Acknowledgments

I would have not been able to have produced this dissertation without the help and support of many people. First and foremost, I wish to thank my advisor, Sean Eddy, for teaching me to think and work as an independent scientist while providing advice and insight where needed. I am sure these skills will serve me well as I move onto postdoctoral research. I am also grateful to all the members, past and present, of my thesis committe – Michael Zuker, Jane Wu, Jan Amend, Barak Cohen, Warren Gish, Mark Johnston, Robi Mitra, and Gary Stormo – for their advice and support in the various turns my research has taken.

Jan deserves especial thanks for letting me use his lab to grow *Pyrococcus furiosus*. His graduate students D’Arcy Meyer-Dombard and Karyn Rogers not only showed me how to do so but welcomed me into the lab as well. I am also grateful to Jim Brown (North Carolina State University) for *Methanocaldococcus jannaschii* cell paste, Frank Robb (Center of Marine Biotechnology) for the initial *P. furiosus* cell culture, and Michael Terns (University of Georgia) for additional *P. furiosus* RNA.

All the members of the Eddy lab were a great source of advice and support during the course of the work. Steve Johnson and Christian Zmasek pulled memories of wet lab from the recesses of their mind when I needed their insight, while Elena Rivas and Robin Dowell gave me much advice on statistics and algorithms. Ziva Misulovin provided invaluable technical assistance in performing the PCR reactions, cloning, and sequencing for the RACE experiments – a monumental task by any measure. Shawn Stricklin and Robin Dowell were kind enough to proofread my thesis on extremely short notice.

Finally, I wish to thank all my friends and family, both here in St. Louis and back in New York, for being an ear to kvetch to during the roller-coaster ride of graduate school.
Contents

Acknowledgments ii
List of Tables v
List of Figures vi
Abstract viii

1 Introduction 1
   Noncoding RNAs ................................................................. 2
   The problem of gene finding .................................................. 4
   Protein coding gene finding .................................................... 5
      Protein coding gene finding de novo ...................................... 6
      Protein coding gene finding by homology .................................. 8
   The Importance of Secondary Structure ...................................... 9
      Finding the secondary structure of a single RNA molecule ................. 11
      Finding the common secondary structure among many RNA molecules ...... 13
      Finding the common secondary structure of a pair of RNA molecules ....... 15
   Noncoding RNA gene finding ................................................. 16
      De novo noncoding RNA gene finding ....................................... 17
      Finding noncoding RNAs by homology ..................................... 18
   Outline of this work ........................................................... 21

2 Noncoding RNA genes identified in AT-rich hyperthermophiles 22
   Abstract .............................................................................. 23
   Introduction .......................................................................... 23
   Materials and Methods ............................................................ 25
      Genomes Used ...................................................................... 25
      GC Content Screens ................................................................ 26
      QRNA Comparative Analysis .................................................. 28
      Growth conditions and RNA extraction ..................................... 29
      Northern blotting .................................................................. 29
      RACE-PCR Analysis .............................................................. 32
      Computational Analysis of the RNAs ...................................... 34
   Results ................................................................................. 34
# List of Tables

2.1 Parameters used in hidden Markov model ........................................... 26  
2.2 Oligonucleotides for northern blots .................................................. 31  
2.3 Oligonucleotides for RACE ................................................................. 33  
2.4 Candidate noncoding RNAs ................................................................. 38  

3.1 Node-state pairs .................................................................................... 65  
3.2 Gap penalties ......................................................................................... 67  

4.1 Oligonucleotides for RACE ................................................................. 92  
4.2 tRNA sequences of various degrees of similarity to RD0500 ................. 94  
4.3 Sequences used for pairwise folding ..................................................... 96  
4.4 *E. coli* RACE results ............................................................................ 97  
4.5 RSEARCH searches with predicted structures ...................................... 102  

5.1 Homologs of *hgcC* ............................................................................. 121
## List of Figures

2.1 The GC content hidden Markov model ..................................... 27
2.2 GC contents of various genomes ........................................... 36
2.3 Northern blots of candidate ncRNAs ...................................... 42
2.4 Expressed RNA loci .......................................................... 43
2.5 Detailed analysis of \textit{hgcC} .............................................. 44
2.6 Detailed analysis of sR9 ..................................................... 46

3.1 Example SCFG architecture .................................................. 64
3.2 Local alignment ............................................................... 68
3.3 RIBOSUM85-60 matrix ....................................................... 75
3.4 Statistics of RSEARCH searches .......................................... 78
3.5 \textit{C. elegans} RNase P .................................................... 83
3.6 Other eukaryotic RNase Ps ................................................ 85

4.1 Evaluation of various folding methods ................................... 100
4.2 Best percentage identity ..................................................... 107
4.3 Homologs of \textit{E. coli} ncRNAs ........................................... 108
4.4 Homologs and structure of tp1 ........................................... 109
4.5 Homologs and structure of t44 ............................................ 110

5.1 Homolog of \textit{hgcG} in \textit{M. jannaschii} ................................ 116
5.2 \textit{hgcA} as a pseudouridylation guide .................................. 117
5.3 \textit{hgcC} and homologs ..................................................... 119
5.4 Multiple sequence alignment of \textit{hgcC} and homologs ............... 120
5.5 Analysis of \textit{hgcB} .......................................................... 123
Abstract

Noncoding RNA genes are genes for which the transcribed RNA is not translated into protein. Instead, the RNA transcript is the functional end product in the cell. Functional specificity is conferred upon many noncoding RNAs by the presence of secondary structure. How noncoding RNA genes can be identified in genomic sequences is an open problem that this dissertation addresses.

I first performed a screen for noncoding RNA genes in the genomes of *Methanocaldococcus jannaschii* (formerly *Methanococcus jannaschii*) and *Pyrococcus furiosus* by exploiting a bias in the GC content of their noncoding RNAs. I also screened for noncoding RNAs in *P. furiosus* using the comparative analysis program QRNA. Expression of nine of these computational predictions as small, stable RNAs has been confirmed on Northern blots and through RACE reactions. One RNA, *hgcC*, appears to be associated with transposition events. The gene *hgcG* from *P. furiosus* is homologous to the pseudouridylation guide RNA Afu-4 in *Archaeoglobus fulgidus*. In no other cases can BLAST find a homolog of any of these RNAs in a non-*Pyrococcus*, non-*Methanocaldococcus* genome.

I then developed a program, RSEARCH, that finds homologs of a single, structured RNA in genomic sequences. RSEARCH finds previously unknown homologs of human RNase P in the eukaryotic genomes *Ciona intestinalis*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Anopheles gambiae*, and *Encephalitozoon cuniculi*. In several tests, queries with structures predicted by the pairwise folding program DYNALIGN perform almost as well as the trusted structure from the literature. Using such predicted structures, RSEARCH finds homologs of several small noncoding RNAs from *Escherichia coli* in other gamma proteobacteria.
RSEARCH finds additional homologs of several of the RNAs I found in *M. jannaschii* and *P. furiosus*. It finds a homolog of *hgcG/Afu-4* in *M. jannaschii* that further supports the hypothesis that this RNA is a pseudouridylation guide RNA. It finds additional homologs of *hgcC* adjacent to a transposase. Finally, it finds that *hgcB* from *M. jannaschii* is homologous to presumably transcribed sequence 5’ of the small subunit ribosomal RNA gene in *Pyrococcus*.  

ix
Chapter 1

Introduction
Noncoding RNAs

Noncoding RNA genes are any genes for which RNA is the functional end product rather than an intermediary messenger. The concept of noncoding RNA dates back to the initial discovery that low-abundance, unstable RNAs, not ribosomal RNA, is the messenger (Brenner et al., 1961; Gros et al., 1961). While many assumed the ribosomal RNA merely served a structural role, the determination of the crystal structure of the large subunit of the ribosome revealed that it is ribosomal RNA, not ribosomal protein, that is present at the site of peptidyl transfer activity (Ban et al., 2000; Nissen et al., 2000). Thus, the noncoding ribosomal RNA functions as an enzyme, though the exact nature of catalysis remains unclear (Nissen et al., 2000; Polacek et al., 2001).

Many other noncoding RNAs have been individually identified in various studies. The bulk of these were identified biochemically. One of the first noncoding RNAs to be identified was transfer RNA (tRNA), found when Hoagland and colleagues noted that specific cytoplasmic RNAs pick up radiolabelled leucine and transfer it to the growing polypeptide chain (Hoagland et al., 1958). The 6S RNA in *Escherichia coli* was first identified as a band on a gel in 1967 (Hindley, 1967). It was not until 2000 that its role as a regulator of RNA polymerase activity was elucidated (Wassarman and Storz, 2000). Biochemically purified ribonuclease P (RNase P) activity from both *E. coli* and *Bacillus subtilis* was found to contain an RNA and a protein subunit, both of which are essential for function (Stark et al., 1978; Gardiner and Pace, 1980). Later experiments showed that the catalytic subunit of this enzyme is the RNA rather than the protein subunit (Guerrier-Takada et al., 1983).

A much smaller number of noncoding RNAs were initially identified genetically. The
genes lin-4 and let-7 in Caenorhabditis elegans are two such genes. lin-4 was first isolated in a mutant screen and characterized as having “abnormal post-embryonic cell lineages” (Horvitz and Sulston, 1980). Using classical epistasis analysis, Ambros later showed that lin-4 is upstream of the genes lin-14 and lin-28 (Ambros, 1989). Later experiments showed that lin-14 is downregulated through its 3’ UTR and that lin-4 is necessary for this regulation (Wightman et al., 1991; Arasu et al., 1991). From this work, Ruvkun and colleagues (correctly) suggested the simple model that the lin-4 gene product directly interacts with the lin-14 UTR (Arasu et al., 1991). In working to clone the lin-4 gene, Lee, Feinbaum, and Ambros rigorously narrowed down the genomic interval of interest and systematically mutated all ORFs there to demonstrate that the lin-4 gene product is not a protein. Instead, they found that the lin-4 gene product is a small noncoding RNA that shows antisense complementarity to the lin-14 3’-UTR (Lee et al., 1993). A later study found and cloned a gene, let-7, that appears to encode a similar small noncoding RNA that functions in a similar manner (Reinhart et al., 2000). Many more of these micro RNAs (miRNAs) have been found in studies that will be discussed later.

One example where both biochemistry and genetics identified a noncoding RNA, but where the RNA was only recognized through biochemistry at first, is the csrB gene in E. coli. In studies on the global regulatory gene csrA in E. coli, Romeo and colleagues isolated recombinant, histidine-tagged CsrA. They found that the CsrA protein binds a previously unknown RNA molecule, CsrB (Liu et al., 1997). When they searched this new RNA against the sequence database, they discovered it was homologous to a locus, aepH, in the bacteria Erwinia carotovora. aepH was identified using EMS mutagenesis as being involved in the regulation of extracellular enzyme levels (Murata et al., 1994). In this study it was
suggested that a short open reading frame (ORF) at the *aepH* locus codes for the protein gene product (Murata et al., 1994). Comparison with the *csrB* sequence strongly argues that *aepH* is the homolog of *csrB* in *E. carotovora* and thus a noncoding RNA (Liu et al., 1997).

These examples are merely illustrative of the various ways noncoding RNA genes have been discovered; they probably come nowhere near representing the quantity and functional diversity of noncoding RNAs that are known (Erdmann et al., 2001; Eddy, 2001; Storz, 2002). Some other known noncoding RNAs to be discussed later deserve mention here. The 7S RNA, now known as the signal recognition particle (SRP) RNA, is involved in targeting secretory and membrane proteins to the appropriate membrane in the cell (Walter and Blobel, 1982; Keenan et al., 2001). Most of the C/D box and H/ACA box small nucleolar RNAs (snoRNAs) guide methylation of 2’ hydroxyl groups and pseudouridylation of uridine bases in target RNA molecules, respectively (Kiss-Laszlo et al., 1996; Ganot et al., 1997). They have now been found not only in eukaryotes but in Archaea as well (Omer et al., 2000; Tang et al., 2002a). In *E. coli*, a large number of small RNAs have been found to act as “riboregulators” that modulate the activity of other genes (Wassarman et al., 1999; Wassarman, 2002).

**The problem of gene finding**

The last decade has seen remarkable advances in genome sequencing. In 1995, the first genome of a free-living organism, that of *Haemophilus influenzae*, was sequenced (Fleischmann et al., 1995). The year 1998 saw the first complete genome sequence of a multicel-
lular organism, that of the nematode *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998). The first draft of the human genome was published in 2001 (International Human Genome Sequencing Consortium, 2001). Currently, the genomes of almost 150 organisms are available in GenBank. With this vast amount of sequence data comes the problem of interpreting it.

One key question we can ask about raw genomic sequence data is “where are the genes?” In other words, we wish to identify those regions of the genome that are transcribed into a functionally significant RNA molecule, whether that RNA is a messenger RNA or a non-coding RNA. This problem is called gene-finding. There are two main types of gene-finding. In *de novo* gene finding, we are given the raw genomic sequence and simply told to find all the genes. While we may know what features genes generally have, we have no guidelines for what a specific gene will look like. In homology-based gene-finding, we already know about one or several members of a family of genes. We wish to find homologs that are evolutionary related to these genes. While techniques exist that address both *de novo* and homology-based noncoding RNA gene-finding, there is much room for improvement. I address both of these questions in this work.

**Protein coding gene-finding**

For a variety of reasons, protein-coding gene-finding techniques are more developed than analogous techniques for finding noncoding RNA genes. Transcription initiation and termination signals are neither as universal nor as easily recognizable as translation initiation signals (start codons) and termination signals (stop codons). Protein-coding genes can be
detected because of a statistical bias in their hexanucleotide frequencies (Fickett and Tung, 1992); no equivalent, generally applicable bias has been observed for noncoding RNAs. Algorithms that take RNA secondary structure into account are more complex, run more slowly, and need more memory relative to the analogous primary sequence algorithms (Sankoff, 1985). Therefore, it will be useful to look at protein-coding gene finding techniques as a point of reference before examining the current state of noncoding RNA gene finding.

**Protein coding gene finding de novo**

Several features of protein-coding genes can be used to find them in genomes *de novo*. One obvious feature is the presence of an open reading frame (ORF). An AUG codon usually marks the start of translation, and there are three stop codons that are almost universally conserved. In eukaryotes, splicing needs to be taken into account but there still cannot be any stop codons within the reading frame of an exon. ORFs are not enough to distinguish coding from noncoding sequence. Many other measures have been tried to distinguish protein-coding from noncoding sequence; hexanucleotide frequencies have been found to work well (Fickett and Tung, 1992). One gene finder that uses a Markov chain model to identify protein-coding regions using the hexamer bias is Genmark (Borodovsky and McIninch, 1993). A similar approach is used to model coding regions in eukaryotic genes in the context of overall eukaryotic gene structure (Burge and Karlin, 1997). As sometimes there is not enough data to reliably train a full sixth-order Markov chain, interpolated Markov models have been used to model a variable order Markov chain depending on the amount of data for each hexamer available (Salzberg et al., 1998; Salzberg et al., 1999; Delcher et al., 1999).
A related approach that has proven to be quite useful is comparative analysis. Rather than examine a single genome, this approach looks at two (or more) genomes together and examines the pattern of conservation and mutation between homologous regions. The program CRITICA combines hexanucleotide frequency analysis with a test for conservation of amino acid sequence sequence in diverged nucleotide sequences between related genomes in order to predict genes in prokaryotic genomes (Badger and Olsen, 1999). In mammals, the programs TWINSCAN, SGP2, and SLAM have been used to predict genes in the mouse genome by reference to the human genome (Guigo et al., 2003; Alexandersson et al., 2003). These programs predict many genes not found by other methods, and many of these predictions were experimentally verified with RT-PCR (Guigo et al., 2003).

While these gene-finding approaches work very well, they do not perfectly model the biological processes that determine what sequences get transcribed and translated into functional proteins. To look at biologically relevant transcripts, Expressed Sequence Tags (ESTs) have been used. ESTs are partial cDNA sequences produced en masse from a given organism, tissue type, and developmental stage. The first set of EST sequences came from the human brain (Adams et al., 1991). These sequences all come from cDNA libraries. For simplicity and to reduce contamination with non-messenger RNA, most such libraries are created by reverse transcription with a poly-dT primer. There are now almost $1.7 \times 10^7$ sequences in dbEST, the division of GenBank that contains EST sequences. (Boguski et al., 1993). These ESTs represent evidence that a genomic sequence is transcribed, and thus may be a functional gene.
Protein coding gene finding by homology

The problem of finding sequences homologous to known proteins can be subdivided into two problems: finding homologs of a single sequence and finding homologs of multiple, related sequences. For the former problem, similarity between proteins is taken as a measure of homology. Similarity is determined by aligning two protein sequences and measuring both the similarity between aligned amino acid residues and taking a penalty for gaps in the sequence. Needleman and Wunsch first proposed an algorithm to do such a global alignment (i.e. the two sequences are aligned end to end) (Needleman and Wunsch, 1970). This solution was later extended to local alignment (finding the best alignment in subsequences of the two sequences) (Smith and Waterman, 1981). These algorithms are relatively slow for searching large databases of sequences. Lipman and colleagues developed heuristic approximations to both the Needleman-Wunsch and Smith-Waterman algorithms to allow rapid searches of sequence databases (Wilbur and Lipman, 1983; Lipman and Pearson, 1985; Pearson and Lipman, 1988; Altschul et al., 1990; Altschul et al., 1997). This work culminated in the widely used FASTA and BLAST programs. These programs all rely on a measure of similarity between two aligned amino acids that is provided by an amino acid substitution matrix. The two most widely used series of matrices are the PAM matrices and the BLOSUM matrices (Dayhoff et al., 1978; Henikoff and Henikoff, 1992). In groundbreaking work, Karlin and Altschul proved that the optimal score of an ungapped local alignment between two random sequences follows the Gumbel distribution (Karlin and Altschul, 1990). Empirically, this appears to be true for gapped local alignment as well, though it has not been proven in these cases (Altschul and Gish, 1996). This distribution can then be used to...
measure statistical significance for a database search using a local alignment algorithm.

The main problem with single sequence alignment is that it does not capture which portions of a protein are highly conserved and which are variable. When one has a multiple alignment of many members of a protein family, it is easy to see which residues are important and which are not. One way to model conserved regions (e.g. the active site) is to define amino acid patterns. These patterns were implemented in the PROSITE database (Gattiker et al., 2002). They suffer from the problem that they only give a yes or no answer; a mismatch at one residue makes the pattern fail to recognize a potentially homologous protein. The profile approach gets around this problem. Profiles are position-specific weight matrices that model the amino acid distribution in the alignment (Gribskov et al., 1987). Profiles can be used to model conserved regions, with a flexible gap scheme used to connect them (Barton and Sternberg, 1990). More recently, hidden Markov models have been adapted to the profile approach (Baldi et al., 1994; Krogh et al., 1994). Finally, the PSI-BLAST program takes an iterative approach, starting the search with a single sequence and building up enough sequences to build a profile (Altschul et al., 1997).

The Importance of Secondary Structure

A key feature of many noncoding RNA genes that can be exploited in genefinding techniques is RNA secondary structure. Unlike DNA, RNA exists as a single-stranded molecule that can fold upon itself and be stabilized with intramolecular base pairs. Doty and colleagues first observed this secondary structure by looking at ultraviolet light absorbence at 258nm (Doty et al., 1959). They suggested that about half of the bases in the RNA
molecules examined were involved in intramolecular base pairs and that these base pairs formed extended helical regions. These ideas were refined a year later, when Fresco, Alberts, and Doty proposed that RNA base pairs are canonical Watson-Crick A-U and G-C pairs (Fresco et al., 1960). Interestingly, they also noted that thermal stability (i.e. melting temperature) of RNA seems to be positively correlated with the G+C content of the RNA, presumably because of the additional hydrogen bond found in a G-C base pair over an A-U base pair. In examining how tRNAs might recognize the genetic code in mRNA, Crick proposed that G-U is also a valid base pair in RNA secondary structure (Crick, 1966).

These ideas were soon put to use analyzing the first sequenced biologically active RNA – the tRNA. When the first tRNA sequence was determined – that of an alanine tRNA from yeast – three possible secondary structures were proposed: one extended stem, two shorter stems, or a cloverleaf shape (Holley et al., 1965). The determination of the tyrosine, serine, and phenylalanine tRNA sequences in yeast one year later revealed that the tRNA folds into the cloverleaf shape (Madison et al., 1966b; Zachau et al., 1966; RajBhandary et al., 1966). Several lines of evidence led to this conclusion. In determining the structure of the tyrosine tRNA, one group tried to maximize the number of base pairs made (Madison et al., 1966a). This resulted in a cloverleaf that looked similar to the cloverleaf Holley proposed. Another group noted that the serine tRNA could fold into either the extended stem or cloverleaf forms. Enzymatic digestion with enzymes that can distinguish single-stranded and double-stranded RNA revealed the cloverleaf was the proper model (Zachau et al., 1966). All 4 tRNA sequences can fold into a very similar cloverleaf shape. This comparative evidence was taken to be the strongest support for the cloverleaf model (Pace et al., 1999). Later structural biology experiments confirmed the cloverleaf structure using
both X-ray crystallography (Kim et al., 1973) and NMR (Shulman et al., 1973). These same four lines of evidence – optimizing a measure of structure, biochemical analysis, comparative analysis, and structural biology techniques – are the methods used today to determine the secondary structure of RNA molecules.

**Finding the secondary structure of a single RNA molecule**

It is often the case that we have a sequence we believe to be a noncoding RNA gene, but do not know its secondary structure. Many of the analytical techniques available, especially for homology-based gene finding, require prior knowledge of the correct secondary structure. Therefore, prediction of RNA secondary structure is an important problem that warrants further discussion. When we only know the sequence of a single RNA molecule and no homologous sequences, we can only use the optimization, biochemical, or structural biological approach. X-ray crystallography or NMR studies can reveal the three-dimensional structure of the RNA, and by extension its secondary structure. The complexity of these approaches means that they are usually reserved for determining more detailed structural information than just the secondary structure. The biochemical approach rests on probing the secondary structure of the RNA molecule in solution, and assuming that this mimics the *in vivo* situation. Enzymatic probes can be used to cleave the phosphate backbone at specific unpaired nucleotides or in double-stranded regions. Chemical probes can also be used to modify specific unpaired bases or cleave the backbone at specific unpaired bases. Cleavage products of end-labelled RNA can be visualized directly on polyacrylamide gels. Both cleavage and base modification locations can be determined via primer extension. The end result of this process is an experimentally determined list of nucleotides that are
paired and unpaired in solution; manual analysis is needed to derive the secondary structure (Moine et al., 1998).

While optimizations on a simple function (such as number of base pairs) can be done by hand, use of a computer program is quicker, helps prevent errors, and allows the use of a more complex objective function. While counting the number of base pairs is a simple objective function, it does not capture the physical reality of secondary structure. Instead, the predicted equilibrium free energy ($\Delta G^0$) can be minimized to find the optimal secondary structure. This approach was first proposed by Tinoco and colleagues, who presented a basic set of energy parameters to be used to calculate the $\Delta G$ of a structure (Tinoco et al., 1971). These energy rules have been revised and improved over the years; the best, most recent set comes from the Turner group (Mathews et al., 1999).

One of the first computer programs to predict RNA secondary structure enumerates all combinations of stems and chooses the structure with the lowest free energy (Pipas and McMahon, 1975). The running time for this approach grows exponentially with sequence length, and is not feasible for RNAs of more than 200 nucleotides or so. To get around this problem, Nussinov proposed a dynamic programming algorithm that finds the optimal secondary structure (Nussinov et al., 1978). This first algorithm optimized the number of base pairs. It was later extended by herself and others to optimize the free energy instead (Nussinov and Jacobson, 1980; Waterman and Smith, 1978; Zuker and Stiegler, 1981). The running time for these algorithms grows cubically with the sequence length ($O(N^3)$).

The energy minimization approach suffers from two fundamental problems. First interactions with proteins or other molecules in the cell may mean that the biologically relevant structure is not the structure with the lowest independent free energy (Williams and Tinoco,
1986). Second, the free energy calculations are only as good as the energy rules, which are known to be inaccurate, with the error on some parameters being 25% or more (Mathews et al., 1999). Therefore, it has been suggested that one should also look at suboptimal solutions, i.e. secondary structures with free energies near the global minimum. Several solutions to this problem have been proposed. Some algorithms find all solutions within a given range that are substantially different from the other solutions presented (Williams and Tinoco, 1986; Zuker, 1989). Others enumerate all possible solutions that are within the given range of the optimum (Waterman, 1983; Waterman and Byers, 1985; Wuchty et al., 1999). Finally, it has been proposed that one can use the partition function to calculate the probability that a given base pair exists among all the possible structures (McCaskill, 1990). Zuker suggests that energy-minimization algorithms should not be thought of as giving the “correct” answer. Rather, they should be a guide to structures for further experimental or phylogenetic examination (Zuker et al., 1991).

**Finding the common secondary structure among many RNA molecules**

A more accurate approach to finding RNA secondary structure is to look at a large number of homologous sequences presumed to have the same secondary structure. Looking for a common secondary structure is considered the gold standard for secondary structure prediction. This method rests on the assumption that homologous RNAs have similar structures (James et al., 1989; Pace et al., 1999). This approach requires the simultaneous construction of a multiple sequence alignment and determination of the common secondary structure. These steps can be performed either manually or automatically. It has been suggested that manual alignment and refinement of the structure works better than any
algorithms developed to date (Pace et al., 1999). While this manual approach is what has been used to determine the secondary structure of many real RNAs, this approach is also quite laborious. While an automated approach may be less accurate, it can also be performed with much less work.

As previously mentioned, the first RNA secondary structure determined by comparative analysis was that of the tRNA (Holley et al., 1965; Madison et al., 1966b; Zachau et al., 1966; RajBhandary et al., 1966; Pace et al., 1999). This result was serendipitous, as it was easily observed that all four sequences fold into the same structure. In contrast, Fox and Woese systematically determined the secondary structure of the 5S RNA molecule, a component of the ribosome, by searching the available 5S sequences for common helices (Fox and Woese, 1975). The secondary structure of the other two RNAs in the prokaryotic ribosome – the 23S and 16S RNAs – were determined in a similar fashion (Woese et al., 1980; Noller et al., 1981). Currently, the standard evidence to accept a helix as being “proven” is that there are at least two compensatory base changes observed in the helix (i.e. there are at least two locations in which the individual nucleotides in a base pair mutate yet conserve the base pair; James et al., 1989). Other structures determined using this approach include those of RNase P (James et al., 1988; Brown et al., 1991; Brown et al., 1996), RNase MRP (Schmitt et al., 1993; Li et al., 2002), telomerase (Romero and Blackburn, 1991; Chen et al., 2000a), and tmRNA (Williams and Bartel, 1996; Felden et al., 1997).

All of these structures required time-consuming manual alignment and stem-finding. A large number of automated procedures for finding phylogenetically conserved secondary structures have also been proposed. Some of these methods detect putative basepairs by looking for potential conserved stems in a multiple sequence alignment (Chiu and
Kolodziejczak, 1991; Gutell et al., 1992; Han and Kim, 1993). Other methods take a
similar approach but add in free energy calculations (Luck et al., 1996; Hofacker et al.,
1998; Juan and Wilson, 1999). Finally, Knudsen and Hein have presented a probabilistic
model that finds a common secondary structure by taking the phylogenetic relationship of
the sequences into account (Knudsen and Hein, 1999). The disadvantage to all of these
methods is that they are highly dependent on the accuracy of the starting alignment. An
incorrect alignment will yield an incorrect structure.

In contrast, several more recent methods do not require an initial alignment. Sankoff
developed an exact algorithm for simultaneous alignment and folding (also known as “struc-
tural alignment”) of \( M \) sequences of length \( N \) (Sankoff, 1985). This algorithm is \( O(N^{3M}) \)
in time and \( O(N^{2M}) \) in space, making it infeasible. One way around this problem is to
use stochastic optimization algorithms to find a locally optimal solution to some objective
function. This has been implemented using both simulated annealing (Kim et al., 1996) and
a genetic algorithm (Chen et al., 2000b). Bouthinon and Soldano encode RNA structures
as “structural patterns” and find those patterns common among the sequences (Bouthi-
non and Soldano, 1999). The program FOLDALIGN performs alignment between pairs of
sequences, looking for a single stem-loop. It then does the multiple alignment through a
greedy algorithm (Gorodkin et al., 1997; Gorodkin et al., 2001b).

**Finding the common secondary structure of a pair of RNA molecules**

This discussion of finding RNA secondary structure is motivated by the need to know the
structure in order to find homologs of a noncoding RNA gene. Often, we do not have a large
number of homologous sequences before undertaking a search, and therefore cannot use the
above methods. Finding a homologous sequence to a given noncoding RNA, however, is often possible using primary sequence search tools similar to those used for protein homology searches (Smith and Waterman, 1981; Pearson and Lipman, 1988; Altschul et al., 1990). If we have two homologous sequences, several techniques exist for finding a common folding among them that are more accurate than trying to fold only a single sequence.

The Sankoff algorithm mentioned earlier can simultaneously align and fold two RNA sequences (Sankoff, 1985). While this algorithm is guaranteed to give the optimal solution, for two sequences it is \( O(N^4) \) in memory and \( O(N^6) \) in time, making it practically infeasible. Several recent papers, however, have described heuristic simplifications or approximations that allow such simultaneous alignment and foldings to be computed. One approach, implemented in the DYNALIGN program, limits the maximum distance between two aligned nucleotides to reduce time and memory requirements (Mathews and Turner, 2002). Within this constraint, DYNALIGN finds the best fold and alignment using dynamic programming. The program FOLDALIGN simplifies the Sankoff algorithm by only allowing a single stem-loop motif (Gorodkin et al., 1997; Gorodkin et al., 2001b). An alternative approach, taken in the program CARNAC, heuristically builds up a common structure in two sequences by finding common stems in the sequence (Perriquet et al., 2003). DYNALIGN tends to overpredict basepairs while CARNAC tends to underpredict them (Perriquet et al., 2003).

Noncoding RNA gene finding

There are many analogs to protein-coding gene finding techniques that have been described for noncoding RNAs. For \textit{de novo} gene finding, transcription initiation and termination
signals, comparative analysis, and reverse transcription and cloning have all been used. Profile techniques have been described for finding noncoding RNA genes by homology. Details of these approaches are described below.

**De novo noncoding RNA gene finding**

In recent years, several approaches have been taken to identify noncoding RNA genes *de novo* in whole genome sequences. Probably the first such study searched for noncoding RNAs in *Saccharomyces cerevisiae* both by searching for RNA polymerase III promoters and for large regions without any predicted protein-coding genes (Olivas et al., 1997). This led to the experimental verification of several new small RNA transcripts. Two studies in *E. coli* computationally searched for promoter and terminator sequences in the same vicinity and proper orientation to find noncoding RNA genes. Both screens experimentally verified expression of some of their candidates (Argaman et al., 2001; Chen et al., 2002). Others have used various, specific base composition statistics and motifs to predict noncoding RNA regions of genomic sequences in a variety of organisms (Carter et al., 2001; Schattner, 2002). An attempt to find RNA genes by looking for regions likely to fold into a more stable secondary structure was not successful (Rivas and Eddy, 2000).

Comparative analysis has also been useful in finding noncoding RNA genes. Searching for conserved, intergenic regions of *E. coli* successfully identified several noncoding RNA genes whose expression was experimentally verified (Wassarman et al., 2001). A more sophisticated approach, implemented in the program QRNA, examines the pattern of mutation between homologous regions and determines if that pattern is most likely caused by conservation of a protein-coding region, an RNA secondary structure (noncoding RNA), or
a random, null model (Rivas and Eddy, 2001). This technique was initially used to successfully identify 11 novel noncoding RNAs in *E. coli* (Rivas et al., 2001). This work will also discuss a use of QRNA.

Experimental approaches have also been used to find noncoding RNA genes. Affymetrix chips were used to interrogate expression at every intergenic locus in *E. coli*. This study identified 317 novel transcripts, though the significance of these transcripts remains unclear (Tjaden et al., 2002). An EST-like approach for small RNAs has also been developed. In this approach, small RNAs are gel-purified and cloned. Clones are then screened to remove tRNAs and ribosomal RNA, and the remaining clones are sequenced. This technique was first used to find noncoding RNAs in mice (Huttenhofer et al., 2001). Since then it has been successfully used in the Archaeon *Archaeoglobus fulgidus*, the plant *Arabidopsis thaliana* and the fruit fly *Drosophila melanogaster* (Tang et al., 2002a; Marker et al., 2002; Yuan et al., 2003). As discussed above, micro RNAs (miRNAs) are small RNAs approximately 22 nucleotides in length first identified in *C. elegans* (Lee et al., 1993; Reinhart et al., 2000). Various groups have cloned small RNAs of this size from *C. elegans* (Lau et al., 2001; Lee and Ambros, 2001; Lim et al., 2003a; Ambros et al., 2003), *D. melanogaster* (Lagos-Quintana et al., 2001), vertebrates (Lagos-Quintana et al., 2001; Mourelatos et al., 2002; Lagos-Quintana et al., 2002; Lagos-Quintana et al., 2003), and plants (Reinhart et al., 2002).

**Finding noncoding RNAs by homology**

There exists a formal algorithm for simultaneously aligning and folding two homologous noncoding RNA sequences (Sankoff, 1985). This algorithm is analogous to the Smith-
Waterman algorithm for protein-coding genes (Smith and Waterman, 1981) but is much more computationally intensive. This algorithm can theoretically be used to search a sequence database for homologs to a given sequence when the structure is unknown, but the computational cost is prohibitive. A recent proposal to constrain the algorithm based on a known “fold envelope” (i.e., knowledge of the secondary structure) looks promising and may allow searching of a sequence database (Holmes and Rubin, 2002).

Most work in homology-based gene finding assumes you have a family of related sequences and know both its secondary structure and conserved positions. A major class of programs searches a database for a conserved, user-defined pattern. These programs are analogous to the PROSITE patterns for proteins (Gattiker et al., 2002). Some programs simply look for sequences that match the specified pattern of primary sequence and secondary structure, and give a yes/no answer for matches (Gautheret et al., 1990; Billoud et al., 1996). These programs suffer from the drawback that rare nucleotides at a position cannot be efficiently modeled. A nucleotide either matches or doesn’t match a pattern; degrees of matching the consensus cannot be quantified. To get around this problem, various programs allow for position-specific scoring of nucleotides in the pattern (Pesole et al., 2000; Gautheret and Lambert, 2001; Macke et al., 2001). A more general method for modeling families of sequences is profile stochastic context-free grammars (SCFGs) (Eddy and Durbin, 1994; Sakakibara et al., 1994; Eddy, 2002b). These fully probabilistic models are the RNA analog to profile hidden Markov models. They take an alignment of sequences with their secondary structure and create a model that can be used to search a sequence database for additional members of the family.

Customized variations of these approaches have been created to find tRNA genes in
nucleotide sequences. One approach is to heuristically search for a pattern that can fold into the tRNA cloverleaf and has the appropriate conserved nucleotides. The first program to do this assigned user-defined scores for each base pair and conserved nucleotide, and reported hits above a specified threshold (Staden, 1980). Other programs that work similarly have also been described (Paolella and Russo, 1985; Shortridge et al., 1986). Better performance was found by introducing weight matrices to score a match to a profile of known tRNA sequences, though this approach also heuristically matches a sequence to the tRNA pattern (Fichant and Burks, 1991). An alternative approach, specific to eukaryotes, is to search for RNA polymerase III promoters and terminators (Paolella and Russo, 1985; Pavesi et al., 1994). The most accurate approach to date is a profile stochastic context free grammar trained off of a large set of tRNA sequences (Eddy and Durbin, 1994). To increase speed, heuristic algorithms (Fichant and Burks, 1991; Pavesi et al., 1994) can be used to produce candidate tRNAs that are then evaluated with the more accurate profile SCFG (Lowe and Eddy, 1997).

Other family-specific models have been developed. A heuristic search approach for group I introns keys off of several specific secondary structure elements (Lisacek et al., 1994). RNA structure patterns and profile SCFGs have been used to find SRP genes (Regalia et al., 2002). A stochastic context-free grammar was used in part to model C/D box methylation guide snoRNAs and find additional such RNAs in *S. cerevisiae* (Lowe and Eddy, 1999). A similar approach was also used to identify C/D box methylation guide RNAs in the Archaea (Omer et al., 2000; Gaspin et al., 2000). Specific genefinders for miRNAs have been written that look for appropriately sized precursor hairpins conserved between two genomes. These have successfully found miRNAs in *C. elegans* (Lee and Ambros, 2001; Lim et al., 2003b;
Ambros et al., 2003) and vertebrates (Lim et al., 2003a).

**Outline of this work**

In this work, I address the issues of both *de novo* and homology based noncoding RNA gene finding. In Chapter 2, I search for noncoding RNAs in the genomes of the hyperthermophiles *Methanocaldococcus jannaschii* (formerly named *Methanococcus jannaschii*; Jones et al., 1983; Bult et al., 1986) and *Pyrococcus furiosus* (Fiala and Stetter, 1986; Maeder et al., 1999). I do this by exploiting a base composition bias peculiar to these two genomes. I also use QRNA to search for noncoding RNAs in *P. furiosus* using the genomes of *Pyroccocus abyssi* and *Pyrococcus horikoshii* as references (Erauso et al., 1993; Gonzalez et al., 1998; Kawarabayasi et al., 1998). After finding and experimentally confirming expression of several novel noncoding RNA genes, I was generally unable to find homologs of them using primary sequence search tools. Therefore, in Chapter 3, I discuss RSEARCH, a search tool I wrote for finding homologs of single noncoding RNA genes (with a secondary structure) in genomic sequences. I demonstrate its utility in several problems where the secondary structure is already known, including the identification of RNase P in several genomes where RNase P was not found previously. I then turn my attention to *E. coli*, where I examine previously identified noncoding RNAs (Hershberg et al., 2003) in more depth (Chapter 4). Finally, in Chapter 5 I move back to the Archaea and use RSEARCH to find additional homologs of the RNAs identified in Chapter 2.
Chapter 2

Noncoding RNA genes identified in AT-rich hyperthermophiles

\[1\]

\[1\]This chapter was co-written with Sean Eddy and Ziva Misulovin, and appears in Klein, Misulovin, and Eddy, *Proc Natl Acad Sci USA* **99**:7542-7547, 2002. I performed the bulk of the work described in this chapter, although Ziva and I did the RACE experiments together.
Abstract

Noncoding RNA (ncRNA) genes that produce functional RNAs instead of encoding proteins seem to be somewhat more prevalent than previously thought. However, estimating their number and importance is difficult because systematic identification of ncRNA genes remains challenging. Here, I exploit a strong, surprising DNA composition bias in genomes of some hyperthermophilic organisms: simply screening for GC-rich regions in the AT-rich *Methanocaldococcus jannaschii* and *Pyrococcus furiosus* genomes efficiently detects both known and new RNA genes with a high degree of secondary structure. A separate screen based on comparative analysis also successfully identifies noncoding RNA genes in *P. furiosus*. Nine of the 30 new candidate genes predicted by these screens have been verified to produce discrete, apparently noncoding transcripts with sizes ranging from 97 to 209 nucleotides.

Introduction

Noncoding RNA (ncRNA) genes are genes for which RNA, rather than protein, is the functional end product. The number and diversity of ncRNA genes is a subject of active research (Eddy, 2001). In principle, the availability of many genome sequences makes it possible to search computationally for novel ncRNA genes. Computational protein gene finders search for ORFs that have certain statistical biases in their nucleotide composition (Borodovsky and McIninch, 1993; Burge and Karlin, 1997; Salzberg et al., 1998). Unfortunately, ncRNA genes have neither ORFs nor (generally speaking) nucleotide composition biases, making ncRNA gene-finding a more formidable problem.
Hyperthermophiles must stabilize double-stranded DNA and RNA against thermal de-
naturation (Grogan, 1998). The simplest stabilization strategy is increased GC content. However, the GC content of hyperthermophile genomes does not correlate with optimal growth temperature (Grogan, 1998; Galtier and Lobry, 1997; Daniel and Cowan, 2000). Hyperthermophiles use various other mechanisms to stabilize their DNA, including increased intracellular ionic concentrations, cationic proteins, and supercoiling (Grogan, 1998; Daniel and Cowan, 2000). Intramolecular RNA secondary structure, however, seems to be partially stabilized by increased hydrogen bonding, as the GC content of ribosomal RNA and transfer RNA genes in hyperthermophiles shows a strong correlation with optimal growth temperature (Galtier and Lobry, 1997). I reasoned that in an AT-rich extreme hyperthermophile, structural RNA genes (i.e., ncRNA genes with a high degree of secondary structure) might be found just by searching for regions of elevated GC content. Such a gene finder would not be able to be generalized. However, one might use novel ncRNAs identified in these unusual genomes to identify homologous RNAs in a variety of other genomes.

Several recent reports describe computationally aided screens for ncRNA genes in Es-
ccherichia coli. Two groups searched for strong promoter and terminator signals appropriately spaced over intergenic regions (Argaman et al., 2001; Chen et al., 2002). This approach obviously requires the genome sequence of an organism for which transcriptional regulation is well understood. Another group used a neural network to classify genomic sequences based on several features, including GC composition (Carter et al., 2001). Two other approaches used a comparative genomics approach, requiring genomic sequence from related organisms as well as that of E. coli. One such approach simply looks for conserved intergenic regions (Wassarman et al., 2001); the other further processed sequence align-
ments of the conserved intergenic regions to decide whether the pattern of mutation was most consistent with a protein-coding gene, an ncRNA gene with secondary structure, or simply random mutation (Rivas et al., 2001). This latter approach has been converted into a general gene finder, QRNA, which can be used for any genome for which additional comparative genomic sequence is available (Rivas and Eddy, 2001).

To date, detailed analysis of the performance of QRNA has been performed only in *E. coli*. Furthermore, comparison of the performance of QRNA with that of an alternative gene finder would prove informative on the trustworthiness of both screens. That is, even though a GC-based screen may work only in unusual organisms, those organisms provide a test bed for further validation of QRNA as a general RNA gene finder. Therefore, I screened for novel ncRNAs by using both the GC content bias and QRNA to compare their performance and results. Here I identify novel ncRNAs in *Methanocaldococcus jannaschii* by using the GC content bias alone and in *Pyrococcus furiosus* by using both the GC bias and QRNA-based comparative analysis. I find that the two screens performed in *P. furiosus* identified nearly exactly overlapping sets of ncRNA genes.

**Materials and Methods**

**Genomes Used**

Fifty-one complete prokaryotic genome sequences in GenBank as of June 21, 2001 were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria. The sequence of *P. furiosus* was downloaded from the Utah Genome Center (http://www.genome.utah.edu/sequence.html) on August 27, 2001. tRNAscan-SE version 1.21 (Lowe and
Table 2.1: Parameters used in the hidden Markov model. \( t_{R,R} \) is always 0.99. Other parameters in the HMM are derived from these values as shown in Figure 2.1.

Eddy, 1997) was used to identify tRNA genes.

### GC Content Screens

A hidden Markov model with two states ("RNA" and "background genome") was used to find GC rich regions (Figure 2.1). The emission probabilities of the genome state were set using the low GC content of the overall genome (\( f_{GC,G} \)), whereas the emission probabilities of the RNA state were set using the higher GC content of the tRNA and rRNA genes (\( f_{GC,R} \); Figure 2.1; Table 2.1). The probability of staying in the RNA state (\( t_{R,R} \)) was always set to 0.99, which assumes that on average an RNA is 100 nucleotides long. The probability of transitioning from the genome state to the RNA state (\( t_{G,R} \)) was set to the number of tRNAs in the genome divided by the length of the genome. The probability of starting in the RNA state (\( t_{S,R} \)) was set to the fraction of nucleotides in the genome known to be in an rRNA or tRNA gene. This parameterization assumes that there are not many more RNAs to be found. The parameters used to derive all the models used are given in Table 2.1.

Standard Viterbi and posterior decoding algorithms were used in conjunction with these models (Durbin et al., 1998). In the Viterbi screen of \( M. \text{jannaschii} \), only nine candidate RNA regions of at least 50 nucleotides were considered; one shorter region was discarded. For the posterior decoding screen, regions of at least 50 nucleotides were se-
Figure 2.1: The hidden Markov model used to find regions of high GC content. The emission parameters $e$ are derived from the GC contents $f_{GC}$, while the relationship between the transition parameters is shown in the model.
lected in which all bases had a posterior probability of the RNA state over a chosen cutoff. The cutoff probabilities were set so that all tRNAs were successfully proposed as GC-rich regions. These cutoffs were 0.130 for P. furiosus, 0.052 for Pyrococcus abyssi, and 0.147 for Pyrococcus horikoshii. Conserved GC-rich P. furiosus candidate ncRNAs were then identified with WU-BLASTN version 2.0 with W=4 (Altschul et al., 1990; Altschul et al., 1997) by requiring a P. furiosus GC-rich region to hit a GC-rich region from both P. abyssi and P. horikoshii with an E-value less than $10^{-5}$. WU-BLASTN can be found at http://blast.wustl.edu/. The source code for the screening program are freely available from ftp://ftp.genetics.wustl.edu/pub/rjklein/hmmgcc.tar.gz.

**QRNA Comparative Analysis**

To perform the comparative analysis by examining the pattern of mutation in the alignments, I first used WU-BLASTN (version 2.0) with default parameters except hspmax = 100000 to compare related genomes. I used the P. furiosus genome as a query against the P. abyssi and P. horikoshii genomes. I kept only alignments with $E < 0.01$, 65-85% identity, and at least 50 nucleotides long for further analysis. Alignments less than 65% identical were considered too dissimilar to be meaningful, while alignments more than 85% identical are known to have a high false positive rate with QRNA (Rivas et al., 2001). These alignments were then analyzed with QRNA 1.1 (Rivas and Eddy, 2001), and a list of candidate ncRNA genes was produced by merging all overlapping P. furiosus regions scoring at least 5 bits. The QRNA program is available at http://www.genetics.wustl.edu/eddy/software/.
Growth conditions and RNA extraction

*M. jannaschii* frozen cell paste was provided by J. Brown (North Carolina State University). These cells were grown in 12-liter batch fermentations in American Type Culture Collection (ATCC) media 2121 at 83 °C with continuous sparging with 60% H<sub>2</sub>/40% CO<sub>2</sub> (vol/vol) and daily replacement with Na<sub>2</sub>S. Cultures were harvested after 2-3 days, approximately during late logarithmic growth. Frozen cell paste was stored at -80 °C until needed. RNA was prepared from cell paste by mortar and pestle lysis and phenol/chloroform extraction. Specifically, the paste was ground in a precooled mortar and pestle with dry ice for 20 minutes. The powder was resuspended in 17mL STE (10mM Tris-HCl, 100mM NaCl, 1mM EDTA) pH 8.0 to which 3mL of 5% SDS, 30% 4-aminosalicylate was added. The tube was mixed, held at 70 °C for 10 minutes, and then extracted three times with phenol pH 4.3/0.2% β-mercaptoethanol followed by one extraction with 24:1 chloroform:isoamyl alcohol. The RNA was recovered by ethanol precipitation followed by resuspension in TE (10mM Tris-HCl, 1mM EDTA) pH 8.0. *P.furiosus* was grown at 95 °C in rich medium containing peptides and maltose, but without sulfur, as described (Adams et al., 2001). Cells were harvested in mid-log phase, and the RNA was extracted as described (Voorhorst et al., 1995).

Northern blotting

Northern blots were performed by running 10μg of total RNA from the given organism alongside 5’-end labeled 100bp (New England Biolabs) and 25bp (Promega) denatured DNA ladders on a 6% denaturing polyacrylamide gel (6% 30:1 acrylamide:bis, 1X TBE, 7M Urea) for 1.5 hours at 400v. The nucleic acids were then electroblotted to a Zeta-
Probe membrane (Bio-Rad) using a semi-dry transfer apparatus at 200mA for 2 hours. Blots were dried and fixed by drying on a gel dryer at 80 °C for 30 minutes followed by ultraviolet cross-linking. Membranes were pre-hybridized in 5X SSC, 2mM Na$_2$HPO$_4$ pH 7.5, 7% SDS, 10X Denhardt’s Solution, 0.1 mg/mL denatured salmon sperm DNA at 55 °C for 2 hours. 10$^6$ cpm of labelled oligonucleotide probe was added per mL of hybridization solution, and hybridization was carried out at 55 °C overnight. Blots were washed twice for 30 minutes each at 55 °C in 3X SSC, 5% SDS, 25mM Na$_2$HPO$_4$ pH 7.5 followed by one wash for 30 minutes at 55 °C in 1X SSC, 1% SDS. Blots were visualized by exposing to a Phosphorimager screen (Molecular Dynamics) for approximately 4 hours. The sequences of the individual oligonucleotides used are provided in Table 2.2; their lengths range from 18-29 nucleotides.

The sensitivity of the northerns were tested with serial dilutions of a homogeneous sample of a single RNA molecule. Plasmid pLAJ-1, which encodes the sequence for gld-1 RNA under the control of a T7 RNA polymerase promoter, (Jones and Schedl, 1995) was digested with XmnI to linearize the plasmid. Reverse transcription was then performed with T7 RNA polymerase. The DNA was then removed with RNase-free DNase, and the resulting RNA was quantitated on a 1.2% agarose gel stained with EtBr by comparing band intensity to the 100bp ladder. Ten-fold serial dilutions of this RNA were then run on a gel for northern blots as above. The blots were hybridized with the oligonucleotide 5' –CATAACGGACGAGATGGCGAT–3'. From this experiment, the northerns were determined to be able to detect 8.8 × 10$^{10}$ molecules without problem. Assuming that, as for *P. furiosus*, 3 × 10$^8$ cells per mL of culture was grown (Fiala and Stetter, 1986), and knowing that an RNA prep of *P. furiosus* yielded 800 μg of RNA from a 1.2L culture, then 3.6 × 10$^{11}$ cells
<table>
<thead>
<tr>
<th>Cand. no.</th>
<th>Forward oligonucleotide</th>
<th>Reverse oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mj1</td>
<td>GCTACTTCTAGCTTCCCCCTATCGG</td>
<td>GACACCCGCTTCCCCCTTTACG</td>
</tr>
<tr>
<td>Mj2</td>
<td>CGGATCTACCCACACAGGCG</td>
<td>ATACAGCTACTCCGCCAGGG</td>
</tr>
<tr>
<td>Mj3</td>
<td>TGATGTAAGCAGCCAACTACG</td>
<td>TGGTGGGGCCCGCTTCTCAG</td>
</tr>
<tr>
<td>Mj4</td>
<td>TTTTACGCGGACCCGTTG</td>
<td>ACGCGGATGCCCCACGGAG</td>
</tr>
<tr>
<td>Mj5</td>
<td>GCTGCCAATCTGTTGTTTCCTG</td>
<td>CGGCTATGGGAGCCTCAACAG</td>
</tr>
<tr>
<td>Mj6</td>
<td>AACACCCTTCTCCTACCGTCCG</td>
<td>CTACCGGCGACGGTAGGGG</td>
</tr>
<tr>
<td>Mj7</td>
<td>CGGCGCTATCGGTCGAGTTG</td>
<td>CCCACATTACCGGCTGAAACCG</td>
</tr>
<tr>
<td>Mj8</td>
<td>CTCCTCCCAGCCAAACGAAG</td>
<td>CGGAGATTTGGAGACAGAG</td>
</tr>
<tr>
<td>Mj9</td>
<td>GGGATGAGCAAGCGGGGAGG</td>
<td>GGGAGCTACAGCTTTGGC</td>
</tr>
<tr>
<td>P1f</td>
<td>CCGGCTTCTCCCGGGGATG</td>
<td>GGGTTCAGATTCCCCGGAGAGGG</td>
</tr>
<tr>
<td>P12</td>
<td>CGCGGCCCAATGACCGGCTTCTG</td>
<td>GGGCTACAGGGCAGTTGGCTAGC</td>
</tr>
<tr>
<td>P13</td>
<td>CGATGAACCGAGTGTCCTCCGCG</td>
<td>CGCCGGAGACGGCTGATACCCG</td>
</tr>
<tr>
<td>P14</td>
<td>CACTTTCATCGGTCGAGTTG</td>
<td>CCCACATTACCGGCTGAAACCG</td>
</tr>
<tr>
<td>P15</td>
<td>TAGGCCACGTTCTTTGGGTGATG</td>
<td>TCACCAGAAGAACCTGAGCCTA</td>
</tr>
<tr>
<td>P16</td>
<td>ACCCCCTCATCGAACCCGCTCGG</td>
<td>TCGGTTGTAAGGCGGTTGGTCG</td>
</tr>
<tr>
<td>P17</td>
<td>CACTCCGGTACGGGGAAGGCACTC</td>
<td>GAGTGCCGTAACGCGGGGTAACCC</td>
</tr>
<tr>
<td>P19</td>
<td>CGGGGAGGACCGCGGACCTGCTG</td>
<td>CCGGGGAGGAGCGGGCTCACAG</td>
</tr>
<tr>
<td>P1fQ1</td>
<td>GAATCCCGCACAGTTCCCTTTTCTCACTC</td>
<td>CAGTGAAATGAACCTTGGGGGACTTC</td>
</tr>
<tr>
<td>P1fQ2</td>
<td>TGTAAATAATCTAAAAGCCTTCTTCAGGGC</td>
<td>GCCCTGAAGGAGGCTTTTAGATTAAACA</td>
</tr>
<tr>
<td>P1fQ3</td>
<td>GAATCCCGCACAGTTCCCTTTTCTCACTC</td>
<td>AGGAAATCTAAAAACATTCGCGGAGAGC</td>
</tr>
<tr>
<td>P1fQ4</td>
<td>CGGATGCGGCTGCAGCTGAGCTGACCCG</td>
<td>TGGGAGCTGCTACTCACTATAGTGGTCTC</td>
</tr>
<tr>
<td>P1fQ5</td>
<td>TGCCAAATCGTACGGGAACTTCTTATTAAATTT</td>
<td>TAATTTAATAAGTTTACCGCGATTTCAG</td>
</tr>
<tr>
<td>P1fQ6</td>
<td>TCGGATTCTGACTTCTTCTTCTCAGCTTG</td>
<td>TGAACCGTGAAGGAGACAAAATTCAG</td>
</tr>
<tr>
<td>P1fQ7</td>
<td>AATCTTTCCGAGTATTAGATATCGC</td>
<td>TTTTCTGACAATCTATACGCTAGACGAAG</td>
</tr>
<tr>
<td>P1fQ8</td>
<td>GCCGGGCTTATACGCTTCTGACACCC</td>
<td>GGTACCCTTACCTGGACGGGGCG</td>
</tr>
<tr>
<td>P1fQ9</td>
<td>AAACCTTTACGCTAAATAAACAGGACC</td>
<td>CACGTTGCTATTTGCTGTCTCAGAG</td>
</tr>
<tr>
<td>P1fQ10</td>
<td>GTGTTCTGCTCCAGAAGTCTTGTCCTACG</td>
<td>CGCTAGACCTTTCTTGGGCAGCACCC</td>
</tr>
<tr>
<td>P1fQ11</td>
<td>GCCAACCCCGCTTTCGACTTGTAGG</td>
<td>CCCCCTACCTTTAGCGGTGTTCCTC</td>
</tr>
<tr>
<td>P1fQ12</td>
<td>GCAAGGCTACAGGAACTAGGAGAAG</td>
<td>CCTCTACCTTCAACTGCTTGGCTTTCG</td>
</tr>
<tr>
<td>P1fQ13</td>
<td>CAGGAGGCGGCTGGCATTGCCCTCG</td>
<td>AGGACGTTGCCGCGCTTCCCTC</td>
</tr>
<tr>
<td>Mj6A</td>
<td>ACCTAAAGCGGATGGTGAGGCTC</td>
<td>GACCTCGGCCAGTGACACCAAACCT</td>
</tr>
<tr>
<td>P78</td>
<td>ACCGTTGGGGAGGTTGCTTCTGTGG</td>
<td>TCGCCGGTCTGCTTTGGGGCAAGTT</td>
</tr>
</tbody>
</table>

Table 2.2: Oligonucleotides in northern blots. The strand listed is the strand of the RNA that the oligonucleotide can detect. In other words, the forward oligonucleotide can detect an RNA whose sequence matches the forward strand of the genome.
yielded 800 μg of RNA. This means that a 10μg loading of RNA on the gel should include the RNA from 4.5 × 10⁹ cells. Thus, RNA that is as rare as 20 copies/cell can easily be detected. As further proof of the sensitivity of the northerns, RNase P RNA from *M. jannaschii*, which is known to exist in low copy number, is easily detected on northern blots (data not shown).

**Rapid Amplification of cDNA Ends (RACE)-PCR Analysis**

Total RNA was purified further by treating with RNase-free DNase (Promega) in the manufacturer-supplied buffer for 30 minutes at 37 °C. RNA was recovered by phenol-chloroform extraction followed by ethanol precipitation. A poly-A tail was added by treating the RNA with *E. coli* poly-A polymerase (GibcoBRL) in a reaction containing 50mM Tris-HCl pH 8.0, 0.25M NaCl, 10mM MgCl₂, 2.5mM MnCl₂, 0.25mM ATP and 0.5 mg/mL BSA for 30 minutes at 37 °C (Sippel, 1973). The resulting polyadenylated RNA was then used to make 5’-RACE-Ready and 3’-RACE-Ready cDNA with the SMART RACE cDNA amplification kit (CLONTECH, Palo Alto, CA) according to the manufacturer’s instructions.

Specific 5’ and 3’ cDNA ends were amplified from the appropriate cDNA pool according to the instructions using “touchdown PCR.” Reactions were scaled down to 25μL. The touchdown PCR amplification program used was as follows: 5 cycles of 94 °C/15s, 72 °C/2min; 5 cycles of 94 °C/15s, 70 °C/15s, 72 °C/2min; 20 cycles of 94 °C/15s, 68 °C/15s, 72 °C/2min. 10μL of each reaction was run on 2.2% agarose gels and bands were visualized with EtBr staining under UV light. Only some amplifications produced visible bands at this stage. 1μL of each reaction, regardless of whether a band was seen or not, was reamplified in a 25μL reaction using the same gene-specific primer and the UPM-Short primer.
Table 2.3: Oligonucleotides for RACE reactions. The GSP1 oligonucleotides are used to find the 5’ end of an RNA. The GSP2 oligonucleotides are used to find the 3’ end.

(5’-CTAATACGACTCACTATAGGC-3’) (0.4μM final concentration). Amplification was performed with HotStar Taq (Qiagen), a final MgCl₂ concentration of 4mM, and the following program: 95 °C/15min; 27 cycles of 95 °C/30s, 68 °C/30s, 72 °C/45s. 10μL of each reaction was visualized on 2.2% agarose gels as before. All reactions had at least one visible band after reamplification. The sequences of the gene-specific oligonucleotides are provided in Table 2.3.

1μL of the fresh PCR product was cloned using the pCRII vector in the TA Cloning Kit (Invitrogen), transformed into competent cells, and plated on X-Gal/IPTG plates using blue/white and ampicillin selection. At least 12 clones were picked for each cDNA end and screened for the presence of an insert with PCR using primers M13-Reverse and M13-Forward. Mini-preps were performed from 5-10 positive clones using the QIAPrep Spin Miniprep Kit (Qiagen) and clones were sequenced with M13 Reverse primer using the Applied Biosystems Big Dye Sequencing Kit version 2. The cDNA ends were then determined from the insert sequence. This RACE-PCR system depends on a poly-C sequence being added to the 3’ end of the cDNA once the reverse transcriptase hits the 5’ end of the RNA. This CCC then anneals to an RNA oligonucleotide in the solution mix to allow
reverse transcription to continue and produce a specific sequence at the 3’ end of the cDNA. This oligonucleotide can also be added spuriously when there is a GGG present in the RNA being reverse transcribed. Therefore, some 5’ ends were identified on the basis of an internal GGG in the ncRNA molecule and were not considered to be true 5’ ends.

Computational Analysis of the RNAs

WU-BLASTN version 2.0 with W=3 was used to search the NCBI nonredundant nucleotide database (version May 16,2001) and a database of all of the available Archaeal genomes in GenBank as of June 21,2001 (Aeropyrum pernix, Archaeoglobus fulgidus, Halobacterium sp. NRC-1, Methanobacterium thermoautotrophicum, M. jannaschii, P. abyssi, P. horikoshii, Sulfolobus solfataricus, Thermoplasma acidophilum, and Thermoplasma volcanium), as well as P. furiosus from August 27,2001. Secondary structure prediction was assisted by MFOLD version 3.1 (Zuker et al., 1999; Mathews et al., 1999).

Results

GC content screen

I first determined the relationship between genomic and ncRNA GC content and optimal growth temperature for 52 prokaryotic genomes available in the summer of 2001 (Grogan, 1998; Galtier and Lobry, 1997; Daniel and Cowan, 2000). The overall GC content of each genomic sequence shows no correlation with optimal growth temperature (Figure 2.2A; correlation coefficient $r = -0.056$). On the other hand, the transfer RNA GC content does clearly correlate with growth temperature (Figure 2.2B; $r = 0.84$). Surprisingly, a large
GC content difference is not restricted to thermophiles, although it is more pronounced there (Figure 2.2C), suggesting that a screen on the basis of GC bias may work in some mesophiles. However, in an AT-rich mesophile, we might expect the protein coding exons to have a GC content that is also higher than background. In such a case, a GC content screen would not be able to distinguish between protein-coding and noncoding RNA gene sequences.

To successfully screen for noncoding RNA genes based on GC content, the GC content of the structured RNAs must be biased away from the GC content from all other sequences. In some hyperthermophiles, this bias is pronounced. In the organism with the largest GC content difference, *M. jannaschii* (Bult et al., 1996), tRNAs are readily apparent because of the GC content difference (Figure 2.2D). I therefore decided to screen *M. jannaschii* for ncRNAs based solely on local GC content.

To define objectively high-GC regions, I parameterized a two-state hidden Markov model and used the Viterbi algorithm to parse the *M. jannaschii* genome. This approach identified 43 different regions. Because Viterbi decoding produces the best (maximum likelihood) assignment of nucleotides to either “RNA” or “background genome,” there is no score associated with individual regions or any cutoff parameters to set. To evaluate sensitivity, I asked what percentage of tRNA genes were correctly identified by overlapping with predicted “RNA” regions; all 37 known *M. jannaschii* tRNAs were identified, as were ribosomal RNA genes, RNase P RNA, and 7S (signal-recognition particle) RNA. After accounting for regions that contain these genes, nine regions at least 50 nucleotides in length remained as candidate ncRNA genes (Table 2.4). To evaluate the false positive rate, I analyzed 1,000 random genome sequences with the same overall G+C composition and length and detected
Figure 2.2: GC content as a basis for finding ncRNA genes. (A) GC content of whole genomes vs. optimal growth temperature. In this and subsequent images, the large square represents *M. jannaschii*, the up triangle *P. abyssi*, the down triangle *P. horikoshii*, and the diamond *P. furiosus*. (B) GC content of tRNA genes vs. optimal growth temperature. (C) Difference in tRNA and genomic GC content vs. optimal growth temperature. (D) GC content of a 1-kb region of the *M. jannaschii* genome containing a tRNA gene calculated in a 100-bp sliding window.
33 GC-rich regions, indicating that the expected number of GC-rich regions detected by chance is about 0.03 per genome.

As I wished to compare the performance of the GC-content ncRNA gene finder with that of QRNA, I needed to consider a set of related organisms in which to do comparative analysis. Although no nearby relative of *M. jannaschii* has yet been completely sequenced, there are three genome sequences available of the AT-rich hyperthermophilic genus *Pyrococcus* – *P. furiosus*, *P. abyssi*, and *P. horikoshii* (Kawarabayasi et al., 1998; Maeder et al., 1999). A GC-content screen with Viterbi parsing was less successful for these genomes; the highest sensitivity observed was in *P. furiosus*, where 67% of tRNAs were identified. Therefore, I decided to include simple comparative analysis in the GC screen. I used a hidden Markov model posterior-decoding algorithm to relax the specificity of the Viterbi screen and identify GC-rich regions with more sensitivity, and then considered only regions from *P. furiosus* that showed significant BLASTN similarity to regions in the other two *Pyrococcus* genomes. The threshold was set such that all 46 known *P. furiosus* tRNAs were found by definition. This screen initially identified 51 conserved GC-rich regions. All of the tRNAs, ribosomal RNAs, RNase P RNA, and 7S RNA were also identified. After accounting for these known ncRNAs, eight regions remained as putative ncRNA genes (Table 2.4). To test the false positive rate, I made a data set for each genome consisting solely of regions identified as protein-coding ORFs at least 200 amino acids long with GLIMMER (Salzberg et al., 1998). I assumed that protein-coding sequences should contain few stable structural RNA regions. The posterior-decoding screen identified no regions in the ORF data set, suggesting that the false positive rate is near 0.
Table 2.4: Candidate ncRNAs. The real size of the gene products given is the size as determined by the ends predicted by the RACE reactions shown in Figure 2.4. A range of sizes indicates that several 3' ends were found at different nucleotides in the same genomic region. V, Viterbi screen; P, posterior decoding with cutoff set to identify all tRNAs plus conservation among GC-rich regions of all three *Pyrococcus* species; Q, QRNA screen; H, homolog of Mj6/*hgcC*. *Located on large extrachromosomal element (ECEL).
Comparative analysis with QRNA

I next wished to see how these results in *P. furiosus* compared with a QRNA screen. I used QRNA to compare alignments between *P. furiosus* and each of the other two *Pyrococcus* genomes and identify putative ncRNA loci (Rivas and Eddy, 2001). This screen identified 60 candidates in comparison to *P. abyssi* and 63 candidates in comparison to *P. horikoshii*. Many of the predicted loci overlap; when the two lists are combined, 73 candidate ncRNA regions are found. Among these, 45 of 46 tRNAs were found, as were the ribosomal RNAs and 7S RNA. After accounting for these known RNAs, 32 candidate regions remained. Many of these were either partially or completely overlapped by protein-coding gene annotation and were eliminated from further consideration. This approach left 17 intergenic regions at least 50 nucleotides in length for further consideration. Four of the 17 correspond to regions identified by the posterior-decoding GC screen (Table 2.4). To estimate the false positive rate, I performed QRNA analysis as before but shuffled the individual columns of each alignment before scoring with QRNA. This analysis resulted in 20 loci being found in comparison to *P. abyssi* and 23 loci being found in comparison to *P. horikoshii*, predicting a false positive rate of approximately 35%. Because 53% of the loci already contained at least one ncRNA characterized before this study, the number of novel ncRNAs was expected to be small.

Experimental analysis of predicted loci

To test whether these candidate loci express a detectable RNA transcript, I used northern blot analysis with strand-specific oligonucleotide probes and RNA from log-phase cultures grown in standard lab conditions. I detected small, stable RNA transcripts from nine
candidate loci in total: four of the nine _M. jannaschii_ candidates, four of the eight _P. furiosus_ conserved GC-rich candidates, and four of the 17 _P. furiosus_ QRNA candidates (Figure 2.3). Three of the five expressed _Pyrococcus_ candidates were found by both screens. Candidate PfQ11 was not found by the GC screen because its GC content is the same as background; candidate Pf3 was not found by QRNA because its identity was above 85% in all windows tested. Although a higher percent identity cutoff would have allowed identification of Pf3, it would also have increased the expected false positive rate to a level where the true positives would be drowned out by noise (Rivas et al., 2001).

Overlapping 5’ and 3’ end fragments of these RNA transcripts were amplified by RACE, cloned, and sequenced. In many cases, a variety of 5’ or 3’ ends were determined for a single locus. Some of these length variations correspond to different bands seen on the northern blots; others may be due to transcripts too rare to be detected on northerns but that can be amplified by PCR. Best 5’ and 3’ ends are given in Figure 2.4. None of these transcripts seem to have any significant coding potential, defined as a start codon (AUG, GUG, UUG) followed by in-frame translation and a stop codon. One corresponds to a known ncRNA, sR9 (see below). I named the other seven GC-rich loci hgcA through hgcG (“high GC”), whereas the RNA identified only through QRNA was named sscA (“secondary structure, conserved”; Figure 2.4).

There are several discrepancies between the RNA lengths predicted from the northern blots and the RACE experiments. In many cases, a RACE product is found that does not correspond to any band on the northern blots. This is neither surprising nor concerning, as it is likely that the PCR-based RACE procedure picks up rare, possibly spurious, transcripts that are invisible on northern blots. In some cases, no RACE product is found for observed
northern blot bands. For hgcC, hgcG, and sscA, there is a northern band that corresponds to the RACE product. The other bands may represent non-specific hybridization between the oligonucleotide and a transcript from another genomic locus. For hgcE, all RACE products were significantly shorter than the major northern band; I believe I could not find the true 5’ end because of an extremely abnormal GC composition (Table 2.4). Finally, for sR9 and hgcF, the RACE product was longer than the observed bands on the northern blot. Such a discrepancy may indicate that the observed northern bands are due to non-specific hybridization.

Homologs of the RNAs

BLASTN searches of GenBank and the available Archaeal genomes failed to identify any significant similarity ($P < 0.005$) between these new genes and any known gene. hgcC shows significant similarity to another region of the M. jannaschii genome (candidate Mj6A) as well as to a region of the P. furiosus genome (candidate Pf8) that is identified as GC-rich in the Viterbi screen but was not pulled out in either comparative screen, because similar sequences do not exist in either P. abyssi or P. horikoshii (Table 2.4). To see whether Mj6A and Pf8 are also biologically relevant, I tested for expression by using oligonucleotide probes to northern blots. The similar region in M. jannaschii shows only weak expression, whereas it appears there are high levels of expression of a 120-nt RNA from the homologous region in P. furiosus (Figure 2.5A). I mapped the 5’ and 3’ ends of both genes and named them hhcA and hhcB (“homolog of hgcC”; Figure 2.5B and C). The similarity between these three loci is obvious to the eye upon looking at a multiple sequence alignment (Figure 2.5D).

When the genomic locus for hhcB was compared with that of syntenic regions in P. abyssi
Figure 2.3: Northern blots of novel ncRNAs. Each pair of blots represents probing with oligonucleotides for RNA on the + or - strand, respectively. On each blot, the leftmost lane is a 100-bp ladder, the center lane is the RNA sample, and the rightmost lane is a 25-bp ladder.
Figure 2.4: The genomic context of each novel ncRNA gene. The left-hand column gives the candidate name from the screen. In the center is an independently scaled schematic of the genomic locus. The black arrow represents the best copy I could find of the ncRNA gene, whereas the gray arrows represent 50 nucleotides of the flanking annotated genes. The numbers above the lines and below the arrows indicate the length of the ncRNA gene as determined by 5’ and 3’ RACE. The other two numbers above the lines are the distances between the gene and its flanking genes, where a number in parentheses indicates the length of the overlap between the two genes. The numbers below the lines represent the start and stop coordinates of the gene. The right-hand column contains the GenBank accession no. for the cDNA sequence.
Figure 2.5: \( hgcC \) and its homologs. (A) Northern blot showing expression of homologs of \( hgcC \) in \( M. jannaschii \) and \( P. furiosus \). (B) Genomic context of \( hhcA \), as in Figure 2.4. (C) Genomic context of \( hhcB \), as in Figure 2.4, except the full length of ORF PF1918 is included. The dashed line indicates a gap not drawn to scale. The shaded region is not present in the syntenic regions of \( P. abyssi \) and \( P. horikoshii \). (D) CLUSTALW alignment between the \( hgcC \), \( hhcA \), and \( hhcB \) loci.
and *P. horikoshii*, it became apparent that *hhcB* and the adjacent ORF PF1918 were an insertion at this locus in *P. furiosus* (or a deletion from the other two genomes) (Figure 2.5C). The orthologs of PF1917 and PF1919 are separated by only eight nucleotides in *P. abyssi* and *P. horikoshii*. PF1918 is identified as a “probable transposase” (E=0.00037) when it is searched against the Pfam database (Bateman et al., 2002). The unexpected phylogenetic distribution of these three homologs along with the association of one with a transposase argues that this RNA species is associated with transposition events, whether by function or by chance. Further evidence presented later supports this hypothesis (Chapter 5).

**C/D box methylation guide RNAs**

The transcript of candidate Pf1 unexpectedly overlapped a C/D box small nucleolar RNA (snoRNA) homolog, sR9 (Omer et al., 2000; Gaspin et al., 2000). The candidate region is upstream of the C/D box consensus sequences of sR9; it forms a putative stem-loop structure that is conserved among the three *Pyrococcus* species and shows covariation (Figure 2.6A). I probed a northern blot and performed 5’ RACE with an oligonucleotide complementary to the C/D box region of sR9 and found that sR9 is present in two abundant forms in the cell—a shorter form similar to other C/D box RNAs in *Pyrococcus* and a longer form that includes the stem-loop structure (Figure 2.6B). Others have proposed that the stem-loop is a pseudouridylation guide RNA (Rozhdestvensky et al., 2003). I have drawn the stem-loop motif to reflect this hypothesis, which will be discussed further below.

*sccA* is also adjacent to a C/D box RNA, sR44, but the bands visible on northern blots (Figure 2.3 and data not shown) suggest that the abundant forms are physically separate
Figure 2.6: Chimeric pseudouridylation and methylation guide RNA sR9. (A) MFOLD predicted secondary structure folding for sR9, with covarying bases in the stem structure noted. The annotations for the C, C’, D, and D’ boxes come from Omer et al., 2000. Boldface sequences are the predicted guides, as described elsewhere (Omer et al., 2000; Rozhdestvensky et al., 2003). (B) Northern blot probed with an oligonucleotide complementary to the shaded region (A).
in vivo. However, evidence from the 5’- and 3’-RACE experiments suggests that sscA is cotranscribed with either sR44 or ORF PF1375, which is the translation elongation factor eF-1α-subunit (Figure 2.4). I have no evidence to either support or contradict cotranscription of all three genes in a single operon.

There are over 50 known Archaeal C/D snoRNA homologs in P. furiosus (Omer et al., 2000; Gaspin et al., 2000). Of these, only sR9, sR44, and four others were detected in my screens; all six of these are either adjacent to or in the intron of another structured ncRNA. All but one have a G+C content of 50-55%; sR40 has a G+C content of 64%. By themselves, the C/D snoRNA homologs seem to have little conserved intramolecular secondary structure. These observations suggest that both of my screens identify only a subset of highly structured ncRNAs and that they fail to reliably detect unstructured ncRNAs.

**H/ACA pseudouridylation guide RNAs**

hgcG is significantly similar to a region of the Archaeoglobus fulgidus genome. In a recent experimental screen for ncRNAs in A. fulgidus, this locus was identified as the second-most abundant transcript (Afu-4), further suggesting that hgcG is a real ncRNA conserved among at least two genera of Archaea (Tang et al., 2002a). Afu-4, along with several other small transcripts identified in A. fulgidus, were suggested to be pseudouridylation guide RNAs (Tang et al., 2002a). In eukaryotes, these RNAs are known to have characteristic H and ACA box motifs, fold into a specific secondary structure, and guide pseudouridylation of a specific nucleotide based on complementarity to its guide sequence (Ganot et al., 1997). In A. fulgidus, the H and ACA motifs are present, the appropriate secondary structure can
form, and the predicted target RNAs are pseudouridylated (Tang et al., 2002a). Intriguingly, 
hgcG can fold into the same secondary structure as Afu-4 (Rozhdestvensky et al., 2003),
though it does not have the H box motif. This suggests that the Archaeal equivalent of the
H and ACA boxes may vary from the eukaryotic consensus.

The *A. fulgidus* pseudouridylation guide RNAs were found to contain a secondary struc-
ture motif called the K-turn (Rozhdestvensky et al., 2003), which is where the L7Ae protein
binds the RNA. Four other of the RNAs identified in this chapter have a K-turn motif and
can fold into a typical pseudouridylation guide structure. On the basis of this evidence,
it was suggested that these RNAs (*hgcA*, sR9, *hgcE*, and *hgcF*), as well as *hgcG* are all
pseudouridylation guide RNAs (Rozhdestvensky et al., 2003). If this hypothesis is correct,
then sR9 is a “chimeric” guide RNA, containing both a pseudouridylation guide and a
methylation guide part. Such a chimeric RNA has been observed in eukaryotes, though in
that case the only transcript found was the full-length chimera (Jady and Kiss, 2001).

**Discussion**

Here I have presented several screens for ncRNAs in the AT-rich hyperthermophiles *M. jan-
naschii* and *P. furiosus*. Each screen identified approximately five previously unidentified
ncRNAs in each organism. Two independent screens in *P. furiosus* produced nearly identi-
cal sets of expressed ncRNA genes. Therefore, I believe these two screens have come close to
saturation for a class of highly structured, conserved ncRNAs. These screens do not identify
ncRNA genes without significant secondary structure, as canonical C/D box methylation
guide RNA sequences were not identified unless they were adjacent to other, highly struc-
tured features (Omer et al., 2000; Gaspin et al., 2000). I cannot exclude the possibility that more nonconserved ncRNA genes or ncRNA genes without significant secondary structure remain to be found.

A QRNA screen of E.coli resulted in an estimate of about 200 structural ncRNA genes in this organism (Rivas et al., 2001), a number that is roughly consistent with the results of other screens (Hershberg et al., 2003). Thirty-six different loci identified in these screens have been experimentally shown on northern blots to express small stable RNAs thus far. These 36 are added to 10 ncRNAs previously known in E.coli (Wassarman et al., 1999). In contrast, I find far fewer new structural ncRNAs in screens of P. furiosus and M. jannaschii. The reasons for this discrepancy remain unclear. One possibility is that the constraints of high temperature environments select against the use of ncRNAs in hyperthermophiles. Another is that E.coli (which has both a genome size and predicted protein-coding gene count about twice that of P. furiosus or M. jannaschii) has more complex regulation and has more regulatory RNAs. It is also possible that these expressed ncRNA transcripts have no significant function, and that these numbers vary greatly from organism to organism because of nonadaptive mechanisms (although conserved RNA structure tends to argue against this). As more screens for ncRNAs are done in more prokaryotes, it should become easier to resolve which hypothesis is correct.

Another open question is that of function. About half the RNAs I found appear to be pseudouridylation guide RNAs, though experimental testing of this hypothesis is necessary. hgcC and its homologs are associated with a transposon, but their relationship to transposition is unknown. Genetic or biochemical studies will be needed to elucidate function. Because M. jannaschii and P. furiosus are not easily manipulable genetic systems
presently (Sowers and Schreier, 1999), finding homologs of these genes in other organisms will be essential to apply reverse genetic approaches (e.g., knockouts).

A related question concerns evolutionary conservation and phylogenetic diversity. Many ncRNA genes, including all those previously known in *M. jannaschii* and *P. furiosus*, are known to exist across at least two of the three domains of life (i.e., ribosomal RNA, tRNA, RNase P, 7S RNA, and C/D box snoRNA homologs). Other ncRNAs are as yet only known in a phylogenetically restricted group. By BLAST analysis, the novel ncRNAs detected here seem to have narrow phylogenetic distributions. With two exceptions discussed above, I did not detect any primary sequence similarity between these novel ncRNAs and other Archaeal genomic sequences, including between the ncRNAs identified in *M. jannaschii* and *P. furiosus*. However, because structural ncRNAs often evolve to conserve structure rather than sequence, it is possible that homologs cannot be detected through simple primary sequence searches. The observation that I could not generally find homologs of the pseudouridylation guide RNAs, which are known to exist in eukaryotes, using primary sequence searching supports this hypothesis. Secondary structure-based search methods may be able to identify homologs (Eddy and Durbin, 1994; Eddy, 2002b). To date, these methods require a trusted secondary structure and multiple sequence alignment with a large number of sequences. Towards this end, I aim to develop a secondary-structure based search method that works with a single sequence and predicted RNA structure.
Data Deposition

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF447575-AF447578 and AF468960-AF468966).
Chapter 3

RSEARCH: Finding homologs of single structured RNA sequences

\[\text{1The bulk of this chapter, excluding the section on metazoan RNase P, was co-written with Sean Eddy and will appear in } \text{BMC Bioinformatics. I performed most of the work described in this chapter, while Sean developed the local alignment model.}\]
Abstract

I have developed a program, RSEARCH, that takes a single RNA sequence with its secondary structure and utilizes a local alignment algorithm to search a database for homologous RNAs. For this purpose, I have developed a series of base pair and single nucleotide substitution matrices for RNA sequences called RIBOSUM matrices. RSEARCH reports the statistical confidence for each hit as well as the structural alignment of the hit. I show several examples in which RSEARCH outperforms the primary sequence search programs BLAST and SSEARCH. I use RSEARCH to find homologs of human RNase P in *Ciona intestinalis, Caenorhabditis elegans, Drosophila melanogaster, Anopheles gambiae*, and *Encephalitozoon cuniculi*. The primary drawback of the program is that it is slow. The C code for RSEARCH is freely available at http://www.genetics.wustl.edu/eddy/software.

Introduction

The inability of BLAST to find many homologs of the noncoding RNA genes identified in the Archaea was not surprising. Two possible hypotheses could explain this. First, there might not be any homologs to these noncoding RNAs. While this is a real possibility, all the other noncoding RNA known in the Archaea (ribosomal RNA, tRNA, RNase P, SRP RNA, and C/D box guide RNAs) do have homologs throughout the Archaea and in the Eukaryotes as well. That these RNAs would be so phylogenetically restricted seems doubtful. Alternatively, BLAST just might not be sensitive enough. Since I looked for noncoding RNAs that have potentially significant secondary structure, it stands to reason that homologs of these RNAs may conserve secondary structure rather than primary sequence (Eddy, 2002a).
For both protein and RNA, homology is most readily inferred at the tertiary structure level. For most proteins and RNAs, however, we only have primary sequence data and do not know the tertiary structure. For RNA, secondary structure confers much functional specificity, and potential folds are readily discernible from the primary sequence. Therefore, we can obtain increased power in homology searching by considering the secondary structure of RNA sequences (Eddy, 2002a).

Such a tool would not only be useful for trans-acting noncoding RNA genes. For some protein-coding genes, secondary structure signals present in the messenger RNA help regulate the gene. Examples of such control elements include the iron-responsive element in genes involved in iron metabolism and the selenocysteine insertion sequence that signals selenocysteine should be incorporated into the amino acid (Hentze et al., 1988; Schlegl et al., 1997; Lambert et al., 2002; Wilting et al., 1997). More recently, riboswitches that directly alter gene expression in response to the concentration of small molecules such as thiamin have been discovered (Miranda-Rios et al., 2001; Stormo and Ji, 2001; Winkler et al., 2002). A program able to find homologs of an RNA secondary structure would be useful in studying these elements.

It is useful to distinguish three classes of alignment algorithms that can be used to find homologs of RNA sequences. The first class only uses primary sequence information to align the query sequence to the target database. Such searches are exemplified by the Smith-Waterman algorithm and its heuristic approximations found in programs like BLAST and FASTA. These sequence alignment programs are $O(N^2)$ in time and memory (Smith and Waterman, 1981; Altschul et al., 1990; Pearson and Lipman, 1988), where $N$ is the length of the sequences being analyzed. The second class consists of a search with a known
RNA structure against a sequence database. Such searches have been implemented with profile stochastic context free grammars (SCFGs) and require $O(N^3)$ memory and $O(N^4)$ time (Eddy and Durbin, 1994; Sakakibara et al., 1994; Durbin et al., 1998; Eddy, 2002b; Gautheret and Lambert, 2001). Alternatively, such searches can be performed by defining an RNA structural pattern, though this approach works best on highly conserved secondary structures, and patterns have to be developed by hand (Gautheret et al., 1990; Billoud et al., 1996; Pesole et al., 2000; Macke et al., 2001). A third type of approach consists of a search with a query sequence with unknown secondary structure, where the algorithm searches over all possible foldings of the query aligned to the target. Sankoff described such an algorithm, which is $O(N^4)$ in memory and $O(N^6)$ in time (Sankoff, 1985).

While it is convenient to distinguish among these three classes of algorithms, the boundaries between them are not absolute. Various constrained versions of the Sankoff algorithm have been published that allow it to run in a reasonable amount of time (Gorodkin et al., 1997; Mathews and Turner, 2002; Holmes and Rubin, 2002). One such algorithm constrains the possible alignments (Mathews and Turner, 2002), while the other two constrain the foldings allowed (Holmes and Rubin, 2002). Holmes and Rubin introduced the idea of a “fold envelope,” which allows the algorithm to be constrained to a subset of folds. It can be argued that the profile SCFG approach to searching a database with an RNA of known structure is the limiting case where the fold envelope includes only one structure.

Three types of scoring functions can be used with these search algorithms. When only a single query sequence is given, log-odds position independent substitution matrices are used to give the alignment scores. These are analogous to the BLOSUM matrices used in protein searches (Henikoff and Henikoff, 1992). In the pattern search approach, a binary
match/doesn’t match scoring function is generally used where all allowed letters at each position are enumerated. This is analogous to PROSITE patterns used to analyze amino acid sequences (Gattiker et al., 2002). Finally, a profile-based scoring scheme can be used where position dependent log-odds scores are derived from the observed frequencies in a multiple sequence alignment. This is analogous to the profile approach used in many protein database search programs, including profile hidden Markov models (Gribskov et al., 1987; Krogh et al., 1994; Eddy, 1998). For RNA sequences, only the pattern approach (Gautheret et al., 1990; Billoud et al., 1996; Pesole et al., 2000; Macke et al., 2001) and the profile approach (Eddy and Durbin, 1994; Sakakibara et al., 1994; Gautheret and Lambert, 2001; Eddy, 2002b) have been described to date.

Here I am specifically interested in the problem of finding structural homologs of a single RNA sequence. Since the alignment algorithm is essentially independent of the scoring system, developing such a tool is just a matter of developing an appropriate pairwise substitution matrix and combining it with one of the aforementioned alignment algorithms. I could, for example, derive a single nucleotide matrix and use it in BLASTN searches. Such a primary sequence search would lose much information, much like doing a BLASTN search for homologs of a protein-coding sequence would. When RNAs have conserved secondary structure, I want to consider the intramolecular base pairs that provide this structure to find homologs optimally (Eddy, 2002a). While using the Sankoff algorithm would be ideal, as we often do not know the correct secondary structure of a single query RNA sequence, its cost in time and memory is so prohibitive as to make it impractical at this time for sequence database searching. Therefore, I have chosen to focus on the case where we know the secondary structure of the query sequence.
Here I describe RSEARCH, a program that, given a query sequence with a known secondary structure, searches a nucleotide sequence database for similar RNAs on the basis of both primary sequence and secondary structure. Its core alignment algorithm is identical to profile SCFG alignment (Eddy, 1996; Durbin et al., 1998; Eddy, 2002b). Since alignments are pairwise, alignments are scored using appropriate pairwise substitution matrices. Furthermore, analogous to BLAST, the program calculates statistical confidence values for all hits (Karlin and Altschul, 1990). It is still quite slow; for the time being, I deal with this problem through brute force by parallelizing the search program for clustered computing using the Message Passing Interface (MPI) library (Pacheco, 1997).

Methods

RIBOSUM substitution matrices

In order to perform database searches with a single, folded RNA sequence query, a 16x16 substitution matrix for scoring aligned base pairs and a 4x4 matrix for single aligned nucleotides are needed. Such matrices should give the log-odds ratio for observing a given substitution relative to background nucleotide frequencies (Altschul, 1991). Specifically, for the 4x4 single nucleotide matrix, the individual scores are given by

$$ s_{ij} = \log_2 \frac{f_{ij}}{g_i g_j} $$

where $i$ and $j$ are the two aligned nucleotides, $f_{ij}$ is the empirically observed frequency of $i$ aligned to $j$ in homologous RNAs, and $g_i$ and $g_j$ are the background frequencies of the individual nucleotides. Similarly, for the 16x16 base pairing matrix, the individual scores
are given by

\[ s'_{ijkl} = \log_2 \frac{f'_{ijkl}}{g_ig_jg_kg_l} \]

where \( i \) is basepaired to \( j \), \( k \) is basepaired to \( l \), \( i \) is aligned with \( k \), and \( j \) is aligned with \( l \). In this case, \( f'_{ijkl} \) is the observed frequency of the two base pairs \( i-j \) and \( k-l \) aligned to each other in homologous RNAs. \( g \) again is the background frequency of the individual nucleotides. Note that \( g \) is used for individual nucleotides and not base pairs; the null model in this case is an identical and independently distributed (i.i.d.) model consisting of unaligned random sequences that do not base pair.

The key question, then, is how to find the values for \( f \), \( f' \), and \( g \) needed to calculate these matrices. The values in \( f \) and \( f' \) are conditional on evolutionary divergence time; a shorter divergence time implies higher scores for identities and lower scores for mismatches. Two methods exist to account for evolutionary divergence time. The first method, used by Dayhoff to construct the PAM matrices, infers a rate matrix from closely related sequences. This rate matrix is then used to calculate an exponential family of matrices at different evolutionary distances (Dayhoff et al., 1978). The second method, used to construct the BLOSUM family of matrices, filters and weights sequences in a multiple sequence alignment to approximate a range around some time point (Henikoff and Henikoff, 1992). Matrices produced using the latter method have been found to perform better (Henikoff and Henikoff, 1993), though it is in dispute whether this is an effect of the algorithm or the underlying data used to generate the matrices (Pearson, 1995). Several evolutionary models and a rate matrix have been published for RNA evolution (Muse, 1995; Tillier and Collins, 1998; Knudsen and Hein, 1999). Because BLOSUM-style matrices are argued to be better for finding distant homology relationships (Henikoff and Henikoff, 1993), I have chosen to forgo
the pre-existing RNA rate matrices and construct BLOSUM-style matrices instead.

The algorithm starts with a structurally annotated alignment of multiple RNA sequences to be used as training data. The consensus secondary structure is mapped onto individual sequences by removing any base pairs from the secondary structure for an individual sequence that align with a gap in that sequence. Sequences are then weighted by grouping all sequences more than a certain percentage identical using single-linkage clustering; all sequences in a group are given equal weights that sum to 1. This is identical to the clustering used in constructing the BLOSUM matrices (Henikoff and Henikoff, 1992). In order to allow for a shorter evolutionary distance than would be allowed by following the BLOSUM algorithm exactly, I added a second percentage identity cutoff not found in the original BLOSUM algorithm. Only pairs of sequences whose percent identity meet or exceed this cutoff are counted at all. It should be noted that this second threshold does not necessarily have to be less than the first, clustering percent identity. If that is the case, then one would be counting weighted pairs within clusters; no intercluster pairs would be counted.

Let each of \(i, j, k, l\) represent a nucleotide (\(1 \leq i, j, k, l \leq 4\)). Then, two triangular count matrices are initialized using \(c_{ij} = 0(1 \leq i \leq j \leq 4), c'_{ijkl} = 0(1 \leq 4i + j \leq 4k + l \leq 16)\), where \(c\) is the count matrix for single-stranded regions and \(c'\) is the count matrix for basepaired nucleotides (an \(ij\) basepair aligned to a \(kl\) basepair). Triangular matrices are used because nucleotide (base pair) X in sequence 1 aligned to nucleotide (base pair) Y in sequence 2 should count the same as nucleotide (base pair) Y in sequence 1 aligned to nucleotide (base pair) X in sequence 2. However, we assume that an X-Y base pair may not be equivalent to a Y-X base pair in the context of the entire RNA molecule and therefore count these pairs separately. A count vector \(d_i = 0(1 \leq i \leq 4)\) is also initialized.
for background nucleotide frequencies. Each pair of sequences is then examined. If the pair does not meet the minimal percent identity criterion, it is skipped and the next pair is examined. Otherwise, the weight of this pairing, $w$ is set to be the average of the weights given to the two individual sequences. (Arguably, this weight should be set to be the product rather than the average of the individual weights. Though I did not fully explore this possibility, preliminary evidence suggests the method of calculating this weight does not appreciably influence performance.) For each aligned base pair $(i, j, k, l)$ in the alignment, $w$ is added to $c_{ijkl}^\prime$, $d_i$, $d_j$, $d_k$, and $d_l$; for all other aligned nucleotides $(i, j)$, $w$ is added to $c_{ij}$, $d_i$, and $d_j$. The counts are then converted to empirical frequencies using:

$$f_{ij} = \frac{c_{ij}}{4} \sum_{i=1}^{4} \sum_{j=1}^{4} c_{ij}$$

$$f_{ijkl}^\prime = \frac{c_{ijkl}^\prime}{4} \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \sum_{l=1}^{4} c_{ijkl}^\prime$$

$$g_i = \frac{d_i}{\sum_{i=1}^{4} d_i}$$

The score matrices $s$ and $s^\prime$ are then calculated using the equations given above.

In order to collect these counts, we need high-quality structure-annotated alignments. I decided to use the small subunit ribosomal RNA alignment from the European Ribosomal RNA Database (Van de Peer et al., 1994). Specifically, the 1995 version of the database was pruned by removing sequences in which either more than 5% of the nucleotides are ambiguous or less than 50% of the base-paired positions are present. The resultant alignment consists of 2492 sequences ranging from 610 to 2305 nucleotides in length. When all pairs of sequences are counted, approximately $2.30 \times 10^9$ aligned single nucleotides and
$1.06 \times 10^9$ aligned base pairs are counted and used to calculate the matrix. I created 170 unique matrices by varying the percent identity level at which clustering occurs and the minimal percent identity for a pair of sequences to be counted. I have chosen to call this series of matrices the RIBOSUM matrices (RIBOsomal rna SUstitution Matrix). The first number in the matrix name is the percent identity used for BLOSUM-style clustering. The second number in the matrix name is the percent identity cutoff below which no pairs were considered.

**Construction of a covariance model from a single RNA query**

For these matrices to be useful, I need a good algorithm to perform alignment between an RNA query and a nucleotide database. Like primary sequence alignment, I need to consider both homologous regions of sequence that align as well as insertion and deletion events that put gaps into the alignment. Unlike primary sequence alignment, I also have to consider the nucleotide correlations within each sequence that make up the secondary structure. This structure can be modeled as a bifurcating tree, with each branch terminating in the loop of a stem-loop. Whatever algorithm I use must unambiguously pair each nucleotide in the query with either a nucleotide in the target or a gap, and vice versa. My algorithm is based on a profile stochastic context-free grammar (SCFG) formulation called a “covariance model.” (Eddy and Durbin, 1994; Durbin et al., 1998; Eddy, 2002b). While this formulation was initially described in the framework of probabilistic modeling of profiles, it can deal with arbitrary, non-probabilistic scores just as well. I therefore use the term “covariance model” to describe both the profile SCFG form of the model (Eddy and Durbin, 1994; Durbin et al., 1998; Eddy, 2002b), and the single-sequence, non-probabilistic form presented here.
A covariance model produces ("emits") a nucleotide sequence. The model consists of a set of interconnected states. The states form a tree-like structure, with the root customarily being drawn at the top. As one moves down the tree, nucleotides are filled in from both the left and the right until they meet in the middle. Each state can emit either no nucleotides, a nucleotide on the left side, a nucleotide on the right side, or a base pair consisting of two nucleotides, one on each side. Bifurcations result in a split in the sequence, with each half being filled in from both sides along one of the two bifurcated branches. The model is traversed by following a series of transitions from one state to the next after each emission. Each transition is governed by a score, and only a limited set of transitions are allowed at all. Given a parameterized covariance model, algorithms exist for searching a database for homologous sequences and aligning the model to hits found in the database (Eddy and Durbin, 1994; Durbin et al., 1998; Eddy, 2002b).

For the present purposes, there are two separable steps in the creation of a covariance model. First, the model architecture needs to be determined from the given secondary structure. It is easiest to think of the model as being composed of modular nodes, where each node contains a characteristic arrangement of states. The node architecture of the model follows from the secondary structure of the query. A sample RNA secondary structure and its corresponding "guide tree" of nodes is given in Figure 3.1. There are eight types of nodes. A ROOT node marks the start of the model. A BIF node marks a bifurcation in the tree, and is always followed by a BEGL node on the left branch and a BEGR node on the right branch. All branches end with an END node. The remaining nodes are match nodes; these nodes represent either a base-pair (MATP) or a single nucleotide (MATL and MATR) in the secondary structure. (For a profile SCFG built from a multiple sequence alignment,
only those positions thought to be conserved by some measure correspond to match nodes; for my single-sequence covariance models, all positions correspond to match nodes.) For consistency, MATL is always preferred over MATR when possible. Each secondary structure yields one and only one model architecture, and a model architecture implies a unique secondary structure. As each node is associated with a static arrangement of states, this architecture also gives the final arrangement of states in the model. A more detailed exposition of this algorithm for profile SCFGs is given elsewhere (Eddy, 2002b) and the C code used to construct a covariance model from a given secondary structure can be found in the modelmaker.c file of the Infernal package (http://infernal.wustl.edu/).

The second step in the algorithm is parameterization of the model. This is best thought of in terms of the possible “node-states,” i.e. the various state types present in each node type. A list of all possible node-states is given in Table 3.1, along with what they signify in the pairwise alignments I am creating here. In a profile SCFG, emission scores are log-odds scores as shown in Table 3.1. When we only have a single sequence, we cannot infer emission probabilities from the data given. I instead set these scores from the log-odds RIBOSUM matrix (Table 3.1). In the MP (match pair), ML (match left), and MR (match right) states, emission scores are log-odds scores for both profiles and single sequence models. In the IL (insert left) and IR (insert right) states, emission scores for a profile SCFG are based on observed nucleotide frequencies in insertions. For my single-sequence model, emission scores are 0 because I assume that the nucleotide distribution in insertions follows the null model.

Transition scores are set for transitions from one node-state to another node-state. In a profile SCFG, the log transition probabilities are derived from the observed frequencies of the various transitions. In the single-sequence case, I derive a transition score using the
Figure 3.1: An example SCFG architecture. The sequence at the top folds into the specified secondary structure. At the bottom, the nodal architecture of the model that would produce this sequence is shown. Shaded triangles represent base pair emitting nodes, and point to the base pair they emit. Open triangles represent single nucleotide emitting nodes, and point to the nucleotide they emit.
<table>
<thead>
<tr>
<th>Node-state</th>
<th>Description</th>
<th>Profile emission score</th>
<th>Single-sequence emission score</th>
<th>Gap class</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT_S</td>
<td>Start of model</td>
<td>0</td>
<td>0</td>
<td>M_cl</td>
</tr>
<tr>
<td>ROOT_JL</td>
<td>Gap in query at left end</td>
<td>log $P(a</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>ROOT_JR</td>
<td>Gap in query at right end</td>
<td>log $P(b</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>BEGL_S</td>
<td>Start of left branch of bifurcation</td>
<td>0</td>
<td>0</td>
<td>M_cl</td>
</tr>
<tr>
<td>BEGR_S</td>
<td>Start of right branch of bifurcation</td>
<td>0</td>
<td>0</td>
<td>M_cl</td>
</tr>
<tr>
<td>BEGR_JL</td>
<td>Gap in query at bifurcation</td>
<td>log $P(a</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>MATP_MP</td>
<td>Matched base pair</td>
<td>log $P(a</td>
<td>v)$</td>
<td>$s'_{abkl}$</td>
</tr>
<tr>
<td>MATP_ML</td>
<td>Match on left side of base pair; gap in target on right</td>
<td>log $P(a</td>
<td>v)$</td>
<td>$s_{aj}$</td>
</tr>
<tr>
<td>MATP_MR</td>
<td>Match on right side of base pair; gap in target on left</td>
<td>log $P(b</td>
<td>v)$</td>
<td>$s_{bj}$</td>
</tr>
<tr>
<td>MATP_D</td>
<td>Two gaps in target, for each side of base pair</td>
<td>0</td>
<td>0</td>
<td>DB_cl</td>
</tr>
<tr>
<td>MATP_JL</td>
<td>Gap in query just after left side of base pair</td>
<td>log $P(a</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>MATP_IR</td>
<td>Gap in query just before right side of base pair</td>
<td>log $P(b</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>MATL_ML</td>
<td>Match to single nucleotide on left</td>
<td>log $P(a</td>
<td>v)$</td>
<td>$s_{aj}$</td>
</tr>
<tr>
<td>MATL_D</td>
<td>Gap in target on left</td>
<td>0</td>
<td>0</td>
<td>DL_cl</td>
</tr>
<tr>
<td>MATL_JL</td>
<td>Gap in query on left</td>
<td>log $P(a</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>MATR_ML</td>
<td>Match to single nucleotide on right</td>
<td>log $P(b</td>
<td>v)$</td>
<td>$s_{bj}$</td>
</tr>
<tr>
<td>MATR_D</td>
<td>Gap in target on right</td>
<td>0</td>
<td>0</td>
<td>DR_cl</td>
</tr>
<tr>
<td>MATR_IR</td>
<td>Gap in query on right</td>
<td>log $P(b</td>
<td>v)$</td>
<td>0</td>
</tr>
<tr>
<td>END_E</td>
<td>End of stem-loop</td>
<td>0</td>
<td>0</td>
<td>M_cl</td>
</tr>
<tr>
<td>BIF_B</td>
<td>Bifurcation</td>
<td>0</td>
<td>0</td>
<td>M_cl</td>
</tr>
</tbody>
</table>

Table 3.1: All possible node-states and their emission scores. $v$ is the current state. $a$ is the nucleotide present in the query on the left, $b$ is the nucleotide present in the query on the right. $j$ is any nucleotide in the target for a single nucleotide alignment, while $k,l$ is a base pair in the target for a base pair alignment. $g$ is the background frequency of a nucleotide and $s$ and $s'$ are the substitution matrices defined in the text. Node-states with an M gap class are in the “mainline” path through the model that the an exact match would follow. Node-states with an IL or IR gap class represent a gap in the query sequence, while node-states with a DL, DR, or DB gap class represent gaps in the target sequence.
standard affine gap penalty formulation. I parameterize the overall penalty for a gap as 
\[ \alpha + \beta n \] where \( \alpha \) is the gap open penalty and \( \beta \) is a gap residue penalty which is multiplied
by the size of the gap. I take half the \( \alpha \) penalty on opening a gap and the other half on
closing it. The \( \beta \) penalty is taken for each residue in a gap. Gaps emitted on both sides
simultaneously (i.e. through a MATP\_D node-state) are taken as two independent gaps. I
also want to use a separate set of gap penalties for gaps within a base-paired region. If both
node-states in a transition are in MATP nodes, the \( \alpha \) parameter is replaced by a different
parameter, \( \alpha' \). Similarly, for transitions from a MATP\_D state to another MATP node, \( \beta \)
is replaced by \( \beta' \). \( \alpha \) and \( \beta \) are used for transitions between base-paired and single-stranded
regions.

To map this formulation onto the covariance model, I classify all node-states into one
of six classes (Table 3.1): M\_cl, for an aligned match or mismatch between the query and
target; IL\_cl, for a gap in the query sequence on the left; IR\_cl, for a gap in the query
sequence on the right; DL\_cl, for a gap in the target sequence on the left; DR\_cl, for a gap
in the target sequence on the right; and DB\_cl, for a gap in the target sequence on both
sides. All classes other than M\_cl represent some sort of gap that requires a gap penalty.
The exact parameterization of transitions between classes is given in Table 3.2. For some
transitions, the gap penalty presented in Table 3.2 represents the sum of penalties for several
different gaps. All gap penalties are multiplied by \(-1\) to get the transition score used in
the model. These four gap parameters are empirically determined as described later. They
are not normalized to have their exponentials sum to 1; therefore the resultant scores using
these models cannot be directly interpreted probabilistically.
Table 3.2: Parameterization of negative transition scores from gap penalties. $\alpha$ and $\beta$ are replaced with $\alpha'$ and $\beta'$ for the specific cases described in the text. The IR to IL transition is never allowed in these models.

<table>
<thead>
<tr>
<th>From class</th>
<th>M_cl</th>
<th>IL_cl</th>
<th>DL_cl</th>
<th>IR_cl</th>
<th>DR_cl</th>
<th>DB_cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>M_cl</td>
<td>0</td>
<td>$\frac{1}{2}\alpha$</td>
<td>$\frac{1}{2}\alpha$</td>
<td>$\frac{1}{2}\alpha$</td>
<td>$\frac{1}{2}\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>IL_cl</td>
<td>$\beta + \frac{1}{2}\alpha$</td>
<td>$\beta$</td>
<td>$\beta + \alpha$</td>
<td>$\beta + \alpha$</td>
<td>$\beta + \frac{3}{2}\alpha$</td>
<td></td>
</tr>
<tr>
<td>DL_cl</td>
<td>$\beta + \frac{1}{2}\alpha$</td>
<td>$\beta + \alpha$</td>
<td>$\beta$</td>
<td>$\beta + \alpha$</td>
<td>$\beta + \frac{1}{2}\alpha$</td>
<td></td>
</tr>
<tr>
<td>IR_cl</td>
<td>$\beta + \frac{1}{2}\alpha$</td>
<td>N.A.</td>
<td>$\beta + \alpha$</td>
<td>$\beta$</td>
<td>$\beta + \frac{3}{2}\alpha$</td>
<td></td>
</tr>
<tr>
<td>DR_cl</td>
<td>$\beta + \frac{1}{2}\alpha$</td>
<td>$\beta + \alpha$</td>
<td>$\beta + \alpha$</td>
<td>$\beta$</td>
<td>$\beta + \frac{1}{2}\alpha$</td>
<td></td>
</tr>
<tr>
<td>DB_cl</td>
<td>$2\beta + \alpha$</td>
<td>$2\beta + \frac{3}{2}\alpha$</td>
<td>$2\beta + \frac{1}{2}\alpha$</td>
<td>$2\beta + \frac{3}{2}\alpha$</td>
<td>$2\beta + \frac{1}{2}\alpha$</td>
<td>$2\beta$</td>
</tr>
</tbody>
</table>

Local alignment searches

The model as described above can only perform global alignment with respect to the query sequence. The model is modified slightly to allow for local alignment as well. Two different types of locality are allowed. The first is called “begin locality,” and resembles local alignment as implemented in the Smith-Waterman algorithm (Smith and Waterman, 1981).

In this case, a penalty – $\text{beginsc}$ – is taken if the alignment begins inside the model, i.e. the states representing the outermost parts of the RNA secondary structure are not included. This is analogous to the convention in the Smith-Waterman algorithm that there is no penalty (score of 0) for a local rather than a global alignment (Smith and Waterman, 1981). Following that convention, the $\text{beginsc}$ penalty is set by default to 0. The second type of locality is “end locality.” In this case, a penalty – $\text{endsc}$ – is taken to allow the subtree of a model below the current state to be ignored, and replaced by an insertion of arbitrary size in the target sequence. Examples of both kinds of locality are shown in Figure 3.2.

These modifications are easily accommodated in the standard scanning algorithm for
Figure 3.2: The two classes of local alignment. Each example shows how the nodal guide tree best aligns to the target sequence. At the bottom is the RSEARCH output for the alignment. On the left is an example of begin locality, while on the right is an example of end locality. The numbers next to the query sequence represent positions relative to the entire query; the numbers next to the target sequence represent positions relative to the subsequence defined in the “Target =” line. Secondary structure is indicated by the < and > characters, which represent an opening and closing nucleotide of a base pair, respectively.
covariance models, which is described in detail elsewhere (Eddy and Durbin, 1994; Durbin et al., 1998). The begin parameter is modeled as a transition from the root state to the consensus states (MATP, MATL, MATR, BIF). The end parameter is modeled as a transition from each of the consensus states (MATP, MATL, MATR, BEGL, BEGR) to a special "EL" (end-local) state, which generates residues at the background residue frequency and thus has a zero score for any subsequence after the transition cost, end, has been paid. (In actuality, version 1.0 of RSEARCH allowed transitioning to any state from the root with a begin penalty, and allows transitioning from any state to EL with an end penalty. More recent versions implement the algorithm as described. This slight difference does not appear to significantly alter performance [data not shown].)

The scanning algorithm takes a covariance model with M states (including B bifurcations), parameterized as described above, as well as a target database sequence of length L. Theoretically, the best alignment could have the nucleotide at position 1 in the database base pair with any nucleotide at position 2 through L. If the database includes large genomic contigs, L could be on the order of tens of megabases, which is much larger than we would expect any RNA to ever be. There is no need to check for base-pairings further apart than the longest RNA we would expect to find. To reduce time and memory requirements, I limit the total length of sequence in the target database for a single hit to a parameter D. Then, only positions 2 through D will need to be checked for a base pair to position 1. D needs to be set small enough for efficient performance but large enough so as not to miss any real homologs. By default, D is set to be two times the query length. The algorithm has a time complexity of \(O((M - B)L + BLD^2)\) and a memory complexity of
A greedy algorithm is used to resolve these scores into a maximally scoring set of $K$ nonoverlapping hits $(i_1, j_1; i_2, j_2; \ldots; i_K, j_K)$ on the target sequence, where $i_x$ and $j_x$ are the starting and ending coordinates of the hit on the target sequence, respectively. Alignments are then determined using the previously reported divide-and-conquer algorithm (Eddy, 2002b). For each hit greater than a specified threshold, the score, alignment positions in the query and the target, the alignment, and E-values and P-values (calculated as described below) are reported.

**Statistics**

In order to determine statistical significance, I need to know what distribution RSEARCH scores follow. Much work has been done on the statistics of primary sequence alignment (Karlin and Altschul, 1990; Mott, 1992; Altschul and Gish, 1996; Pearson, 1998; Olsen et al., 1999; Altschul et al., 2001; Bailey and Gribskov, 2002). All these approaches rest on the proposition, proven for the ungapped case and empirically true for the gapped case, that local alignment scores follow the Gumbel distribution (Karlin and Altschul, 1990; Gumbel, 1958). For a specific query sequence, the expected number of hits ($E$) with score greater than or equal to a given score ($x$) is given by the formula $E = K Ne^{-\lambda x}$, where $N$ is the size of the database and $K$ and $\lambda$ are characteristic parameters dependent on the query sequence and the base composition of the database. (It should be noted that this equation is often seen written as $E = KM Ne^{-\lambda x}$, where $M$ is the size of the query sequence. As I have chosen to recalculate $\lambda$ and $K$ for each individual query sequence, I have incorporated the $M$ parameter into my $K$.) This formula can also be written as $E = e^{-\lambda(x-\mu)}$, where $\mu = \frac{\log KN}{\lambda}$. The probability ($P$) that a score greater than or equal to a given score ($x$) is
observed by chance is then given by \( P = 1 - e^{-E} = 1 - \exp(KN e^{-\lambda x}) \). Thus, calculating the E-value and P-value for a given score is simple provided a reasonable procedure for determining \( \lambda \) and \( K \) is found.

In the absence of a theory for the distribution of gapped structural alignment scores, I have chosen to determine \( K \) and \( \lambda \) empirically through maximum likelihood fitting of a Gumbel distribution to the score histogram obtained from alignment to random, simulated sequences. A large number (usually 1000) of i.i.d. sequences of length \( 2 \times D \) (where \( D \) is the maximum possible length of a hit) are generated. The G+C content of these sequences are set as described below. The query is searched against each random sequence, and the best score is recorded in a histogram. A maximum-likelihood method is then used to determine \( \lambda \) and \( \mu \) for a database of length \( 2 \times D \) from these data (Lawless, 1982; Mott, 1992; Bailey and Gribskov, 2002; ftp://ftp.genetics.wustl.edu/pub/eddy/papers/evd.pdf). I can then calculate \( K \) using the formula \( K = e^{\mu / D} \).

I initially created the random sequences using an i.i.d. model assuming a single, fixed G+C content for all sequences. As will be described below, this proved to be inadequate, as many databases have heterogeneous G+C contents. I then randomly choose a G+C content for each random sequence based on the distribution of G+C contents in the target database. I determine the G+C content in the target database measured in adjacent, non-overlapping windows of 100nt each, and use the distribution of these contents to select randomly a G+C content for each random sequence. For some databases where the range of frequently observed G+C contents is large, one pair of values for \((\lambda, K)\) is not enough to accurately calculate E values. To allow for multiple values of \((\lambda, K)\), partition points in the G+C content distribution can be set. For \( N \) partition points, the distribution is divided into \( N+1 \)
bins, and $\lambda$ and $K$ are calculated for each bin. For instance, if a partition point of 50 is set, $\lambda$ and $K$ are first calculated for random sequences with G+C contents sampled from the portion of the G+C content distribution with G+C content $< 50\%$, and $\lambda$ and $K$ are then calculated again with G+C content sampled from the part of the distribution where the G+C content is $\geq 50\%$. Then, for a given database hit, the G+C content of the sequence of the hit is calculated and used to select the appropriate $\lambda$ and $K$ for calculating statistics. Thus, if partitions are used, the rank order of hits based on score and rank order of hits based on statistics may be different.

**Implementation and parallelization**

RSEARCH was implemented in C. Version 1.0 was used for all experiments reported here. Timings and benchmarks reported were performed on a 1 GHz Pentium III Linux workstation with the Mandrake distribution, using the Intel C compiler version 6.0 with options “-O3 -static -mp1 -xK” to compile the program. Because the RSEARCH algorithm is time-consuming, I also implemented a data-parallel version of RSEARCH using the Message-Passing Interface (MPI) (Pacheco, 1997). Source code is available from the lab web site at: http://www.genetics.wustl.edu/eddy/software and is available free of charge under the terms of the GNU General Public License (GPL).

**Data sets and parameters**

Several different data sets were used for testing and analysis, as described below. Sequence and structures for ribonuclease P RNA were taken from the RNase P database (Brown, 1999). Signal Recognition Particle (SRP) sequences and structures were taken from the
SRP database (Gorodkin et al., 2001a). Three different human SRP sequences appear in the database. I chose to use sequence A, which corresponds to the originally sequenced RNA molecule. (This sequence was taken from GenBank accession X01037, but has two nucleotides that are different from the current GenBank version X01037.1). I used an Asn tRNA from *Archaeoglobus fulgidus* (GenBank AE001087.1, positions 4936-5008) (Klenk et al., 1997) with the structure proposed by tRNAscan-SE (Lowe and Eddy, 1997). For a representative yeast (*S. cerevisiae*) tRNA, I took the genomic sequence of the Ala tRNA originally sequenced by Holley (Holley et al., 1965; Goffeau et al., 1997) (GenBank accession Z28265.1, positions 1117-1189). The precursor to the *C. elegans* miRNA mir-40 was also used (Lau et al., 2001) (GenBank accession AL110499.1, positions 17411-17507). Unless otherwise noted, the full length of each gene was used as the query sequence.

Three different databases were used for searches. The yeast genome was downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/NCBI_genome_source/*.fsa and dated August 29, 2001 (Goffeau et al., 1997). The database of 11 Archaeal genomes was previously described (Klein et al., 2002). The *Arabidopsis thaliana* genome was downloaded from ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES (Arabidopsis Genome Initiative, 2000).

For testing BLAST, WU-BLAST 2.0MP, dated October 20, 2002, was used with the -W3 and -kap options (http://blast.wustl.edu)(Altschul et al., 1990). For SSEARCH (an implementation of the full Smith-Waterman algorithm), version 3.4t05 was used with default parameters (Smith and Waterman, 1981; Pearson, 1991).
Results

Optimal parameter set

I first asked what set of parameters – matrix, gap penalties, beginsc, and endsc – would be optimal to use as the defaults in RSEARCH. To assess this, I decided to choose the set of parameters that gives the lowest minimum error rate for a set of two test searches. The minimum error rate is defined as the minimal possible sum of false positives and false negatives for a search taken over all possible cutoff scores. The first search I used was the genomic copy of the alanine tRNA from *S. cerevisiae* folded using tRNAscan-SE searched against the yeast genome to identify the 295 tRNAs present there. The second search I used was *M. jannaschii* RNase P searched against a database of 11 Archaeal genomes to identify the 11 RNase P homologs found there. As doing the real searches for all the parameters I wanted to test would have been computationally infeasible, I estimated the false negative rate in many cases by searching a smaller database and extrapolating to the size of the full database. To abbreviate the yeast tRNA search, I took chromosome VII as a proxy for the whole genome. For the RNase P search, I created a smaller database of similar G+C content. After several rounds of iterative trial and error optimizing different parameters, I decided to use RIBOSUM85-60 as the default matrix with $\alpha = 10.00$, $\beta = 5.00$, $\alpha' = 0.00$, $\beta' = 15.00$, $\text{beginsc} = 0.00$, and $\text{endsc} = -15.00$. I might have been able to derive a more robust parameter set had I used a more comprehensive set of tests, but the long running time required by RSEARCH makes such an approach infeasible.

RIBOSUM85-60 has several characteristics typical of these matrices (Figure 3.3). It consists of two matrices – one 16 x 16 for base pair substitutions and the other 4 x 4 for single
Figure 3.3: The RIBOSUM85-60 matrix. The 16x16 matrix is used to get scores for aligning base pairs. The 4x4 matrix is used to get scores for aligning single-stranded regions. Positive scores are shaded.

nucleotide substitutions. In the single nucleotide substitution matrix, the A-A identity has a score (2.22) much larger than the other single nucleotide identities. This suggests that conserved As are especially common in single stranded regions of 16S ribosomal RNA.

Unlike typical nucleotide or amino acid substitution matrices, not all values on the identity diagonal are positive. This reflects the specificity of base pairing. Canonical Watson-Crick and G-U pairs are observed much more often than non-canonical pairs. Since non-canonical pairs occur less often than expected on the basis of individual nucleotide probabilities, the log-odds score for these pairs aligned to themselves is negative. Second, substitution of one canonical pair for another usually gives a positive score (e.g. A-U to C-G has a score of 1.47). Therefore, the RIBOSUM matrices resemble what I intuitively assume a good base pairing substitution matrix would look like.

I compared the minimum error rates at these parameter choices to the performance of BLAST and SSEARCH on the same search problems. For the problem of finding yeast
tRNAs using the alanine yeast tRNA as a query, the minimum error rate for the BLAST search was 194, while SSEARCH gave a minimum error rate of 223. The minimum error rate observed using RSEARCH was 50. Instead of the default matrices in BLAST and SSEARCH, I also tried other matrices and gap penalties, both made in a similar fashion to RIBOSUM85-60 and as suggested by others (States et al., 1991). None of these changes resulted in a significant improvement in performance for either BLAST or SSEARCH. For the *M. jannaschii* RNase P search, both BLAST and SSEARCH give a minimum error rate of 4, while RSEARCH gives a minimum error rate of 2. These tests indicate that RSEARCH, using secondary structure, is capable of outperforming primary sequence search programs.

To help insure that the above results were not the result of overtraining on those two specific searches, I performed similar tests with another tRNA and RNase P query sequence. I first asked how well an Asn tRNA from *A. fulgidus* could find the 494 tRNAs present in the database of Archaea genomes. The minimum error rates for BLAST and SSEARCH were 305 and 373, respectively. The RSEARCH search had a minimum error rate of 66. I also used the *P. furiosus* RNase P sequence to search the database of Archaeal genomes for homologs. The minimum error rate for BLAST was 6 and for SSEARCH was 5. The minimum error rate using RSEARCH was a perfect 0. These data reinforce my conclusion that RSEARCH can outperform primary sequence search programs.

**Statistics**

Calculation of minimum error rates requires prior knowledge of all homologs of the query sequence in the database. As I wish to use RSEARCH to search a database when such information is still unknown, I need a method for evaluating the statistical significance of a
hit. I assumed that RSEARCH scores would follow the Gumbel distribution, just as scores from primary sequence search programs like BLAST do (Karlin and Altschul, 1990; Altschul and Gish, 1996; Pearson, 1998). I therefore asked whether the scores produced by a search of random sequence do in fact follow the Gumbel distribution. The distribution of scores from one such search of random sequences is shown in Figure 3.4a. It is clear from these plots that the score distribution more closely fits the Gumbel distribution than the normal Gaussian distribution.

I next assessed whether, on average, the E-values reported are an accurate reflection of the false positive rate. I examined six searches of the Archaeal genome database with *M. jannaschii*, *P. furiosus*, *E. coli*, *B. subtilis*, *S. cerevisiae*, and *H. sapiens* RNase P sequences as the queries. For the six searches, I then computed the average observed E-value (observed number of false positives) at various calculated E-value cutoffs. If the statistical model is correct, I expect the calculated E-value cutoff to be equal to the average number of observed false positives scoring better than the cutoff. I first calculated E-values using random sequences with a fixed G+C content of 45.8%, which is the overall G+C content of the Archaeal database. Under this model, there were 246 ± 257 false positives at an E-value cutoff of 1. Therefore, this statistical model was inadequate.

Looking more closely at the data led me to hypothesize that the statistical method was failing because the target database consists of a heterogeneous population of sequences with widely varying G+C contents. I first tried correcting for this by randomly picking a G+C content for each random sequence used in the simulation to calculate $\lambda$ and $K$. This G+C content was picked from the distribution of G+C contents observed in the database. With this model, there were 8 ± 8 false positives at an E-value cutoff of 1. While the
Figure 3.4: a Distribution of scores for a search against random sequences. I searched a database of 10,000 random sequences of 10,000 nucleotides each with a GC composition of 50% using the precursor to the *C. elegans* miRNA mir-40 as the query (Lau et al., 2001). I took the best score found for each of the 10,000 sequences in the database and plotted their distribution. I then calculated the mean and standard deviation and plotted the Gaussian distribution for those values. I also calculated $K$ and $\lambda$ for the Gumbel distribution and plotted that distribution. b Average observed number of hits with E-value less than a cutoff versus reported E-value for searches of various RNase P queries against database of Archaeal genomes. E-values were computed using partition points of 40% and 60% G+C content.
average number of false positives is closer to that predicted by the E-value, and the standard
deviation is much smaller, we wished to improve the statistics even further. Since the G+C
content distribution of the database has a large variance, I decided to partition the G+C
distribution into 3 bins: one for G+C contents less than 40%, one for contents between 40% and 60%, and one for G+C contents greater than 60%. I calculated separate values of \( \lambda \) and \( K \) for each of these bins. With this statistical model, there are \( 2 \pm 3 \) false positives at an E-value of 1. Observed E-values between 1 and 100 never deviate significantly from the computed E-value (Figure 3.4b), especially for observed E-values less than 10. Therefore, this statistical model was used for further searches of the Archaeal database. Since partitions are only necessary for databases with a large variance, and since the optimal partitions vary from database to database, the default statistical model in RSEARCH is to calculate a single \( \lambda \) and \( K \) without using any partitions.

**Examples of Performance**

I then wished to assess how well RSEARCH would perform in additional realistic scenarios. To study this, I chose an RNA molecule which was not part of the training set at all – the Signal Recognition Particle (SRP) RNA. I tested a variety of SRP query sequences against several database genomes. Each test was designed to look across phylogenetic domains or kingdoms. In each case, I compared its performance to BLAST and SSEARCH. In some cases, RSEARCH performed as well as these primary sequence search programs. In one rare case, using *Pyrococcus horikoshii* SRP as the query, SSEARCH and BLAST outperformed RSEARCH. Some examples where RSEARCH does outperform primary sequence searches are given below.
In one example, I searched for the 11 SRP genes in the Archaeal genomes using SRP from the Eubacteria \textit{B. subtilis} as the query. No hits with an E-value less than 10 were observed with BLAST. SSEARCH found 13 hits at an E-value cutoff of 10, three of which were true homologs and 10 of which were false positives. No hits were observed with an E-value less than 0.05 using SSEARCH. In contrast, 16 hits with an E-value less than 10 were observed with RSEARCH, six of which are true homologs. Two of these true positives, but none of the false positives, had an E-value less than 0.05 (E=0.0064 for \textit{M. jannaschii} and E=0.0067 for \textit{A. fulgidus}).

If I instead use \textit{H. sapiens} (a eukaryote) SRP as the query to find homologs in the Archaeal genomes, BLAST found seven hits with an E-value less than 10, none of which are true homologs. SSEARCH found nine hits with an E-value less than 10, only one of which was a true homolog. SSEARCH did not find any hits with an E-value less than 0.05. RSEARCH, on the other hand, found four hits, two of which are true homologs, with an E-value less than 10. The two true homologs, but not the two false positives, had E-values less than 0.05 (E= 0.0067 for \textit{Methanobacterium thermoautotrophicum} and E=0.0081 for \textit{A. fulgidus}).

I then searched the genome of the plant \textit{A. thaliana} with \textit{H. sapiens} (an animal) SRP. There are at least eight copies of SRP in the genome; I take a significant hit to any of these eight copies as indicative of an ability to find SRP (Regalia et al., 2002). Neither BLAST nor SSEARCH can find any of these copies with an E-value less than 10. In contrast, several copies of SRP can be found using RSEARCH, with the most significant hit having an E-value of $9.6 \times 10^{-6}$. Taken together, these data suggest that if I knew about either \textit{H. sapiens} or \textit{B. subtilis} SRP, I would be able to find SRP genes in distantly related genomes.
in other phylogenetic domains or kingdoms with confidence using RSEARCH, but not with either SSEARCH or BLAST.

**RNase P in non-vertebrate metazoans**

Since RSEARCH performed so well in finding already known distant homologies, I next asked if it could answer an open question in biology. RNase P is a ribonucleoprotein enzyme that catalyzes maturation of the 5’ ends of tRNA molecules (Stark et al., 1978; Gardiner and Pace, 1980). Since the discovery that its RNA subunit has catalytic properties (Guerrier-Takada et al., 1983), much work has focused on elucidating the structure and structure-function relationship in this molecule. The first metazoan RNase P was biochemically purified from *H. sapiens* (Bartkiewicz et al., 1989). Since the primary sequence of RNase P diverges quite rapidly, Southern blot analysis could not find homologs of the RNA subunit of human RNase P in any non-mammalian organisms tested (Bartkiewicz et al., 1989). Biochemistry and PCR have been used to identify many other vertebrate RNase P sequences and build a consensus secondary structure (Doria et al., 1991; Altman et al., 1993; Eder et al., 1996; Pitulle et al., 1998). In building a consensus vertebrate RNase P structure, Pitulle and colleagues noted that the vertebrates represent just a small fraction of metazoan organisms. Despite the availability of several invertebrate genome sequences (The *C. elegans* Sequencing Consortium, 1998; Adams et al., 2000; Holt et al., 2002; Dehal et al., 2002), no invertebrate RNase P sequences are currently in the RNase P database (Brown, 1999).

I decided to try to use RSEARCH to find invertebrate and other eukaryotic RNase P sequences. Using *H. sapiens* RNase P as the query, I first searched the genome of *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998) using RSEARCH,
SSEARCH, and BLAST with a word size of 3. Neither BLAST nor SSEARCH were able to find a statistically significant hit \((E < 0.05)\) in the \(C.\ elegans\) genome. In contrast, RSEARCH finds a statistically significant hit \((E=0.00014)\) on chromosome I (Figure 3.5a). RSEARCH suggests a secondary structure for \(C.\ elegans\) RNase P based upon the optimally scoring alignment between the two sequences. Upon close examination, and with knowledge of the conserved primary sequence and secondary structure features of eukaryotic RNase P (Chen and Pace, 1997; Pitulle et al., 1998; Frank et al., 2000), I modified the secondary structure in several areas to produce what I believe to be the true secondary structure of \(C.\ elegans\) RNase P (Figure 3.5b). In helices P1 and P8, I moved a gap from within a stem to outside a stem. I shifted the alignments of P3a, P10/11 and P12 to make better stems. I added base pairs to helices P9 and P3a. I created a secondary structure for P3b, which was unaligned in an end locality region. Finally, I noted that proposed structure for helices P2 and P19 were quite weak. Moreover, the only observable consensus to conserved region IV (Chen and Pace, 1997), which is known to be just 5’ of the closing half of helix P2, overlaps the proposed helices P2 and P19. Therefore, I removed helix P19, which allowed me to create a stronger helix P2 that is just 5’ of conserved region IV. In support of this decision, P19 is known to be missing from some non-vertebrate eukaryotic RNase P homologs (Frank et al., 2000). I also found a copy of RNase P in \(C.\ briggsae\) using RSEARCH, which is 92% identical to the \(C.\ elegans\) copy. RSEARCH does not suggest the presence of P19 in \(C.\ briggsae\), further supporting its absence in \(C.\ elegans\) (data not shown). As this sequence contains all of the conserved primary sequence and secondary structure elements of eukaryotic RNase P, I believe this represents the true RNase P sequence from \(C.\ elegans\) (Frank et al., 2000).
Figure 3.5: The *C. elegans* homolog of RNase P.  

**a** Output from the search of human RNase P versus the *C. elegans* genome. The suggested helix P19, removed in my final structure, is in bold.  

**b** Proposed secondary structure for *C. elegans* RNase P, modified from the RSEARCH alignment as described in the text. The sequences in a lighter gray represent regions that were unaligned due to end locality. The sequences in boldface are the universally conserved regions (Chen and Pace, 1997).
With this success, I then went on to search the genomes of *Ciona intestinalis* (Dehal et al., 2002), *Drosophila melanogaster* (Adams et al., 2000), *Anopheles gambiae* (Holt et al., 2002), and *Encephalitozoon cuniculi* (Katinka et al., 2001). Again, neither BLAST nor SSEARCH detected a statistically significant hit ($E < 0.05$) in any of these genomes. RSEARCH finds a statistically significant hit in each of the four genomes, uncorrected for multiple testing (Figure 3.6). (Two of the hits remain statistically significant after such a correction.) I adjusted the structures of all four sequences. The resultant structures all have the conserved elements of eukaryotic RNase P (Figure 3.6). The conserved regions are all in the right place, and the pseudoknot between the single-stranded region at CR I and CR V can form as previously described (Pitulle et al., 1998). The most variation was observed in helix P12, which was always unalignable. This helix is known to be the most variable in vertebrate RNase P; helix P12 sequences from different vertebrates are often unalignable (Pitulle et al., 1998). The other common variations were in P19 and P3, also known to be variable helices (Pitulle et al., 1998). A BLASTN search of GenBank revealed that two of the sequences—those from *C. elegans* and *D. melanogaster*—were already deposited by Steve Marquez and Norm Pace in GenBank, but not published, as RNase P from those two organisms (AF434764.1 and AF434763.1, respectively). This further supports the hypothesis that these sequences represent the true RNase P sequences.

**Timings**

As mentioned previously, the time complexity of the scanning algorithm in RSEARCH is $O((M - B)LD + BLD^2)$. $D$ is set to be $2M$ by default, and I assume that in the unrealistic worst case, every position in the query structure represents a bifurcation. Then,
Figure 3.6: Secondary structure of other eukaryotic RNase P sequences found with RSEARCH. Boldface and light gray letters have the same meaning as in Figure 3.5b.  

- **a** A. gambiae ($E = 0.00046$)  
- **b** C. intestinalis ($E = 1.3 \times 10^{-6}$)  
- **c** D. melanogaster ($E = 0.044$)  
- **d** E. cuniculi ($E = 0.046$)
the worst-case running time of the scanning algorithm is $O(NM^3)$, for a query of length $M$ and database of length $N$, though actual running time will be less based on the number of bifurcations. Calculation of the statistics, which is $O(M^4)$, takes an additional amount of time. Therefore, for a large database where $M \ll N$, the algorithm scales linearly with the size of the database but as the cube of the length of the query sequence. It takes 2.9 CPU days to search *E. coli* SRP (113 nt) against the $2.1 \times 10^7$ nucleotide Archaeal database. In contrast, the *P. furiosus* RNase P sequence (330nt) requires 38 CPU days to search the same database. These searches would take 26 CPU years and 340 CPU years respectively to search the non-redundant nucleotide database of GenBank ($6.9 \times 10^9$ nucleotides). Actual running times can be reduced by using a large-scale clustered computing facility. Actual running times for the above searches on a parallel cluster are 33 minutes for finding homologs of *E. coli* SRP in the Archaea (128 CPUs), and 7.4 hours for finding homologs of *P. furiosus* RNase P in the Archaea (124 CPUs). Therefore, use of RSEARCH is currently practical only when a large multiprocessor computing facility is available.

**Discussion**

Here I have presented RSEARCH, a program for finding homologs of a single RNA sequence given its secondary structure. RSEARCH extends previous profile SCFG implementations in three ways, each of which contributes to its superior performance over BLAST and SSEARCH (Durbin et al., 1998; Eddy, 1996). First, RSEARCH allows the use of a single sequence as a query, by incorporating a substitution matrix and gap penalties to set the parameters of the covariance model. Second, RSEARCH includes local alignment. Third,
RSEARCH includes empirically derived values for statistical significance. Combined, these improvements make RSEARCH a useful tool for finding homologs of biologically important RNAs.

There are three areas in which future development efforts should be focused to improve RSEARCH’s performance. First, the quality of the substitution matrix influences the performance of the program. Here I built the matrix using only a single class of RNA molecules and chose the best matrix based on only two sample tests. Using additional classes of RNA molecules for both building the matrix and choosing the best default may improve RSEARCH’s performance. Alternative algorithms for clustering and weighting sequences should also be explored. Finally, an exponential family of matrices (like the PAM matrices) rather than an empirical family (like the BLOSUM matrices) may be worth considering as well. The rate matrix of Knudsen and Hein would be useful in this approach (Knudsen and Hein, 1999).

Second, RSEARCH is quite slow. Many searches are infeasible on a single CPU. I worked around this problem by performing searches in parallel using a clustered computing environment. This solution is not ideal due to the resources required for such an environment. Advances in computing technology will gradually make more and more searches practical on a single workstation; a new workstation purchased today is two to three times as fast as the machines used in this paper. More importantly, heuristic improvements to RSEARCH may speed it up significantly, just as BLAST and FASTA are significant speed improvements to the Smith-Waterman algorithm.

Finally, the requirement that the secondary structure of the query sequence is known must be addressed. Even a one base pair misprediction can significantly alter the results
of the search (data not shown). This is not a problem if one is searching for homologs of an RNA sequence whose structure is well established (e.g., tRNA, RNase P, SRP). As RNA secondary structures are established through the sequencing of many homologs and comparative analysis (James et al., 1989), there is less need for a program that can handle a single sequence query rather than a large sequence family in these cases. The power of RSEARCH comes from being able to do searches when we only know of a single member of an RNA sequence family (e.g., novel noncoding RNA genes recently discovered in *E. coli* (Hershberg et al., 2003) and various Archaea (see Chapter 2). In these cases, ideally we would like to be able to accurately predict secondary structure starting only with a single sequence. Recent work shows promise in simultaneously aligning and folding a pair of RNA sequences (Sankoff, 1985; Holmes and Rubin, 2002; Mathews and Turner, 2002; Perriquet et al., 2003). These algorithms predict structure more accurately than single-sequence RNA folding algorithms. Many RNA gene-finding approaches take advantage of comparative data. Close homologs of novel RNAs can often be found by primary sequence search programs. These homologs can then be used in a pairwise RNA folder to get a structure for the query sequence. Improvements in such algorithms and an understanding of how best to predict the folding of a query sequence for RSEARCH should allow us to use RSEARCH to find homologs of these novel RNAs.
Chapter 4

Homologs of small RNAs from \textit{E. coli} \textsuperscript{1}

\textsuperscript{1}Ziva Misulovin performed the RACE reactions described in this chapter under my supervision
Abstract

To date, at least 55 noncoding RNAs have been found in E. coli. The function and phylogenetic distribution of most of these ncRNAs remains unclear. To determine the boundaries of the QRNA-predicted ncRNA transcripts, I used 5’ and 3’ RACE. I used the program RSEARCH to locate several additional homologs of various E. coli ncRNAs in other gamma proteobacteria, given the E. coli ncRNA and a structure provided by the program DYNALIGN.

Introduction

As discussed in Chapter 2, it has been estimated that there are approximately 200 small, structural ncRNA genes (excluding tRNAs) in the 4 megabase genome of E. coli (Rivas et al., 2001). There are currently 55 small RNAs for which at least some evidence of expression in E. coli exists (Hershberg et al., 2003). Of these, 10 are well characterized RNAs that were found in individually (Wassarman et al., 1999). Of the remaining 45 ncRNAs, expression of 36 has been demonstrated on Northern blots (Wassarman et al., 2001; Argaman et al., 2001; Rivas et al., 2001; Chen et al., 2002). This is in stark contrast to the five to ten small ncRNA genes found in the approximately 2 megabase genomes M. jannaschii and P. furiosus (Chapter 2). While it is unclear why this is the case, understanding the function of the RNAs from E. coli may shed light onto this issue.

A critical starting point for functional analysis is to have a good multiple alignment and consensus secondary structure. Such a structure provides a basis for experimental analysis. Furthermore, knowing the phylogenetic range of a noncoding RNA may help predict its
function. Homologs of these 46 *E. coli* RNAs have been systematically found using BLAST (Hershberg et al., 2003). For many of these RNAs, homologs were not found outside of the enterobacteria. With RSEARCH, it may be possible to find additional homologs and more accurately define the phylogenetic range of these RNAs. Additional homologs will also allow improved multiple alignments and secondary structure predictions.

From Chapter 3, it is clear that RSEARCH can find homology relationships invisible to primary sequence search programs if the correct secondary structure is known. As discussed in the introduction, the most accurate secondary structure predictions are derived from a large number of homologous sequences. The secondary structures for tRNA, SRP, and RNase P used in Chapter 3 come from just such a collection. The single-sequence model used in RSEARCH is therefore not strictly necessary in these cases; similar success could probably be found using a profile stochastic context free grammar with good priors.

A tool such as RSEARCH is essential when seeking homologs of an RNA for which very few sequences are available. For such cases, the question of how to determine the secondary structure immediately arises. With only a single sequence, either the biochemical approach (Moine et al., 1998) or the computational optimization approach, such as that implemented in MFOLD (Zuker et al., 1999), can be used. If instead of a single sequence, we can find at least one additional homolog using a primary sequence search tool, then we can attempt to fold the two sequences simultaneously to get a better prediction (Mathews and Turner, 2002; Perriquet et al., 2003). Pairwise folding may be especially appropriate for this application as it is the conserved stems that must be properly predicted so that RSEARCH can find them in additional sequences. All of these folding techniques involve some level of error. The questions are which of these techniques minimizes that error, and how does that error
translate into the sensitivity of a search with RSEARCH.

In this chapter, I further characterize the small RNAs of *E. coli* through both RACE analysis and homology searching for RSEARCH. In the process, I test various computational pairwise RNA folding methods to see what method produces the best RSEARCH query. The additional homologs found give more insight into the secondary structure of these RNAs and their phylogenetic range.

**Materials and Methods**

**RACE reactions**

Total RNA was isolated from log-phase cultures of *E. coli* grown on LB medium at 37 °C as previously described (Rivas et al., 2001). The RNA was treated with DNase, polyadenylated, and reverse transcribed to make cDNA for RACE reactions as in Chapter 2. RACE reactions and sequencing was performed as in Chapter 2, using primers specific for the *E. coli* genes. The oligonucleotides used are listed in Table 4.1.

<table>
<thead>
<tr>
<th>Cand. no.</th>
<th>GSP1 (5' end)</th>
<th>GSP2 (3' end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tpke11</td>
<td>AAATTCCCCCTTCGCCCCGTGTCAAGTA</td>
<td>TACTGACACGCGGAGGGAATTT</td>
</tr>
<tr>
<td>tp2</td>
<td>ATATAAAAAGCGGCAACTAAGCTTAG</td>
<td>AGACAGGTCTCGGTTAGGTCGG</td>
</tr>
<tr>
<td>t44</td>
<td>AAAGTGAATGTCGCTTCCGTCA</td>
<td>TGGACCGGAGCGACATCTCATT</td>
</tr>
<tr>
<td>tpe7</td>
<td>ACCTCTCAGTGGCTAGCTCAGTA</td>
<td>CTTCAACAGCTAAGGCCATAGAT</td>
</tr>
<tr>
<td>tpke70</td>
<td>GCCCCAGGTAATAACAGGGCTTTT</td>
<td>AAAAGCCCTGTATTACCCGTCG</td>
</tr>
<tr>
<td>tp1</td>
<td>CACCAGTGCCGATATTGCCGCCGATAGC</td>
<td>CGCTTGCTATCGGCAATATACGG</td>
</tr>
<tr>
<td>tke1</td>
<td>ATATCAAAGTTGACCGCAAGGCAC</td>
<td>TGCTTGCTATCGGCAATATACGG</td>
</tr>
<tr>
<td>tp8</td>
<td>CCGAGCGTTATTAACAGCAATCAGA</td>
<td>TCTGATTGCTGTTAATAGC</td>
</tr>
<tr>
<td>tpk1</td>
<td>TGCCCGGTGACCGGATGTTTTTTT</td>
<td>AAAAAACATCGCGGTACCCG</td>
</tr>
<tr>
<td>k4</td>
<td>CTTTTTCCGATGTACATCGCCCTTTCC</td>
<td>GGAAAGCGGCTACATGCAAAAGG</td>
</tr>
<tr>
<td>tpk2</td>
<td>ACACACCGCTTGTTCGTCGAGGCG</td>
<td>CGGCTAAAGGAAAAACAGG</td>
</tr>
</tbody>
</table>

Table 4.1: Oligonucleotides for RACE reactions. The GSP1 oligonucleotides are used to find the 5’ end of an RNA. The GSP2 oligonucleotides are used to find the 3’ end.
**Programs used**

For BLAST searches, WU-BLASTN version 2.0 was used with options -W3 -kap. MFOLD version 3.1 with default parameters was used for single-sequence RNA foldings (Zuker et al., 1999; Mathews et al., 1999). The program DYNALIGN was used for pairwise foldings (Mathews and Turner, 2002). Unless otherwise stated, the parameters used were $\Delta G_{gap}^0 = 1.8$, $M = 10$, and single-nucleotide insertions allowed. The program CARNAC was also used for pairwise foldings, using the default parameters (Perriquet et al., 2003). RSEARCH version 1.1 was used for all experiments described here.

**Determining best folding method**

RNase P sequences and structures were taken from the RNase P database (Brown, 1999) downloaded on March 13, 2002.

The sequences for tRNAs RD0260 and RD0500 came with the DYNALIGN distribution (http://rna.chem.rochester.edu/dynalign.html; Mathews and Turner, 2002). These are the same as GenBank AJ316554.1 and K00170.1, respectively. Other tRNAs sequences for testing the best pairwise identity to use were found by searching all Archaeal genomes with BLAST and finding full-length hits of varying percent identity. The tRNAs chosen are listed in Table 4.2.

To test the pairwise foldings of the various RNAs used to test RSEARCH in Chapter 3, a BLASTN search (-W3 -kap) was first performed to find the appropriate RNA. For the alanine yeast tRNA, this was on Chromosome III, coordinates 123642-123571. For the *A. fulgidus* tRNA, this was from the *Halobacterium NRC-1* genome, coordinates 115152-115224. *M. jannaschii* RNase P was folded with *A. fulgidus* RNase P (genome coordinates
86285-86042). For *P. furiosus* RNase P, *M. thermoautotrophicum* nucleotides 117462-1173863 were used as the homologous sequence.

To simulate a search of GenBank for homologs of *Desulfovibrio desulfuricans* RNase P, each *D. desulfuricans* RNase P structure was searched and scored against the 470 RNase P sequences in the database (Brown, 1999). An E-value for each one was calculated assuming the entirety of the NCBI non-redundant nucleotide database, as downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt.Z on January 20, 2003, was searched, using the algorithm described in Chapter 3.

**Gamma proteobacteria searches**

A database of 12 gamma proteobacteria genomes was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria. The genomes of *Buchnera* sp, *Escherichia coli* K12, *E. coli O157H7*, *E. coli O157H7:EDL933*, *Haemophilus influenzae*, *Pasteurella multocida*, *Pseudomonas aeruginosa*, *Vibrio cholerae*, and *Xylella fastidiosa* were downloaded on June 21, 2001. The genomes of *Salmonella typhi*, *Salmonella typhimurium* LT2, and *Yersinia pestis* were downloaded on December 10, 2001.

To predict the secondary structure of RNAs from *E. coli*, I performed pairwise com-

<table>
<thead>
<tr>
<th>GenBank Accession Number</th>
<th>Coordinates</th>
<th>Species</th>
<th>Percent identical to RD0500</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP000996.2</td>
<td>44823-44895</td>
<td><em>Thermoplasma volcanium</em></td>
<td>60%</td>
</tr>
<tr>
<td>AP000992.2</td>
<td>248961-249033</td>
<td><em>Thermoplasma volcanium</em></td>
<td>69%</td>
</tr>
<tr>
<td>U67528.1</td>
<td>1221-1296</td>
<td><em>Methanocaldococcus janaschii</em></td>
<td>74%</td>
</tr>
<tr>
<td>AE010266.1</td>
<td>2512-2589</td>
<td><em>Pyrococcus furiosus</em></td>
<td>79%</td>
</tr>
<tr>
<td>AE005139.1</td>
<td>121-193</td>
<td><em>Halobacterium NRC-1</em></td>
<td>91%</td>
</tr>
</tbody>
</table>

Table 4.2: List of tRNA sequences used to evaluate the best percent identity between two RNAs to use for pairwise folding.
parisons using DYNALIGN. The comparative sequences I used are given in Table 4.3. In some cases, structure was predicted without comparative evidence using MFOLD. All RSEARCHes of the gamma proteobacteria database were performed using a partition point of 50 for calculating statistics.

Results

RACE Reactions

The methods used to screen for ncRNAs in E. coli do not precisely identify the 5’ and 3’ ends of the RNA molecule (Hershberg et al., 2003). While one could guess at the bounds of the RNA molecule to define a query sequence for homology searching, knowing the biologically accurate bounds presumably would result in more accurate searches. Only one of the E. coli RNA screens present data on the precise ends of the sequences (Argaman et al., 2001). Before proceeding to see if RSEARCH can find additional homologs of these E. coli RNAs, I wished to define precisely the bounds of the 11 RNAs identified and shown to be expressed in the QRNA screen from our lab (Rivas et al., 2001). To do this, we performed 5’- and 3’-RACE reactions similar to those used to find the ends of the Archaeal RNAs in Chapter 2.

The RACE reactions identified the 5’ and 3’ ends of seven out of the 11 RNAs (Table 4.4). In doing so, I discovered that the oligonucleotides used in the Northern blots for tke1 (Rivas et al., 2001) were mislabelled; tke1 is actually on the minus (-) strand of the genome. This result was verified by repeating the Northern blot analysis with the same and new oligonucleotides (data not shown). For one gene, tp2, I was only able to identify
<table>
<thead>
<tr>
<th>Gene</th>
<th>Genome</th>
<th>GenBank Accession</th>
<th>Coordinates</th>
<th>E-value</th>
<th>Percent Identity</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>H. influenzae</td>
<td>L24083.1</td>
<td>906076-906273</td>
<td>0.0034</td>
<td>71</td>
<td>15</td>
</tr>
<tr>
<td>Spot42</td>
<td>V. cholerae</td>
<td>AE003852</td>
<td>106079-106130</td>
<td>1.4e-11</td>
<td>86</td>
<td>10</td>
</tr>
<tr>
<td>csrF</td>
<td>Y. pestis</td>
<td>AY590842</td>
<td>784156-784782</td>
<td>6.8e-05</td>
<td>60</td>
<td>15</td>
</tr>
<tr>
<td>daF</td>
<td>E. coli</td>
<td>AY590843</td>
<td>116578-116573</td>
<td>1.7e-05</td>
<td>96</td>
<td>10</td>
</tr>
<tr>
<td>daA</td>
<td>S. typhi</td>
<td>AY533842</td>
<td>2033254-2033271</td>
<td>1.1e-09</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>oxyN</td>
<td>S. typhi</td>
<td>AY533842</td>
<td>3668216-3668264</td>
<td>1.1e-10</td>
<td>83</td>
<td>15</td>
</tr>
<tr>
<td>sip</td>
<td>P. pseudocoxa</td>
<td>AE004439.1</td>
<td>656280-656393</td>
<td>2.4e-06</td>
<td>75</td>
<td>10</td>
</tr>
<tr>
<td>10841</td>
<td>MFOIDL</td>
<td>IS063/tke8</td>
<td>S. typhimurium</td>
<td>AY590646</td>
<td>1745788-1745872</td>
<td>4.70E-10</td>
</tr>
<tr>
<td>10992</td>
<td>MFOIDL</td>
<td>IS128</td>
<td>S. typhi</td>
<td>AY590646</td>
<td>1933951-1934359</td>
<td>6.90E-30</td>
</tr>
<tr>
<td>10814</td>
<td>MFOIDL</td>
<td>IS145</td>
<td>V. cholera</td>
<td>AE003853</td>
<td>784158-7843955</td>
<td>1.6e-14</td>
</tr>
<tr>
<td>10867</td>
<td>IS062</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>106078-106183</td>
<td>1.1e-11</td>
<td>86</td>
</tr>
<tr>
<td>11592</td>
<td>MFOLD</td>
<td>csrB</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>1177026-1177357</td>
<td>2.7e-09</td>
</tr>
<tr>
<td>11602</td>
<td>MFOIDL</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>1177026-1177357</td>
<td>2.7e-09</td>
<td>69</td>
</tr>
<tr>
<td>11612</td>
<td>MFOIDL</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>2323864-2323955</td>
<td>1.6e-14</td>
<td>97</td>
</tr>
<tr>
<td>11628</td>
<td>MFOLD</td>
<td>oxyS</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>3668216-3668264</td>
<td>1.1e-10</td>
</tr>
<tr>
<td>11638</td>
<td>MFOLD</td>
<td>csrB</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>1177026-1177357</td>
<td>2.7e-09</td>
</tr>
<tr>
<td>11647</td>
<td>MFOLD</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>2323864-2323955</td>
<td>1.6e-14</td>
<td>97</td>
</tr>
<tr>
<td>11657</td>
<td>MFOLD</td>
<td>oxyS</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>3668216-3668264</td>
<td>1.1e-10</td>
</tr>
<tr>
<td>11667</td>
<td>MFOLD</td>
<td>csrB</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>1177026-1177357</td>
<td>2.7e-09</td>
</tr>
<tr>
<td>11677</td>
<td>MFOLD</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>2323864-2323955</td>
<td>1.6e-14</td>
<td>97</td>
</tr>
<tr>
<td>11687</td>
<td>MFOLD</td>
<td>oxyS</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>3668216-3668264</td>
<td>1.1e-10</td>
</tr>
<tr>
<td>11697</td>
<td>MFOLD</td>
<td>csrB</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>1177026-1177357</td>
<td>2.7e-09</td>
</tr>
<tr>
<td>11707</td>
<td>MFOLD</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>2323864-2323955</td>
<td>1.6e-14</td>
<td>97</td>
</tr>
<tr>
<td>11717</td>
<td>MFOLD</td>
<td>oxyS</td>
<td>Y. pestis</td>
<td>AE003853</td>
<td>3668216-3668264</td>
<td>1.1e-10</td>
</tr>
</tbody>
</table>

Table 4.3: Homologs of the E. coli RNAs used in folding with DYNALIGN. The percent identity and E-values reported are for a BLAST search of the database of gamma proteobacteria genomes with the given gene as the query. M is the M parameter in DYNALIGN. When more than one homolog is listed for a given gene, the first sequence listed was used for all experiments unless otherwise stated. MFOIDL was used in cases where there was no BLAST hit, the predicted gene region was too long for efficient folding with DYNALIGN, or any hits were subjectively too similar for DYNALIGN to provide an advantage over MFOIDL.
the 5’ end; all of the attempts to sequence the 3’ end failed. Furthermore, the heptamer AACCGTG was observed between the 5’ vector sequence and sequence that matches the tp2 genomic locus. For the remaining three genes – tpke11, tpke70, and k4 – I was unable to successfully sequence either end of the gene. RACE results for tpk1 (sraF) and tpk2 (sraK) had been published previously (Argaman et al., 2001). My results exactly matched those for the 5’ ends of tpk1 and tpk2. The 3’ ends I determined for both tpk1 and tpk2 are upstream of the end reported by Argaman et al. For both genes, I was only able to get one usable 3’ sequence, and the 3’ ends reported by Argaman et al. yield a full length closer to that observed on the Northern blots. Therefore, I believe the 3’ ends previously reported (Argaman et al., 2001) to be correct.

<table>
<thead>
<tr>
<th>Gene</th>
<th>5’ end</th>
<th>3’ end</th>
<th>Size from RACE</th>
<th>Size on northerns (Rivas et al., 2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tpke11</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
<td>370</td>
</tr>
<tr>
<td>tp2</td>
<td>122939</td>
<td>N.D.</td>
<td>N.D.</td>
<td>60; 120</td>
</tr>
<tr>
<td>t44</td>
<td>189712-3</td>
<td>189836-47</td>
<td>124-136</td>
<td>135</td>
</tr>
<tr>
<td>tpe7</td>
<td>1762792-3</td>
<td>1762726-37</td>
<td>56-68</td>
<td>67</td>
</tr>
<tr>
<td>tpke70</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
<td>40</td>
</tr>
<tr>
<td>tp1</td>
<td>2651875-7</td>
<td>2652175-7</td>
<td>299-303</td>
<td>300</td>
</tr>
<tr>
<td>tke1</td>
<td>2689360</td>
<td>2689212-2689215</td>
<td>146-49</td>
<td>150; 180</td>
</tr>
<tr>
<td>tp8</td>
<td>3192881</td>
<td>3192737; 3192772</td>
<td>110; 145</td>
<td>110; 140</td>
</tr>
<tr>
<td>tpk1</td>
<td>3236015</td>
<td>3236161(*)</td>
<td>146</td>
<td>120; 180</td>
</tr>
<tr>
<td>k4</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
<td>200</td>
</tr>
<tr>
<td>tpk2</td>
<td>4048616</td>
<td>4048816(*)</td>
<td>200(*)</td>
<td>250</td>
</tr>
</tbody>
</table>

Table 4.4: 5’ and 3’ coordinates of the E. coli RNAs as determined by RACE. All coordinates are for version M52 of the E. coli K12 genome. The sizes observed on the Northern blots is taken directly from Rivas et al., 2001. (*) The 3’ ends for for tpk1 and tpk2 are probably inaccurate. Argaman et al., 2001 report the 3’ end for tpk1 as 3236203 and the 3’ end as for tpk2 as 4048860. These give lengths of 189 and 245, respectively, which more closely match the lengths observed on the Northern blots.
Best folding method

In order to use these *E. coli* ncRNAs as RSEARCH queries, the secondary structure needs to be known. For ten of the genes, secondary structures have been predicted based on a variety of methods (Wassarman et al., 1999). In many cases, BLAST only finds a small number of homologs of these genes. Therefore, a secondary structure prediction method that requires at most several homologs needs to be found. A program that folds a single sequence based on thermodynamic parameters, such as MFOLD (Zuker et al., 1999; Mathews et al., 1999), may do a good enough job. It is often possible to find relatively close homologs of a noncoding RNA using primary sequence search programs. Therefore, I also tried programs that fold two related RNA sequences simultaneously – DYNAALIGN (Mathews and Turner, 2002) and CARNAC (Perriquet et al., 2003).

MFOLD and CARNAC come with default parameters, which I did not modify. DYNAALIGN does not come with default parameters; in fact, the initial description of the program notes that different parameters are ideal for folding different RNA molecules (Mathews and Turner, 2002). The three parameters that can be varied are $\Delta G^0_{\text{gap}}$, the gap penalty expressed as a free-energy penalty; $M$, the maximum offset between an aligned position in two sequences; and whether or not single-nucleotide gaps are allowed. If a nucleotide is at position $i$ in sequence A, it can only be aligned with nucleotide $j$ in sequence B if $(i - M) < j < (i + M)$. This parameter allows DYNAALIGN to run much quicker than the full dynamic programming algorithm. Because DYNAALIGN does not come with default parameters, I tried a variety of parameters with DYNAALIGN.
Choosing the optimal method

To choose the optimal folding method, I used two test RNAs whose “correct” secondary structure is known. For each RNA, I folded it with a variety of methods and parameters and asked how well it performed in a search for homologous RNAs. I chose two RNAs as my test RNAs: RNase P from *D. desulfuricans* compared to RNase P from *G. sulfurreducens*, and tRNA RD0500 from *Halobacterium volcanii* compared to phage T5 tRNA RD0260. For the RNase P RNA, I chose these two RNAs because they were used for evaluation in the papers that presented CARNAC and DYNALIGN, respectively (Perriquet et al., 2003; Mathews and Turner, 2002).

I asked what percentage of RNase P sequences from the RNase P database (Brown, 1999) would I find with an E-value less than 0.05 in a simulated search of GenBank (Figure 4.1a). DYNALIGN with a large gap penalty performs very well, whether or not single nucleotide insertions are allowed. CARNAC appears to outperform DYNALIGN slightly. One of the MFOLD structures performed almost as well as the pairwise foldings, though many MFOLD foldings performed much worse. The CARNAC folding and best DYNALIGN foldings come very close to matching the performance of the trusted RNase P structure. For each folding of tRNA RD0500, I searched the database of Archaeal genomes used in previous chapters, and asked what minimum error rate was observed (Figure 4.1b). DYNALIGN with a high $\Delta G_{gap}^0$ clearly outperforms the other folding methods and performs as well as searching with the true tRNA structure. Based on these data, I decided to use the parameter $\Delta G_{gap}^0 = 1.8$ with single nucleotide insertions allowed for future DYNALIGN foldings.
Figure 4.1: Evaluation of various folding methods. **a.** Percentage of RNase Ps found in simulated search of GenBank using RSEARCH and *D. desulfuricans* RNase P as query. **b.** Minimum error rate in search for tRNAs in Archaeal genomes using various methods to fold RD0500 with RD0260. Note that in panel a, a higher value on the y-axis means better performance, while in panel b a lower value on the y-axis means better performance.
Optimal percent identity

To examine the effect of the choice of the second sequence for the pairwise folding algorithms, I identified a series of Archaeal tRNAs that vary in their percent identity to RD0500 over the full length of the RNA (Table 4.2). I folded RD0500 using both DYNALIGN and CARNAC compared to each of these tRNAs. I repeated the search of Archaeal genomes with each structure, and assessed performance using the minimum error rate. For CARNAC, foldings with a second RNA that is approximately 70-80% identical results in reasonable performance (Figure 4.2). The DYNALIGN structures folded with $\Delta G_{\text{gap}}^0 = 1.8$ and single nucleotide insertions allowed match or even beat the performance of the real structure when the percent identity between the two RNAs is approximately 65-75% (Figure 4.2). Similar performance is seen with other high values for $\Delta G_{\text{gap}}^0$; low values for $\Delta G_{\text{gap}}^0$ result in a high minimum error rate (data not shown). All DYNALIGN folds made with a tRNA that was less than 65% identical or more than 75% identical had a minimum error rate of at least 200 (data not shown).

Further tests

To get a sense for how RSEARCH would perform if the correct secondary structure is not known a priori, I repeated some of the searches from Chapter 3 using predicted secondary structures for the query. I folded each molecule with both CARNAC and DYNALIGN, and determined the minimum error rate (Table 4.5). The correct structure generally performs best in the searches. Given a choice between DYNALIGN and CARNAC, DYNALIGN usually performs much better but on one occasion CARNAC outperformed DYNALIGN. Based on these data, it is clear that DYNALIGN can be used to predict the structure for
an RSEARCH query when the “true” secondary structure is not known.

New homologs of *E. coli* RNAs

I next used this tandem DYNALIGN/RSEARCH approach to search for additional homologs of the *E. coli* small RNAs in other gamma proteobacteria genomes. Homologs for each RNA were found using BLAST (Table 4.3). DYNALIGN was then used to fold these RNAs. In some cases, MFOLD was used instead (Table 4.3). These folded RNAs were then used as queries in RSEARCH searches against my database of gamma proteobacteria genomes. I wished to see if I could find biologically relevant homologs that the previous search missed (Hershberg et al., 2003). As I was using a smaller database than previously, I repeated the BLAST searches to make sure that any new significant hits found were not the result of a smaller database size.

BLAST easily found homologs of tmRNA and RNase P in each genome; these genes were removed from further consideration. While most genes tested have clear homologs in the *Salmonella*, only about half have an observable homolog in *Y. pestis* and almost no homologs can be found in any additional genomes (Figure 4.3). In only six cases were

<table>
<thead>
<tr>
<th>Query</th>
<th>Target</th>
<th>Minimum Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Real structure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DYNAALIGN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CARNAC</td>
</tr>
<tr>
<td>Yeast Ala tRNA</td>
<td>Yeast genome</td>
<td>50</td>
</tr>
<tr>
<td><em>A. fulgidus</em> Asn tRNA</td>
<td>Archaeal genomes</td>
<td>66</td>
</tr>
<tr>
<td><em>M. jannaschii</em> RNase P</td>
<td>Archaeal genomes</td>
<td>2</td>
</tr>
<tr>
<td><em>P. furiosus</em> RNase P</td>
<td>Archaeal genomes</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.5: RSEARCH searches performed using the real structure and various predicted structures. The data for the searches with the real structure was taken directly from Chapter 3.
homologs found outside of the enterobacteria (*E. coli*, *Salmonella*, *Y. pestis*, and *Buchnera* in this data set).

**Proposed secondary structure for t44 and tp1**

From the DYNALIGN predicted structure and its RSEARCH alignment to other homologous sequences, it is possible to construct a multiple sequence alignment and present a consensus secondary structure. Such an alignment may give phylogenetic evidence for stems when base pairs are conserved even though individual nucleotides are not. I built such alignments for several of the RNAs whose ends I mapped with RACE.

The sequence of tp1 is approximately 300 nucleotides long (Table 4.4). Homologs were found in *E. coli* O157:H7 and the two *Salmonella* species; neither BLAST nor RSEARCH finds a homolog that the other one does not (Figure 4.3). The structure predicted by DYNALIGN is well-conserved among the four sequences (Figure 4.4a). The structure consists of seven stems (Figure 4.4b). Of the seven stems, three have one covarying basepair and one has eight covarying positions (Figure 4.4b).

The sequence of t44 is 135 nucleotides long in *E. coli* K12. (Table 4.4) DYNALIGN predicted a secondary structure with two stems in it. There are homologs to this structure in *E. coli* O157:H7, the two *Salmonella* genomes, *Y. pestis*, *V. cholera*, and *P. aeruginosa*. The *Y. pestis* and *P. aeruginosa* homologs were only found with RSEARCH (Figure 4.3). Upon closer examination of the RSEARCH alignments, it became apparent that only the second stem is conserved among all these species; the alignments to the first stem were not convincing. If I then search the gamma proteobacteria genome database with t44 containing
only this second stem, the same hits are significantly found as well as a hit in \textit{H. influenzae}.

Outside of the \textit{Salmonella} sequences, the region of similarity is only to the 3’ end of the sequence which contains this stem. This stem is well-supported by covariation analysis (Figure 4.5). It has been proposed that t44 might be an attenuator since it is found just upstream of, and in the same orientation as, an annotated protein coding gene (Rivas et al., 2001). This hypothesis cannot be discounted as all copies of t44 identified here are in the same relative orientation and location to \textit{rpsB} (30S ribosomal protein S2) and its homologs.

\section*{Discussion}

In this chapter, I further studied noncoding RNAs from \textit{E. coli}. I mapped the ends of several genes first identified in a QRNA screen (Rivas et al., 2001) and found homologs of these genes with RSEARCH. In doing so, I demonstrated that a program that simultaneously folds two RNA sequences can be used to produce a query structure for RSEARCH.

Using RSEARCH with DYNALIGN-predicted structures allowed many homology relationships to be discerned. Overall, this approach found several more homologs than BLAST did, though there were cases where BLAST found homologs that RSEARCH missed. Several reasons could explain this finding. First, not all of the RNAs may have a secondary structure. Many of the screens used to find the noncoding RNAs in \textit{E. coli} looked for conservation and transcription initiation and termination signals rather than any features associated with secondary structure (Argaman et al., 2001; Wassarman et al., 2001; Chen et al., 2002). Such RNAs may not have secondary structure, and thus RSEARCH would give no advantage over BLAST. Second, the folding method used may be too inaccurate.
This suggests a need for improved computational predictions of RNA folding. Third, use of DYNALIGN required me to choose to which homolog I would compare the *E. coli* K12 RNA. In many cases, I was forced to choose between a more similar (and less informative) full-length hit or a less similar (and more informative) partial length hit. As DYNALIGN requires the full length of both RNAs, I was forced to choose the less informative full-length hit. An algorithm that can accurately fold a single RNA sequence by comparison to several homologous sequences that may or may not be full-length would be ideal.

The results of these searches in the gamma proteobacteria were surprising in that many of these RNAs seem to be found in only a subset of the gamma proteobacteria. One possible explanation for this is that the RNAs are too short for homologs to be found. Consistent with this hypothesis, tmRNA and RNase P, which are both over 300nt long, were found in all gamma proteobacteria, as well as other bacterial genomes, by using BLAST (This chapter; Hershberg et al., 2003). The other two RNAs found in the most gamma proteobacteria – 6S and SRP – are not among the longest RNAs tested (183 and 138 nucleotides, respectively). Therefore, for at least some *E. coli* noncoding RNAs, it is possible to find homologs throughout the gamma proteobacteria. These RNAs may be less rapidly evolving than the RNAs for which fewer homologs were found. Subjectively, it appears that RNAs with homologs outside of the enterobacteria tend to be more similar to their *S. typhi* homologs. If this is the case, then it implies that our ability to find distant homologs of noncoding RNA genes is still limited by current methods.

Alternatively, many of these RNAs may only be found in *E. coli* and its closer relatives. Their function may be non-existent or performed by proteins in other organisms. Since the vast majority of RNAs are only found in the enterobacteria, it may be that the enterobac-
teria in general, and *Escherechia* and *Salmonella* in particular, have made extensive use of noncoding RNA mechanisms in their evolution. *Y. pestis* would make a good organism in which to investigate this possibility further. Several well-studied RNAs from *E. coli* (*dsrA, micF, oxyS, dicF*) as well as about half of the recently found RNAs do not have obvious homologs in *Y. pestis*. Approaches similar to those taken to find RNAs in *E. coli* could be used to screen the *Y. pestis* genome. Some of the *E. coli* RNAs may show significant similarity to a subset of *Y. pestis* sequences selected for encoding small RNAs even if this similarity is insignificant in the context of a set of 12 bacterial genomes. Furthermore, experimental approaches could be used to determine the molecular basis of the pathways homologous to those in which the well-studied RNAs from *E. coli* can be found. Such experiments would reveal if a dissimilar homolog of these RNAs are in *Y. pestis* or if the role of these RNAs is performed by proteins there. Such experiments should shed light onto whether the number of noncoding RNAs found in *E. coli* is typical or unusual.
Figure 4.2: The minimum error rate in a search for tRNAs in the Archaeal genomes using DYNALIGN ($\Delta G_{gap}^0 = 1.8$, single nucleotide insertions allowed) and CARNAC with tRNAs of different percent identity to RD0500.
Figure 4.3: Summary of homologs of the 46 ncRNAs from *E. coli*. A black box means the given gene was found in the given genome by both BLAST and RSEARCH with an E-value less than 0.001. A gray box with a B means that only BLAST found a homolog of the gene in the genome; a gray box with an R means that only RSEARCH found a homolog of the gene in the genome. At least one significant hit was always observed in *E. coli* K12. Any gene for which a homolog was found in one strain of *E. coli* O157:H7 was also found in the other one.
Figure 4.4: Multiple sequence alignment and secondary structure of tp1.  

a. Multiple sequence alignment of tp1 from *E. coli* K12 and O157:H7, *S. typhi*, and *S. typhimurium*.

The copies from the two O157:H7 strains were identical. An “X” above the column means the proposed base pair does not exist in all sequences; An “*” means covariation is observed in the proposed basepair.  

b. Proposed secondary structure for *E. coli* K12 tp1
Figure 4.5: Multiple sequence alignment and secondary structure of t44.  

a. Multiple sequence alignment of the 3' end of t44. The copies from all *E. coli* strains were identical.

An “X” above the column means the proposed base pair does not exist in all sequences; An “*” means covariation is observed in the proposed basepair.  

b. Proposed secondary structure for *E. coli* K12 t44.
Chapter 5

Homologs of the RNAs from the Archaea
Abstract

Using RSEARCH, I find several homologs of the RNAs previously identified in *M. jannaschii* and *P. furiosus*. Additional homologs of the pseudouridylation guide RNAs *hgcA* and *hgcG* are found, supporting their proposed structure and suggesting a way to test their function experimentally. Many homologs of *hgcC* are found associated with several known transposons in the Archaea, providing further support to the hypothesis that *hgcC* is associated with transposition. The *M. jannaschii* RNA *hgcB* is found to be homologous to a region of upstream of the 16S ribosomal RNA in *Pyrococcus*; the functional significance of this relationship remains unknown.

Introduction

One question that sparked the development of RSEARCH is whether it would be possible to find more homologs of the RNAs I discovered in *M. jannaschii* and *P. furiosus*. In most cases, BLAST was unable to find significant homologs of those RNAs in other, non-*Methanocaldococcus* non-*Pyrococcus* genomes. This raised the question of whether homologs of these noncoding RNA genes can be found throughout the Archaea or whether they might be restricted to certain taxa.

The ability to find homologs also bears indirectly on the question of function. Molecular genetic techniques are powerful tools for dissecting the function of genes for which we know the sequence. Ideally, such an analysis would be performed in an organism in which targeted gene knockout is possible. Such techniques do not exist for either *M. jannaschii* or *P. furiosus*. Among the Archaea, the most developed genetic systems are *Halobacterium*
(Peck et al., 2000) and *Methanococcus maripaludis* (Stathopoulos et al., 2001).

In this chapter, I use RSEARCH to find homologs of the RNAs I found in Chapter 2. For the four presumed pseudouridylation guide RNAs in *P. furiosus*, the previously published secondary structure (Rozhdestvensky et al., 2003) can be used as the query for an RSEARCH search. For the other RNAs, structures will need to be predicted.

**Methods**

**Datasets and programs**

MFOLD, DYNAALIGN, CARNAC, and RSEARCH were used as described in Chapter 4. The database of Archaeal genomes used in this chapter consists of the database first used in Chapter 2 along with five additional genomes. These genomes were downloaded from the NCBI on May 14, 2003 and are: *Methanopyrus kandleri* AV19 (GenBank NC_003551.1), *Methanosarcina mazei* Goe1 (NC_003901.1), *Methanosarcina acetivorans* C2A (NC_003552.1), *Sulfolobus tokodaii* (NC_003106.2), and *Pyrobaculum aerophilum* (NC_003364.1). Unfinished genome sequence of *Methanococcus maripaludis* was downloaded from [http://www.genome.washington.edu/UWGC/methanococcus/Methanococcus.cfm](http://www.genome.washington.edu/UWGC/methanococcus/Methanococcus.cfm) on April 3, 2003. The *M. maripaludis* genome was preprocessed with dust and searched separately from the database of the other Archaeal genomes. All RSEARCH searches of the Archaeal genomes database were done with partition points of 40% and 60% GC for calculating statistics. No partition points were used for the searches of *M. maripaludis*. Hits were considered significant if the E-value was less than 0.05.
INFERNAL searches

For the iterative INFERNAL searches, I used version 0.55 of the INFERNAL package (Eddy, 2002b). I started with a structurally annotated alignment of sequences. I then built a profile SCFG with cmbuild. I searched the Archaeal genome database with cmsearch and a score cutoff of 22.5 bits and a window size of 140. I then retrieved all of the sequences and made a new multiple sequence alignment using cmalign. I repeated the process until the search converged on finding only the exact sequences used to build the model.

Results and Discussion

Homologs of the pseudouridylation guide RNAs

To find homologs of pseudouridylation guide RNAs in the Archaea, I searched the Archaeal genome database with the previously published structures for the putative pseudouridylation guide RNAs from P. furiosus (sR9, hgcE, hgcF, and hgcG) and A. fulgidus (Afu-4, Afu-46, Afu-52, and Afu-190; Tang et al., 2002a; Rozhdestvensky et al., 2003). The A. fulgidus RNAs Afu-190, Afu-46, and Afu-52 only found themselves in the database. The P. furiosus RNAs sR9, hgcE, and hgcF only found themselves and their orthologs in the other two Pyrococcus genomes. The search with Afu-4 found itself, its homolog in the three Pyrococcus genomes (Tang et al., 2002 and Chapter 2), and a new homolog in M. jannaschii (Figure 5.1a). The search with hgcG surprisingly did not find Afu-4, but it did find the same homolog in M. jannaschii.

Closer examination of the alignments between Afu-4, hgcG, and this new hit in M. jannaschii revealed that the region of similarity is restricted to the third stem of Afu-4
and \textit{hgcG}. This is the same region that is similar between Afu-4 and \textit{hgcG} by BLAST (Chapter 2) and is predicted to guide a pseudouridylation modification in 23S ribosomal RNA that is present in eukaryotes and \textit{E. coli} (Rozhdestvensky et al., 2003). This sequence can be folded into a pseudouridylation guide structure similar to Afu-4 and \textit{hgcG}, and also exhibits the K-turn motif (Rozhdestvensky et al., 2003) (Figure 5.1b). The predicted guide sequence can guide pseudouridylation of the homologous uridine residue in \textit{M. jannaschii} 23S ribosomal RNA (Figure 5.1c).

This RNA was not found in the GC content screen of \textit{M. jannaschii} (Chapter 2). The GC content of the RNA is 52\%, which suggests why that screen missed it. If instead of the Viterbi screen, I use posterior decoding and look at the probability each nucleotide is in a GC-rich region, the proposed sequence of the RNA is clearly more GC rich than the surrounding sequence (data not shown).

Using the predicted structure of the \textit{M. jannaschii} homolog of \textit{hgcG}, RSEARCH finds a homolog in \textit{M. maripaludis} (E=0.00036). This structure conserves the guide sequence but does not appear to make a strong K-turn (Figure 5.1d). It is not clear whether this sequence can function as a pseudouridylation guide RNA.

\textit{hgcA} from \textit{M. jannaschii} was also reported to be a pseudouridylation guide RNA. It can be folded into the canonical fold, and contains a K-turn and ACA motif. (Figure 5.2a). It is predicted to guide modification of the 16S rRNA (Figure 5.2b). 5’ of the pseudouridylation guide stem is another stem. This stem does not appear to be a pseudouridylation guide; it has neither the K-turn nor the ACA motif (Figure 5.2a). BLAST finds a homolog of \textit{hgcA} in \textit{M. maripaludis} (E=3.2e-06; Figure 5.2c). This homolog conserves the guide sequence and can also make a K-turn and ACA motif. This guide sequence can guide pseudouridylation
Figure 5.1: Pseudouridylation guide RNA homologous to Afu-4/hgcG in *M. jannaschii*. This was found by searching the Archaeal genome database with both Afu-4 (E=8.0e-05) and hgcG (E=0.0097). a. RSEARCH alignment between Afu-4 and its *M. jannaschii* homolog. b. Secondary structure of the proposed RNA. The guide sequence is in boldface, the K-turn is shaded, and the ACA box is boxed. c. The guide sequence paired with its target in *M. jannaschii* 23S ribosomal RNA. d. Secondary structure of the RNA in *M. maripaludis*. 
of the homologous site in *M. maripaludis* 16S rRNA (Figure 5.2d). Searching the Archaeal genome database with RSEARCH and the proposed structure for *hgcA* does not reveal any new significant hits.

Figure 5.2: *hgcA* is a pseudouridylation guide RNA. a. Secondary structure of *hgcA*. The guide sequence is in boldface, the K-turn is shaded, and the ACA box is boxed. b. The guide sequence of *hgcA* paired with its target in *M. jannaschii* 16S rRNA. c. The secondary structure of the *M. maripaludis* homolog to *hgcA*. d. The *M. maripaludis* guide sequence and its target

**Homologs of the *hgcC* family**

I next searched for homologs of *hgcC* and *hhcB*. These two genes, from *M. jannaschii* and *P. furiosus*, respectively, were found in Chapter 2 to have a similar sequence and a similar size. I folded these two sequences in comparison to each other with my default parameters for DYNALIGN. These parameters were unable to predict a folding for these two sequences, so
I instead used CARNAC. For both \textit{hgcC} and \textit{hhcB}, CARNAC produced a folding consisting of three stem-loop motifs (Figure 5.3a,b).

Using the structures that CARNAC predicted for \textit{hgcC} and \textit{hhcB}, I searched the Archaeal genome database. As expected, both sequences found each other as well as \textit{hhcA} from \textit{M. jannaschii}. Significant homologs ($E < 0.05$) were also found in \textit{P. horikoshii} with \textit{hhcB} as the query and in \textit{Halobacterium} and \textit{S. tokodaii} with \textit{hgcC} as the query. Looking at these alignments revealed that only the outer two stems appear to be well conserved (Figure 5.3c,d). Using manual refinements of the alignment between and structure of these sequences, I made a profile SCFG using INFERNAL (Figure 5.4). I then iteratively searched the Archaeal genome database as described above. The searches converge after 14 rounds to a set of 56 sequences (Table 5.1). Nineteen of these 56 sequences are adjacent to or overlap an annotated protein coding gene that is called a transposase by the PFAM database (Bateman et al., 2002). In 14 of these cases, the transposase family is Transposase 2, the same family to which the transposase adjacent to \textit{hhcB} belongs (Chapter 2). At least nine of these 56 sequences are in or adjacent to a previously described Archaeal insertion sequence (IS) element (Brugger et al., 2002). Of these nine, seven are from the IS605/ISC1316 family (Table 5.1). All of the sequences in the \textit{S. tokodaii} genome described as a Tn-like element or truncated Tn-like element overlap a sequence homologous to \textit{hgcC} (Kawarabayasi et al., 2001).

The IS605 transposon family has a complex structure in which the IS200 transposase is adjacent to another transposase (Mahillon and Chandler, 1998). Intriguingly, IS200 is known to be regulated by two RNA stem-loop motifs located upstream of the transposase gene. One of these acts as a terminator to block read-through transcription from other
Figure 5.3: Structure of *hgcC* and some of its homologs.  

**a.** Secondary structure of *hgcC* predicted by CARNAC. Further analysis suggests that the middle stem is not conserved.  

**b.** Secondary structure of *hhcB* predicted by CARNAC.  

**c.** Secondary structure of a homolog of *hgcC* found in *P. horikoshii* (coordinates 565628-565498). This locus was first found by RSEARCH with *hhcB* as the query (E=0.023).  

**d.** Secondary structure of a homolog of *hgcC* found in *S. tokodaii* (coordinates 2562672-2562555). This locus was first found by RSEARCH with *hgcC* as the query (E=0.00099).
Figure 5.4: Multiple sequence alignment and proposed secondary structure of hgcC and some of its homologs found by RSEARCH. Ph, *P. horikoshii*. Halo, *Halobacterium* sp. pNRC200; St, *S. tokodaii.*
Table 5.1: All homologs of hgcC found by iterative searching with INFERNAL. Overlapping or adjacent annotated proteins called a transposase by PFAM are noted and the relative orientation of the protein (single-line arrow) and the homolog of hgcC (double-line arrow) is also shown. Overlapping previously identified transposable elements are also listed. (Table 2 of Brugger et al., 2002 and Table 3 of Kawarabayasi et al., 2001).
nearby promoters and the other occludes the ribosome-binding site (Beuzon et al., 1999). It is not clear whether the independent RNA transcript I identified is related to this motif, especially as the RNA I found is usually located downstream of the transposase (Table 5.1). It is also unclear whether the Archaeal transposases annotated as being in the IS605/IS1316 family show similarity to the IS200 transposase or the adjacent transposase (Mahillon and Chandler, 1998; Brugger et al., 2002).

**Homologs of the remaining RNAs**

Two RNAs from *M. jannaschii* (*hgcB* and *hgcD*) remain to be tried as queries, as does one RNA from *P. furiosus* (*sscA*). BLAST can find homologs of *sscA* in *P. abyssi* and *P. horikoshii*; BLAST cannot find homologs of the other two RNAs. I used the *P. abyssi* homolog to fold *sscA* using both DYNALIGN and CARNAC. Neither of these structures could find any significant hit in the Archaeal genome database other than the ortholog in the three *Pyrococcus* genomes. Since there is no comparative sequence for *hgcB* and *hgcD*, I folded both of these sequences using MFOLD. None of the *hgcD* structures found any significant hits in the Archaeal genome database.

One of the the MFOLD structures of *hgcB* significantly hit a locus in all three *Pyrococcus* genomes (Figure 5.5a-c). The E-values for the hit are 0.0077 (*P. furiosus*), 0.00040 (*P. abyssi*) and 0.0017 (*P. horikoshii*). The match is local with respect to the *M. jannaschii* structure. Only two stems are involved in the match; the rest of the sequence is not (Figure 5.5a). The *Pyrococcus* sequences are around 72% GC. The hits remain significant upon recalculation of statistics for this GC content (data not shown). Therefore, the proposed homology relationship is not an artifact due to a skewed GC content.
Figure 5.5: Analysis of hgcB from M. jannaschii. a. Secondary structure of hgcB predicted by MFOLD that found homologs in the Pyrococcus genomes. The darker sequence is the section homologous to sequences from Pyrococcus. b. Alignment of hgcB to its homolog in the P. furiosus genome. c. Structure of the homologous region in P. furiosus as predicted by RSEARCH. d. Structure of the homologous region in P. furiosus as pre-ribosomal RNA. Stem B and the 16S processing stem are shown (Kjems and Garrett, 1990). The darker sequence is the portion homologous to hgcB. The two arrows point to the putative cut sites in the bulge-helix-bulge motif (Tang et al., 2002b).
The three hits are to syntenic loci that are just upstream of the small subunit ribosomal RNA gene of *Pyrococcus*. Previous studies have found conserved elements present in most Archaeal ribosomal RNA primary transcripts (Kjems and Garrett, 1990). Two of the most conserved features are a large processing stem with two bulge-helix-bulge motifs on opposite strands and helix B, which is just upstream of the small subunit ribosomal RNA processing stem. Both of these features can be found in the *Pyrococcus* sequence at the ribosomal RNA locus (Figure 5.5d). One of the stems from *hgcB* appears to be homologous to helix B; the homolog to the other stem can be formed in *Pyrococcus* after the processing stem is disrupted by the excision of the 16S sequence (Figure 5.5c). Searching GenBank with the *P. furiosus* homolog to *hhcB* finds the ribosomal RNA operon from *Thermococcus celer* (accession M93359.1). The optimal alignment between the *hhcB* structure and this sequence would not be significant in the context of the Archaeal genomes database (data not shown).

**Conclusions and Future Directions**

In this thesis, I have found a variety of noncoding RNA genes in genomic sequences using several approaches. I found nine novel genes by screening for regions of high GC content in *M. jannaschii*, conserved regions of high GC content in *P. furiosus*, and conserved regions in *P. furiosus* whose pattern of evolution suggests a structured RNA molecule. Expression of small, stable RNAs from these loci were verified on Northern blots, and the ends were mapped using RACE. I developed a program, RSEARCH, to find homologs of a single, structured RNA sequence. This program was used to find homologs of RNase P in various eukaryotic genomes and homologs of a variety of small RNAs from *E. coli* in other gamma
proteobacteria. Finally, I identified additional homologs in the Archaea of some of the genes I had previously discovered.

There are three major future directions in which this research can go. One area of future research is improvement of the algorithms for finding noncoding RNA genes in genomic sequences. A second line of inquiry is to utilize the techniques described here, as well as other techniques, to find noncoding RNA genes in other genomes. Finally, much specific research can be done on each of the RNAs in order to elucidate fully their function.

**Improved algorithms**

Improvements can be made in some of the algorithms described here. Specifically, the improvements are centered around RSEARCH. One area for improvement is the matrices used in RSEARCH. As discussed in Chapter 3, an expanded data set for building the matrices, alternative clustering and weighting procedures, and using an exponential family of matrices (like the PAM amino acid substitution matrices) all may improve RSEARCH's performance. The work of Knudsen and Hein, who developed an RNA instantaneous rate matrix may be especially useful in this regard (Knudsen and Hein, 1999). Heuristic approximations could provide RSEARCH with massive speed increases without much loss of power, just as FASTA and BLAST speed up the Smith-Waterman algorithm without much loss of power (Pearson and Lipman, 1988; Altschul et al., 1990). Finally, improved secondary structure prediction algorithms should improve RSEARCH’s performance greatly when the secondary structure of the query sequence is not well-established. A secondary structure prediction program explicitly designed to fold a single sequence based on both thermodynamic parameters and alignment to homologous sequences where the ends of the homologous sequence may not be
known would be ideal for this situation.

**Searching other genomes**

This dissertation defines an approach for discovering novel noncoding RNA genes. A computational screen is first applied to the genome of interest to generate a list of candidates. Candidates are then tested experimentally through Northern blot analysis. The precise ends of the small, stable RNAs identified in this way are then mapped using RACE. Finally, secondary structures are predicted and RSEARCH is used to find homologs of these RNAs.

The GC content screen is probably not applicable to many other genomes. It only works when there is a large difference between the genomic GC and RNA GC content. Such a difference appears to be consistently present only in AT-rich hyperthermophiles. Other approaches to find noncoding RNA genes *de novo* can be used instead. One such approach is the comparative analysis program QRNA (Rivas and Eddy, 2001). Following the same paradigm outlined above, it has been used to find noncoding RNA genes in *E. coli* (Rivas et al., 2001; Chapter 4) and *S. cerevisiae* (McCutcheon and Eddy, 2003). Such screens in other prokaryotes may be especially informative, as they will help reveal whether the variety of small RNAs in *E. coli* is atypical or the norm.

**Specific function of the Archaea RNAs**

Finally, various experimental techniques can be used to further elucidate the function of the Archaeal RNAs identified in this work. While it appears that five of these RNAs are pseudouridylation guide RNAs, there is no direct evidence in support of this hypothesis.
Confirmation that the predicted target sequences for each of the putative pseudouridylation guide RNAs are actually pseudouridylated is the obvious first experiment. Directly demonstrating that a specific RNA guides a specific modification is more complicated. The ideal evidence is to show the loss of a modification when the putative guide RNA is knocked out (Lowe and Eddy, 1999). Targeted knockout technology does not exist for either \textit{M. jannaschii} or \textit{P. furiosus}. Target knockout technology does exist for \textit{M. maripaludis} (Stathopoulos et al., 2001), and such an approach could be tried on the homologs of \textit{hgcA} and \textit{hgcG} predicted to be in \textit{M. maripaludis}. Several of the C/D box methylation guide RNAs from \textit{P. furiosus} were shown to be able to guide methylation of \textit{Xenopus laevis} ribosomal RNA when injected into \textit{Xenopus} oocytes after appropriate changes were made to the guide sequences (Speckmann et al., 2002). A similar approach may work for the pseudouridylation guide RNAs. Identification of additional pseudouridylation guide RNAs will allow further elucidation and refinement of the secondary structure and conserved motifs of these RNAs in the Archaea.

For \textit{hgcC} and its homologs, functional characterization involves dissecting the relationship between this small RNA and the transposons with which it associates. The first question to be asked is whether or not these homologous sequences are expressed as a small, stable RNA. If they are expressed, mapping their exact ends via RACE would provide useful information as well. The other major question is whether the RNA-coding sequence is part of the transposon or whether the transposon prefers to hop into sequences where this RNA-coding sequence already exists. Answering this question should also provide insight into why the \textit{hgcC} family of RNAs appears to be adjacent to transposases in only some cases. In order to examine this question, an active transposon containing the RNA would
need to be found and studied in vitro. This may be difficult, as the IS200 transposase is known to transpose very infrequently (Mahillon and Chandler, 1998).

For the remaining RNAs, there are no obvious hypotheses about what their function may be. For \textit{hgcB}, expression analysis (\textit{e.g.} Northern blots, RACE) can be performed on the homologous region of the \textit{P. furiosus} pre-ribosomal RNA. Such analysis would reveal whether or not the homologous region is expressed as a small, stable RNA and what its exact ends are. It would also reveal whether this RNA, like the 16SD RNA in \textit{A. fulgidus}, contains sequence from both sides of the intervening 16S rRNA (Tang et al., 2002b). Further functional characterization of this RNA, as well as \textit{hgcD} and \textit{sscA}, will require biochemical analysis. Identifying what molecules in the cell each RNA physically interacts with may provide insight into their function. Similar approaches can also be taken to understand the function of the RNAs found in \textit{E. coli}. 


