comparative genomic approach to snoRNA gene identification. Paper 2 describes an algorithm for locating yeast pseudouridylation guide snoRNAs, results from a genomic scan by an implementation of the algorithm are discussed. Paper 3 describes phylogenetic or comparative

genomic approaches for further analyses of results from the genomic scan algorithm described in Paper 2. Results from applying this technique are discussed.

The authors of paper 2 (Edvardsson *et al.*, 2003) are Sverker Edvardsson, myself (Paul Gardner), Anthony Poole, Mike Hendy, David Penny and Vincent Moulton. Edvardsson and myself are joint first authors. Edvardsson developed the majority of the code with assistance and much discussion with myself (and our fellow authors). I did the majority of the data-gathering and analysis of the primary and secondary structure elements of our training data-set. The original manuscript was drafted by myself which has been completely re-written by successive rounds of proof-reading by the authors. The other authors have operated in a largely supervisory capacity. The supplementary material (Tables 3.1-3.5, referred to in paper 2 as tables 1-5) are appended after paper 2.

Paper 3 discusses the use of comparative genomics to further reduce false-positives and increase confidence in the genomic scan for H/ACA snoRNAs detailed in Paper 2. This paper is currently solely authored by myself but others may be added as their contributions are included.

3.2 *Paper 2*, A search for H/ACA snoRNAs using predicted MFE secondary structures

Author: Sverker Edvardsson, Paul P. Gardner, Anthony M. Poole, Michael D. Hendy, David Penny, & Vincent Moulton

Year: 2003

Journal: *Bioinformatics*

Volume: 19

Number: 7

Page: 865-873

BIOINFORMATICS

Vol. 19 no. 7 2003, pages 865–873
DOI: 10.1093/bioinformatics/btg080

A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction

Sverker Edvardsson[1,†], Paul P. Gardner[2,4,†], Anthony M. Poole[3,4], Michael D. Hendy[2,4], David Penny[3,4] and Vincent Moulton[5,*]

[1]Department of Information Technology, Mid Sweden University, S-851 70, Sundsvall, Sweden, [2]Institute of Fundamental Science, [3]Institute of Molecular BioScience, [4]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand and [5]The Linnaeus Centre for Bioinformatics, Uppsala University, BMC Box 598, S-751 24, Uppsala, Sweden

Received on June 7, 2002; revised on November 1, 2002; accepted on November 26, 2002

ABSTRACT

Motivation: Noncoding RNA genes produce functional RNA molecules rather than coding for proteins. One such family is the H/ACA snoRNAs. Unlike the related C/D snoRNAs these have resisted automated detection to date.

Results: We develop an algorithm to screen the yeast genome for novel H/ACA snoRNAs. To achieve this, we introduce some new methods for facilitating the search for noncoding 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 generalized to enable screening of other eukaryote genomes. We find that use of primary sequence alone is insufficient for identifying novel H/ACA snoRNAs. Only the use of secondary structure filters reduces the number of candidates to a manageable size. From genomic context, we identify three strong H/ACA snoRNA candidates. These together with a further 47 candidates obtained by our analysis are being experimentally screened.

Contact: vincent.moulton@lcb.uu.se

Supplementary Information: Tables 1–5 referred to in the text can be downloaded from http://RNA.massey.ac.nz/fisher/

INTRODUCTION

The number of genes identified that code for noncoding RNAs is growing rapidly (Eddy, 2001; Erdmann *et al.*, 2001; Meli *et al.*, 2001). While labor-intensive molecular biological approaches have been successful in identifying noncoding RNAs (Hüttenhofer *et al.*, 2001; Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee and Ambros, 2001), it is preferable to carry out initial RNA gene prediction *in silico*, as is common with protein-coding genes, e.g. (Delcher *et al.*, 1999).

Standard search methods such as BLAST (Altschul *et al.*, 1990) have been used in comparative searches of bacterial genomes for novel RNAs (Argaman *et al.*, 2001; Rivas *et al.*, 2001; Wassarman *et al.*, 2001) and in searches for novel small regulatory RNAs in animals and invertebrates (Pasquinelli *et al.*, 2000). In addition, programs for RNA gene finding are available; for example, the programs `tRNAscan-SE` (Lowe and Eddy, 1997), `QRNA` (Rivas and Eddy, 2001; Rivas *et al.*, 2001), and `RNAMotif` (Macke *et al.*, 2001) have been successfully applied in whole genome searches for novel RNAs.

The importance of primary sequence for the finding of new RNAs is clear, and was employed heavily in a comparative search for noncoding RNAs in *E.coli* (Rivas *et al.*, 2001). However, in general, standard homology searches are not suitable for finding RNAs. Thus successful searches have tended to use techniques such as neural networks (Carter *et al.*, 2001), pattern-based descriptors (Macke *et al.*, 2001) and covariance models (Eddy and Durbin, 1994; Lowe and Eddy, 1997, 1999) which incorporate RNA secondary structure information.

In this paper we investigate an alternative approach for incorporating secondary structure information into RNA searches. Secondary structure is amenable to mathematical analysis making minimum free-energy (MFE) structure prediction using algorithms such as dynamic programming possible. In consequence programs such as `VIENNA` (Hofacker *et al.*, 1994) and `Mfold` (Zuker *et al.*, 1999) can quite accurately predict secondary structure. Even so, Rivas and Eddy (2000) determined that a general search for noncoding RNAs in genomes

*To whom correspondence should be addressed.
† Both authors contributed equally to this work.

using MFE structure stability alone is unlikely to succeed since background noise is too high.

However, in (Collins *et al.*, 2000) the discovery of an RNase P candidate in the maize chloroplast genome was detected using an *ad hoc* combination of comparative genomics and MFE structure comparison. Encouraged by this result, we developed the RNA shape comparison techniques described in (Moulton *et al.*, 2000) and incorporated them into an algorithm that we present here which screens the budding yeast *Saccharomyces cerevisiae* genome (Goffeau *et al.*, 1996) for H/ACA snoRNAs. Our method is similar to that used by Lowe and Eddy (1999) in their successful computational screen of the *S.cerevisiae* genome for the related C/D snoRNAs, which employed a probabilistic model as opposed to MFE structure prediction.

METHODS

Our search strategy for novel snoRNAs in the *S.cerevisiae* or yeast genome uses known H/ACA snoRNAs to form primary and secondary structure models. Then we make a sequential search for novel snoRNAs in both directions of the yeast genome, passing candidate sequences obtained with the primary structure search through various secondary structure filters. The sequences that pass through all of these filters are then scored using both primary and secondary structure information.

Training data set

SnoRNAs (small nucleolar RNAs) are named because of their localization to the eukaryote cell nucleolus. They fall into two families, the C/D box family and the H/ACA box family (reviewed in Weinstein and Steitz, 1999). Within the H/ACA family there is significant conservation of predicted MFE secondary structures, but very limited conservation of primary sequence Ganot *et al.*, 1997a,b).

The H/ACA box family guide site-specific isomerization of rRNA (Ni *et al.*, 1997; Ganot *et al.*, 1997a), whereby uridine (U) is converted to pseudouridine (Ψ) (reviewed by Ofengand and Fournier, 1998), see Figure 1. To date, 44 pseudouridines have been identified on yeast rRNAs (Table 1) and 17 H/ACA snoRNAs have been shown to guide 21 of these (Ofengand and Fournier, 1998; Samarsky and Fournier, 1999). Based on this data we suspect that perhaps 10–20 yeast H/ACA snoRNAs have yet to be identified.

We obtained a dataset of 16 yeast H/ACA snoRNA sequences from the Yeast SnoRNA Database (Samarsky and Fournier, 1999). These had been identified primarily by biochemical techniques (Ganot *et al.*, 1997a; Ni *et al.*, 1997) and are provided with demonstrated or predicted locations for H and ACA motifs and rRNA interactions. The sequences flanking the pseudouridylation sites in rRNA are obtained from (Ofengand and Fournier, 1998)

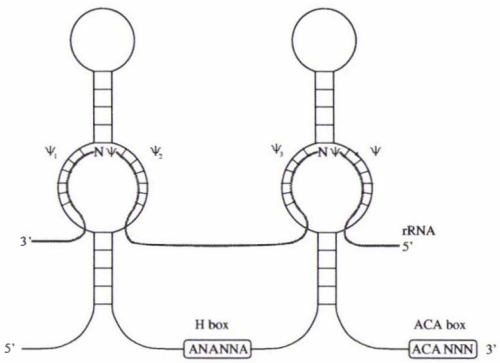


Fig. 1. Schematic of the consensus primary and secondary structural elements of the H/ACA box snoRNA. Note the hairpin-hinge-hairpin-tail secondary structure and the internal loop structures termed the pseudouridylation pockets (Ganot *et al.*, 1997b). The interaction of these pockets with rRNA is also shown. Ψ_i refers to the parts of the snoRNA that are complementary to the rRNA.

where information regarding pseudouridines in yeast rRNA is presented. We did not include snR9, snR30 or snR37 in our training data. For snR9, no capacity for guiding pseudouridylation has been assigned, and snR30 is involved in rRNA cleavage, not pseudouridylation. snR37 is 386nt long and does not compare well with the snoRNAs in the training set.

Primary structure search

The primary structure search algorithm sequentially identifies parts of the yeast genome harboring various primary structural motifs, separated as detailed in Figure 2. The algorithm first searches for an H-box. This motif is a sequence of the form $AN_1AN_2N_3N_4N_5$ with $N_i \in \{A, U, C, G\}$, $N_1 \neq C$, $N_3 \neq G$, and either $N_4 = A$ or $N_5 = A$. Once a candidate H-box is identified, it is scored using a probabilistic model that we constructed using the snoRNA dataset. In particular, we compute a similarity score between the putative H-box and each of the known H-boxes (presented in Table 2) using the frequencies of nucleotides at positions $(N_1N_2N_3N_4N_5)$ (presented in Table 3). The similarity between the putative H-box and each known H-box is computed as follows; the two sequences are placed one above the other, matches are given a score of 200, mismatches are scored according to the nucleotide frequencies at positions $(N_1N_2N_3N_4N_5)$ (e.g. if the putative sequence has a G in position N_1 which mismatches it is scored 81.25) and the scores are added, in a similar fashion to the profile matrix method used by PSI-BLAST (Altschul *et al.*, 1997). If the maximum

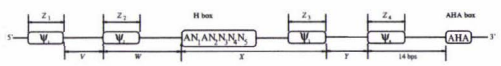


Fig. 2. Primary structure model used to search for putative snoRNAs consisting of an H-box, an ACA-box (here denoted AHA—see text) and four regions of complementarity to the rRNA subsequences flanking some pseudouridylation site on rRNA (denoted by $\Psi_1\Psi_2$ and $\Psi_3\Psi_4$). Our model requires: $X + Y + 14 \leqslant 142$; $16 \leqslant X \leqslant 70$; $Y \geqslant 30$; $3 \leqslant Z_3, Z_4 \leqslant 10$; $Z_3 + Z_4 \geqslant 9$; $20 \leqslant V \leqslant 100$; $11 \leqslant W \leqslant 17$; $3 \leqslant Z_1, Z_2 \leqslant 10$; $Z_1 + Z_2 \geqslant 9$.

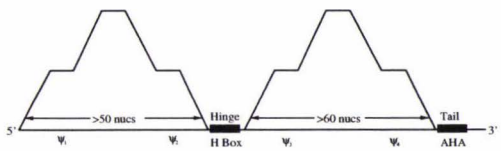


Fig. 3. Secondary structure model of H/ACA snoRNA. It consists of two 'mountains' with widths as indicated.

of these similarities exceeds the threshold value of 800 (obtained using a leave-one-out analysis), the H-box is accepted and this similarity score is recorded for the H-box. In addition, 200 is added to the similarity in case a complete match is obtained between the putative H-box and an H-box that occurs at least twice for the known snoRNAs (e.g. snR189 and snR34). Although such an H-box would be accepted without this bonus, the addition is made since the similarity is used later when scoring the final candidates.

After locating a high-scoring H-box, the algorithm searches downstream for Ψ_3 and Ψ_4 motifs. These are two sequences that are almost complementary to the sequences flanking a pseudouridylation site in the yeast rRNA, see Figure 1 (these motifs are listed in Table 1). Similar complementary motifs were also employed by Lowe and Eddy (1999) in their search for C/D snoRNAs. To look for a putative Ψ_3 motif, a known Ψ_3 motif is directly compared with the yeast genome. The comparison is considered a match if either the sequences are identical or there is at most one wobble, where a wobble corresponds to a C or an A in Ψ_3 lining up to an U or a G in the genome, respectively. The wobble corresponds to a non-canonical base pairing between the H/ACA snoRNA and the rRNA. Such pairings occur for the known snoRNAs. The same comparison is performed for the Ψ_4 motif. The lengths of the Ψ_3, Ψ_4 motifs (Z_3 and Z_4 in Figure 2), which were inferred by analyzing the snoRNA dataset, are required to be between three and ten bases, and the sum of their lengths must always exceed 8. If Ψ_3, Ψ_4 motifs are found in the correct locations (given by X and Y in Figure 2), then the algorithm continues to search for the ACA sequence. To reduce any confusion from now on we denote this sequence by AHA, where H can equal A, U or C. The AHA box is exactly 14 bases from the beginning of the Ψ_4 motif, a distance that is conserved for all known yeast snoRNAs (Ganot *et al.*, 1997b) and, if found, the complete H-AHA region is passed to the secondary structure filters described in the next section. Failure to locate a downstream motif in the above procedure in general results in a continuation of the sequential search for another H-box.

Secondary structure filters

The H-AHA region identified by the primary structure filter is passed through several secondary structure filters to reduce false positives.

A secondary structure model for yeast H/ACA snoRNA was derived using MFE structure prediction (Zuker and Steigler, 1981). For each known snoRNA two sequences consisting of the H-AHA region together with a sequence of length 100 or 120 bases upstream from the H-box were formed and then folded using the `RNAfold` function of the `VIENNA` v. 1.4 package (Hofacker *et al.*, 1994). The option '`no dangling ends`', improved the folds. Upstream lengths of 100 and 120 gave a good signal, even though these do not correspond exactly to those for the known snoRNAs. The dynamic length was necessitated because the 5′ end of a putative snoRNA sequence cannot be determined *a priori* in the yeast sequence.

The resulting structures were represented by mountain plots (see Moulton *et al.*, 2000), which are based on the representation of Hogeweg and Hesper (1984). This type of plot allows a simple connection between primary and secondary structure. The mountain plot consists of the points with x-coordinate k corresponding to the kth nucleotide and y-coordinate y_k equaling the number of base-pairs enclosing this nucleotide (see Figure 3). When we compare structures whose underlying sequences have different lengths, we normalize the corresponding mountain plots, scaling the x-coordinates to lie between 0 and 1 and the y-coordinates so that the total area under the graph equals one. In practice, mountain plots are represented by the vector containing the y-coordinates y_k corresponding to each nucleotide k, whereas normalized mountains plot are represented by vectors of a suitably large fixed length N, that contain the y-coordinates y_i of the normalized mountain plot at x-coordinates $\frac{i}{N}$, $1 \leqslant i \leqslant N$. To obtain these normalized vectors we employed splines.

Good similarity was observed between the normalized mountain plots of the known snoRNA dataset (Figure 4). The significant common structural features were incorporated into a secondary structure model consisting of two 'mountains' separated by a hinge region, the position of

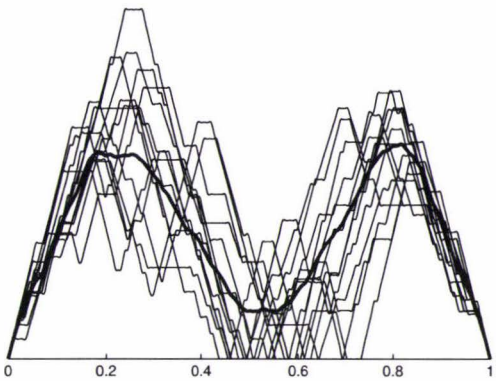


Fig. 4. Normalised mountain plots (L=100) for the 16 yeast snoRNAs. The thick line represents the mean structure for the whole snoRNA dataset.

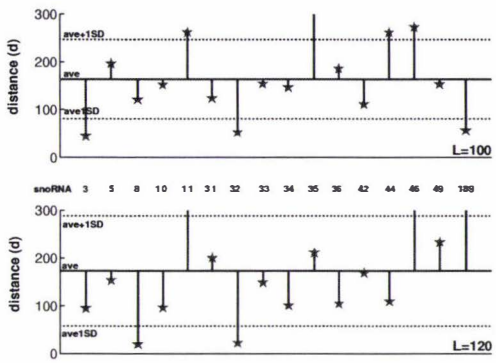


Fig. 5. Distance d from the 16 known snoRNAs to the mean snoRNA structure. The dotted lines represent the standard deviations (±1 SD). A candidate structure will pass if $d < 300$ for either $L = 100$ or $L = 120$.

which roughly corresponds to the H-box (Figure 3). As a preliminary coarse filter, the sequence comprising of the H-AHA region identified previously, with either $L = 100$ or 120 upstream bases, is folded. The resulting mountain plot is accepted only if it has a local minimum (corresponding to the hinge position) within ±11 bases of the H-box, the height of this minimum is at most 4 above the H-box, the width of the left mountain exceeds 50 bases (fulfilled automatically for the right mountain, see Figure 2), and above Ψ_3 and Ψ_4 the graph is high enough (>4) and also non-zero between these two motifs.

Those candidates displaying these coarse criteria are then passed through more-sensitive filters. The first filter computes a squared distance from its normalized mountain plot to a mean snoRNA structure ($d = \sum_{i=1}^{N}(y_i - \overline{y}_i^L)^2$, where y_i is the normalized structure and $\overline{y}_i^L$ is the mean normalized snoRNA structure taken over the training dataset). Figure 4 displays this mean snoRNA structure for the case $L = 100$. The distance d between a known snoRNA and the mean snoRNA is typically about 150 (see Figure 5) so that low values of d are not expected for candidate snoRNAs. Even though distances for candidate snoRNAs are expected to be about the same as for known snoRNAs (see Figure 5), a candidate snoRNA is still allowed to pass through this filter if $d < 300$.

A second filter uses the observation that known snoRNA structures whether obtained using the old (v1.3) or new (v1.4) folding parameters (Mathews *et al.*, 1999) provided in the `VIENNA` package were similar—a property that we did not observe in general for random sequences (data not shown). This may be because a small perturbation in parameters does not significantly change stable secondary structures. We implemented a stability filter that compares normalized mountain plots generated for candidate snoRNA sequences using both the old and new folding parameters. In particular, we compute the distances d_{old} and d_{new} for the 'old' and 'new' normalized mountain plots. Only candidate snoRNAs satisfying $|d_{old} - d_{new}| < 300$ are accepted.

Scoring the output

The last stage computes a score based on both primary and secondary structure for each candidate snoRNA. A score for the AHA-box is added to the H-box similarity score described earlier. The H in the AHA-box is scored according to: $A = 6.25$, $U = 18.75$, $C = 75$ and G is not allowed (based on frequencies from the training dataset; see Table 2). The scores are added as described in the section above and then transformed into a number $0 \leqslant P_1 \leqslant 1$.

As part of the score we also computed three other quantities P_2, P_3 and P_4 defined as follows (see Figure 2 and Table 4). If $X \leqslant 40$ then we put $P_2 = 1$, else $P_2 = 0.5$. Furthermore, if $66 \leqslant X + Y \leqslant 100$ then we put $P_3 = 1$, else $P_3 = 0.5$. The score is also based on the performance of the secondary structure. The average distance $\overline{d}$ is computed for the training dataset. For a putative snoRNA if $|d - \overline{d}| > \overline{d}$, then we put $P_4 = 0$, else $P_4 = (\overline{d} - |d - \overline{d}|)/\overline{d}$. Thus, the closer d is to the average $\overline{d}$, the higher the score. The putative snoRNA is only accepted if both $P_1 > 0.8$ and $P_4 > 0.65$ hold.

The total score P_{tot} of the candidate snoRNA is then computed using the formula

$$P_{tot} = 100\left[\frac{w_1P_1 + w_2P_2 + w_3P_3 + w_4P_4}{w_1 + w_2 + w_3 + w_4}\right],$$

where w_1=10, w_2=2, w_3=1, and w_4=2. The values of the weights w_i were obtained by optimization using the Nelder–Mead method (see e.g. Kelley, 1999) on the training dataset. Only if $P_{tot} > 70$ is the structure accepted. All snoRNAs in the training dataset satisfy $P_{tot} \gg 70$ (see Figure 6).

Final processing

In case we target snoRNAs also having the complementary pair Ψ_1 and Ψ_2, a special procedure is called. This looks for the motifs Ψ_1 and Ψ_2, that fulfill the criteria $V \geqslant 20$ (the majority have V in the range 30-40) and $11 \leqslant W \leqslant 17$ (the majority have $W = 14$)—see Figure 2 and Table 4.

RESULTS

We have implemented the strategies and filters described above in a C program (`Fisher`). This is available via electronic mail [sverker.edvardsson@mh.se].

Ψ-pair assignments

In the known yeast H/ACA snoRNA dataset, no two snoRNAs have been demonstrated to guide the same pseudouridylation (Table 1). However, of the snoRNAs in our dataset, only 13 of 22 assigned Ψ-pairs perform the corresponding pseudouridylation (see Table 2) and snR3 is potentially capable of more than one pseudouridylation at the 3′ pocket ($\Psi_3\Psi_4$). We therefore examined the known snoRNAs for redundancy, using our primary structure engine to search for all $\Psi_1\Psi_2$- and $\Psi_3\Psi_4$-pairs within these (see Table 5). We used the following constraints: $25 \leqslant V \leqslant 45$ and $13 \leqslant W \leqslant 16$ (see Figure 2). Our algorithm locates all the assigned Ψ-pairs except 39, corresponding to snR34 (Table 2). The reason is the unusually large distance $W = 38$. A potential stem involving $24 + 2$ bases lies between snR34's Ψ_2 and the H-box. Despite this feature, it is reasonable to assume that functionally important spatial determinants are preserved (W. Decatur, J. Ni, and M. Fournier, pers. commun.). This is the only such situation known to exist for the yeast snoRNAs. Our examination of sequence complementarity between the rRNA and the $\Psi_1\Psi_2$ and $\Psi_3\Psi_4$ sequence pairs in the 5′ and 3′ pseudouridylation pockets of the known H/ACA snoRNAs reveals extensive potential for functional redundancy (Table 5). For instance, pseudouridylation of U_{1056} in the 25S rRNA subunit (23 in Table 1) is guided by snR44 (Ganot *et al.*, 1997a; Samarsky and Fournier, 1999), yet our analysis (Table 5) suggests that snR31, snR33, snR36 and snR49 are also potentially capable of guiding this pseudouridylation. Furthermore, we find that many of the known H/ACA snoRNAs can potentially guide more than 2 pseudouridylations.

A test scan through a randomized genome sequence

In order to investigate the performance of our search strategy with respect to false positives we created a randomized test genome sequence. To conserve the approximate frequencies of A, U, C and G, our test genome was created by copying a sequence of length 540 000 bases from the yeast genome. The sequence was shuffled using an algorithm that preserves dinucleotide frequencies (Workman and Krogh, 1999; Altschul and Erikson, 1985). For a sequence of length N, this is performed by randomly selecting pairs of triplets of the form XQY and XPY and then exchanging Q and P. This is repeated $10N$ times. We then added all 13 snoRNAs that have $\Psi_3\Psi_4$-pairs (Table 2) to create the final test genome. A complete scan through this genome took about a day on an AMD Athlon 1.4 Ghz which was reasonable for testing purposes.

The total number of hits obtained by the primary search was 66 600, which demonstrates that it is unrealistic to only consider the primary motifs of H/ACA snoRNAs. However, several of these were actually at the same H-box position. This redundancy occurs since the search can locate several different $\Psi_3\Psi_4$-pairs and AHA-boxes. For each H-box we only kept hits with highest primary score (i.e. $100(w_1P_1 + w_2P_2 + w_3P_3)/(w_1 + w_2 + w_3)$). After the initial secondary structure filters have been applied, we are left with 15 428 hits.

The scores P_{tot} for these hits are plotted in Figure 6. The squares indicate the scores obtained for the known snoRNAs, which were planted within the first 54 000 bases of the test genome. Out of the 15 428 hits, 2397 have total scores greater than 70. We observe in Figure 6 that the snoRNAs have scores well above most of the other hits. The snoRNA with the lowest score was ranked 192. Thus, to hit all of the known snoRNAs we need to accept 179 false positives. The final requirement that both ($P_1 > 0.8$) and ($P_4 > 0.65$) hold simultaneously, further reduced the number of false positives to 96. Thus, out of the 15 428 distinct hits, 96 false positives remained, giving a performance of $(15\,415 - 96)/15\,415 = 99.4\%$ (searching the reverse complemented test genome gave 99.3%).

Unfortunately, snR8 does not satisfy ($P_1 > 0.8$) and ($P_4 > 0.65$). Of course, this last filter could be relaxed in order to hit snR8, but then we would need to deal with many more false positives. After considerable testing, we concluded that the balance between the number of false positives and false negatives was acceptable.

Screening the yeast genome with Fisher

The yeast genome is approximately 12 Mb, which is about 20 times larger than our test genome, and we must search it

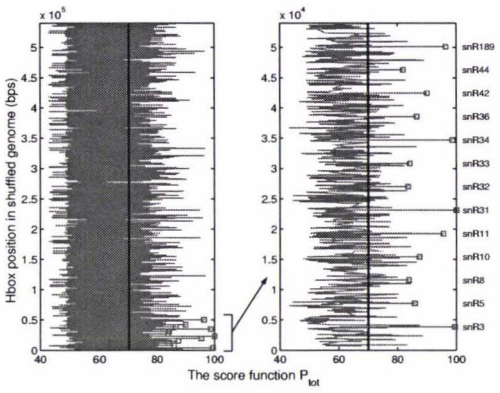


Fig. 6. The total score (P_{tot}) for the hits in the shuffled test genome. The left figure shows the whole test genome consisting of 540 000 bases. The snoRNAs were planted amongst the first 54 000 bases. This part is enlarged on the right. The total scores for the 13 Ψ_3 Ψ_4-snoRNAs are marked with squares. For a normal search we only accept hits with scores above 70 (marked with the bold line).

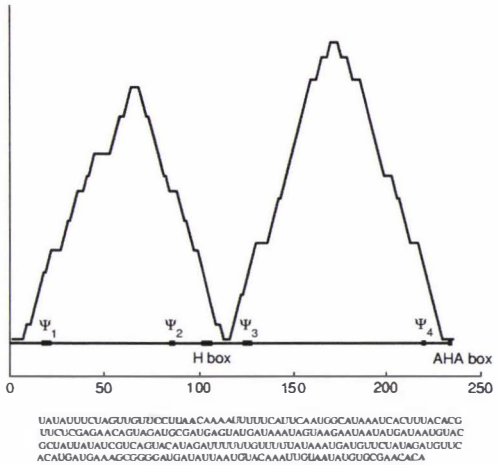


Fig. 7. An example of a typical hit in the yeast genome. This particular hit has both a $\Psi_1\Psi_2$ (32) and a $\Psi_3\Psi_4$-pair (2). Reading left to right, the bold motifs in the above sequence are: Ψ_1, Ψ_2, H-box, Ψ_3, Ψ_4 and AHA-box.

in both directions. Thus, from the results above we expect `Fisher` to yield perhaps ten quite highly ranked novel snoRNAs and about 4000 false positives.

In order to decrease this rather large number of expected false positives we created a reduced yeast genome sequence of approximately 3.5 Mb. This consisted of the NotFeature.fasta file (produced by removing all regions corresponding to ORFs listed in the yeast ORFs files), obtained by ftp from the *Saccharomyces* Genome Database (Cherry *et al.*, 2001), together with the known introns in yeast obtained from the Ares lab Yeast Intron Database version 2.0 (Davis *et al.*, 2000). This not only reduced the expected number of false positives to about 1000, but also saved significant CPU time (run time was about one week).

Instead of the approximately 1000 false positives/novel snoRNAs that we expected for the reduced yeast genome, we in fact found 579. These candidates were further examined and reduced in number by considering their scores and performing some manual processing, such as checking primary and secondary structures. We also checked the high ranking candidates regarding their genomic context. Amongst the 579 candidates, we only found 31 snoRNA structures having both a $\Psi_1\Psi_2$- and a $\Psi_3\Psi_4$-pair. To create a list of 50 candidates for experimental screening, we also added another 19 of our most interesting Ψ_3 Ψ_4-candidates.

We now discuss these 50 hits in more detail. In Figure 7 we present an example of a putative snoRNA that `Fisher` located in the yeast genome. The motifs: Ψ_1 Ψ_2, H-box, Ψ_3 Ψ_4 and the AHA-box are marked in bold. Its Ψ-pairs are Ψ_1 Ψ_2=32 and Ψ_3 Ψ_4=2 (see Table 1); these have not been previously assigned to any known snoRNA. The distances between the motifs are $V = 61$, $W = 17$, $X = 27$ and $Y = 91$ (see Figure 2). The secondary structure, that exhibits the typical double mountain, is also displayed in Figure 7. Encouragingly, the highest scoring candidates showed a clear over-representation of Ψ-pairs that are not assigned to known snoRNAs, whereas hits with lower scores more often had Ψ-pairs that are already assigned to known snoRNAs.

Both the 50 candidates and the known snoRNAs are broadly distributed on the yeast genome, with all chromosomes possessing either known snoRNAs or candidates, or both. Chromosome XV is notable in that it carries the genes for four known snoRNAs, and is also the chromosome with the largest number of candidates located along its length. Most of our top candidates are located in chromosomes XII-XVI.

Three of our candidates were found to have especially interesting genomic locations. Two are located in the introns of the genes for the yeast ribosomal proteins, RPL43A and RPS11A (both genes contain one intron only). We consider this to be a strong indication that these two candidates are indeed snoRNAs, since the majority of

intronic snoRNAs are found in the introns of ribosomal protein genes (Maxwell and Fournier, 1995) and, in yeast, all intronic snoRNAs except one are in ribosomal or ribosome-associated proteins (Samarsky and Fournier, 1999). An examination of orthologous ribosomal proteins in other organisms revealed no additional information, though (Higa *et al.*, 1999) have demonstrated that the human and mouse *Rps11* genes house U35 (a C/D family snoRNA) in the third intron (the yeast *RPS11A* gene has only one intron). A third candidate was located in the ORF coding for the snoRNP U3 protein MPP10. This candidate is not intronic, and completely overlaps the coding sequence. This arrangement has been recently demonstrated for the C/D family snoRNA U86, in yeast (Filippini *et al.*, 2001). Given this demonstration of completely overlapping snoRNA-protein coding genes, and the fact that the host gene for our candidate is also involved in snoRNA-dependent rRNA processing, we consider this hit to be a good candidate for a *bona fide* snoRNA. This suggests that future genomic searches may require the entire genome sequence.

DISCUSSION

We have presented an algorithm for searching the yeast genome for H/ACA snoRNAs. It is reasonably fast and can be tuned to produce a manageable number of good candidates.

The method we describe could in principle be applied to any family of RNAs with low level conserved sequence and well-conserved secondary structure. However, it might be that it works well for H/ACA snoRNAs since the corresponding structure is quite simple; more studies need to be made to determine whether the method works for more complex structures. In any case, some of the methods we have developed might still be usefully incorporated into existing search strategies.

Two issues warranting discussion are the number of false hits and overtraining. For the test genome described in the results section our strategy had an performance of 99.4%, but this also required the introduction of one false negative (snR8). Since we did not include snR9, snR30, or snR37 in our training data, it is likely that our approach will not hit all known H/ACA snoRNAs, but it will hopefully recover most, as per the computational screen for yeast C/D snoRNAs (Lowe and Eddy, 1999). We are as yet unaware how iteration (adding verified candidates to the training dataset) will affect the ability of our method to identify new H/ACA snoRNAs, and it is not possible to predict how many iterations will be required to recover the majority of H/ACA snoRNAs in yeast. However, since it was found that the known snoRNAs were close to the top in the candidate list for the test genome, partial screening might be expected to effectively recover the majority of additional H/ACA snoRNAs.

In terms of immediate application of our algorithm to other organisms, the human genome provides an important data set for which genome sequence and a sizeable number of characterized H/ACA snoRNAs is available (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). Preliminary work on the known human H/ACA snoRNAs indicates greater H-box and secondary structural homogeneity than for yeast. However, since the human genome is about 250 times longer than the yeast genome and since time complexity for folding a sequence with n bases is $O(n^3)$, the search may become too slow if too much secondary structure filtering is required. This could be offset by, for example, adapting the scanning algorithms described in Rivas and Eddy (2000) or by parallelizing the search. Perhaps more importantly with regards to folding, the accuracy of MFE structure prediction can depend quite heavily on the length of the subsequence of the genome that is being folded. Even so, we emphasize that the ability of the predicted structures to provide signal for discovery of new RNA family members is more important than their correctness.

In conclusion, as additional sequence data for yeasts becomes available (Souciet *et al.*, 2000), it should be possible to not only identify known snoRNAs in other yeasts using BLAST (Cliften *et al.*, 2001; Cervelli *et al.*, 2002), but also to evaluate a list of candidates by genome comparison. This has two implications. First, preliminary evidence that a candidate is a snoRNA can be gathered bioinformatically, as opposed to using labor-intensive experimental screening. Second, we can potentially reverse our approach and establish the site of pseudouridylation. While this does not replace the importance of experimentally determining the position of pseudouridylation, it does mean that our methods can in principle be applied in reverse order in cases where there is comparative data available but no experimentally determined pseudouridylation sites. We are currently developing this strategy, together with a comparative pseudouridylation map for rRNA alignments that may aid in assigning confidence to H/ACA snoRNAs identified by comparative genome analysis.

ACKNOWLEDGEMENTS

This work was supported by The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), The Swedish Research Council (VR), and the New Zealand Marsden Fund. Thanks to M. Fournier and W. Decatur for sharing unpublished data and for helpful discussions. Thanks are also due to Alicia Gore for assistance in examining the genomic context of candidates and discussions, and Linus Sandegren for assistance with data gathering and analysis at an early stage of this project.

REFERENCES

Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.

Altschul,S.F. and Erikson,B.W. (1985) Significance of nucleotide-sequence alignments—a method for random sequence permutation that perserves dinucleotide and codon usage. *Mol. Biol. Evol.*, **2**, 526–538.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.*, **29**, 3928–3938.

Cervelli,M., Cecconi,F., Giorgi,M., Annesi,F., Oliverio,M. and Mariottini,P. (2002) Comparative structure analysis of vertebrate U17 small nucleolar RNA (snoRNA). *J. Mol. Evol.*, **54**, 166–179.

Cherry,J.M., Ball,C., Dolinski,K., Dwight,S., Harris,M., Matese,J.C., Sherlock,G., Binkley,G., Jin,H., Weng,S. and Botstein,D. (2001) '*Saccharomyces* Genome Database' (downloaded 11/1/2001 from ftp://genome-ftp.stanford.edu/pub/yeast/yeast_NotFeature/) (visited 18/2/2002 http://genome-www.stanford.edu/Saccharomyces/).

Cliften,P.F., Hillier,L.W., Fulton,L., Graves,T., Miner,T., Gish,W.R., Waterson,R.H. and Johnston,M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.

Collins,L., Moulton,V. and Penny,D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194–204.

Davis,C.A., Grate,L., Spingola,M. and Ares,Jr,M. (2000) Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.*, **28**, 1700–1706. http://www.cse.ucsc.edu/research/compbio/yeast_introns/currentDBv2/intronsAround.fa (downloaded 11/1/2001).

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.

Eddy,S.R. (2001) Noncoding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Erdmann,V.A., Barciszewska,M.Z., Szymanski,M., Hochberg,A., de Groot,N. and Barciszewski,J. (2001) The non-coding RNAs as riboregulators. *Nucleic Acids Res.*, **29**, 189–193.

Filippini,D., Renzi,F., Bozzoni,I. and Caffarelli,E. (2001) U86, a novel snoRNA with an unprecedented gene organization in Yeast. *Biochem. Biophys. Res. Comm.*, **288**, 16–21.

Ganot,P., Bortolin,M.L. and Kiss,T. (1997a) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.

Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997b) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, **11**, 941–956.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.

Higa,S., Yoshihama,M., Tanaka,T. and Kenmochi,N. (1999) Gene organization and sequence of the region containing the ribosomal protein genes RPL13A and RPS11 in the human genome and conserved features in the mouse genome. *Gene*, **240**, 371–377.

Hofacker,I.L., Fontana,W., Bonhoeffer,S. and Stadler,P.F. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–188.

Hogeweg,P. and Hesper,B. (1984) Energy directed folding of RNA sequences. *Nucleic Acids Res.*, **12**, 67–74.

Hüttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.

International Human Genome Sequencing Consortium (2001) Initialsequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kelley,C.T. (1999) *Iterative Methods for Optimization, Frontiers in Applied Mathematics*, SIAM, 18, Philadelphia.

Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.

Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.

Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAmotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.

Mathews,D.H., Sabina,J., Zucker,M. and Turner,H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Maxwell,E.S. and Fournier,M.J. (1995) The small nucleolar RNAs. *Annual Reviews of Biochemistry*, **35**, 897–934.

Meli,M., Albert-Fournier,B. and Maurel,M.C. (2001) Recent findings in the modern RNA world. *Int. Microbiol.*, **4**, 5–11.

Moulton,V., Zuker,M., Steel,M., Pointon,R. and Penny,D. (2000) Metrics on RNA secondary structures. *J. Comp. Biol.*, **7**, 277–292.

Ni,J., Tien,A.L. and Fournier,M.J. (1997) Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell.* **89**, 565–573.

Ofengand,J. and Fournier,M.J. (1998) The pseudouridine residues of ribosomal RNA: number, location, biosynthesis, and function. In Grosjean,H. and Benne,R. (eds). *Modification and Editing of RNA*. ASM Press, pp. 229–253.

Pasquinelli,A.E. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.

Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in *E.coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.

Samarsky,D.A. and Fournier,M.J. (1999) A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **27**, 161–164. http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html.

Souciet,J.-L. *et al.* (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett.*, **487**, 3–12.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.

Weinstein,L.B. and Steitz,J.A. (1999) Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.*, **11**, 378–384.

Workman,C. and Krogh,A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, **27**, 4816–4822.

Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide in RNA biochemistry and biotechnology. In Barciszewski,J. and Clark,B.F.C. (eds), *NATO ASI Series*. Kluwer Academic Publishers.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.

3.2.1 Supplementary material: A search for H/ACA snoRNAs using predicted MFE secondary structures

Table 3.1:
Regions that are reverse complementary to those flanking the yeast rRNA pseudouridylation sites (see Figure 1). Column one contains our labelling system for Ψ-pairs. The locations of rRNA uridines targeted for modification by H/ACA snoRNAs are given in column 2, see (Samarsky & Fournier, 1999) Positions are numbered according to the rRNA sequences in the *Saccharomyces* Genome Database (Cherry *et al.*, 2001). The reverse complementary motifs of the rRNA are tabulated in columns three and four (Ψ-pair). Parts of these motifs (including possible wobbles) are expected to be found in snoRNA pseudouridylation pockets. The last column gives predicted or demonstrated Ψ-pairs for the known snoRNAs. Ψ-pair 39 (U_{2826}) might be incorrectly assigned to snR34 (W=38), see Results section. Ψ-pair 27 (U_{2133}) is probably also incorrectly assigned to snR3, see Table 2.

Ψ-pair label	Ψ-site	Ψ_1 (Ψ_3)	Ψ_2 (Ψ_4)	snoRNA
18S Subunit				
1	U_{106}	AACGAUAACU	UUUAAUGAGC	snR44
2	U_{120}	GGAACUAUCA	UAAACGAUAA	
3	U_{211}	UUUUUUAUCU	UAAAUACAUC	
4	U_{302}	AUAGGGCAGA	UUUGAAUGAA	
5	U_{466}	UAUUUAUUGU	CUACCUCCCU	snR189
6	U_{632}	CCCAAAGUUC	CUACGAGCUU	
7	U_{759}	UUGAACACUC	AUUUUUUCAA	
8	U_{766}	GCCUGCUUUG	CACUCUAAUU	
9	U_{999}	GACGGUAUCU	UCAUCUUCGA	snR31
10	U_{1181}	GAGUCAAAUU	GCCGCAGGCU	
11	U_{1187}	CGUGUUGAGU	AAUUAAGCCG	snR36
12	U_{1191}	UCCCCGUGUU	GUCAAAUUAA	snR35
13	U_{1290}	AUCACUCCAC	ACUAAGAACG	
14	U_{1415}	GUUAUUGCCU	AACUUCCAUC	
25S Subunit				
15	U_{776}	CUCUUACUCA	UCCAUCCGAA	
16	U_{960}	CUGCUAUCCU	GGGAAACUUC	snR8
17	U_{966}	GAGCUUCUGC	UCCUGAGGGA	
18	U_{986}	UUACCUCAUA	ACUGAUACGA	snR8
19	U_{990}	CGCUUUACCU	UAAAACUGAU	snR49
20	U_{1004}	AACCUCUAAU	UUCGCUUUAC	snR5
21	U_{1042}	UAAAGUUUGA	AUAGGUCAAG	snR33
22	U_{1052}	CUUACAUAUU	AAGUUUGAGA	
23	U_{1056}	ACUUCUUACA	UUUAAAGUUU	snR44
24	U_{1110}	AUGGCCCACU	AAGCUCUUCA	
25	U_{1124}	UCUGCUUACC	AAAUGGCCCA	snR5
26	U_{2129}	AUUAGACAGU	GAUUCCCCUU	snR11
27	U_{2133}	UUUAAUUAGA	GUCAGAUUCC	snR3?
28	U_{2191}	GCACUGGGCA	AAUCACAUUG	snR32
29	U_{2258}	AAGAGAGUCA	GUUACUCCCG	
30	U_{2260}	UUAAGAGAGU	UAGUUACUCC	
31	U_{2264}	UACCUUAAGA	GUCAUAGUUA	snR3
32	U_{2266}	GCUACCUUAA	GAGUCAUAGU	
33	U_{2314}	CUCGUUAAUC	UUCAUGCGCG	
34	U_{2340}	AGAUAGUAGA	GGGACAGUGG	
35	U_{2349}	GGUUUCGCUA	UAGUAGAUAG	
36	U_{2351}	GUGGUUUCGC	GAUAGUAGAU	
37	U_{2416}	AACUAGAGUC	GCUCAACAGG	
38	U_{2735}	UCAUGGUUUG	UUCACACUGA	snR189
39	U_{2826}	UGACUGCCAC	GCCAGUUAUC	snR34?
40	U_{2865}	GACAUCGAAG	UCAAAAAGCA	snR46
41	U_{2880}	AUGAUAGGAA	GCCGACAUCG	snR34
42	U_{2923}	UUAGUGGGUG	CAAUCCAACG	snR10
43	U_{2944}	AAACCCAGCU	CGUUCCCUAU	snR37
44	U_{2975}	AGGGUAAAAC	ACCUGUCUCA	snR42

Table 3.2: The snoRNA training dataset. The numbering convention for the demonstrated (d) and predicted (p) Ψ-pair entries are defined in Table 1. For snR3 the Ψ-pair 27p is predicted in the Yeast SnoRNA Database (Samarsky & Fournier, 1999) and has not been demonstrated. For snR34 the Ψ-pair 39p may be an incorrect assignment – see Results section.

		H box		
snoRNA	$\Psi_1\Psi_2$	$AN_1AN_2N_3N_4N_5$	$\Psi_3\Psi_4$	*AHA*
snR3		*AGAUCAA*	*27p, 31d*	*AUA*
snR5	*25d*	*AGACCAA*	*20d*	*ACA*
snR8	*16d*	*AGAGCAA*	*18d*	*AUA*
snR10		*AGAACAA*	*42d*	*ACA*
snR11		*AGAUAAA*	*26p*	*ACA*
snR31		*AGAUUAA*	*9d*	*ACA*
snR32		*AGAUAGA*	*28d*	*ACA*
snR33		*AGAUUGA*	*21d*	*ACA*
snR34	*39p*	*AGAAUAA*	*41d*	*ACA*
snR35	*12p*	*AGAUCAU*		*ACA*
snR36		*AAAACAA*	*11d*	*AUA*
snR42		*AGAUAAA*	*44d*	*ACA*
snR44	*1p*	*AUAUUUA*	*23p*	*AAA*
snR46	*40d*	*AAAUUAA*		*ACA*
snR49	*19p*	*AGAUUAU*		*ACA*
snR189	*5p*	*AGAAUAA*	*38p*	*ACA*

Table 3.3: Frequencies of nucleotides used to score H-boxes (see text).

%	N_1	N_2	N_3	N_4	N_5
A	12.50	25.00	18.75	81.25	87.50
U	6.25	62.50	43.75	6.25	12.50
C	0	6.25	37.50	0	0
G	81.25	6.25	0	12.50	0

Table 3.4: Distances for demonstrated or predicted Ψ-pairs. See Figure 2 for definition of *V*, *W*, *X* and *Y*. Multiple distances for X and Y are possible for snR3, snR5, snR8, snR11 and snR44. Since X+Y is conserved, multiple hits occur for various Ψ_3 motifs.

snoRNA	*V*	*W*	*X*	*Y*	*snoRNA*	*V*	*W*	*X*	*Y*
snR3	-	-	*27*	*48*	*snR5*	*41*	*14*	*40*	*51*
	-	-	*38*	*37*		*41*	*14*	*45*	*46*
snR8	*37*	*15*	*21*	*71*		*41*	*14*	*50*	41
	37	*15*	*33*	*59*		*41*	*14*	*56*	*35*
snR11	-	-	*32*	*43*		*41*	*14*	*69*	*22*
	-	-	*36*	*39*	*snR10*	-	-	*25*	*73*
	-	-	*53*	*22*	*snR31*	-	-	*26*	*70*
snR32	-	-	*24*	*35*	*snR33*	-	-	*32*	*33*
snR34	-	-	*28*	*44*	*snR35*	*29*	*16*	-	-
snR36	-	-	*23*	*40*		*26*	*16*	-	-
snR42	-	-	*61*	*59*	*snR44*	*38*	*14*	*32*	*45*
snR46	*37*	*14*	-	-		*38*	*14*	*47*	*30*
snR49	*32*	*14*	-	-	*snR189*	*33*	*15*	*26*	*60*

Table 3.5: The computed results of a Ψ-pair search for the known snoRNAs in our dataset. See Table 1 for the definition of our Ψ-pair entries (1-14 represent the pseudouridylation sites on the small rRNA subunit, 15-44 the large subunit). Underlined entries correspond to pseudouridylation sites in the Yeast SnoRNA Database (Samarsky & Fournier, 1999); The letters *p* and *d* stand for *predicted* and *demonstrated* assignments, respectively. Bracketed entries have no corresponding $\Psi_1\Psi_2$-pair.

snoRNA	$\Psi_1\Psi_2$	$\Psi_3\Psi_4$
snR3	*2, 3*	*12, 27p, 31d*
snR5	*25d*	*8, 20d, 33, (38)*
snR8	*16d*	*18d*
snR10	-	*(42d)*
snR11	*26*	*11, (26p), (33)*
snR31	*21*	*1, (9d), 23*
snR32	-	*(28d)*
snR33	*1, 4, 14, 20, 23, 33*	*21d, 30*
snR34	-	*(12), (31), (41d)*
snR35	*12p*	-
snR36	*1, 14, 17, 23, 24*	*11d*
snR42	*3*	*44d*
snR44	*1p*	*23p*
snR46	*40d*	*9, 14, 17*
snR49	*2, 3, 19p*	*4, 23*
snR189	*5p*	*38p*

3.3 *Paper 3 (Draft)*, Locating H/ACA snoRNAs using a combination of comparative genomics and MFE structure prediction

Author: Paul P. Gardner

Year: 2003

Introduction

To date the location of non-coding RNA (ncRNA) genes using computational tools has proved to be a difficult problem. However, specific cases have yielded interesting results, for example, tRNAs with tRNAscan-SE (Lowe & Eddy, 1997*a*), C/D box snoRNAs with snoscan (Lowe & Eddy, 1999), and a general algorithm using base composition statistics in the hyper-thermophile *M. jannaschi* (Klein *et al.*, 2002; Schattner, 2002). The difficulties inherent in ncRNA gene identification are due to the fact that ncRNAs generally conserve a secondary structure more than a primary structure. Proteins, on the other hand strongly conserve a primary sequence. However, secondary structure prediction has proved to be a computationally complex problem and the accuracy of any inference decreases with increasing sequence length. In addition, to scan for a specific ncRNA, a secondary structure model must be known *a priori*, which requires several biochemically characterised (and preferably aligned) examples (Eddy, 2001; Eddy, 2002).

Comparing homologous genes from a variety of carefully selected organisms provides an extra dimension of information. The detection of neutral mutations with respect to homologues genes can prove very informative. For example, a mutation (approximately) every third base is a strong indicator of a protein coding gene, and mutations which conserve an underlying secondary structure are a strong indicator for structural RNA genes. In order to use comparative genomics for the identification of non-coding RNAs, candidate species for genome sequencing projects must be selected which are not too close (and therefore unlikely to yield any significant information) but are not too divergent (and therefore signal is likely to be lost in noise) (Cliften *et al.*, 2001; Cliften *et al.*, 2002). Using this concept should vastly reduce the number of false positives likely to be the result of any computational screen for ncRNAs. In fact, some progress has been made in this direction by Sean Eddy's group (Rivas & Eddy, 2000*c*; Rivas & Eddy, 2001; Rivas *et al.*, 2001).

In this paper we study characteristic H/ACA box small nucleolar RNAs (snoRNAs) elements in a group of partially sequenced (2 to 4-fold shotgun coverage) *Saccharomyces* genomes (Cliften *et al.*, 2001). We use a modified version of `Fisher`

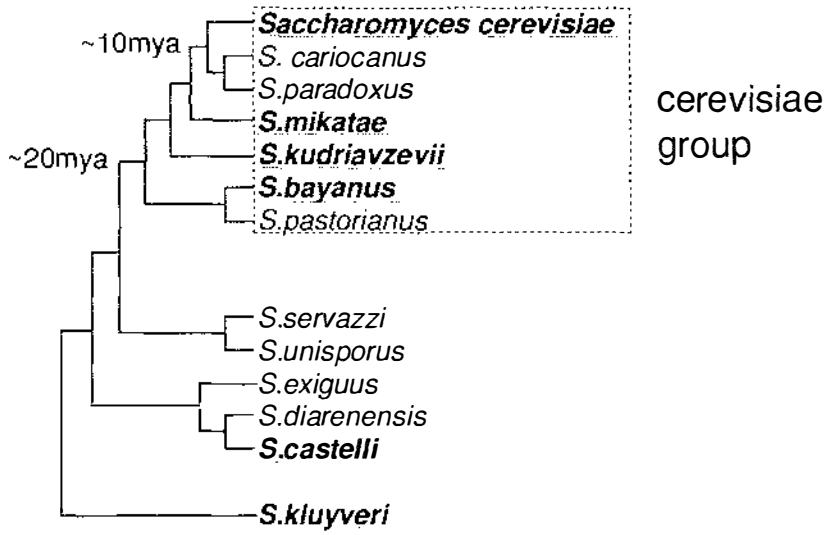


Figure 3.1: *Saccharomyces* Phylogenetic Tree (Kurtzman & Robnett, 1998).

(Edvardsson *et al.*, 2003) to locate candidate H/ACA snoRNA genes in *S. cerevisiae*, then use the heuristic string matching algorithm, `BLAST` to identify homologous genes in genomic data for *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, *S. kluyveri* (see figure 3.1 for a phylogeny). Unfortunately, the *Saccharomyces* sequence data is not freely available for download, however a limited BLAST server has been implemented which we can access.

Methods and Results

Comparative analysis of known *S.cerevisiae* H/ACA snoRNAs

In order to test whether the use of comparative sequence analysis to filter candidate snoRNAs was viable, a study of the known snoRNAs has been made. The list of known snoRNAs used in Edvardsson *et al.*, (2003)[1], has been updated, two newly discovered *S.cerevisiae* snoRNAs labeled here as snR161 and snNOG2. snNOG2 is intronic to a GTPase protein called NOG2, so for want of a better name we have dubbed it snNOG2 (W. Decatur and M. Fournier, pers. commun.).

BLAST has been used to identify snoRNA homologues in each of the newly sequenced *Saccharomyces* species (see figure 3.1). It is known that the best (lowest) BLAST score is not necessarily the closest phylogenetic match (Koski & Golding, 2001). However, in each of the results shown in table 3.6 there was just one low scoring match. Hence, it is unlikely that the results of Koski and Golding (2001) are a factor in this work.

[1]see section 3.2.

As the results in table 3.6 show, the majority of the known snoRNAs are well conserved, particularly across the cerevisiae group (also known as the *sensu stricto* complex in the literature (James *et al.*, 1997)) (see figure 3.1). The essential snR10 sequence is particularly well conserved, snR33 however is not, homologous snR33 genes were only discovered in *S.kudriavzevii* and *S.bayanus*.

It is interesting to note that according to the phylogeny shown in figure 3.1 *S.mikatae* and *S.cerevisiae* share a common ancestor more recently than the rest of the sequenced *Saccharomyces* species, yet `BLAST` results for the snoRNAs suggest *S.kudriavzevii* is closer to *S.cerevisiae*. This indicates that either there are gaps in the *S.mikatae* sequences, or that the number of snoRNA genes has been reduced in *S.mikatae* since the *cerevisiae* and *mikatae* lineages diverged.

The resultant homologous sequences were aligned using `clustalW` (Thompson *et al.*, 1994), and `RNAalifold` (Hofacker *et al.*, 2002) was used to infer a consensus secondary structure using a combination of free-energy and comparative information (see figure 3.6). As an example an alignment of snR36 is shown in figure 3.2 and annotated secondary structures of snR34 and snR36 are shown in figure 3.3. The alignment shows that the variable positions of the H and ACA sequence motifs are not conserved even between the different *Saccharomyces* species. In a comparative study of vertebrate U17 snoRNA Cervelli *et al.* (2002) note that the nucleolar localisation signal `AHA`, is not conserved in several species of turtle, instead it appears as a `GCA`, yet the mature U17 is still functional.

The hairpin-hinge-hairpin-tail secondary structure motif is well conserved by the majority of the snoRNAs (see figure 3.6). Of particular note is the secondary structure of snR34, which as discussed in Edvardsson *et al.* (2003) (see page 57 of this document) has a stem between the region of complementarity with rRNA Ψ_2 and the H box (labeled "*Stem Insert*" in figure 3.3). This stem increases the nucleotide distance between Ψ_2 and the H box from a well conserved 14 (± 2) in the other snoRNAs to 38 in snR34, additionally the stem is conserved and contains neutral mutations.

	cerevisiae group			out-group		
	S.mikatae	*S.kudriavzevii*	*S.bayanus*	*S.castelli*	*S.kluyveri*	
snR3	-	9.0e-37	5.3e-33	4.9e-05	0.0048	(-)
snR5	6.3e-20	1.4e-33	8.9e-32	8.7e-15	6.5e-06	(4.0e-05)
snR8	7.8e-34	2.2e-34	1.4e-31	1.0e-14	3.5e-07	(-)
snR10	6.5e-45	2.2e-46	4.6e-44	4.8e-23	6.3e-26	(3.6e-23)
snR11	-	3.2e-46	4.4e-26	9.4e-13	-	(-)
snR31	3.9e-38	2.7e-32	-	1.3e-12	6.6e-06	(0.0047)
snR32	7.6e-25	7.0e-29	1.1e-27	7.1e-09	1.2e-06	(0.0053)
snR33	-	1.3e-27	1.4e-27	-	-	(-)
snR34	7.0e-37	1.8e-35	-	1.1e-12	7.1e-17	(7.6e-13)
snR35	-	8.5e-31	5.6e-28	1.3e-14	4.5e-07	(0.00043)
snR36	1.5e-22	3.4e-20	-	3.8e-13	2.4e-07	(6.4e-10)
snR42	2.4e-17	3.6e-41	4.0e-43	6.7e-11	5.7e-08	(3.1e-07)
snR44	3.7e-37	8.8e-34	1.9e-31	7.5e-11	1.7e-09	(-)
snR46	1.6e-28	2.6e-26	2.6e-30	1.4e-05	1.5e-05	(0.075)
snR49	5.3e-23	1.2e-22	9.8e-23	8.3e-09	7.4e-05	(7.6e-07)
snR161	1.3e-22	9.7e-20	1.4e-17	-	2.8e-05	(-)
snR189	1.6e-30	5.2e-29	3.8e-29	-	1.6e-08	(-)
snNOG2	2.0e-43	1.6e-47	3.3e-45	4.6e-17	5.7e-17	(1.2e-41)

Table 3.6: BLAST E-values (the expected number of sequence matches of this quality (Altschul *et al.*, 1990)) between *S.cerevisiae* and other *Saccharomyces* are shown for pair-wise alignments between the 18 known snoRNAs and their respective homologues genes (when present). The 5 *Saccharomyces* species sequenced by Cliften *et al.*, 2001 were used here. Absence of any hits of significance is indicated with a "-".
To investigate the feasibility of improving the quality of the results a second search of the *S.kluyveri* genome with the *S.castelli* results was made. The results are shown in brackets. This approach met with limited success, only snR36 and snR49 had more significant hits than the original search with *S.cerevisiae* sequences. This was not too suprising as both the *S.cerevisiae* and *S.castelli* lineages diverged from the *S.kluyveri* lineage at approximately the same time (see figure 3.1).

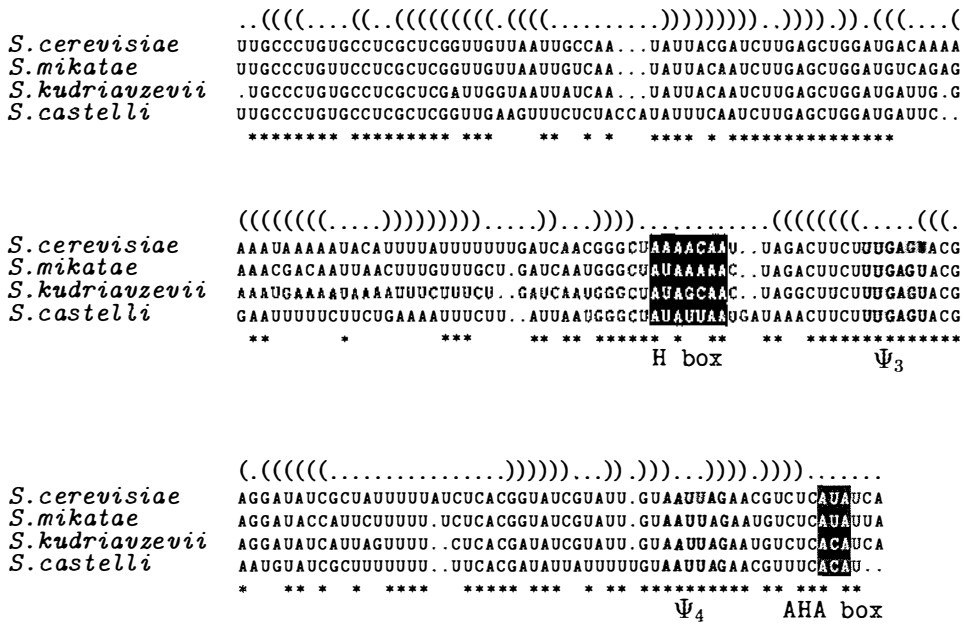


Figure 3.2: An alignment of homologous snR36 genes (Cliften *et al.*, 2001), homologous sequences were identified using BLAST (Altschul *et al.*, 1990) and aligned using CLUSTALW (Thompson *et al.*, 1994). The RNAalifold secondary structure is shown in dot-bracket notation above the alignment. Conserved positions in the alignment are indicated with a '*' below. .

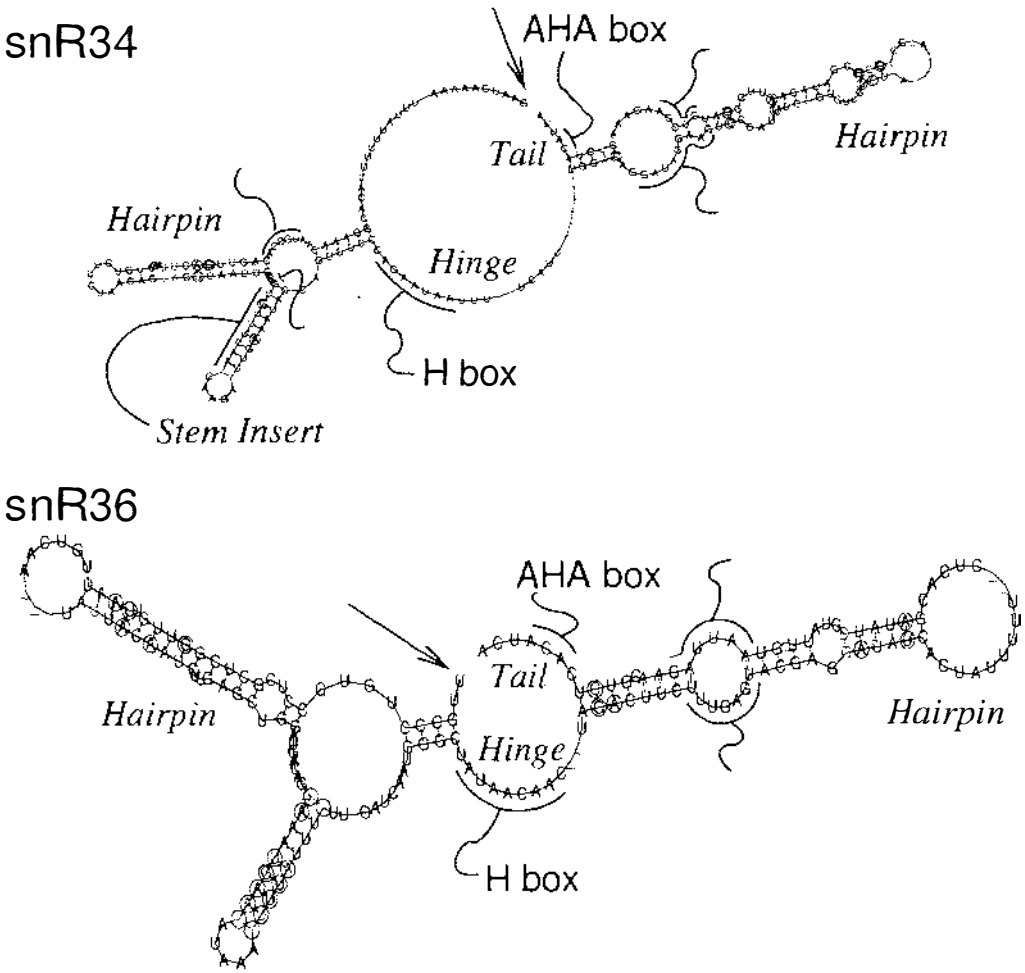


Figure 3.3: Inferred structures for snR34 and snR36. Circled nucleotides indicate a neutral mutation with respect to secondary structure. A base-pair with both nucleotides circled shows that compensatory mutations have maintained this base-pair. Characteristic structural elements of snoRNAs are the hairpin-hinge-hairpin-tail secondary structures (labeled in *italics*), the regions of complementarity to rRNA:Ψ_1/Ψ_2 and Ψ_3/Ψ_4, which reside in the same bulges (pseudouridylation pocket), the H box in the hinge region and the AHA box in the tail.

Comparative analysis of the 50 `Fisher` candidates

Encouraged by the comparative analysis of the known snoRNAs, we have used this approach to study the 50 candidates mentioned in Edvardsson *et al.*, (2003). Three candidates of particular interest were the two intronic to the ribosomal proteins (RPL43A and RPS11A) and the candidate completely overlapping the coding region of the snoRNP U3 protein MPP10. Unfortunately neither of these "interesting" candidates has been conserved, the host-genes of the intronic candidates are conserved yet the introns are mutated far more than one would expect for a snoRNA based upon the earlier study. The candidate overlapping MPP10 is also not conserved which is surprising as this sequence is thought to encode a protein in *S.cerevisiae.*

The results for the conserved candidates are shown in table 3.7 and secondary structures in figure 3.7. Only the cerevisiae group were used for this study as these have been shown in the previous section to be sufficiently close (in evolutionary terms) to locate homologous snoRNA genes, yet sufficiently divergent for gene-neutral mutations to appear. Comparing secondary structures of the known and candidate snoRNAs we note that many of the candidate snoRNAs do not conserve the classical hairpin-hinge-hairpin-tail secondary structure. In fact only candidates 35 and 37 maintain this structure. But upon analysis of the position of primary motifs within the secondary structures (see figure 3.4), our confidence in these candidates decreases. The predicted regions of complementarity to rRNA (Ψ_{1-4}) do not occur within the same bulge as they do in the majority of known snoRNAs (compare with figure 3.3). In fact the predicted H box isn't in the hinge region in either of the candidates, as it is in most of the known snoRNAs. This suggests that `Fisher` is not yet selective enough and requires further training.

`Fisher ver0.1` and comparative genomics

As a result of the analysis of the 50 likely candidates in the previous section an updated version of the `Fisher ver0.0` snoRNA search algorithm has been implemented (`Fisher ver0.1`). The new algorithm does not weight H/AHA box primary information as highly as the previous version. The original version modelled an H box as: $AN_1AN_2N_3N_4N_5$ ($N_i \in \{A,U,C,G\}$). With $N_1 \neq C$, $N_3 \neq G$, and either $N_4 = A$ or $N_5 = A$. The updated version no longer restricts the N_1 and N_3 positions in the motif or the H position of the AHA box. In addition the heights of the regions of complementarity (Ψ_1/Ψ_2 and Ψ_3/Ψ_4) are restricted to the same height ($\pm 3 nucs$) in the mountain plot of the MFE secondary structure. This forces candidates to have regions of complementarity in the same bulge and/or stem of

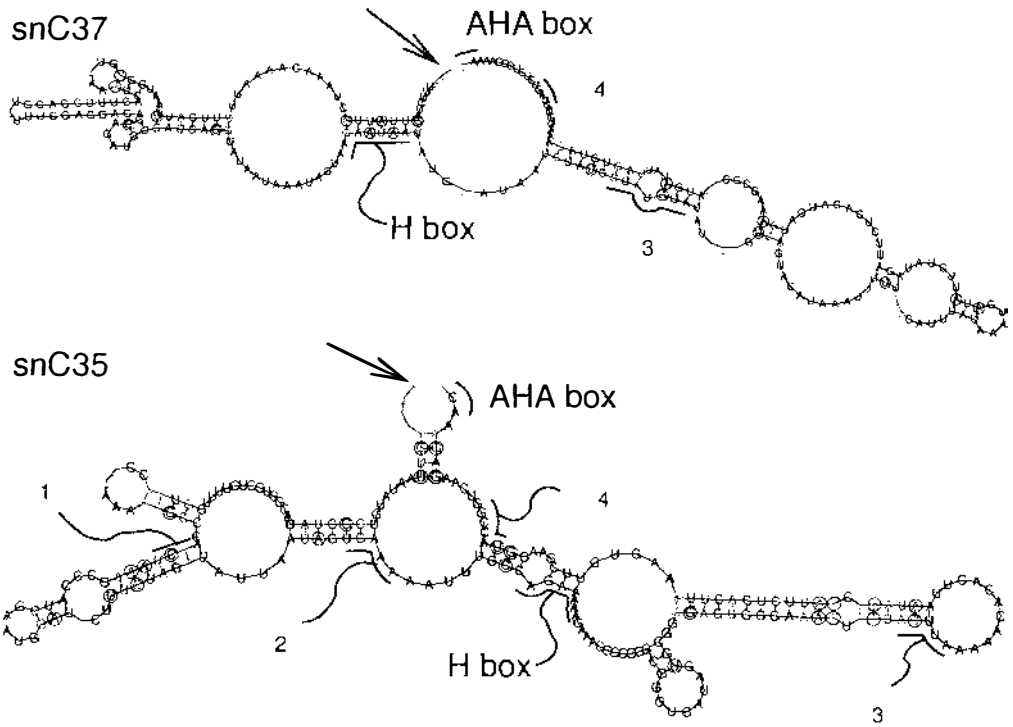


Figure 3.4: Inferred structures for sno-candidates 35 and 37. Circled nucleotides indicate a neutral mutation with respect to secondary structure. A base-pair with both nucleotides circled indicate compensatory mutations have maintained this base-pair. The positions of the primary motifs Ψ_{1-4}, H & AHA boxes are indicated.

---

the inferred secondary structure. In addition the filters restricting secondary and primary structures to score above a threshold ($P_1 > 0.8$ and $P_4 > 0.65$, see page ??) were removed. `Fisher ver0.1` was run upon the not-feature+introns *S.cerevisiae* data set and produced thousands of candidate snoRNAs. Candidate sequences with matching regions in *S.kudriavzevii* with E-values less than 10^{-10} and length greater than 150 were used for further analysis. These matching regions in *S.mikatae* and *S.bayanus* were identified for the conserved candidates if they existed. These were folded using `RNAalifold`, and those fitting the classical hairpin-hinge-hairpin-tail snoRNA shape were considered extremely likely candidates.

After analysing over 75% of the candidates identified by `Fisher ver0.1` only one candidate (snC53) passed all the filters for sno-likeness. This candidate is displayed in figure 3.5 and matches many of the characteristics of the known snoRNAs. However, when the sequence was used to probe 20μg of total *S.cerevisiae* RNA extract, it was not detected at the predicted size as a PCR product using a northern

candidate	*S.mikatae*	*S.kudriavzevii*	*S.bayanus*
snC3	4.7e-19	9.3e-16	2.8e-17
snC11	7.1e-06	3.4e-10	-
snC15	4.0e-23	2.6e-22	1.6e-17
snC16	5.4e-14	1.9e-12	-
snC22	3.9e-16	4.1e-18	5.1e-14
snC27	1.0e-17	8.1e-21	1.8e-20
snC35	2.9e-25	6.4e-20	2.3e-22
snC37	3.0e-20	4.0e-26	2.3e-18
snC38	3.1e-22	7.6e-27	5.5e-25
snC39	6.2e-21	7.6e-14	1.5e-13
snC42	7.1e-17	5.8e-15	-
snC43	3.4e-16	8.4e-10	1.6e-08
snC45	1.3e-26	1.4e-23	4.0e-29
snC46	8.3e-32	-	5.0e-25
snC47	3.6e-29	1.6e-28	4.3e-28

Table 3.7: Comparative sequence analysis of the 50 `Fisher ver0.0` candidates. BLAST E-values for pair-wise alignments of the 15 conserved candidates and their homologous in *S.mikatae*, *S.kudriavzevii* and *S.bayanus* are shown. The absence of any homologues is indicated by a "-".

blot analysis. Whereas the known snoRNAs that were used as a positive control were detected (A. Idicula, pers. commun.). Which means that based upon biochemical evidence, it is unlikely that snC53 is a novel snoRNA.

Discussion

It is pleasing to note that the hairpin-hinge-hairpin-tail secondary structure of H/ACA box snoRNAs is conserved in general. This was consistently detected by using alignments of the 4 cerevisiae group *Saccharomyces*: *S.cerevisiae*, *S.mikatae*, *S.kudriavzevii* and *S.bayanus* (where sequences were available). This increases the likelihood of success if a comparative genomic screen for H/ACA box snoRNAs is employed. Exceptions that did not conserve the canonical secondary structure were snR11 and snNOG2. snR11 consistently fails to form a stem between the H box to Ψ_3 motifs, yet still functions according to biochemical evidence (Samarsky & Fournier, 1999). snNOG2 displays a secondary structure closer to a hairpin-hinge-hairpin-hinge-hairpin-tail motif (see the structure labeled snNOG2 in figure 3.6) and therefore is not detected by genomic scans for the classical H/ACA snoRNA secondary structure. The other newly verified snR161 on the other hand was detected by `Fisher 0.0` but was conferred a poor score due to its primary structure and consequently wasn't included in the list of 50 candidates.

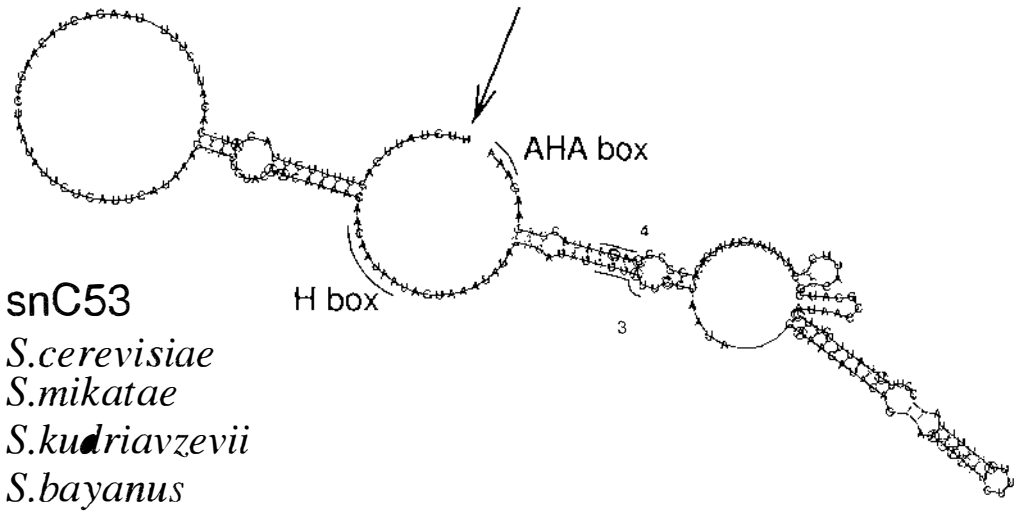


Figure 3.5: The extremely likely candidate identified by `Fisher ver0.1` and endorsed by comparative genomic analysis. It resides on *S.cerevisiae* chromosome IV between an ORF of "unknown function" (YDL159W-A) and a putative RNA helicase of the DEAD box family (DHH1). Homologues were located in *S.mikatae*, *S.kudriavzevii* and *S.bayanus*. This is the most likely candidate to date based upon conserved secondary structure and primary snoRNA motifs.

Upon reflection, this project has suffered from one basic limitation: we have assumed the characteristic snoRNA primary and secondary motifs are conserved by the yet-to-be-discovered-snoRNAs. Yet both the new snoRNAs, snR161 and snNOG2, show that this assumption is false, and has meant that novel snoRNAs were been neglected by this strategy. In addition, the characteristic primary and secondary structures of the snoRNAs is not a very selective signal for genomic scans. Particularly the `A`-rich primary motifs `ANANNAA` and `AHA` in a genome which has 60% `A+U` content. Hence, by chance alone we expect approximately 3 million subsequences of the yeast genome to match the H box. Perhaps a future successful screen for snoRNAs will follow the approach of Rivas & Eddy, 2001, by classifying aligned sequences derived from related species as either coding (protein or structural RNA) or non-coding on the basis of neutral mutations with respect to a gene-product. In fact at least one novel *S.cerevisiae* snoRNA has been located by these techniques (McCutcheon & Eddy, 2003). However, the snoRNA found by McCutcheon & Eddy's study was also discovered by `Fisher ver0.1`, but was in the 25% of candidates that had not undergone the comparative analysis phase.

snR3
S.cerevisiae
S.kudravzevii
S.bayanus

snR5
S.cerevisiae
S.kudriavzevii
S.bayanus
S.castellii

snR8
S.cerevisiae
S.mikitae
S.kudriavzevii
S.bayanus
S.castelli

snR10
S.cerevisiae
S.mikitae
S.kudriavzevii
S.bayanus
S.castelli
S.kluyveri

snR11
S.cerevisiae
S.kudriavzevii
S.bayanus
S.castelli

snR31
S.cerevisiae
S.mikitae
S.kudriavzevii
S.castelli

snR32
S.cerevisiae
S.mikitae
S.kudriavzevii
S.bayanus

snR33
S.cerevisiae
S.kudriavzevii
S.bayanus

snR34
S.cerevisiae
S.mikatae
S.kudriavzevii
S.castelli

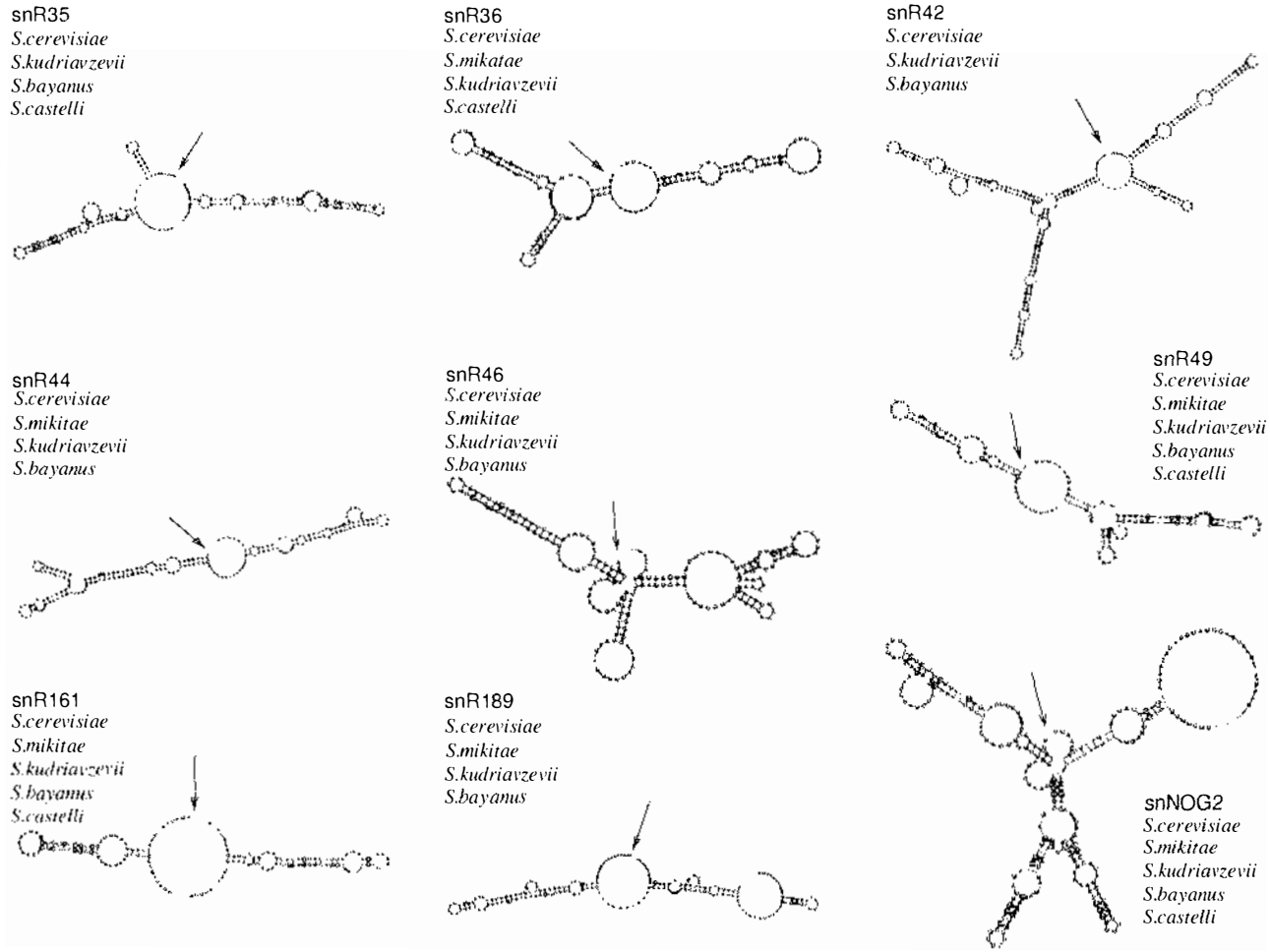
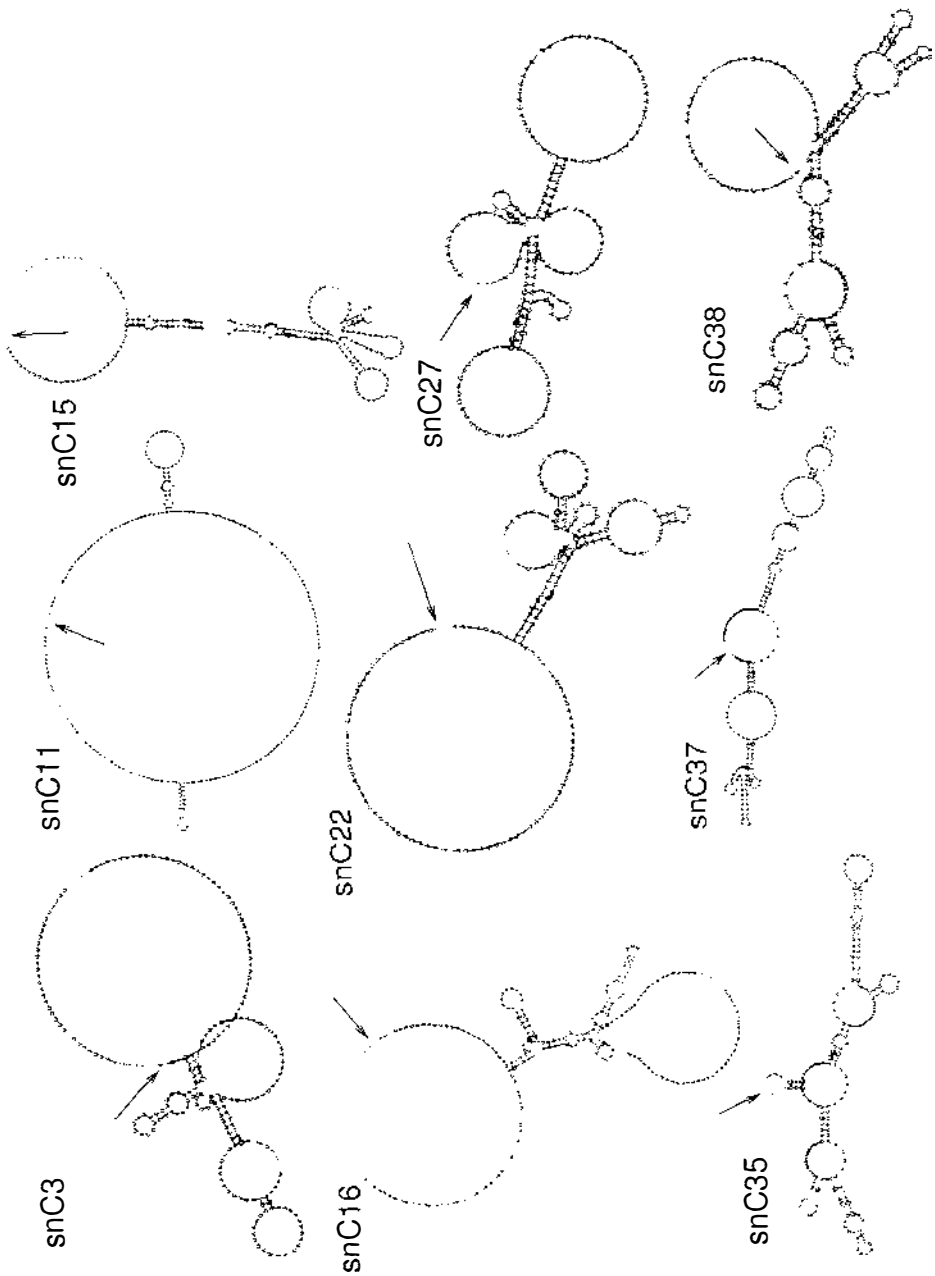


Figure 3.6:
Above figures show secondary structures predicted from alignments of homologous H/ACA snoRNA genes. The structures were inferred using RNAalifold (Hofacker *et al.*, 2002). Where available homologous genes from the following yeasts were used: *S. cerevisiae*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, *S. kluyveri* (Cliften *et al.*, 2001). The 5′ end of each sequence is indicated with an arrow.


snC15
snC27
snC38
snC11
snC22
snC37
snC3
snC16
snC35

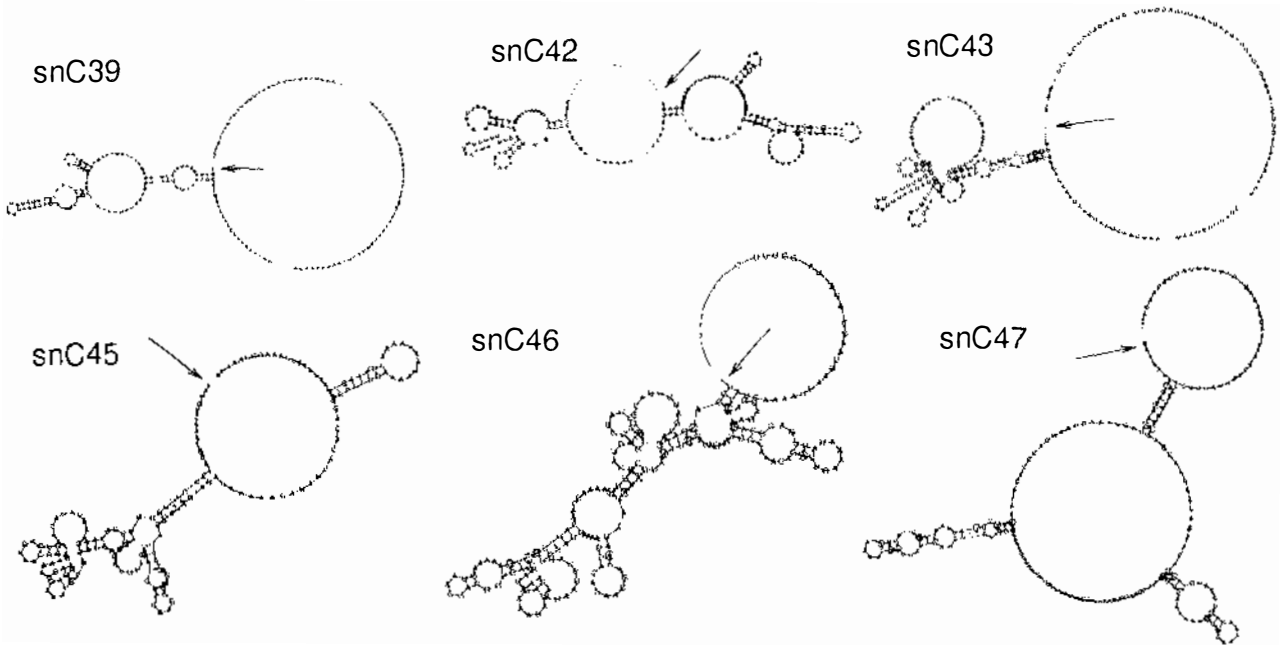


Figure 3.7:
Above figures show secondary structures predicted from alignments of the 15 conserved candidate snoRNA genes.
The structures were inferred using RNAalifold (Hofacker *et al.*, 2002). The yeasts used for this study were *S. cerevisiae*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* (Cliften *et al.*, 2001). The 5′ end of each sequence is indicated with an arrow.

Postscript

4.1 Future Directions

Comparative Genomics

Computational scans for RNA coding genes can be enhanced by using comparative genomic techniques to identify neutral mutations with respect to a gene product, thus increasing confidence in a prediction. This has been shown in Rivas *et al.*, 2001, in a test upon *E. coli.* Ideally a similar approach could be used to identify H/ACA specifically. A combination of a `Fisher` (Edvardsson *et al.*, 2003) and `QRNA` (Rivas & Eddy, 2001) type analysis that takes alignments as input rather than a genome has a much greater chance of success than our current work. However, this project relies heavily upon the availability of *Saccharomyces* sequence data. A group based in France have made sequences from 13 Hemiascomycetous yeast species freely available (Feldmann, 2000), one of which is *S.bayanus* which may be close enough in evolutionary terms to use for a comparative analysis.

Significance of Secondary Structure Signal for RNA Gene Finding

When using computational genomic screens to locate RNA genes often the most discriminatory signal comes from an inferred secondary structure (Lowe & Eddy, 1997*a*; Edvardsson *et al.*, 2003). But are all secondary structure signals equal? For example is a clover-leaf (tRNA) shape a better signal than a dual stem (H/ACA snoRNA) shape? I suspect that the dual-stem is a more common shape than the clover-leaf, therefore a genomic search for clover-leaf shapes is likely to contain less false positives than a search for the dual-stem.

I propose to use `RNAfold` (or equivalent) (Hofacker *et al.*, 1994) to fold subsequences of a fixed length from a sequenced genome. Course-grained shapes and a

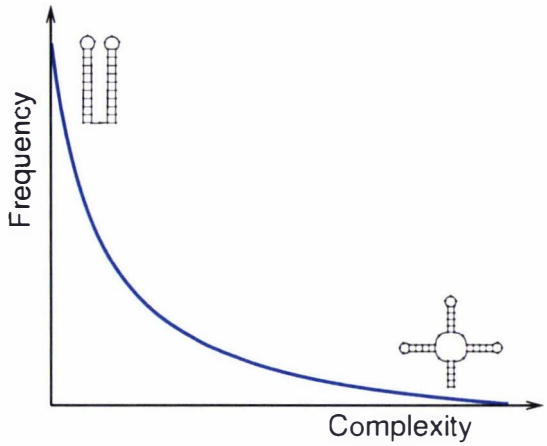


Figure 4.1: We hypothesise that the complexity of an RNA secondary structure is inversely proportional to its frequency in genomic sequences.

yet-to-be-developed complexity measure can then be used to empirically classify the resultant structures (see figure 4.1). It is possible that previous work by Meyer and Giegerich (2002) will prove useful for this project.

`MIfold`: reliably inferring RNA structure using mutual information

Current implementations that predict RNA secondary structure from an alignment using mutual information measures (introduced in chapter 1 (equations 1.8 and 1.9, page 19) and in equations 4.1 and 4.2), are limited in that a structure must be known *a priori* and are susceptible to noise. This is due to the fact that the mutual information is generally evaluated over all possible base-pairs rather than just the canonical pairs (see equations 4.1 & 4.2).

$$H(m,n) = \sum_{B_i,B_j} f_{m,n}(B_i,B_j) \times \log_2\{h_{mn}\}. \tag{4.1}$$

$$h_{mn} = \frac{f_{m,n}(B_i,B_j)}{f_m(B_i) \times f_n(B_j)}. \qquad B_i \in \{\mathtt{A},\mathtt{U},\mathtt{C},\mathtt{G}\} \tag{4.2}$$

A toolbox called `MIfold` is being developed for the mathematical package, `MATLAB`. To reduce noise, a modified mutual information measure is used where $H(m,n)$ is only evaluated over the canonical base-pairs: $(B_i,B_j) \in \{(\mathtt{A},\mathtt{U}),(\mathtt{U},\mathtt{A}),(\mathtt{C},\mathtt{G}),(\mathtt{G},\mathtt{C}),(\mathtt{G},\mathtt{U}),(\mathtt{U},\mathtt{G})\}$. A mountain-plot is constructed using $maxMI(m) = \max_n(H(m,n))$ values, with the further restriction that $maxMI(m)$ exceeds a threshold T, i.e. $maxMI(m) > T$ (see figure 4.2).

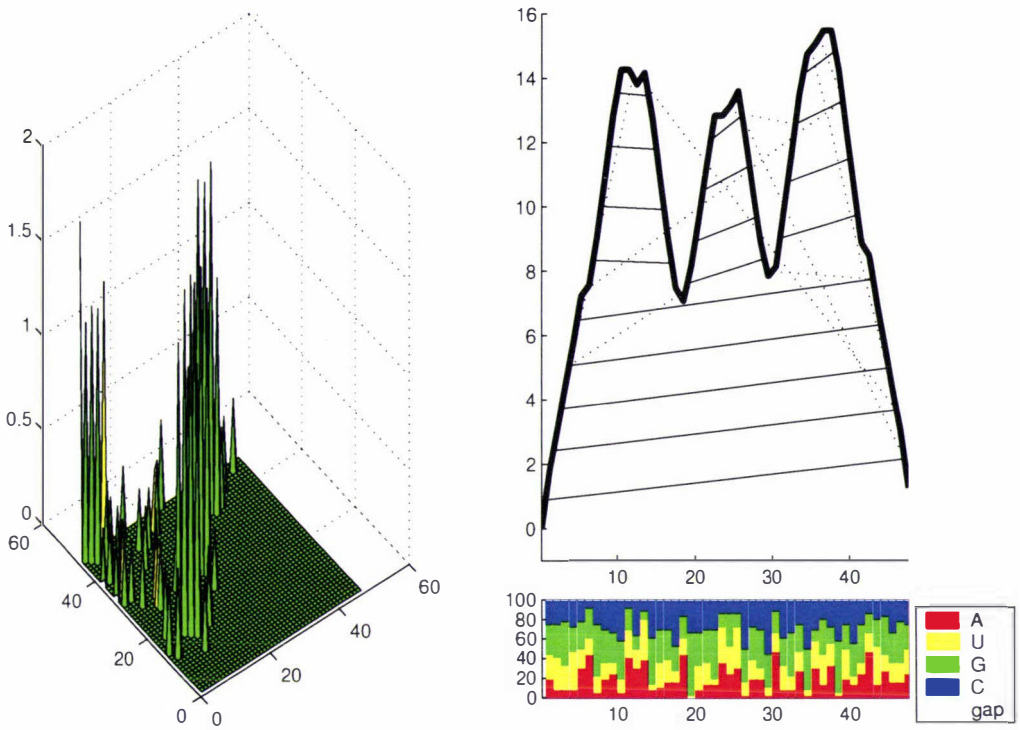


Figure 4.2: A figure generated by `MIfold` (E. Freyhult, pers. commun.) from 40 sequences produced by RNAinverse (available with the `VIENNA` package (Hofacker *et al.*, 1994)), all the sequences in the alignment have the clover-leaf as an MFE structure. The figure on the left shows $H(m,n)$, spikes show alignment positions m and n where $H(m,n)$ is greater than the threshold 0.3. The mountain-plot on the right is incremented by: $\mathrm{sgn}(n-m) \times (\max_n(H(m,n))) \iff maxMI(m) > 0.3$. Below the mountain-plot is a colourmap showing the relative frequencies of each nucleotide in each position of the alignment.

Bibliography

Akmaev, V. R., Kelley, S. T. & Stormo, G. D. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics,* **16** (6), 501–512.

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215**, 403–410.

Antao, V. P. & Tinoco, I. (1992) Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Research,* **20** (4), 819–824.

Beitz, E. (2000) TeXshade: shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics,* **16**, 135–139.

Borer, P., Dengler, B., Tinoco, I. & Uhlenbeck, ●. (1974) Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Evolution,* **86**, 843-853.

Cech, T. (1986) RNA as an enzyme. *Scientific American,* **255** (5), 64.

Cech, T. (1987) The chemistry of self-splicing RNA and RNA enzymes. *Science,* **236**, 1532–1539.

Cervelli, M., Cecconi, F., Giorgi, M., Annesi, F., Oliverio, M. & P., M. (2002) Comparative structure analysis of vertebrate U17 small nucleolar RNA (snoRNA). *J. Mol. Evol.,* **54**, 166–179.

Cherry, J. M., Ball, C., Dolinski, K., Dwight, S., Harris, M. Matese, J., Sherlock, G., Binkley, G., Jin, H., Weng, S. & Botstein, D. (2001). *Saccharomyces* genome database. *ftp://genome-ftp.stanford.edu/pub/yeast/yeast_NotFeature/.*

Cliften, P., Hillier, L., Fulton, L., Graves, T., Miner, T. & Gish, W. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.,* **11**, 1175–1186.

Cliften, P., Hillier, L., Fulton, L., Graves, T., Miner, T. & Gish, W. (2002). *Saccharomyces* genome sequencing at the GSC. World Wide Web. http://genome.wustl.edu/projects/yeast/index.php.

Crick, F. (1968) The origin of the genetic code. *J. Mol. Biol.,* **38**, 367–379.

Dennis, C. (2002) The brave new world of RNA. *Nature,* **418**, 122–124.

Dirheimer, G., Keith, G., Dumas, P. & Westhof, E. (1995) *RNA: Structure, Biosynthesis and Function.* American Society for Microbiology, Washington, DC pp. 93–126.

Doudna, J. & Cech, T. (2002) The natural chemical repertoire of natural ribozymes. *Nature,* **418**, 222–228.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic models of protein and nucleic acids.* Cambridge University Press.

Eddy, S. (2001) Noncoding RNA genes and the modern RNA world. *Nat. Rev. Genet.,* **2**, 919–29.

Eddy, S. (2002) Computational genomics of noncoding RNA genes. *Cell,* **109**, 137–140.

Eddy, S. & Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.,* **22**, 2079–2088.

Edvardsson, S., Gardner, P., Poole, A., Hendy, M., Penny, D. & Moulton, V. (2003) A search for pseudouridylation guide snoRNAs using predicted MFE secondary structures. *Bioinformatics,* **19** (7), 865–873.

Eigen, M., McCaskill, J. & Schuster, P. (1988) Molecular quasispecies. *J. Phys. Chem.,* **92**, 6881–6891.

Eigen, M., McCaskill, J. & Schuster, P. (1989) The molecular quasi-species. *Adv. Chem. Phys.,* **75**, 149–263.

Feldmann, H. (2000) Génolevures - a novel approach to 'evolutionary genomics'. *FEBS Letters,* **487** (1), 1–2.

Fontana, W. (2002) Modelling 'evo-devo' with RNA. *BioEssays,* **24**, 1164–1177.

Gardner, P., Holland, B., Hendy, M., Moulton, V. & Penny, D. (2003) Optimal alphabets for an RNA-world. *Proceedings of the Royal Society of London, Series B.,* **270** (1520), 1177–1182.

Gesteland, R. & Atkins, J., eds (1993) *The RNA World.* Cold Spring Harbor, NY, first edition,, Cold Spring Harbor Laboratory Press.

Gesteland, R., Cech, T. & Atkins, J., eds (1999) *The RNA World.* Cold Spring Harbor, NY, second edition,, Cold Spring Harbor Lab Press.

Giegerich, R., Haase, D. & Rehmsmeier, M. (1999) Prediction and visualization of structural switches in RNA. *Pacific Symposium on Biocomputing,* **4**, 126–137.

Gilbert, W. (1986) The RNA world. *Nature,* **319**, 618.

Goldberg, D. E. (1989) *Genetic algorithms in search, optimization and machine learning.* Addison-Wesley, Reading, MA.

Gorodkin, J., Heyer, L., Brunak, S. & Stormo, G. (1997) Displaying the information contents of structural RNA alignments. *Bioinformatics,* **13**, 583–586.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Schuster, P. & Stadler, P. F. (1996) Analysis of RNA sequence structure maps by exhaustive enumeration. *Monatshefte für Chemie,* **127** (4), 355–374.

Hamming, R. (1950) Error detecting and error correcting codes. *Syst. Tech. J.,* **29**, 147–160.

Henderson, R. & Tweten, D. (1996). Portable batch system: external reference specification. Technical report NASA Ames Research Center.

Hofacker, I., Fekete, M. & Stadler, P. (2002) Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology,* **319** (5), 1059 1066.

Hofacker, I. L., Fontana, W., Bonhoeffer, S. & Stadler, P. F. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie,* **125**, 167–188.

Hofacker, I. L., P., S. & F., S. P. (1998) Combinatorics of RNA secondary structures. *Discrete Applied Mathematics,* **88**, 207–237.

Hogeweg, P. & Hesper, B. (1984) Energy directed folding of RNA sequences. *Nucleic Acids Res.,* **12**, 67–74.

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor.

Jadhav, V. & Yarus, M. (2002) Coenzymes as coribozymes. *Biochimie,* **84** (9), 877–888.

James, S., Cai, J., Roberts, I. & Collins, M. (1997) A phylogenetic analysis of the genus *Saccharomyces* based on 18S rDNA gene sequences: description of *Saccharomyces kunashirensis* sp. nov. and *Saccharomyces martinae* sp. nov. *Int. J. Syst. Bacteriol.,* **47**, 453–460.

Jeffares, D. C., Poole, A. M. & Penny, D. (1998) Relics from the RNA world. *Journal of Molecular Evolution,* **46**, 18–36.

Joyce, G. (2002) The antiquity of RNA-based evolution. *Nature,* **418**, 214–221.

Kernighan, B. W. & Ritchie, D. M. (1988) *The C Programming Language.* Second edition,, Prentice Hall, Englewood, New Jersey.

Kierzek, R., Caruthers, M., Longfellow, C., Swinton, D., Turner, D. & Freier, S. (1986) Polymer-supported RNA synthesis and its application to test the nearest-neighbor model for duplex stability. *Biochemistry,* **25**, 7840–7846.

Kim, S. H., Suddath, G. J., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H. J., Seeman, N. C. & Rich, A. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science,* **185**, 435–439.

Klein, R., Misulovin, Z. & Eddy, S. (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences,* **99**, 7542–7547.

Koski, L. B. & Golding, G. B. (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution,* **52** (6), 540–542.

Kruger, K., Grabowski, P., Zaug, A., Sands, J., Gottschling, D. & Cech, T. (1982) Self-splicing RNA - auto-excision and auto-cyclization of the ribosomal-RNA intervening sequence of *Tetrahymena. Cell,* **31** (1), 147–157.

Kurtzman, C. & Robnett, C. (1998) Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology,* **73** (4), 331–371.

Le, S. & Zuker, M. (1991) Predicting common foldings of homologous RNAs. *Journal of Biomolecular Structure and Dynamics,* **8**, 1027–1044.

Lesk, A. (1991) *Protein Architechture: A Practical Approach.* Oxford University Press.

Lowe, T. (2002). GtRDB: the genomic tRNA database. World Wide Web. http://rna.wustl.edu/GtRDB/.

Lowe, T. M. & Eddy, S. R. (1997*a*) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research,* **25** (5), 955–964.

Lowe, T. M. & Eddy, S. R. (1997*b*) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research,* **25** (5), 955–964.

Lowe, T. M. & Eddy, S. R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science,* **283** (5405), 1168–1171.

Mac Dónaill, D. (2002) A parity code interpretation of nucleotide alphabet composition. *Chem. Comm.,* **18**, 2062–2063.

Man, K., Tang, K. & Kwong, S. (1999) *Genetic Algorithms.* Springer-Verlag London Limited.

Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA. *Journal of Molecular Biology,* **288** (5), 911–940.

Mathews, D.H. Sabina, J. Z. M. & Turner, H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.,* **288**, 911–940.

Mattick, J. & Gagen, M. (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.,* **18**, 1611–1630.

McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers,* **29**, 1105–1119.

McCutcheon, J. & Eddy, S. (2003). Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. Unpublished preprint. `www.genetics.wustl.edu/eddy/publications`.

Message Passing Interface Forum (1994) MPI: a message-passing interface standard. *Journal of Supercomputing Applications,* **8** (3/4).

Meyer, C. & Giegerich, R. (2002) Matching and significance evaluation of combined sequence-structure motifs in RNA. *Zeitschrift fur physikalische chemie-international journal of research in physical chemistry and chemical physics,* **216** (2), 193–216.

Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research,* **29** (13), 2850–2859.

Moulton, V., Gardner, P., Pointon, R., Creamer, L., Jameson, G. & Penny, D. (2000*a*) RNA folding argues against a hot-start origin of life. *Journal of Molecular Evolution,* **51**, 416–421.

Moulton, V., Zuker, M., Steel, M., Pointon, R. & Penny, D. (2000*b*) Metrics on RNA secondary structures. *Journal of Computational Biology,* **7** (1-2), 277–292.

Mount, D. (2001) *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press.

Nissen, P., Hansen, J., Ban, N., Moore, P. & Steitz, T. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science,* **289** (5481), 920–930.

Ofengand, J. & Fournier, M. (1998) *Modification and Editing of RNA.* ASM Press pp. 229–253.

Orgel, L. (1968) Evolution of the genetic apparatus. *J. Mol. Biol.,* **38**, 381–393.

Pace, N., Smith, D., Olsen, G. & James, B. (1989) Phylogenetic comparative analysis and the secondary structure of ribonuclease. *Gene,* **82**, 65–75.

Papanicolaou, C., Gouy, M. & Ninio, J. (1984) An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. *Nucleic Acids Research,* **12**, 31–44.

Parsch, J., Braverman, J. & Stephan, W. (2000) Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics,* **154** (2), 909–921.

Piccirilli, J., Krauch, T., Moroney, S. & Benner, S. (1990) Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature,* **343**, 33–37.

Poole, A., Jeffares, D. & Penny, D. (1999) Early evolution: prokaryotes, the new kids on the block. *BioEssays,* **21**, 880–889.

Poole, A. M., Jeffares, D. C. & Penny, D. (1998) The path from the RNA world. *Journal of Molecular Evolution,* **46**, 1–17.

Rivas, E. & Eddy, S. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics,* **2**, 8.

Rivas, E. & Eddy, S. R. (2000*a*) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics,* **16** (7), 583–605.

Rivas, E. & Eddy, S. R. (2000*b*) The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics,* **16** (4), 334–340.

Rivas, E. & Eddy, S. R. (2000*c*) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics,* **16** (7), 583–605.

Rivas, E., Klein, R., Jones, T. & Eddy, S. (2001) Computational identification of noncoding RNAs in *e. coli* by comparative genomics. *Curr. Biol.,* **11**, 1369–73.

Samarsky, D. & Fournier, M. (1999) A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae. Nucleic Acids Res.,* **27**, 161–164.

Sankoff, D., Morin, A. & Cedergren, R. (1978) The evolution of 5S RNA secondary structures. *Canadian Journal of Biochemistry,* **56**, 440–443.

Schattner, P. (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Research,* **30** (9), 2076–2082.

Schultes, E. & Bartel, D. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science,* **289**, 448–452.

Schultes, E. A., Hraber, P. T. & LaBean, T. H. (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *Journal of Molecular Evolution,* **49**, 76–83.

Schuster, P. (1993) RNA based evolutionary optimization. *Origins of Life and Evolution of the Biosphere,* **23**, 373–391.

Schuster, P. (2000) *Evolutionary Dynamics – Exploring the Interplay of Accident, Selection, Neutrality, and Function.* New York: Oxford University Press.

Schuster, P., Fontana, W., Stadler, P. & Hofacker, I. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London, Series B.,* **255**, 279–284.

Seffens, W. & Digby, D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research,* **27** (7), 1578–1584.

Shapiro, B. & Zhang, K. (1990) Comparing multiple RNA secondary structures using tree comparisons. *CABIOS,* **6**, 309–318.

Shapiro, B. A. (1988) An algorithm for comparing multiple RNA secondary structures. *CABIOS,* **4**, 381–393.

Shi, H. & Moore, P. B. (2000) The crystal structure of yeast phenylalanine tRNA at 1.93 a resolution: a classic structure revisited. *RNA,* **6**, 1091–1105.

Spirin, A. (2002) Omnipotent RNA. *FEBS Letters,* **530** (1-3), 4–8.

Stryer, L. (1995) *Biochemistry.* Fourth edition,, W. H. Freeman and Company.

Sundaralingham, M. & Rao, S., eds (1975) *Structure and Conformation of Nucleic Acids and Protein-Nucleic Acid Interactions.* University Park Press pp. 12–13.

Szathmáry, E. (1991) Four letters in the genetic alphabet: a frozen evolutionary optimum? *Proceedings of the Royal Society of London, Series B.,* **245**, 91–99.

Szathmáry, E. (1992) What is the optimal size for the genetic alphabet? *Proceedings of the National Academy of Sciences, USA,* **89**, 2614–2618.

Szostak, J. W. (1993) Ribozymes - evolution *exvivo. Nature,* **361** (6408), 119–120.

Thompson, J., Higgins, D. & Gibson, T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* **22**, 4673–4680.

Tinoco, I., Uhlenbeck, O. & Levine, M. (1971) Estimation of secondary structure in ribonucleic acids. *Nature,* **230**, 362–367.

Tuerk, C. & Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science,* **249**, 505–510.

Waterman, M. (1995) *Introduction to Computational Biology: Maps, sequences and genomes.* Chapman and Hall. Chapter 13.

White, H. B. (1976) Coenzymes as fossils of an earlier metabolic state. *Journal of Molecular Evolution,* **7**, 101–104.

Witwer, C., Rauscher, S., Hofacker, I. & Stadler, P. (2001) Conserved RNA secondary structures in *Picornaviridae* genomes. *Nucleic Acids Research,* **29** (24), 5079–5089.

Woese, C. (1967) *The genetic code.* Harper and Row, New York.

Woese, C., Gutell, R., Gupta, R. & Noller, H. (1983) Detailed analysis of the higher-order structure of the 16S-like ribosomal ribonucleic acids. *Microbiology Review,* **47**, 621–669.

Woese, C. & Pace, N. (1993) *The RNA World.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY pp. 91–117.

Workman, C. & Krogh, A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.,* **27**, 4816–4822.

Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers,* **49** (2), 145–165.

Zuker, M. (2000) Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology,* **10** (3), 303–310.

Zuker, M. & Le, S. (1990) Common structures of the 5' non-coding RNA in enteroviruses and rhinoviruses. thermodynamical stability and statistical significance. *Journal of Molecular Biology,* **216**, 729–741.

Zuker, M., Mathews, D. & Turner, D. (1999) *Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide in RNA biochemistry and biotechnology.* NATO ASI Series, Kluwer Academic Publishers.

Zuker, M. & Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research,* **9**, 133–148.

Appendix I

Title: RNA Folding Argues Against a Hot-Start Origin of Life

Author: Vincent Moulton, Paul P. Gardner, Robert F. Pointon, Lawrence K. Creamer, Geoffrey B. Jameson, David Penny.

Year: 2000

Journal: *Journal of Molecular Evolution*

Volume: 51

Page: 416-421

Appendix II

Software used for the presentation of this thesis:

LaTeX 2ε	www.latex-project.org
MatLab	www.mathworks.com
Debian (Linux)	www.debian.org
Emacs	www.emacs.org
Xfig	www.xfig.org
TeXshade	http://homepages.uni-tuebingen.de/beitz/tse.html
Protein Explorer	http://proteinexplorer.org

Software used to gather results for this thesis:

gcc (GNU C compiler)	http://gcc.gnu.org
-Revolver	email: P.P.Gardner@massey.ac.nz
-RiboRace	email: P.P.Gardner@massey.ac.nz
-Fisher	email: Sverker.Edvardsson@mh.se
Message Passing Interface (MPI)	www-unix.mcs.anl.gov/mpi
Portable Batch System (PBS)	www.openpbs.org
Vienna ver 1.4&1.5	www.tbi.univie.ac.at/~ivo/RNA/
-RNAfold	
-RNAalifold	

Index