ABSTRACT

RNA SECONDARY STRUCTURE DETECTION PROGRAMS WITH AN EMPHASIS ON COVARIANCE MODELS

by
Justin Slotman

RNA secondary structure prediction requires a different approach from traditional alignment methods. Functional RNAs often have their secondary structure better conserved than their primary structure. Covariance models, probabilistic models that utilize stochastic-context-free grammars, are one approach. CMs allow for homology to be detected where purely sequence-based methods would fail. A background on CMs is given, as well as a background of the major classes of non-coding RNAs (ncRNAs). Comparisons are made between some CM-using tools (the Infernal suite and CMfinder) and some other RNA secondary structure tools (CARNAC, miRNAminer, Pfold, Mfold) as well as between Infernal and the primary alignment tool BLAT. CMfinder and Infernal are also compared against each other. RNA secondary structure databases, mainly Rfam and miRBase, are used to provide sequence and alignment data.

RNA SECONDARY STRUCTURE DETECTION PROGRAMS WITH AN EMPHASIS ON COVARIANCE MODELS

by

Justin Slotman

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

January 2009

APPROVAL PAGE

RNA SECONDARY STRUCTURE DETECTION PROGRAMS WITH AN EMPHASIS ON COVARIANCE MODELS

Justin Slotman

Dec. 10, 2008

Jason T. ~~J.~~ L. Wang, Ph.D., Thesis Advisor — Date
Professor, Department of Computer Science, NJIT

Dec. 10, 2008

Dimitrios Theodoratos, Ph.D., Committee Member — Date
Associate Professor, Department of Computer Science, NJIT

Dec. 10. 2008

Guiling Wang, Ph.D., Committee Member — Date
Assistant Professor, Department of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Justin Slotman

Degree: Master of Science

Date: January 2009

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2009

- Bachelor of Art in Biology,
  Rutgers University, Camden, New Jersey, 2006

- Bachelor of Art in Creative Writing,
  The Johns Hopkins University, Baltimore, Maryland, 1997

Major: Bioinformatics

To my family

ACKNOWLEDGMENT

I would like to express my appreciation to Dr. Jason T. L. Wang, who served as my thesis advisor and basically kept me on enough of an even keel that I could finally get this thing finished. Special thanks are given to Drs. Dimitri Theodoratos and Grace Wang for serving on my thesis committee.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Objective

The objective of this thesis is to provide a review of the problem of RNA secondary structure prediction, and then contrast differing methods of RNA secondary structure prediction, with an emphasis on comparing methods that use stochastic context-free grammars (SCFGs) with methods that do not. The first part of the objective will be accomplished by delving into basic RNA science and computer science, followed by a survey of the selected databases and software tools selected for this thesis. The second part of the objective will be accomplished by running a similar data set on a number of software tools, some web-based and some requiring local installation.

1.2 Introduction

The problem of RNA secondary structure prediction involves a few factors. Primary structure does not imply secondary structure, so prediction methods that align nucleotide sequences are not necessarily useful for secondary structure prediction [1]. Secondary structure is also better conserved in an evolutionary sense than primary structure. That is, nucleotide substitutions or deletions will more often than not leave an RNA's secondary structure unaffected. Typical alignment methods such as BLAST can prove unsatisfactory, as in some cases related RNAs can have below 60% to 70% primary sequence identity in common, despite only having tens of millions of years to separate them [2]. Common secondary structures can thus be used to show that two RNAs are related, when an alignment of their sequences would not have made that clear. Different methods from the usual primary sequence alignment methods are therefore called for when attempting secondary structure prediction [3].

One method involves the use of covariance models [1, 4 (pp. 279-299)]. These are probabilistic, mathematical models that describe an RNA's primary and secondary sequence simultaneously. They can be used for secondary structure prediction, multiple sequence alignment (so they can assist with primary and secondary sequence-related problems), and finding similar matches to a target RNA within databases. They are intended to find RNA homology where sequence alignments alone would not work as well.

This thesis will investigate the application of covariance models to RNA homology detection in comparison with more traditional alignment tools and some specialized detection tools. Background information on the biology of RNA and some computer science theory will be discussed first. Then there will be a survey of selected software tools that are being utilized in this thesis—some that use covariance models (CMs), some that do not. Following that will be a brief discussion of the RNA databases that were used in this thesis. Finally there will be some comparisons between the software tools using a similar data set from one of the databases.

CHAPTER 2

BACKGROUND

2.1 RNA Background

2.1.1 History of RNA

RNA is an extremely important biological molecule. Once thought to be a mere messenger molecule for the information contained in DNA, it is now known to be both an information carrier and an enzymatically active molecule. Some theorists believe it is the original biological molecule—this is the so-called “RNA world” theory, which attempts to explain why RNA has informational and enzymatic functions [5]. According to the theory the dual functions of RNA were later taken over by DNA (for information storage) and protein (for catalysis and enzymatic activity.) But, crucially, RNA retains both functions in modern organisms [6].

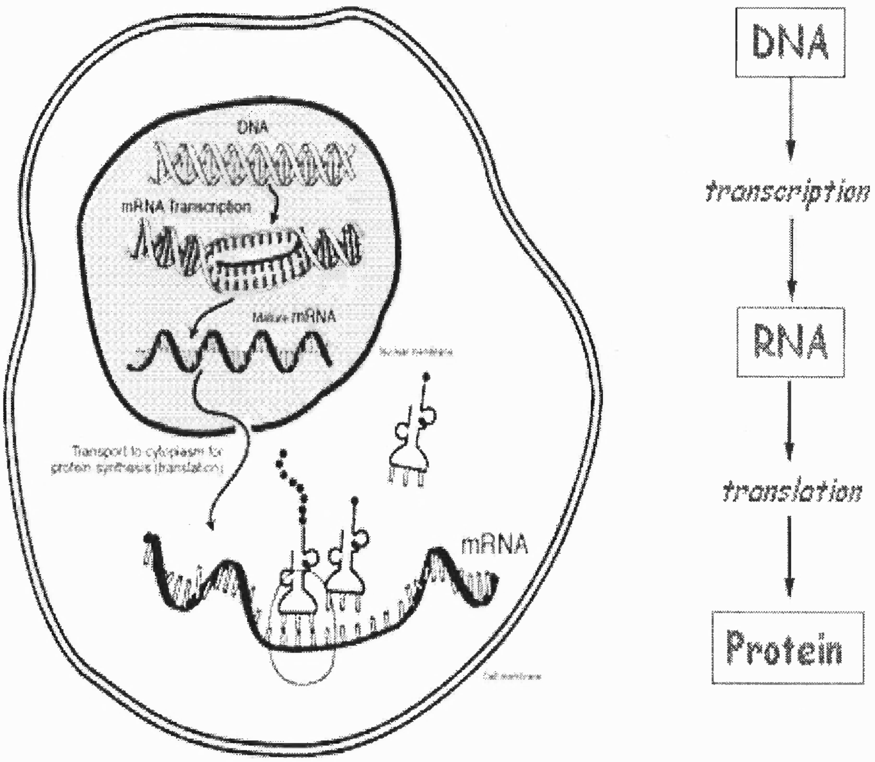


Figure 2.1 The Central Dogma.
http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

RNA is a major part of the central dogma of molecular biology, which says information cannot flow from protein to nucleic acid (or to protein.) Information must go from nucleic acid to nucleic acid or protein [7, 8]. The flow of information in the cell is generally described as a three step process. First, there is replication which is the process of DNA copying itself. Next is transcription, which is the process of a portion of DNA being copied onto a string of RNA (specifically messenger RNA, which will be addressed later.) Finally there is translation, which is the process where the information in RNA is used to create protein. Translation is assisted by transfer RNA and ribosomal RNA (ribosomes being the organelles where individual proteins are constructed [9, ch. II.6].) In terms of the central dogma RNA's roles are limited to those mentioned above:

information transfer agent and translation assistant. The true picture, though, is more complicated than that.

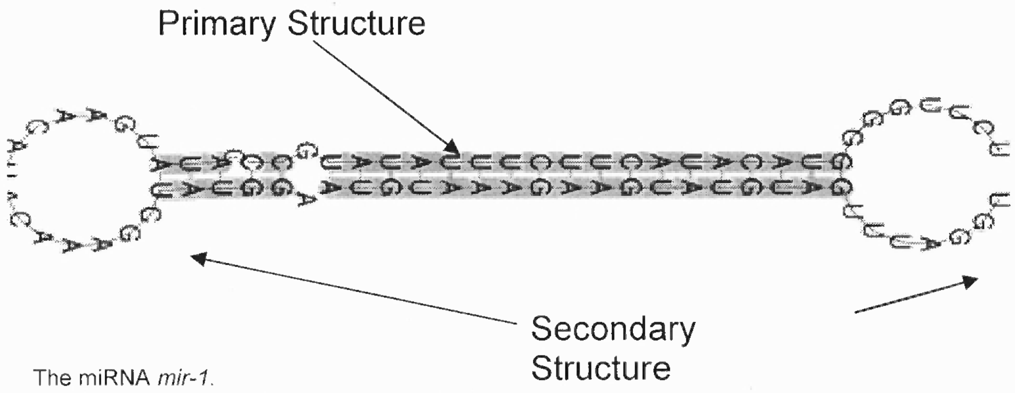
RNA has a much more active role in all facets of cell life than was previously realized, including regulation and gene expression. The emerging science of epigenetics has made this plain [10]. Epigenetics is the study of heritable traits that do not involve a change in DNA sequence, and in that sense represents a challenge to the central dogma. To be fair, the central dogma (as defined by Crick) specifies that information cannot flow from protein to nucleic acid, not that information cannot flow from one nucleic acid (RNA) to another (DNA.) But epigenetics does at least question the primacy of DNA in the central dogma. On the RNA side of things, epigenetics has shown RNA to influence heredity in a few ways: via RNA interference, which selectively hinders gene expression at the transcription and/or translation stages [11]; via the methylation (adding or substituting a methyl group) of certain DNA sites [12, 13]; and by activating and/or degrading some RNAs [14]. The epigenetic abilities of RNA have challenged prevailing notions of RNA's role in the cell.

There is also the case of reverse transcription [15]. This is how RNA viruses copy themselves onto cellular DNA. Using the enzyme reverse transcriptase these viruses created double-stranded DNA from single-stranded RNA, and use other enzymes to insert their genetic material into the host's genome.

2.1.2 RNA Background: Structure

At this point it would be useful to elucidate the classification of RNA structure. Primary structure refers to the sequence of nucleic acid residues that make up the RNA molecule. RNA is composed of four residues: guanine (G), adenine (A), cytosine (C), and uracil (U). Uracil is the residue that distinguishes RNA from DNA; DNA has thymine (T) instead of uracil. It is thought that the major reason DNA took over the bulk of the information-carrying properties across all organisms is thymine, as it is strong protector of sequence information from damage [16, pp. 544]. Primary structure is synonymous with primary sequence.

Secondary structure refers to RNA's tendency to fold into a number of small subunits [9, ch. II.6]. These subunits usually contain regions of helical structure and of loop structure. Helical structure involves Watson-Crick base-pairing—that is, U binding with A and C with G. Loops are, as the name implies, areas of curving into circular or semi-circular shapes. They can perhaps be thought of as the areas of RNA secondary structure that are not base-paired.



The miRNA *mir-1*.
http://rfam.sanger.ac.uk/family?acc=RF00103

Figure 2.2 Primary versus secondary structure.

Typically loops are of four types: hairpin, or stem-loops, which have a single loop on the end of a helical region; internal loops, where base-pairing is interrupted on two strands so a loop is formed between two helical regions; bulge loops, where base-pairing is interrupted on one strand and forms a hairpin-like shape attached to the rest of the RNA via that single strand; and multibranch loops, which can have more than two regions of helical structure attached to them [17, pp. 144-145]. There are also pseudoknots, a type of tertiary structure that will be discussed later (it is problematic for most secondary structure prediction methods.) Variations on the themes of loops and helices are the main way of distinguishing different families of RNAs.

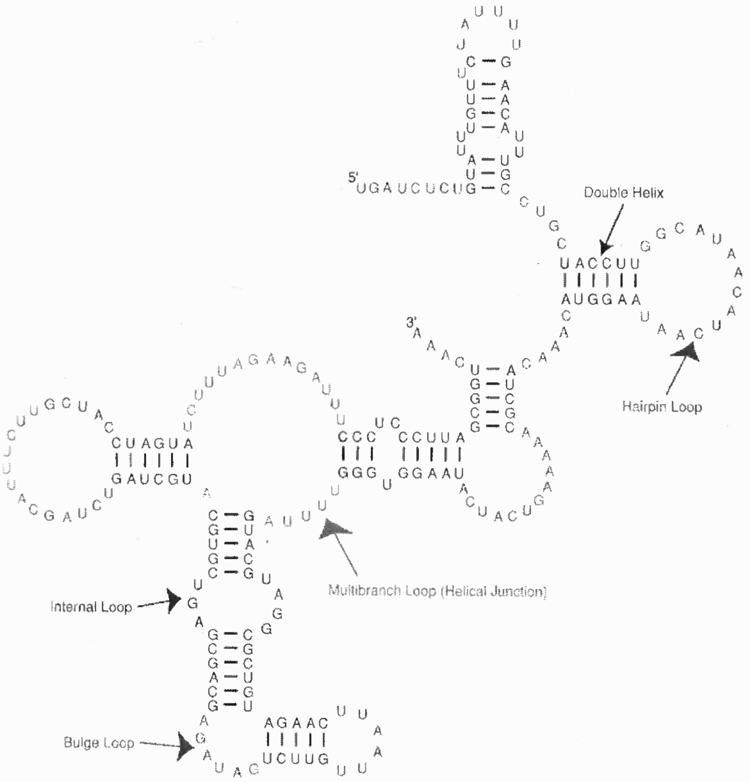


FIGURE 6.2 The RNA secondary structure of the 3′ UTR from the *D. sucinea* R2 element (Lathe & Eickbush, 1997; Mathews et al., 1997). Base pairs in nonhelical regions, known as loops, are colored by type of loop.

Figure 2.3 Types of RNA secondary structure. From page 145 of Baxevanis' *Bioinformatics*, 3rd edition.

2.1.3 RNA Background: Types of RNA

There are quite a few types of RNA. This section attempts to represent the major recognized types of RNA.

Messenger RNA (mRNA): These are the RNAs that function in translation by providing sequence-based blueprints for protein synthesis [9, ch. II.6]. They are formed within the nucleus during transcription. The initial RNA transcript is called the primary transcript [16, pp. 644], or precursor mRNA (pre-mRNA) [9, ch. II.6]. This undergoes processing and cleavage within the nucleus. Pre-mRNA is composed of sections called exons, which are expressed (i.e., used to construct proteins), and introns, which are not expressed. Introns are excised enzymatically. The exons are processed by the addition of a 5' methylguanosine cap and a poly-A tail, molecular markers that will protect the mRNA from degradation outside of the nucleus. The mature mRNA then leaves the nucleus and heads over to the ribosomes to be translated. It is quickly degraded post-translation [18], in part by other RNAs (miRNAs and siRNAs, discussed later.) The secondary structure of the information-carrying portion of mRNA does not have a predictable secondary structure that lends itself to the problem-solving methods discussed in this thesis, and is mentioned in passing. The functional portions of mRNA, the cis-regulatory elements, are generally analyzed separately from the mRNA strands they are attached to, and are discussed in the remainder of this section.

Transfer RNA (tRNA): These are the RNAs which are responsible for the movement of single amino acids (the building blocks of proteins) to the ribosomes. They are small molecules, about 70 to 90 nucleotides long. There are 61 tRNAs for each of the 61 codons. (Codons are the three-nucleotide sequences that specify which of the twenty amino acids are to be used at a given position during protein synthesis. The sequential order of the codons specifies the order in which amino acids should be added to a growing protein.) The genetic code is known to be degenerate—that is, some amino acids are represented by more than one codon. Alanine, for example, is represented by four codons, while typtophan only has one [19, ch. 5.5.1]. Likewise any particular cell can function without 61 distinct tRNAs; in fact a cell can function with only 31 types of tRNA [9, ch. II.6]. This phenomenon, called wobble base-pairing, is thought to explain why many of the alternate codons for an amino acid only differ in their third nucleotide—many tRNAs are constructed such that they only require to provide an accurate match on their first two codon positions. Transfer RNAs are formed in a similar way to mRNAs, with precursor sequences being subject to intron removal via splicing and post-processing alterations. This latter processing is why tRNA often has nucleic acids residues besides the usual A, C, G, and U, such as ribothymidine [19, ch.29.1] and inosine [20, ch. 3.10.3]. 10-15% of tRNA nucleic acid residues are transformed into the nontraditional bases during post-processing [16, pp.646].

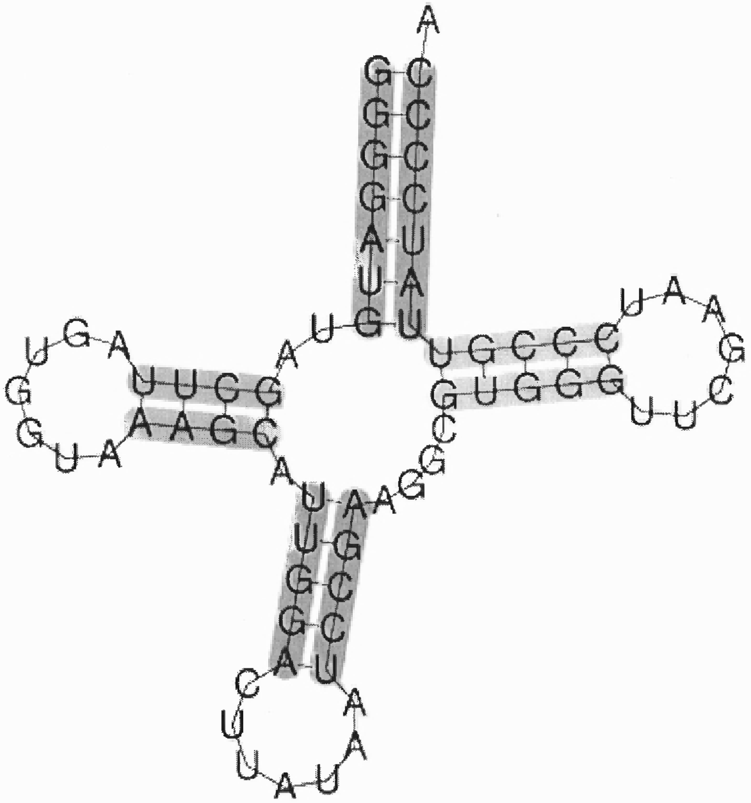


Figure 2.4 Transfer RNA. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00005

Ribosomal RNA (rRNA): Ribosomal RNA is that portion of a ribosome that is RNA; the rest of the ribosome is protein. As ribosomes are the site of protein synthesis and thus are generally found in vast quantities in the cell (in the millions in the typical eukaryotic cell [9, ch. II.6]) vast quantities of rRNA is also required. This is accomplished by having multiple copies of RNA genes on the transcription sites. rRNA also has an entire subcompartment of the nucleus devoted to its production, the nucleolus. The DNA strands that code for rRNA protrudes from the nucleolus, where multiple rRNAs are generated and then processed into ribosomal subunits which are already part RNA and part protein. There is a large subunit and a small subunit which fit together to form the mature ribosome. The active site of the mature ribosome is entirely RNA [9, ch. II.6], another example of RNA's enzymatic abilities.

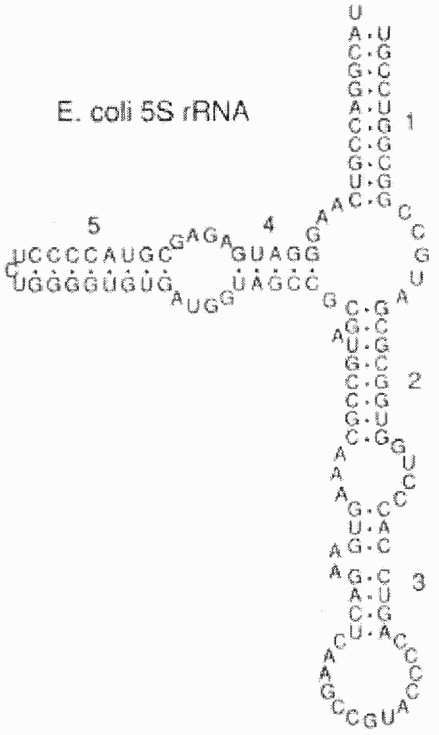


Figure 2.5 5s rRNA. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00001

Non-coding RNAs (ncRNAs): Non-coding RNAs is a broad term for any class of RNA that isn't an information carrier. Essentially, any RNA that is not mRNA is an ncRNA. The term is of more recent origin, though, and came into prominence with the discovery of new classes of RNA outside of the traditional functional RNAs, tRNA and rRNA. The bulk of non-coding RNAs were once considered to be "junk," simple waste byproducts of cell metabolism [21]. But they are part of the ongoing reevaluation of the role of RNA in the cell, and are introduced below.

Ribozymes: Ribozymes are simply RNAs that function like enzymes. Some sources classify the ribosome itself as a ribozyme [9, ch. II.6], as RNA is responsible for both the enzymatic activity and the secondary structure of the ribosomes. Another example would be ribonuclease P, which is involved in tRNA cleavage. Ribonuclease p does have a protein component, but, like with ribosomes, the protein does not appear to play a role in catalysis or seconday structure [22]. There are also examples of precursor RNAs self-cleaving themselves [23] (their introns remove themselves and leave processed RNA strands behind.) Related to ribozymes are the **riboswitches**, small untranslated segments attached to mRNA. An mRNA with a riboswitch has the ability to regulate its own activity (a translation effect) and the gene that coded for it (a transcription effect.) Riboswitches are common in bacteria, though at least one eukaryotic ribtch has been identified (the TPP-binding THI element [24].)

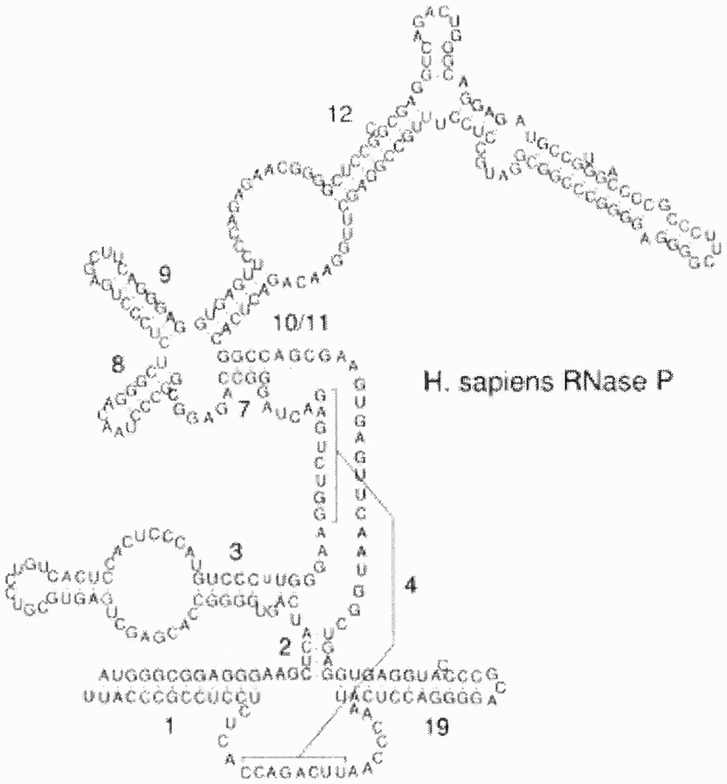


Figure 2.6 Human RNase P. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00009

Cis-regulatory (or cis-acting) elements: In general cis regulation is regulation that has a local effect. Trans regulation is regulation that has a distant effect. In biological terms, a cis-regulatory element is a strand of RNA (or protein, but RNA is the subject of interest here) produced by DNA that is intended to regulate a gene on the same strand as the cis-regulatory producing DNA [25, ch. 5]. These RNAs attach to binding sites on the DNA and influence transcription. Others play a role in RNA replication.

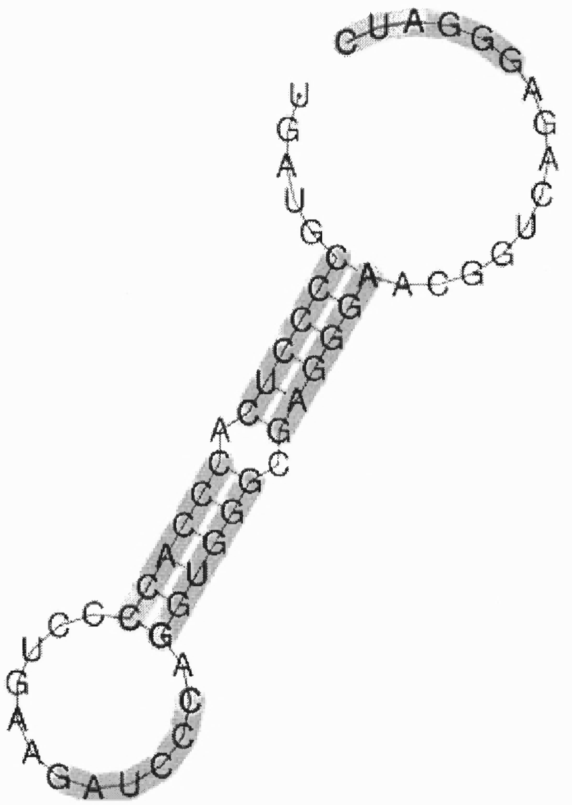


Figure 2.7 Antizyme RNA frame shift element. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00381

Micro RNAs (miRNAs): Micro RNAs are (as the name implies) tiny RNAs, usually 21 to 23 nucleotides in length. They are formed from precursors about 50 to 80 nucleotides in length. They have a few established roles. One is to regulate gene expression by binding to mRNAs and preventing translation [26]. They also can specify mRNA cleavage sites, setting up degradation pathways for mRNA [27]. They may also have the ability to methylate complementary genomic sites [28].

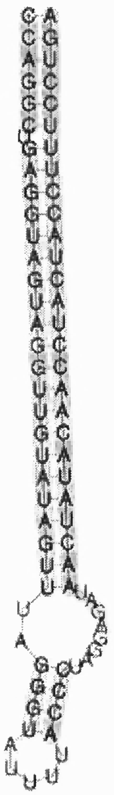

G C A G G C U G A G G U A G U A G G U U G U A U A G U U
A G U C C U U U C C A U C A U C C A A C A U A U C A A
U A
G G G U
A U U U U
A C C C

Figure 2.8 Let-7 microRNA precursor. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00027

Small interfering RNAs (siRNAs): These are about 20 to 25 nucleotides in length, and function in the RNA interference of gene expression as part of the RNAi pathway [52]. Like miRNAs, they are formed from precursors that are cut down to form mature siRNAs [29]. The siRNAs precursors are **shRNAs**, or **small hairpin RNAs,** due to their distinctive shape (they are a stem with a loop on the end, a pure hairpin form.) Curiously they do not appear do have much of a presence in the RNA databases, perhaps due to the sameness of their structure. There appears to be significant interest in these as a medical application, as they can silence gene expression, which is perhaps another reason why there is not much in the RNA databases that concerns them. There are shRNA/siRNA-specific databases, such as The MIT/ICBP siRNA Database [51].

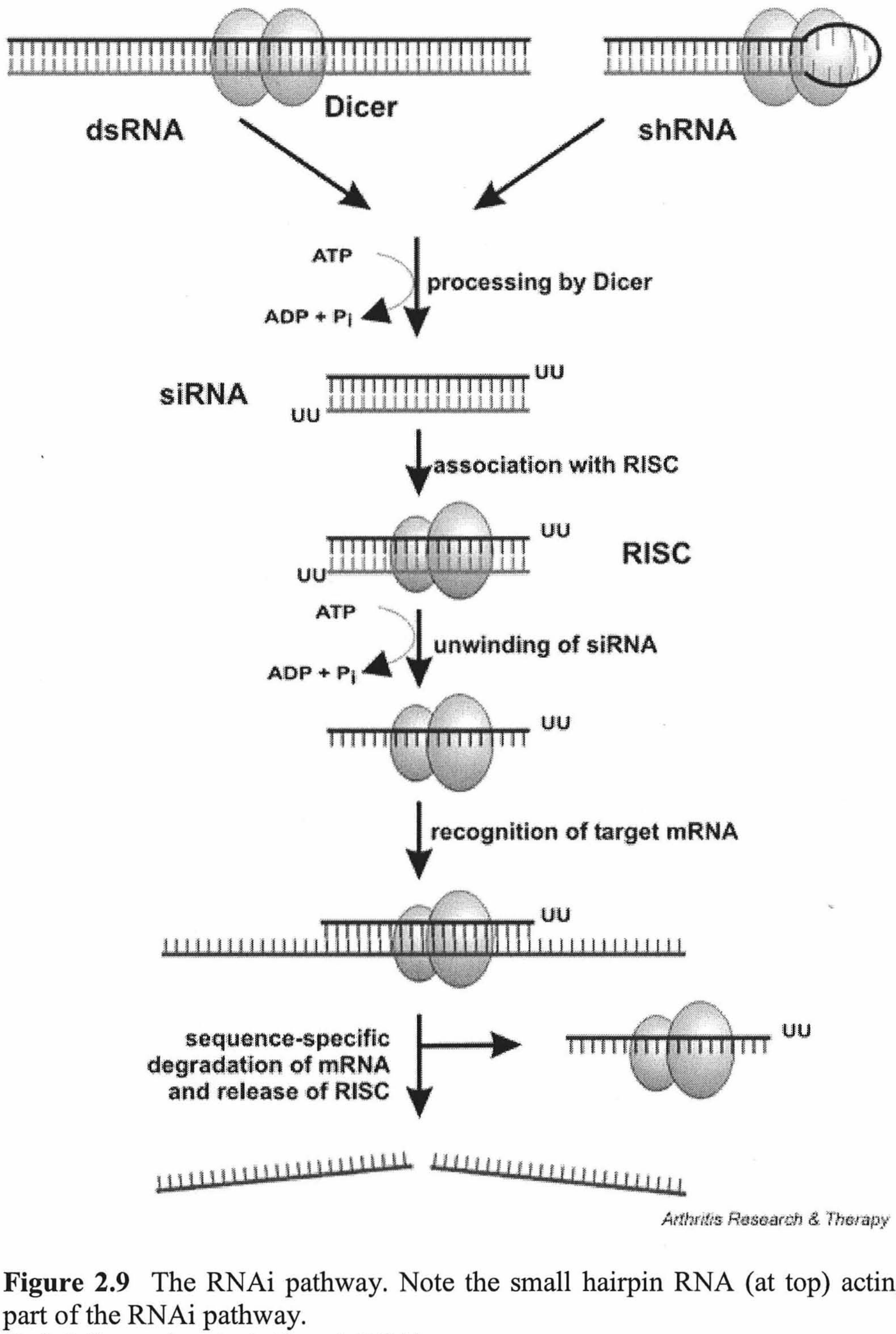


Figure 2.9 The RNAi pathway. Note the small hairpin RNA (at top) acting as part of the RNAi pathway.
http://arthritis-research.com/content/figures/ar1168-1.jpg

Small nuclear RNAs (snRNAs): The small nuclear RNAs are a bit bigger than the miRNAs, about 100 to 300 nucleotides long. They are active as regulators [30] and splicing agents [19, ch. 28.3.3]. A major class of the snRNAs are the **small nucleolar RNAs (snoRNAs)**. These aid the nucleolus in ribosomal creation. They form RNA-protein complexes called small nucleolar ribonucleoproteins (snoRNPs) and act via methylation and pseudouridylation [31] (which is the isomeration of uracil residues; it converts uridine to pseudouridine.)

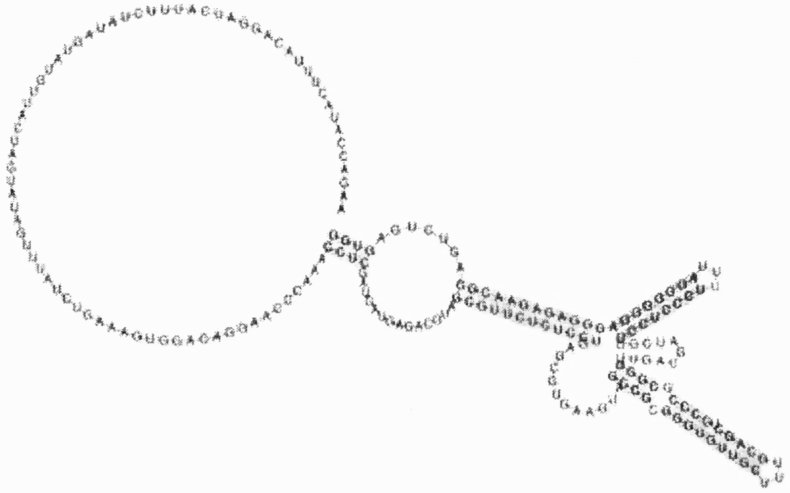


Figure 2.10 The snoRNA U3. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00012

Telomerase RNAs: This is another type of RNA that binds with proteins to form an enzymatic complex, in this case telomerase. Telomeres are stretches of repetitive DNA at the end of chromosomes that protect the chromosomes from sequence loss.

Telomerase is a reverse transcriptase that restores the telomeres to the ends of chromosomes [9. ch. II.5]. The telomerase RNA is a complementary copy of the telomere, and it is an example of RNA being used as a blueprint for DNA. The RNA subunit of telomerase is considered to be a snoRNA [32].

Pseudoknots: Pseudoknots are not functional RNAs themselves, nor are they a type of RNA; they are included here because they are an important consideration in secondary structure prediction. They are a type of RNA tertiary structure. (Tertiary structures in general are units of secondary structure that are formed by hydrogen bonding and can be grouped into classes or domains [19, ch. 3.4].) The base pairing within pseudoknots does not follow typical grammatical rules. Consequently pseudoknots are quite difficult for most secondary structure methods to detect [33].

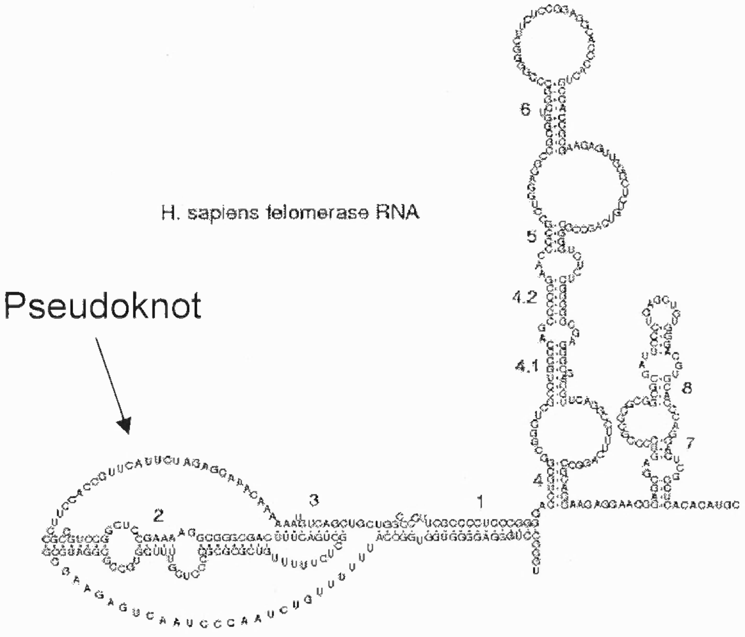


Figure 2.11 Vertebrate telomerase. Also an example of a pseudoknot. From Rfam.
http://rfam.sanger.ac.uk/family?acc=RF00024

2.2 Computer Science Background

2.2.1 Basic Computer Science

There are a few basic computer science concepts that should be introduced at this point. An algorithm is a sequence of instructions that must be performed to solve a well-formulated problem [34, p. 7]. This is how computer programs accomplish their work, and the "well-formulated problem" stipulation is especially important within bioinformatics. A common algorithm within bioinformatics in general and RNA secondary structure prediction in particular is dynamic programming. Simply put, a dynamic programming algorithm is one that breaks a problem into smaller problems, and those problems into smaller problems until a small enough problem is reached that a series of them can be solved much more quickly than approaching the original problem directly. This can lead to huge complexity if the algorithm is not designed elegantly [34, pp. 43-44].

2.2.2 Grammars

Another useful computer science concept is the grammar. In CS terms, a grammar is a set that describes all the possible words or statements in a language. Grammars are traditionally organized into the Chomsky hierarchy, which includes (going from the lowest to the highest level) unrestricted (phase structure) grammars, context-sensitive grammars, context-free grammars, and regular grammars [4, p. 237]. Each grammar has a particular automaton that recognizes it. Automata in CS are abstract computational devices that describe individual grammars.

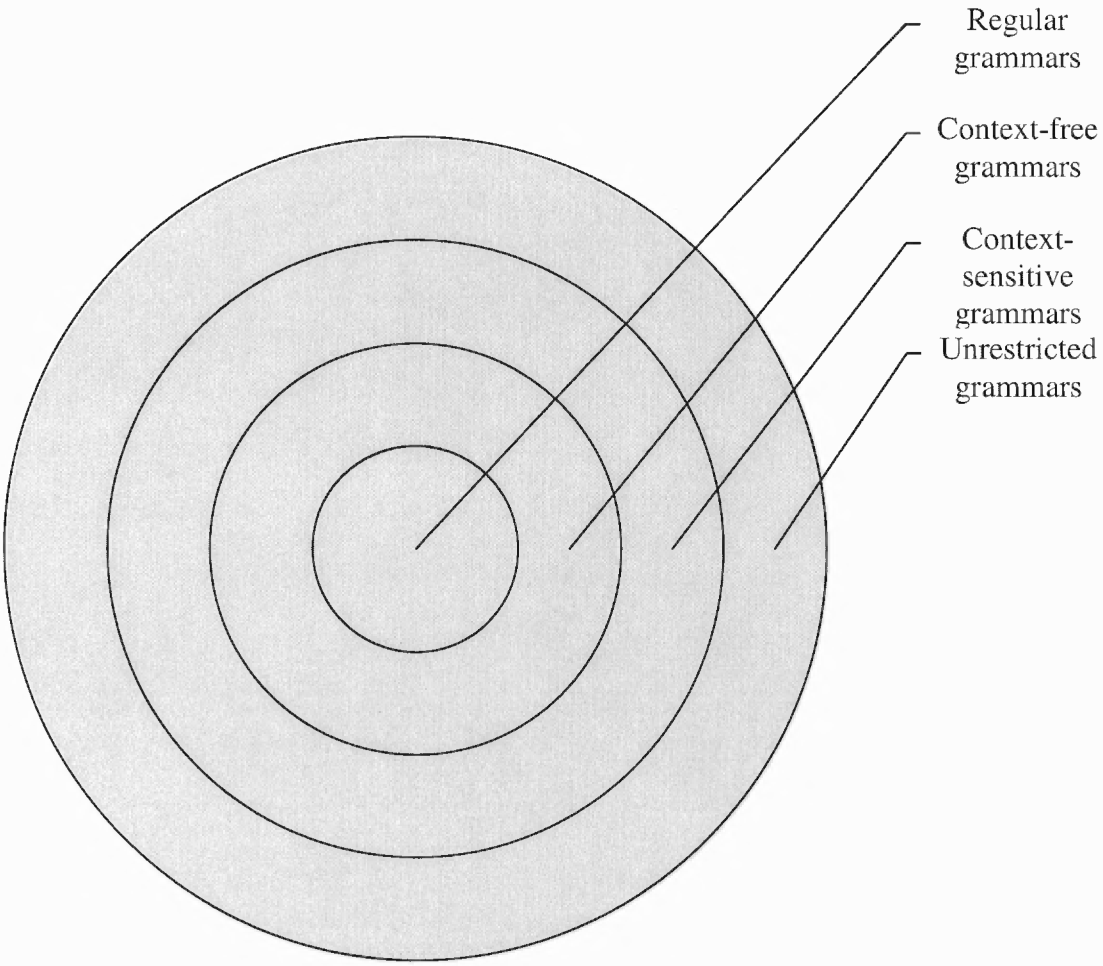


Figure 2.12 The Chomsky hierarchy.

Some of these grammars have specific bioinformatic applications. Regular grammars, for example, generate sequence from left to right, and thus are useful for modeling primary sequence [4, pp. 238-243]. The context-free grammars were originally designed to describe natural languages [35, p. 77]; they have rules that allow the grammar to make correlations between the ends of sentences. This turns out to have value in predicting RNA secondary structure, where sequence differences may not imply secondary structure differences.

All of the Chomsky hierarchical grammars have a stochastic form as well [4, p. 250]. A stochastic grammar is a probabilistic grammar where characters are given scores based on a consensus understanding of how the language is supposed to work. These stochastic grammars are very useful for biological analysis, since there are numerous grammatical exceptions in the "language" of DNA and RNA. A probabilistic model can account for these exceptions and still find related "words," or homologues in biological terms. (A homologue is a characteristic common to different organisms due to shared ancestry. Homology refers to the study and detection of homologues.) For example, many alignment methods use sequence profiles that contain enough specificity to find distantly related family members, despite perhaps large evolutionary distances between them. Hidden Markov models (HMMs) are a widely-used type of stochastic grammar [4, p. 252].

Covariance models are another type of stochastic grammar-based profile; in particular they are profiles of stochastic context-free grammars (SCFGs). Their main advantage over HMMs is that they can be used to predict secondary structure. According to a book co-authored by one of their main proponents, Sean Eddy, they are the "SCFG analogue of profile HMMs" [4, p. 287]. They specify a repetitive tree-like SCFG architecture, and are detailed, complex probabilistic models.

CHAPTER 3

SOFTWARE REVIEW

3.1 Introduction

This is a review of the software used in this thesis. Software tools were selected based on if they were RNA secondary structure related, and to give multiple points of comparison between CM-based programs and other methodologies. This is not intended to be a broad survey of all available secondary structure-related tools.

3.2 Software Based on Covariance Models (CMs)

3.2.1 The Infernal Suite

Infernal is a suite of programs written in C for Unix/Linux; after some time as a "beta" program its had a 1.0 release in June 2008. As of November 2008 it is up to Version 1.0rc4; the version used in this project is Version 1.0rc3 [36, 54]. It contains seven individual programs: cmalign, cmbuild, cmcalibrate, cmemit, cmscore, cmsearch, and cmstat. In total the suite allows a user to start with an RNA multiple alignment, create a CM-based profile for it and use the new profile to discover homologues in existing data. Not all the programs will be discussed equally, as the path from alignment to homologues does not involve all of them.

Infernal's cmbuild program takes in an annotated alignment and returns a profile CM. Alignments must be in Stockholm format to be accepted. Stockholm format requirements include a well-defined header and sequence gaps represented with dashes or dots. It also can include some secondary structure information. The CM it returns is a mathematical model, and is not intended to be human-interpretable.

```
# STOCKHOLM 1.0
ABAQ02000001/1286001-1285924
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGAAAUUGAAAUAAAAAACCCGAUGCGCAGAUCAUCGGGUU
CAUUUCA
AAJZ01000001/4863026-4863103
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGAAAUUGAAAUAAAAAACCCGAUGCGCAGAUCAUCGGGUU
CAUUUCA
AE005174/2649880-2649955
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGAAAUUGAAAU..AAAACCCGAUGCGCAGAUCAUCGGGUU
CAUUUCA
CP000247/1882367-1882444
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGCAAUUGAAAUAAAAAACCCGAUGCGCAGAUCAUCGGGUU
CAUUUCA
BA000007/2574669-2574744
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGAAAUUGAAAU..AAAACCCGAUGCGCAGAUCAUCGGGUU
CAUUUCA
AAJX01000031/11199-11122
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGAAAUUGAAAUAAAAAACCCGAUGCGCAGAUCAUCGGGGU
CAUUUCA
AAJU01000021/17257-17180
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGCAAUUGAAAUAAAAAACCCGAUGCGCAGAUCAUCGGGUU
CAUUUCA
AAJV01000029/42474-42551
GGGAAACUUUAUUGCUGAUGCCACCCGCCGCGAAAUUGAAAUAAAAAACCCGAUGCGCAGAUCAUCGGGGU
CAUUUCA
#=GC SS_cons
<<....>>...<<<<....<<.....>>.>>>>...<<<<<<....<<<<<<<<<.......>>>>>>>>>
.>>>>>>
#=GC RF
gGgAAACuuuAucGcugAuGccAcccgCcgCgaAAuuGaaauAAAAacccGauGcgcAgAuCauCggguu
cauuuCa
//
```

Figure 3.1 A Stockholm format alignment.

```
INFERNAL-1 [1.0rc3]
NAME      let_7_seed.sto-1
STATES    253
NODES     54
ALPHABET 1
ELSELF    -0.08926734
WBETA     1e-07
NSEQ      14
EFFNSEQ   1.472
CLEN      83
BCOM      cmbuild let_7_seed.cm let_7_seed.sto.txt
BDATE     Sat Oct 11 16:54:22 2008
CCOM      cmcalibrate  -s 1225657789 let_7_seed.cm
CDATE     Sun Nov  2 15:29:49 2008
NULL      0.000  0.000  0.000  0.000
PART      1        0    100
E-LC      0       0.68283    -5.60319     2.48993     1500000       282592
0.003981
E-GC      0       0.46639   -14.30641    -2.33744     1500000        99616
0.003764
E-LI      0       0.66449    -4.31426     3.73357     1500000       236387
0.004759
E-GI      0       0.53707    -8.17005     1.68969     1500000        74781
0.005015
E-LV      0       0.63456    -1.81019     5.34813    17010000       119815
0.010648
E-GV      0       0.58663    -2.98714     6.62317    17010000       119412
0.003561
E-LF      0       0.68117     1.71673     8.38697    17010000       119956
0.010635
E-GF      0       0.61161     0.10646     9.32712    17010000       119630
0.003555
      FT-LC    35  0.99500  10000  1500000  0
```

Figure 3.2 A portion of a covariance model. This is the let-7 seed CM used in this thesis.

Infernal's cmbuild output contains information on alignment size and some statistical information on how the model was constructed. Using the CM new RNA sequences can be aligned to the model using cmalign. RNA families can be categorized in this fashion. Single or multiple sequences (in FASTA format) can be used. The cmcalibrate program can then be used to tweak the model so the activities of cmsearch work faster. Using cmcalibrate is recommended to improve search time and sensitivity [36, p. 11].

Infernal's cmsearch, then, uses CMs to find RNA family homologues. It reads in a "database" which is any FASTA-formatted RNA sequence under investigation. It returns a score, an alignment, a predicted secondary structure in dot-parenthesis format, information on how highly or weakly residues are conserved, and information on how the score was obtained.

To summarize, the homologue discovery process with Infernal is as follows: using a Stockholm formatted alignment to produce a CM; using the CM and cmsearch to find homologues; using cmalign to align new RNA sequences to the consensus structure. Infernal is used by the RFAM database (discussed below) in this way to maintain its RNA family distinctions and to add more as they are discovered.

```
CM: let_7_seed.sto-1
>ref|NW_047799.2|Rn8_WGA2323_4:14134978-14155107

  Plus strand results:

 Query = 1 - 83, Target = 10026 - 10101
 Score = 60.60, E = 7.074e-16, P = 1.117e-19, GC =  45

          <<<<<<-<<<<<<<<<<<<<<<<<<<<<---<<<<_______>>>>---------->>>>
        1 ccaGGaUgAGGuAGuAGguuGuauaGUuuuagGGcuaaaauagCCcauuaGGAGAUaACu
60
          C AGG UGAG UAG AGG:UGUA:AGUUU+  G::     + ::C   + GGAGAUAACU
    10026 CCAGGCUGAGGUAGUAGGUUGUAUAGUUUAGAGUU-----ACAACA--AGGGAGAUAACU
10078

          >>>>>>>>>>>>>>>>>>>>>>>
       61 auaCaacCUaCUaCCUuuCCugg 83
          :UACA:CCU CUA CUU CCU G
    10079 GUACAGCCUCCUAGCUUUCCUUG 10101
```

Figure 3.3 Partial cmsearch output. The query sequence is preceded by the 1 and the 61; the found sequence is preceded by 10026 and 10079. Note the E value after the score near the top (under the line that starts with Query.) See Appendix B for a full example of cmsearch output.

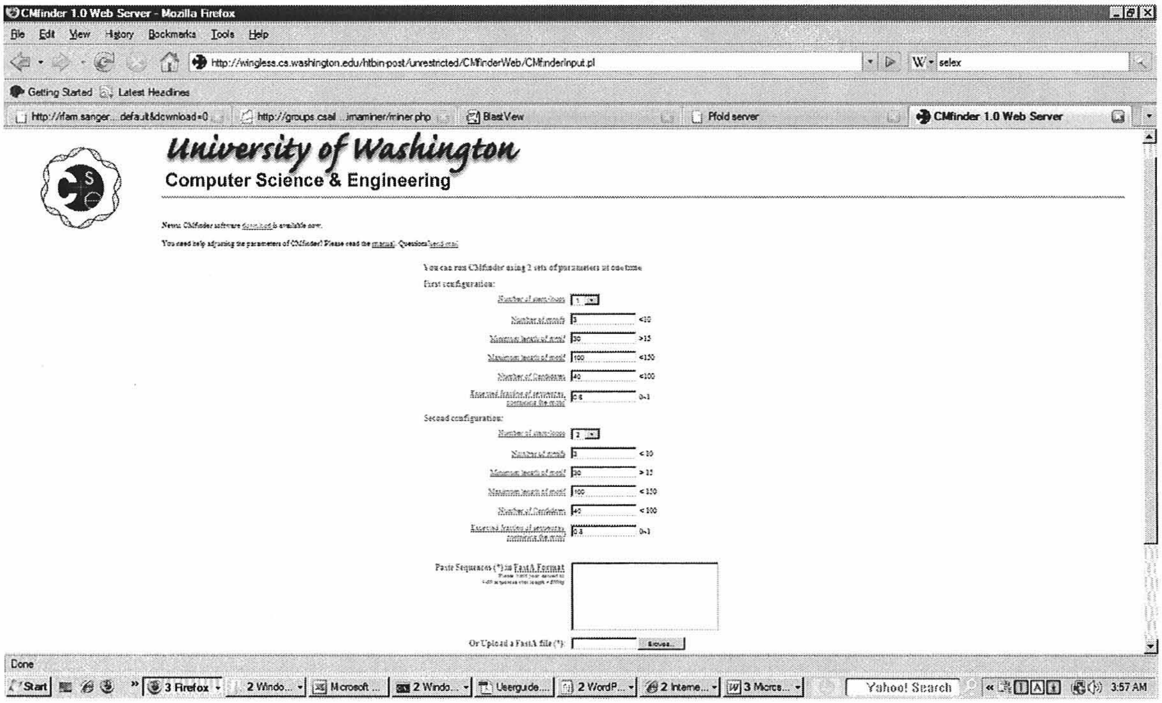
Special mention should be made of cmemit, which is not part of the Infernal discovery process but has the useful function of reading a CM and returning unaligned sequence data in FASTA format. It outputs 10 by default, though that number can be changed. Note that the outputted sequences will be simulated biological data; while FASTA-formatted data is used to create a CM, a CM cannot be unpacked to return the original data. The remaining Infernal modules are experimental or of limited practical value. Cmscore outputs statistics that could be useful in the further honing of Infernal's algorithms. Cmstat returns statistics on covariance models, with more information being returned on calibrated versus non-calibrated models.

3.2.2 CMfinder

CMfinder is a web-based tool [60] that returns results via e-mail. Its input is a FASTA sequence; its output is CM-based profiles and motifs in Stockholm format. It is described by its authors as an adaptation of the DNA motif-finding tool MEME for use with CMs and RNA secondary structure [37]. It uses an "expectation matrix" to score possible secondary structure matches after a covariance model has aligned and identified them. Customizable features include adjustments for motif length and the number of candidates to search for. There is also an option for number of stem loops expected in the RNAs entered. CMfinder allows two sets of parameters to be applied to the same dataset at the same time.

Like Infernal, CMfinder returns covariance models from sequence data. Unlike Infernal, it breaks the sequence data down into a set of CMs, and a corresponding set of

Stockholm-formatted files. Unfortunately, the CMs it generates are in Infernal 0.55 format [38], which are not accepted by 1.0 Infernal releases.


CMfinder 1.0 Web Server - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://wingless.cs.washington.edu/htbin-post/unrestricted/CMfinderWeb/CMfinderInput.pl
selex
Getting Started Latest Headlines
BlastView
Pfold server
CMfinder 1.0 Web Server
University of Washington
Computer Science & Engineering
Browse...
Done
Start
3 Firefox
2 Windo...
Microsoft ...
2 Windo...
Userguide...
2 WordP...
2 Interne...
3 Micros...
Yahoo! Search
3:57 AM

News: CMfinder software download is available now.

You need help adjusting the parameters of CMfinder? Please read the manual. Questions? send mail

You can run CMfinder using 2 sets of parameters at one time

First configuration:

Number of stem-loops	1	
Number of motifs	3	<10
Minimum length of motif	30	>15
Maximum length of motif	100	<150
Number of Candidates	40	<100
Expected fraction of sequences containing the motif	0.8	0-1

Second configuration:

Number of stem-loops	2	
Number of motifs	3	< 10
Minimum length of motif	30	> 15
Maximum length of motif	100	< 150
Number of Candidates	40	< 100
Expected fraction of sequences containing the motif	0.8	0-1

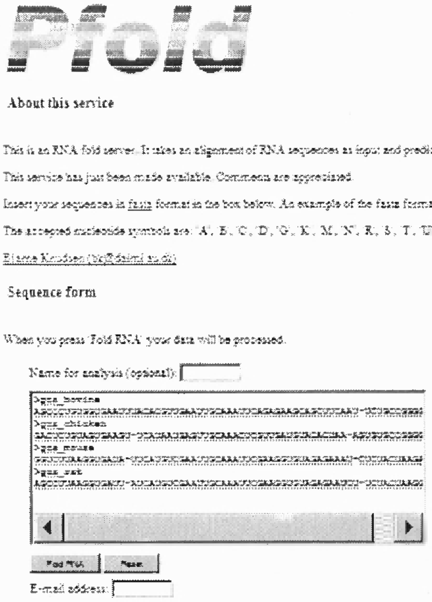
Paste Sequences (*) in FastA Format
[illegible]

Or Upload a FastA file (*):

Figure 3.4 The Cmfinder main page.
http://wingless.cs.washington.edu/htbin-post/unrestricted/CMfinderWeb/CMfinderInput.pl

3.2.3 Pfold

Pfold is not a CM-based method, but it does use SCFGs in concert with probabilistic models to predict secondary structure, so this may be the appropriate section to mention it. It is a web-based tool that takes in sequence(s) (a maximum of 40 with a maximum length of 500; this is a limitation of its current server, rather than the underlying algorithm) and returns both a common structure for all the sequences and a structure for each individual sequence in dot-parenthesis format. Pfold uses a method that applies an evolutionary model in addition to using SCFGs, unlike CMs, which do not take phylogeny into account [39, 40].



About this service

This is an RNA fold server. It takes an alignment of RNA sequences as input and predicts a common structure for all sequences. The algorithm is described in the references below.

This service has just been made available. Comments are appreciated.

Insert your sequences in fasta format in the box below. An example of the fasta format is shown in the box. At the moment, the server has a maximum of 40 sequences, and an alignment length of 500.

The accepted nucleotide symbols are: 'A', 'B', 'C', 'D', 'G', 'K', 'M', 'N', 'R', 'S', 'T', 'U', 'V', 'W', 'X' and 'Y', as used in the extended nucleotide alphabet. Gap signs, '-', are also accepted along with the above letters in lower case.

Bjarne Knudsen (bk@daimi.au.dk)

Sequence form

When you press 'Fold RNA' your data will be processed.

Name for analysis (optional):

```
>gsa_bovine
[illegible]
>gsa_chicken
[illegible]
>gsa_mouse
[illegible]
>gsa_rat
[illegible]
```

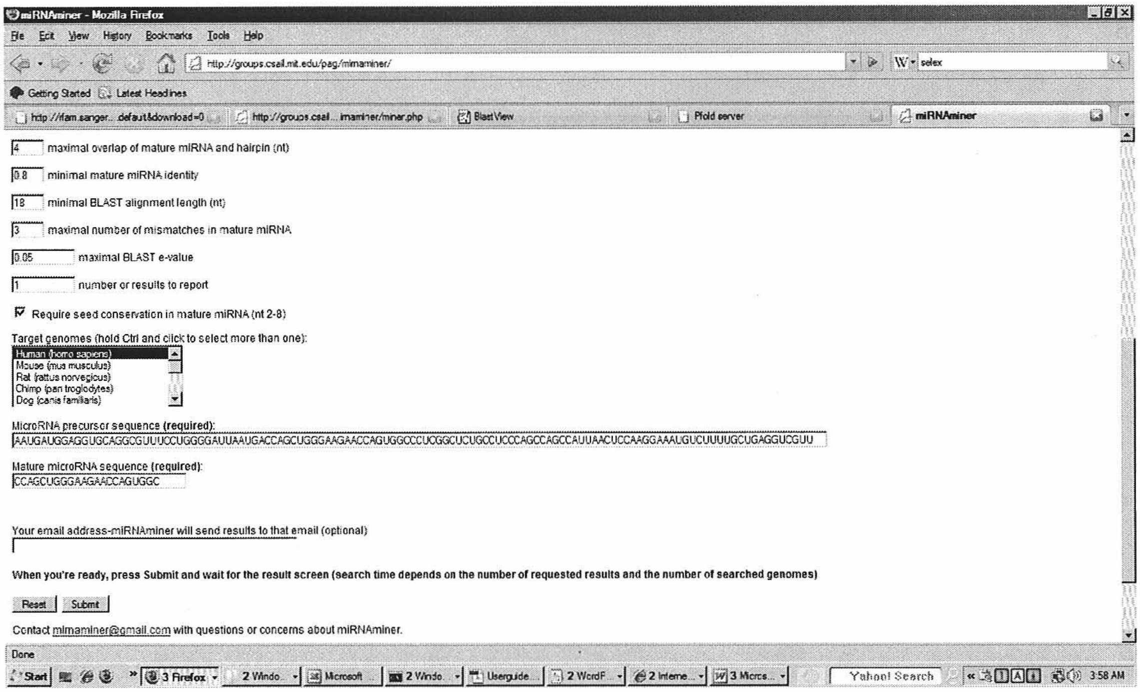
E-mail address:

Figure 3.5 The Pfold web server.
http://www.daimi.au.dk/~compbio/rnafold/

3.3 Other Software Methods

3.3.1 MiRNAminer

MiRNAminer is another web-based tool, written in Java [41]. It is intended for searching for miRNA homologues in animals. The input is plain RNA sequence, but it also has to include the mature miRNA sequence that's being targeted. The output returns a score, matching sequence(s) and secondary structure in dot-parenthesis format. At present miRNAminer is limited to 10 metazoan genomes. It begins with a BLAST search, then uses e-values to sort out potential miRNAs. (BLAST is a common primary sequence alignment tool, discussed briefly in Section 3.3.2.)


miRNAminer - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://groups.csail.mit.edu/pag/mirnaminer/
selex
Getting Started Latest Headlines
http://rfam.sanger... default&download=0
http://groups.csail... imaminer/miner.php
BlastView
Pfold server
miRNAminer
Done
Start
3 Firefox
2 Windo...
Microsoft
2 Windo...
Userguide...
2 WordF...
2 Interne...
3 Micros...
Yahoo! Search
3:58 AM

4 maximal overlap of mature miRNA and hairpin (nt)

0.8 minimal mature miRNA identity

18 minimal BLAST alignment length (nt)

3 maximal number of mismatches in mature miRNA

0.05 maximal BLAST e-value

1 number or results to report

☑ Require seed conservation in mature miRNA (nt 2-8)

Target genomes (hold Ctrl and click to select more than one):

Human (homo sapiens)
Mouse (mus musculus)
Rat (rattus norvegicus)
Chimp (pan troglodytes)
Dog (canis familiaris)

MicroRNA precursor sequence (required):

AAUGAUGGAGGUGCAGGCGUUUCCUGGGGAUUAAUGACCAGCUGGGAAGAACCAGUGGCCCUCGGCUCUGCCUCCCAGCCAGCCAUUAACUCCAAGGAAAUGUCUUUUGCUGAGGUCGUU

Mature microRNA sequence (required):

CCAGCUGGGAAGAACCAGUGGC

Your email address-miRNAminer will send results to that email (optional)

When you're ready, press Submit and wait for the result screen (search time depends on the number of requested results and the number of searched genomes)

Reset Submit

Contact mirnaminer@gmail.com with questions or concerns about miRNAminer.

Figure 3.6 The miRNAminer main page. Partial screenshot.

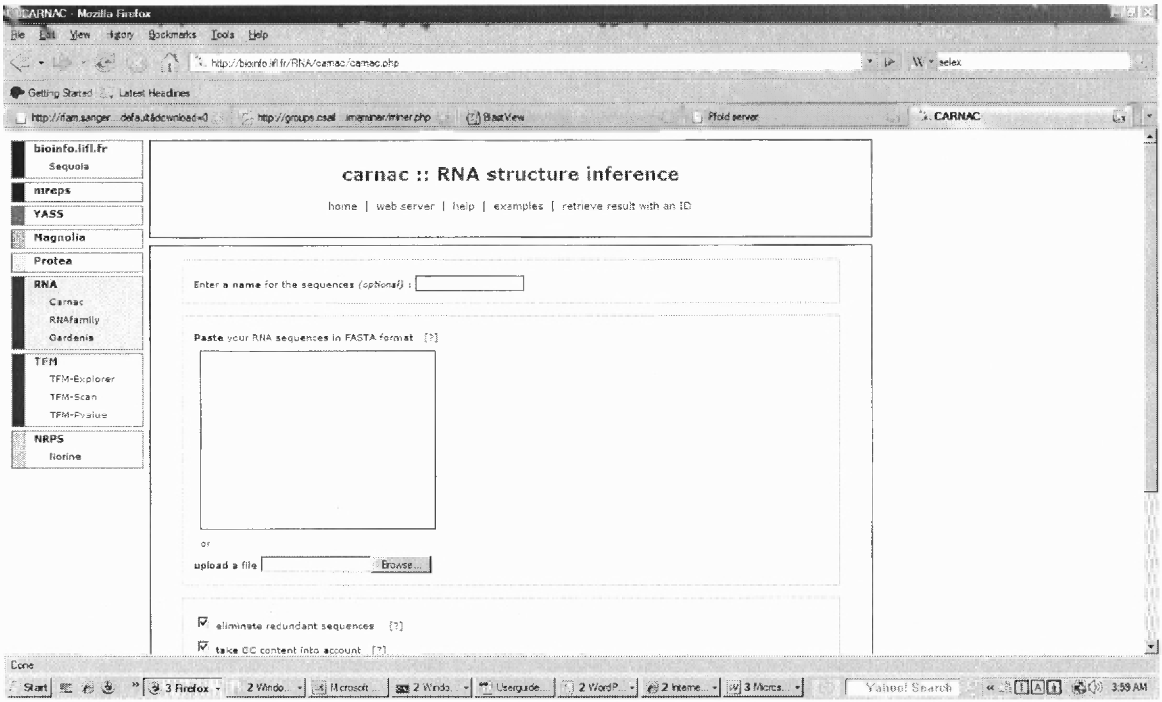
3.3.2 BLAST and BLAT

BLAST is a widely-used primary sequence tool, based on an algorithm of the same name [42]. It uses a local alignment method and uses matrices to score. It sacrifices sensitivity for speed. As noted above, it is often used a step in other software programs.

A recent and functionally similar tool is BLAT, the BLAST-like alignment tool [43]. It is a major size and speed improvement over BLAST, and the featured search at the UCSC Genome Browser [64]. BLAT was used in this project since it was of more recent origin than the well-understood BLAST, and may serve as a more interesting point of comparison with a CM-based method.

3.3.3 CARNAC

CARNAC has both a web version and a Linux version; it is written in C. The input is FASTA sequence, which does not have to be aligned. Its output is a text “CT file” which contains secondary structure information that can be visualized by Naview or RNAfamily, in addition to a .jpg of an RNA’s secondary structure and a dot-bracket file with the same information. It uses a heuristic algorithm that finds all possible secondary structure stems, then sorts through these to find the more probable stems. At least on the web server, sequences must be less than 80 nucleotides, and at least two must be submitted together, as CARNAC attempts to find a consensus secondary structure [44].


CARNAC - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://bioinfo.lifl.fr/RNA/carnac/carnac.php
selex
Getting Started Latest Headlines
http://rfam.sanger...default&download=0
http://groups.csail...imaminer/miner.php
BlastView
Mfold server
CARNAC
bioinfo.lifl.fr
Sequoia
mreps
YASS
Magnolia
Protea
RNA
Carnac
RNAfamily
Gardenia
TFM
TFM-Explorer
TFM-Scan
TFM-Pvalue
NRPS
Norine
home | web server | help | examples | retrieve result with an ID
Browse...
Done
Start
3 Firefox
Yahoo! Search
3:59 AM

carnac :: RNA structure inference

Enter a name for the sequences (optional) :

Paste your RNA sequences in FASTA format [?]

or

upload a file

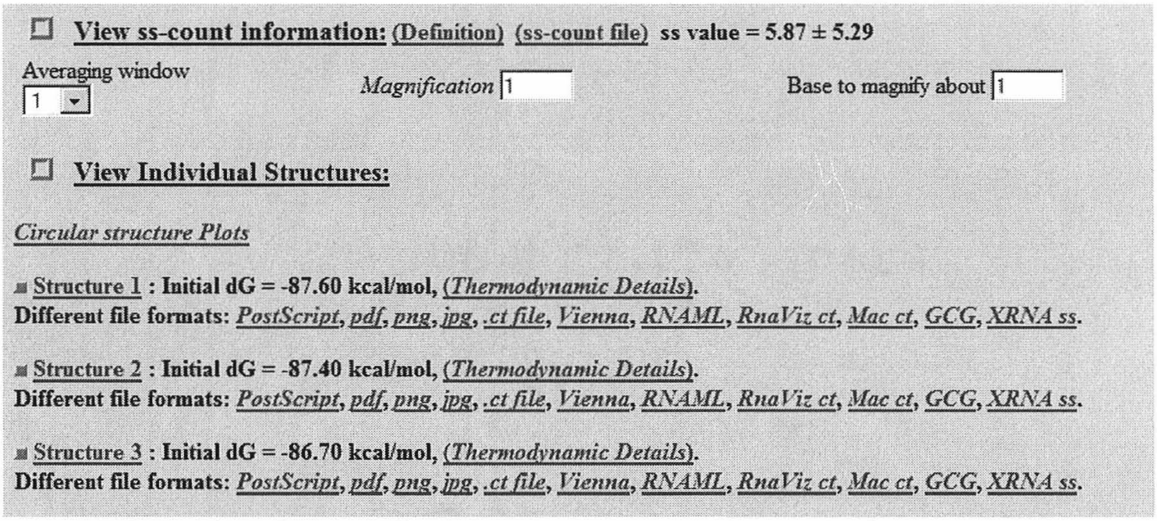
☑ eliminate redundant sequences [?]

☑ take GC content into account [?]

Figure 3.7 The CARNAC web server.
http://bioinfo.lifl.fr/RNA/carnac/index.php

3.3.4 Mfold

Mfold [45, 46] is another web tool, useful for folding single RNAs, also available in downloadable cross-platform versions. It uses a minimum energy algorithm. Sequences do not have to be in any format to be folded. There are a variety of customizable options for Mfold: number of foldings returned, loop angles, maximum distance between paired bases. Output file options are also plentiful, with visualizations returned in PostScript, pdf, png, and a variety of RNA visualization programs' native formats.



□ **View ss-count information:** (Definition) (ss-count file) **ss value = 5.87 ± 5.29**

Averaging window 1

Magnification 1

Base to magnify about 1

□ **View Individual Structures:**

Circular structure Plots

■ Structure 1 : **Initial dG = -87.60 kcal/mol,** (*Thermodynamic Details*).
Different file formats: *PostScript*, *pdf*, *png*, *jpg*, *.ct file*, *Vienna*, *RNAML*, *RnaViz ct*, *Mac ct*, *GCG*, *XRNA ss*.

■ Structure 2 : **Initial dG = -87.40 kcal/mol,** (*Thermodynamic Details*).
Different file formats: *PostScript*, *pdf*, *png*, *jpg*, *.ct file*, *Vienna*, *RNAML*, *RnaViz ct*, *Mac ct*, *GCG*, *XRNA ss*.

■ Structure 3 : **Initial dG = -86.70 kcal/mol,** (*Thermodynamic Details*).
Different file formats: *PostScript*, *pdf*, *png*, *jpg*, *.ct file*, *Vienna*, *RNAML*, *RnaViz ct*, *Mac ct*, *GCG*, *XRNA ss*.

Figure 3.8 Partial screenshot of Mfold output.
http://www.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi

CHAPTER 4

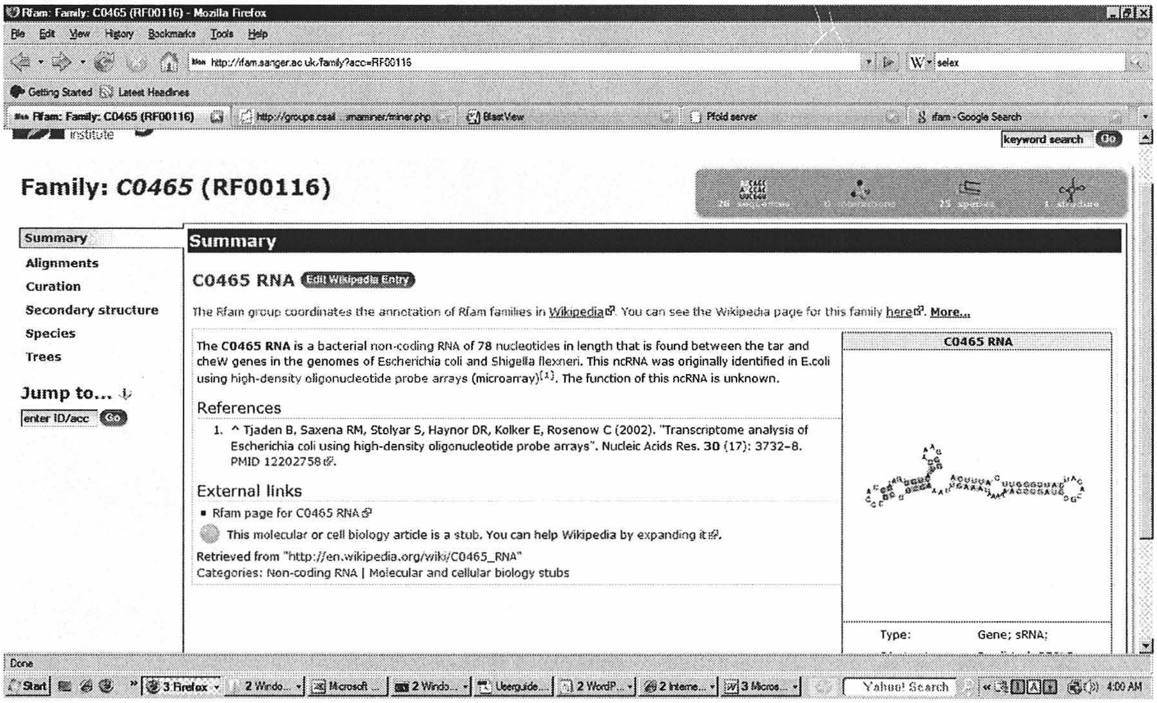
DATABASE REVIEW

4.1 Introduction

This chapter is an overview of the repositories of RNA secondary structure information used in this thesis. Some were used extensively (miRBase and particularly Rfam) and others in a more functional way (RmotifDB and the various genome browsers.) One is simply an interesting case (RNA STRAND.)

4.2 Rfam

Rfam [47] is an RNA secondary structure database. Broadly, it contains two types of data: hand-curated families of ncRNAs taken from published sequence alignments, which are called “seed” alignments; and those same familes with additional representatives aligned by Infernal’s cmalign program, the “full” alignments [36, p.13]. The covariance model used to generate the alignments is also available, as well as a secondary structure diagram and a direct link to the family’s Wikipedia page (to be specific, each family’s Wikipedia entry is also displayed within each family’s Rfam page.) Rfam does not specialize in one type of family, and is meant to be a broad collection of ncRNAs. It contains ncRNAs and their genes, cis-regulatory elements, and self-splicing RNA families.



Family: *C0465* (RF00116)

Summary

C0465 RNA

The Rfam group coordinates the annotation of Rfam families in Wikipedia. You can see the Wikipedia page for this family here. More...

The **C0465 RNA** is a bacterial non-coding RNA of 78 nucleotides in length that is found between the tar and cheW genes in the genomes of Escherichia coli and Shigella flexneri. This ncRNA was originally identified in E.coli using high-density oligonucleotide probe arrays (microarray)[1]. The function of this ncRNA is unknown.

References

1. ^ Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C (2002). "Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays". Nucleic Acids Res. **30** (17): 3732–8. PMID 12202758.

External links

- Rfam page for C0465 RNA

This molecular or cell biology article is a stub. You can help Wikipedia by expanding it.

Retrieved from "http://en.wikipedia.org/wiki/C0465_RNA"
Categories: Non-coding RNA | Molecular and cellular biology stubs

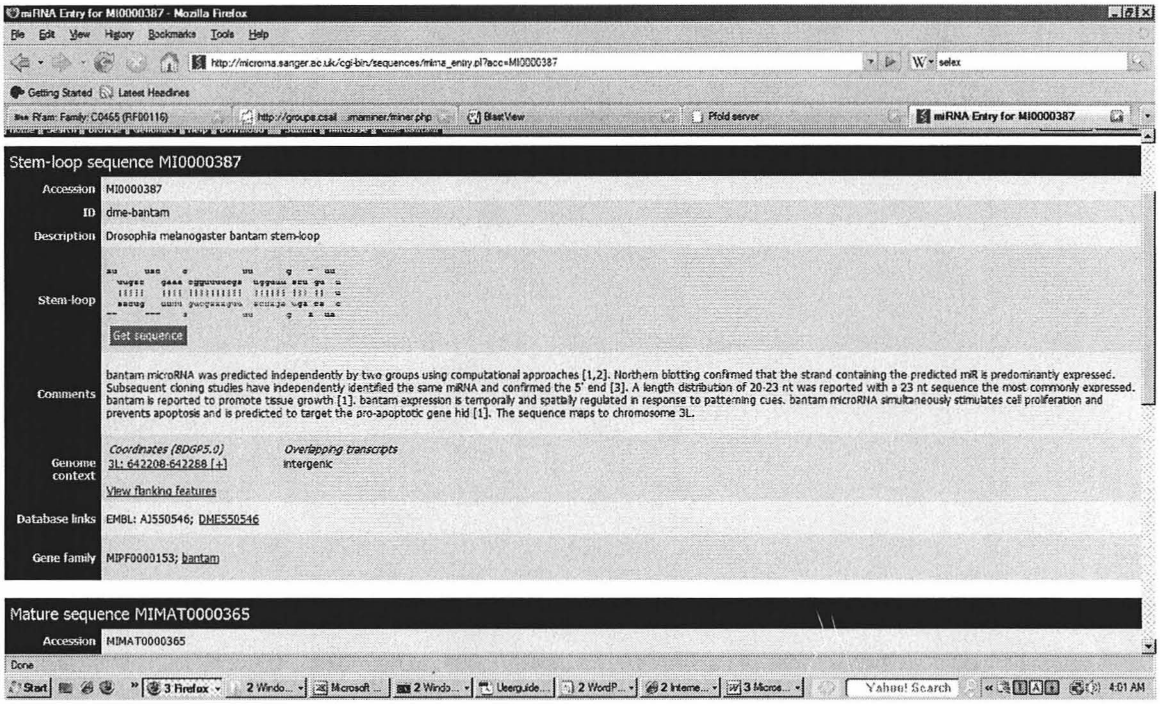
C0465 RNA

Type:	Gene; sRNA;

Figure 4.1 An Rfam page. This is the ncRNA C0465.
http://rfam.sanger.ac.uk/family?acc=RF00116

4.3 MiRBase

MiRBase is a specialty database [48, 49, 50, 68] with crosslinks between it and Rfam. Rfam miRNA families lead to family pages in miRBase. Unlike Rfam, miRBase contains both mature miRNA sequences and their precursors. It also acts as a clearinghouse for miRNA information, as it assigns its own set of accession numbers to each miRNA, and includes information on the genes each miRNA targets. It has a wide variety of search options, including accession numbers, keywords, organism, genomic location, and supports sequence searches (which made verifying Infernal and miRNAminer results a simple process.)


miRNA Entry for MI0000387 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://microrna.sanger.ac.uk/cgi-bin/sequences/mirna_entry.pl?acc=MI0000387
selex
Getting Started
Latest Headlines
Rfam: Family: C0465 (RF00116)
http://groups.csail ...mamner/miner.php
BlastView
Pfold server
miRNA Entry for MI0000387
Get sequence
Done
Start
3 Firefox
2 Windo...
Microsoft ...
2 Windo...
Userguide...
2 WordP...
2 Interne...
3 Micros...
Yahoo! Search
4:01 AM

Stem-loop sequence MI0000387

Accession	MI0000387
ID	dme-bantam
Description	Drosophia melanogaster bantam stem-loop
Stem-loop	
Comments	bantam microRNA was predicted independently by two groups using computational approaches [1,2]. Northern blotting confirmed that the strand containing the predicted miR is predominantly expressed. Subsequent cloning studies have independently identified the same miRNA and confirmed the 5' end [3]. A length distribution of 20-23 nt was reported with a 23 nt sequence the most commonly expressed. bantam is reported to promote tissue growth [1]. bantam expression is temporally and spatially regulated in response to patterning cues. bantam microRNA simultaneously stimulates cell proliferation and prevents apoptosis and is predicted to target the pro-apoptotic gene hid [1]. The sequence maps to chromosome 3L.
Genome context	*Coordinates (BDGP5.0)* *Overlapping transcripts* 3L: 642208-642288 [+] intergenic View flanking features
Database links	EMBL: AJ550546; DMES50546
Gene family	MIPF0000153; bantam

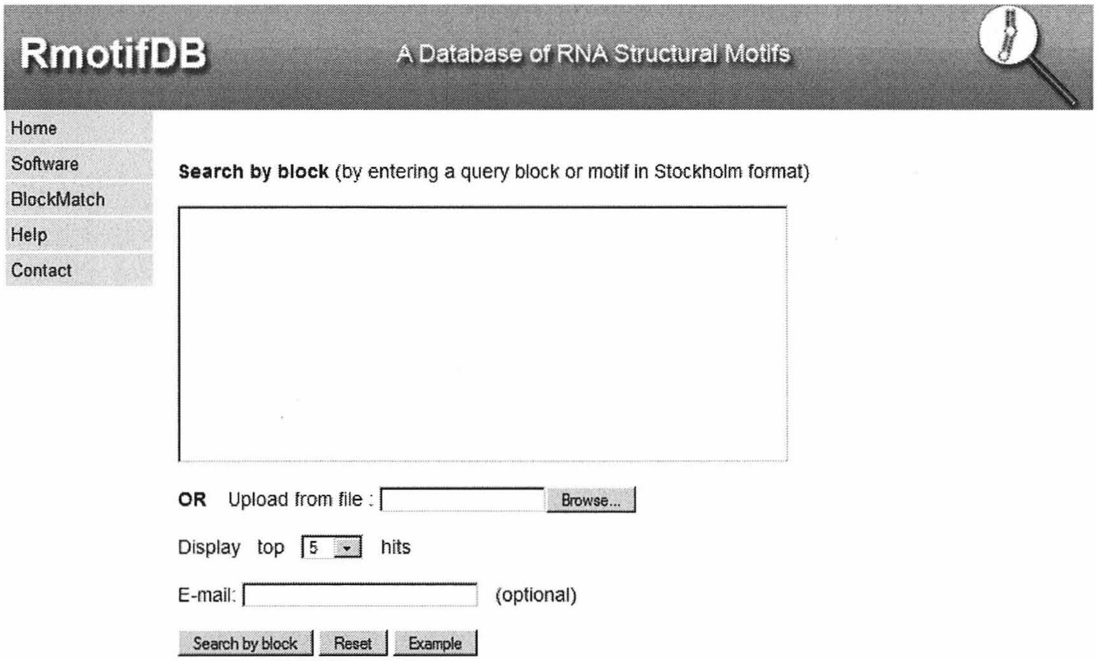
Mature sequence MIMAT0000365

Accession	MIMAT0000365

Figure 4.2 A miRBase page. This is bantam miRNA from *D. melanogaster*.
http://microrna.sanger.ac.uk/cgi-bin/sequences/mirna_entry.pl?acc=MI0000387

4.4 RmotifDB

RmotifDB is a database of RNA structural motifs [65, 66]. It is intended to mirror Rfam's releases, and as of this writing contains all 603 Rfam seed alignments [67]. RmotifDB was chosen for this project as allows for searches in Stockholm format (that is in fact the only input format it accepts) and it was thought this would provide a simple way to verify the output of CMfinder's CMs. Stockholm input is either inputted or uploaded via the search page. There are no customization options for the searches other than controlling the number of hits displayed (minimum of 5, maximum of 20.)



Search by block (by entering a query block or motif in Stockholm format)

OR Upload from file :

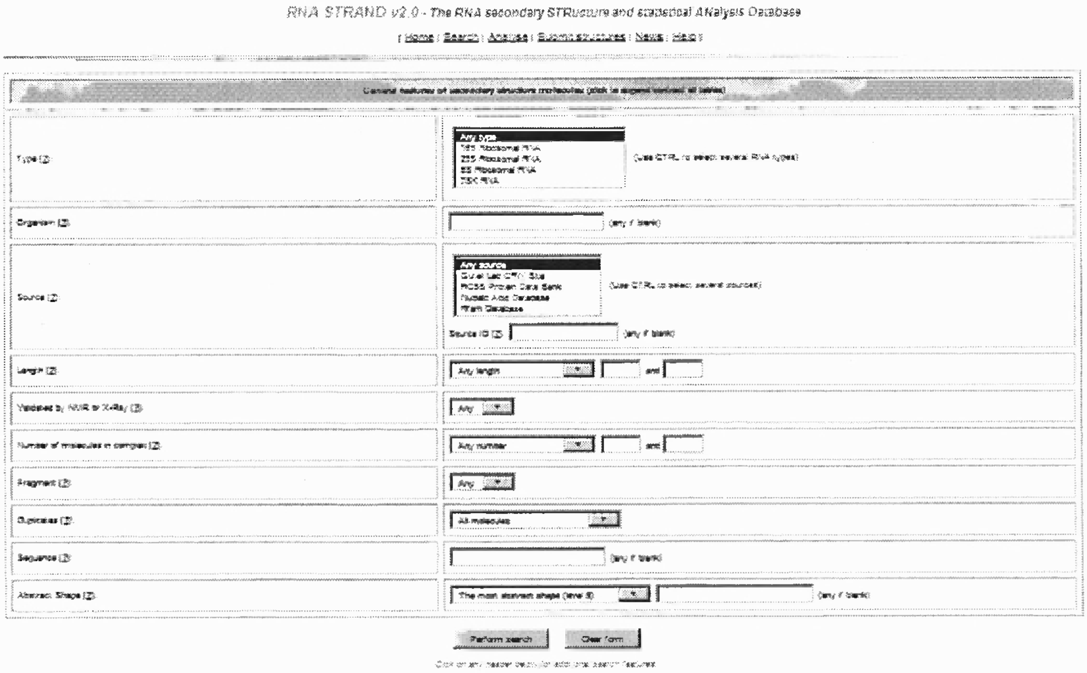
Display top 5 hits

E-mail: (optional)

Figure 4.3 The RmotifDB search screen.
http://datalab.njit.edu/bioinfo/index.html

4.5 RNA STRAND

RNA STRAND—the RNA secondary STRucture and statistical Analysis Database—is a recent attempt (the paper that introduced it was published in August, 2008 [53]) at a curated database of RNA secondary structures. It is interesting because it is actually a database of databases; all its entries are drawn from outside sources which are standardized and entered into RNA STRAND. Its constituent databases include Rfam, the RCSB Protein Data Bank, the Comparative RNA Web Site, the tmRNA database, the Sprinzl tRNA Database, the RNase P database, the SRP Database, and the Nucleic Acid Database. The criteria for inclusion in the database is stringent; only 19 families found at Rfam are included, out of 607 in the 8.1 Rfam release, as many of those families have computationally predicted secondary structures (as opposed to laboratory verified secondary structures.) Searching can be done via type of RNA, source, sequence, length, even whether a secondary structure was verified by NMR or x-ray. However, it does lack a keyword search function, which limited its use in this thesis. But it does seem to be off to a promising start.



RNA STRAND v2.0 - The RNA secondary STRucture and statistical ANalysis Database

| Home | Search | Analyse | Submit structures | News | Help |

General features of secondary structure molecules (click to expand/contract all tables)

Type (?)	Any type 16S Ribosomal RNA 23S Ribosomal RNA 5S Ribosomal RNA 7SK RNA (Use CTRL to select several RNA types)
Organism (?)	(any if blank)
Source (?)	Any source Gutell Lab CRW Site RCSB Protein Data Bank Nucleic Acid Database Rfam Database (Use CTRL to select several sources) Source ID (?) (any if blank)
Length (?)	Any length and
Validated by NMR or X-Ray (?)	Any
Number of molecules in complex (?)	Any number and
Fragment (?)	Any
Duplicates (?)	All molecules
Sequence (?)	(any if blank)
Abstract Shape (?)	The most abstract shape (level 5) (any if blank)

Figure 4.4 The RNA STRAND search page.
http://www.rnasoft.ca/strand/search.php

4.6 UCSC Genome Browser, ENSEMBL, and NCBI Genome

These are three of the main repositories of genetic data [54, 55, 56]. All three allow full genome sequence searching and genomic data downloads. The latter function was their primary use in this project, as they allow fine control over sequence data downloads (NCBI, for example, allows you to identify a target sequence, and then download user-specified flanking areas with it of any base pair length.) The UCSC and ENSEMBL browsers also house the BLAT software.

CHAPTER 5

METHODS

5.1 Introduction

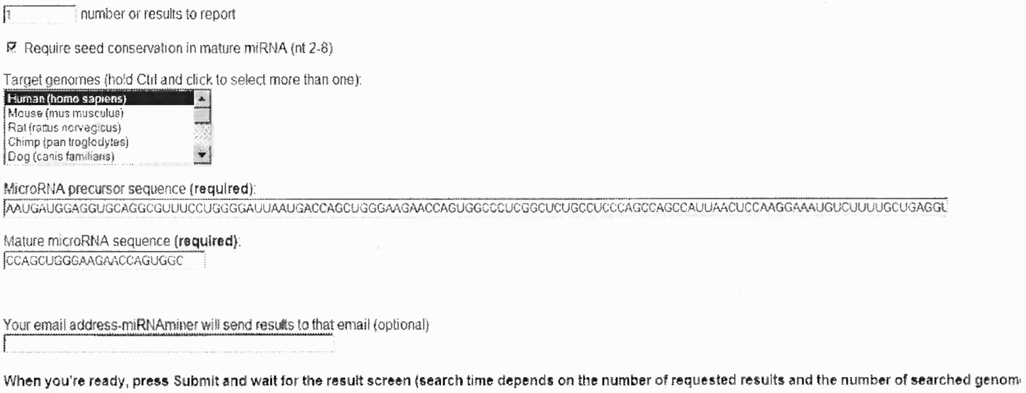
This chapter outlines the methods used in this thesis. The logic as to why certain tools and methods were grouped together is explained here. There are also descriptions of how each of the tools is used.

5.2 Infernal and miRNAminer

In general comparisons were made between programs that received and returned similar inputs and outputs. Infernal and miRNAminer both accept sequence data as inputs and return possible homologues, so they seemed to form a natural and simple point of comparison. MiRNAminer searches are limited to the eleven genomes on its server, so comparisons between the two would have to involve only those organisms.

The miRNAminer searches were made first. A representative miRNA was chosen, the let-7 family, simply because it was one of the first miRNAs discovered and (not coincidentally) it has the lowest accession number among miRNAs in Rfam. The precursor and mature let-7 sequences were found at miRBase. The let-7 of *C. elegans* was chosen; let-7 was originally identified in *C. elegans*, and the *C. elegans* genome is part of miRNAminer's genome set (so there would be at least one homologue found for certain.)

As mentioned above, miRNAminer is a web-based tool. Searches are made by entering matched miRNA precursor and mature miRNA sequences. Sequences must be plain sequences without any kind of header or additional characters; the latter will cause the input to be rejected. There are a number of parameters that can be changed, including maximum RNA folding energy, precursor sequence length, BLAST options, and various minimal matching percentages. The defaults are described as stringent on the webpage. Default parameters were used for the let-7 searches, with one exception: number of results to report (per genome) was changed from 1 to 100. This was done because Infernal's cmsearch tends to produce a lot of results with default parameters (and throughout this project default parameters were used when possible, just for the sake of simplicity) and there would thus be a larger set of points of comparison between the two programs. MiRNAminer is supposed to optionally return results via e-mail, but this did not appear to be working when the searches were run. It is simple enough to save the web results as an HTML file, though.



1 number or results to report

☑ Require seed conservation in mature miRNA (nt 2-8)

Target genomes (hold Ctrl and click to select more than one):

- Human (homo sapiens)
- Mouse (mus musculus)
- Rat (rattus norvegicus)
- Chimp (pan troglodytes)
- Dog (canis familiaris)

MicroRNA precursor sequence (**required**):
AAUGAUGGAGGUGCAGGCGUUUCCUGGGGAUUAAUGACCAGCUGGGAAGAACCAGUGGCCCUCGGCUCUGCCUCCCAGCCAGCCAUUAACUCCAAGGAAAUGUCUUUUGCUGAGGU

Mature microRNA sequence (**required**):
CCAGCUGGGAAGAACCAGUGGC

Your email address-miRNAminer will send results to that email (optional)

When you're ready, press Submit and wait for the result screen (search time depends on the number of requested results and the number of searched genom

Figure 5.1 Partial screenshot of the miRNAminer search screen.
http://groups.csail.mit.edu/pag/mirnaminer/

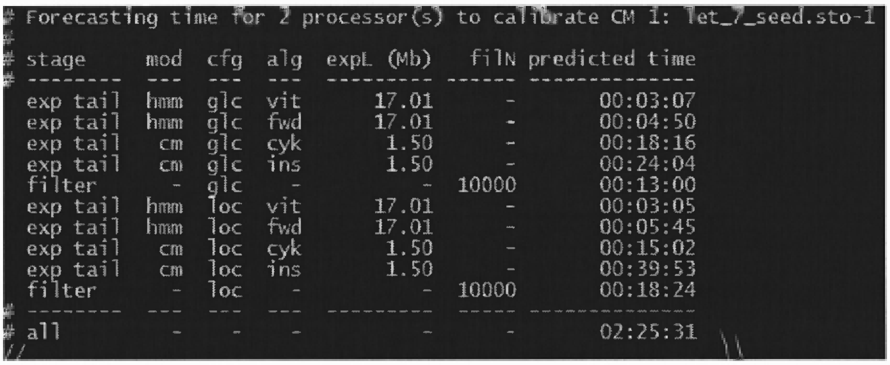
The miRNA searches using the Infernal suite illustrate the process of working with Infernal fairly well. The first step is creating a covariance model. Rfam has the covariance models that were used to assemble the full set of family members for each RNA family ready for download, but as of this writing they were all created with the .57 release of Infernal which are incompatible with Infernal 1.0 releases. (It is unclear what has changed as of the 1.0 releases that rendered pre-1.0 covariance model files unusable.) This project is using Infernal 1.0rc1, so a new let-7 CM needed to be created. Now cmbuild requires an alignment to be in Stockholm format to build a CM, which does appear to be something of a drawback, as databases that give results in Stockholm format do not appear to be common. For obvious reasons Rfam does have Stockholm alignments but it should be mentioned this is a potential limitation to the Infernal system. The Infernal suite is entirely command-line based so CM building and all searches were done from a Linux terminal in the Cygwin environment (which had been installed on a Windows Vista computer.) Cmbuild commands take the following form:

```
cmbuild [-options] <cmfile output> <alignment file>
```

Where <cmfile output> is a user-specified file and <alignment file> is the CM. Generally the output file should be of the form *.cm, but the Infernal programs don't need a correct filename to use a CM. Two CMs for the let-7 family were created with cmbuild: one using the seed alignment from Rfam, the other using the full alignment. Using two CMs was meant to test if there were any significant differences between using the curated seed alignment from Rfam and the full alignment which contains additional computationally predicted let-7 family members.

A recommended step in CM creation is calibration using cmcalibrate. This will allow a CM to generate E-values in addition to the bit scores cmsearch produces. The Infernal manual says that E-values are the preferred way to score a potential homologues

[36, p.40], and calibration is thus necessary if a CM is going to be used for searches. Calibration, though, is a lengthy process, taking a few hours on the dual-processor PC used in this project. There is an option to check a CM before it is calibrated to see how long the process will last. Calibration is lengthy but in addition to improving scoring it will reduce the time cmsearch takes to operate.



```
# Forecasting time for 2 processor(s) to calibrate CM 1: let_7_seed.sto-1
#
# stage      mod  cfg  alg  expL (Mb)   filN predicted time
# --------   ---  ---  ---  ---------  ----- --------------
  exp tail   hmm  glc  vit      17.01      -       00:03:07
  exp tail   hmm  glc  fwd      17.01      -       00:04:50
  exp tail    cm  glc  cyk       1.50      -       00:18:16
  exp tail    cm  glc  ins       1.50      -       00:24:04
  filter       -  glc    -          -  10000       00:13:00
  exp tail   hmm  loc  vit      17.01      -       00:03:05
  exp tail   hmm  loc  fwd      17.01      -       00:05:45
  exp tail    cm  loc  cyk       1.50      -       00:15:02
  exp tail    cm  loc  ins       1.50      -       00:39:53
  filter       -  loc    -          -  10000       00:18:24
# --------   ---  ---  ---  ---------  ----- --------------
# all          -    -    -          -      -       02:25:31
//
```

Figure 5.2 A cmcalibrate prediction screen.

As only two covariance models were calibrated the results of cmcalibrate's predicted times versus actual times may not be of the greatest value. They are noted for the record in Table 5.1.

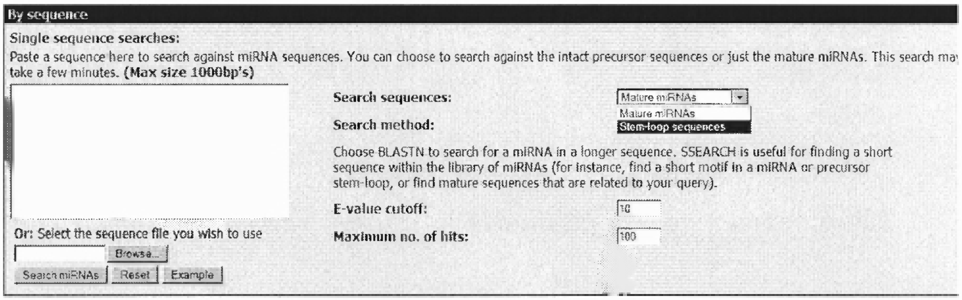
Table 5.1 Predicted Versus Actual Times For Cmcalibrate

Covariance Model	Predicted Time	Actual Time
Let-7 seed alignment	2:33:43	2:23:28
Let-7 full alignment	2:48:47	2:33:44

The next step is searching for homologues with the calibrated CM. Cmsearch produces quite a few hits on default settings, so searching lengthy stretches of genomic data would produce too many hits to sift through. There was also the time factor to consider, as CM searches tend to take significant time to run. A straight comparison with miRNAminer by searching for homologues on whole genomes, or even one genome, seemed untenable. Therefore, CM searching was done on the genomic regions that contained let-7 homologues, as identified in the miRNAminer results. MiRNAminer helped in this regard, as its results page includes links to the location of potential homologues at *ENSEMBL* and the UCSC Genome Browser. This made downloading the homologue-containing region a simple process. A 1000 bp flanking region was added on each side of the homologue identified by miRNAminer, and the subsequent block of sequence was copied into a text file for cmsearch to search. The attempt here was to create a balance between giving cmsearch a bit of a challenge for finding the homologues miRNAminer had identified, while at the same time minimizing the amount of data generated by cmsearch.

Now the default settings for cmsearch use a local alignment—that is, the algorithm allows only a part of the CM to match some subsequence of the data being searched [36, p.14]. This is considered to be a more sensitive search setting, since it does not take a match on the entire CM to produce a match. But "glocal" alignment, the other setting, may be the more accurate setting, as it attempts to match the entire CM with a subsequence. (The word glocal is used to differentiate from true global alignment, which would align the CM with the entire sequence, not just a subsequence. This would not be desirable when the sequence under investigation is thousands of bps long.) Thus, for the purposes of comparison cmsearch was ran twice, once with the glocal option turned on, and once under default settings (local alignment.) Four sets of results were thus produced using cmsearch, two for each of the calibrated CMs. (For full examples of miRNAminer and cmsearch output, please see Appendices A and B.)

In an attempt to confirm to some degree how well both Infernal and miRNAminer had done with their searches, miRBase was searched for their output. MiRBase has a sequence-based search option that uses BLASTN to find matches. This search option was used to test both programs' results, to determine that they were finding let-7 family members, and to identify potential differences between the two programs as well. The miRBase BLASTN search has a limit of 1000 bps, which was not an issue for the data produced by miRNAminer and cmsearch. It also gave the option of searching against either mature miRNAs or the stem-loop precursors. As the data produced was of the precursors the stem-loop option was chosen.



By sequence

Single sequence searches:

Paste a sequence here to search against miRNA sequences. You can choose to search against the intact precursor sequences or just the mature miRNAs. This search ma take a few minutes. **(Max size 1000bp's)**

Search sequences: Mature miRNAs

Search method:

Choose BLASTN to search for a miRNA in a longer sequence. SSEARCH is useful for finding a short sequence within the library of miRNAs (for instance, find a short motif in a miRNA or precursor stem-loop, or find mature sequences that are related to your query).

E-value cutoff: 10

Maximum no. of hits: 100

Or: Select the sequence file you wish to use

Figure 5.3 The miRBase search screen.
http://microrna.sanger.ac.uk/sequences/search.shtml

5.3 CMfinder and CARNAC

CMfinder and CARNAC are both secondary structure-predicting tools available both as a web version and as downloadable software. More importantly for the purpose of drawing comparisons between CM-using software and other methods, they both take multiple unaligned sequences as an input and attempt to return a common structure as the output.

The problem would be finding a dataset that would be accepted by both programs' web versions. CMfinder needs a minimum of four sequences and a maximum of about 60, each with a maximum length less that 500 bp. CARNAC needs two sequences at minimum, with no minimum bp restrictions (though each line of input sequence has a maximum of 80 characters, longer sequences would have to be entered as multiple lines.) Thanks to Rfam the problem of finding a dataset was minimized, as it contains information about average family length from the "Browse by family name" section of the database. (Curiously this information was removed from the corresponding section in Rfam 9.0 [63], but remains on the 8.1 release [62] of the website.)

In terms of this project there were additional concerns with the assembly of the dataset. As the background section explored the major types of RNA, the dataset should reflect those types (leaving out miRNA, which was explored in the previous section.) To that end examples from the following types of RNAs were located on Rfam: cis-regulatory elements, riboswitches, ribozymes, and snRNA/snoRNAs. Also located at Rfam were tRNA (there is only one entry for tRNA at Rfam, as the secondary structure does not vary among the tRNAs that correspond to codons) and the three types of rRNA. Not considered were mRNA, whose secondary structure is outside the scope of this project, and the siRNAs/shRNAs, where there does not appear to be a lot of information about, or perhaps interest in, their secondary structures (they have no Rfam entry, nor does mRNA.) The dataset was assembled on simple ground: Rfam searches for the desired type of RNA, and then the first family (ranked by accession number) that met the shared needs of CARNAC and CMfinder was chosen. This led to the final dataset of the antizyme RNA frameshifting stimulation element (Antizyme FSE, a cis-regulatory element), the PreQ1-I riboswitch (PreQ1), the hairpin ribozyme (Hairpin), and Pyrococcus C/D box snoRNA (Pyrococcus C/D). This was in addition to the three types of rRNA and tRNA. Finally—with the caveat that pseudoknots tend to pose a problem for secondary structure prediction programs—telomerase was added into the mix, with ciliate telomerase as the representative. Seed alignments were used. For those families with more than 60 members in the seed alignment (the upper bound for CMfinder) the first 60 members were used.

The web versions of CMfinder [60] and CARNAC [61] both have simple-to-use user interfaces. Lines of FASTA-formatted sequences can be directly pasted into a search box. Both programs also allow for the upload of sequence data from an external file. CMfinder, however, allows for greater customization of searches than CARNAC. CARNAC has three options: eliminate redundant sequences, take GC content into account, and allow isolated stems (with an accompanying warning that this may slow the processing of the sequences.) CMfinder allows the use to control numbers of stem-loops, motifs, and candidates, as well as the minimum and maximum length of the motifs and the expected fraction of sequences containing the motifs. It also runs two sets of configurations simultaneously on the dataset, so it produces two sets of results as well. Additional options are merging motifs and removing redundant motifs. Both tools were run with default parameters. For CARNAC this meant eliminating redundant sequences and taking GC content into account. For CMfinder this meant two configurations, each with three motifs, a minimum motif length of 30 and a maximum of 100, 40 candidates, and 0.8 for the expected fraction of sequences containing the motif. The only difference between the two configurations under the default parameters is that the first configuration uses one stem-loop, the second uses two. The option to merge motifs is also left checked in the default situation. It should also be noted that both CARNAC and CMfinder retain results on their servers for some time afterwards; CARNAC allows them to be retrieved with an identification number, while CMfinder automatically sends an e-mail with a link to results. The CARNAC results retrieval proved useful, as there appeared to be a problem with forwarding from the "results processing" screen to the results screen. Luckily entering the identification number proved that CARNAC had completed its task.

5.4 Pfold and Mfold

For the sake of convenience the dataset used with CMfinder and CARNAC was also used on Pfold and Mfold. The Pfold server [57] allows for no customization of searches, and FASTA sequences should be aligned before they are used with Pfold. The Mfold server [58] also has a very simple interface, but, as noted above, has a number of customization options. The difference between Mfold and Pfold (and also CARNAC) is that Mfold is a single-sequence RNA folder; it does not attempt to find a consensus sequence between sequences, nor does it allow more than one sequence on a single run. (There is a multiple sequence version of Mfold [59], but it also does not predict a consensus structure.)

Data collection was the same as with CMfinder and CARNAC: FASTA sequences found at Rfam. The only difference was in the number of sequences used. Pfold has an upper bound of 40 sequences with a maximum length of 500 bp, so for those Rfam families with more than 40 members the first 40 members were used. For Mfold the first member in each family was used.

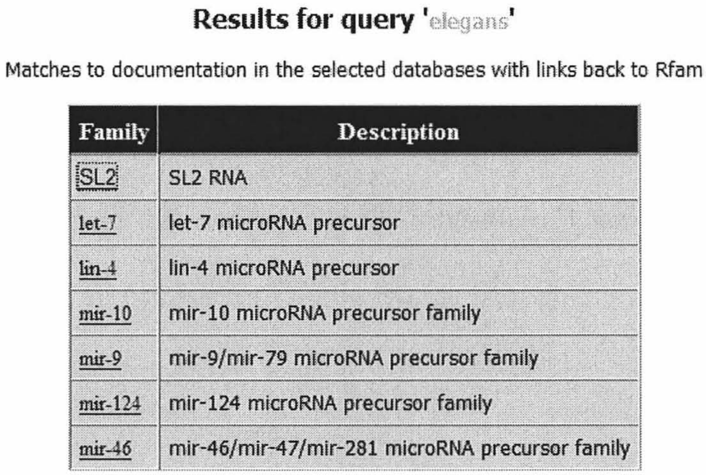
5.5 Infernal and CMfinder

There was some interest in comparing the CM-based methods of Infernal and CMfinder against each other. (Pfold was not used here as it is more of a visualization tool and does not generate CMs.) There were already a set of results involving Infernal from the miRNAminer comparisons. The simplest option seemed to be to run the CMs CMfinder generated on those same datasets. This necessitated an installation of an earlier version of Infernal, as CMfinder produced CMs in a pre-Infernal 1.0 release format, and on a different machine (a dual-processor laptop running Ubuntu Hardy Heron) as running two Infernal installations on the same machine could prove problematic. Infernal had to be used since CMfinder has no ability to use a CM on its own (which became problem with the intended comparison between it and CARNAC, as detailed in the Results section.) So the comparisons will basically illustrate the differences between the CMs cmbuild creates and those CMfinder creates.

5.6 Infernal (Via Rfam) and BLAT

A few tests were run comparing Infernal and BLAT. The interest here was to compare the Infernal suite's alignment abilities with a more traditional alignment tool. Now Rfam contains a number of alignments that start with a seed alignment, which is then used with cmalign to produce a full alignment [36, p.13]. So there are a number of Infernal-generated alignments already available for use. A decent point of comparison, then, was thought to be comparing these results with similar data and results found with BLAT.

Now BLAT searches on a number of genomes, but only one genome at a time (or at least that is a limitation of the version housed at the UCSC Genome Browser [52], which was used in this section.) In the interests of keeping the searches manageable it was decided to use *C. elegans* as the search example as it is a well-understood, simple organism that generated a manageable number of hits with a keyword search on Rfam (8.1 release.) It is also a genome available to BLAT. These turned out to be mostly a set of microRNA precursors, including the let-7 miRNA precursor used earlier (see Figure 5.4.)



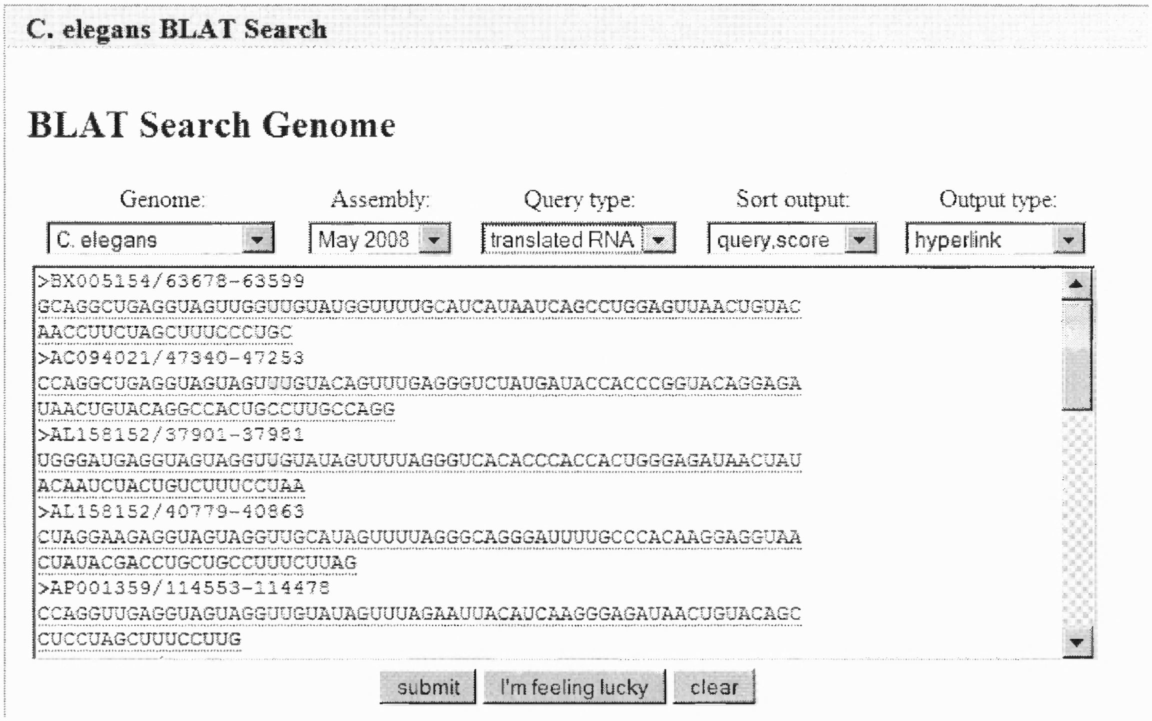
Results for query 'elegans'

Matches to documentation in the selected databases with links back to Rfam

Family	Description
SL2	SL2 RNA
let-7	let-7 microRNA precursor
lin-4	lin-4 microRNA precursor
mir-10	mir-10 microRNA precursor family
mir-9	mir-9/mir-79 microRNA precursor family
mir-124	mir-124 microRNA precursor family
mir-46	mir-46/mir-47/mir-281 microRNA precursor family

Figure 5.4 Rfam keyword search used to find BLAT dataset.

BLAT searches are simple from a user's standpoint. There are drop-down boxes to select the desired genome, the version of the genome, the query type (whether the input is DNA, RNA, or protein, or if BLAT is supposed to guess), sort order of the results and output type (hyperlink or pcl, which appears to be a type of printer format.) The default options for *C. elegans* were used: the May 2008 version of the genome, and query type guessed by BLAT. Searches were then redone with one difference: the query type was changed to translated RNA. This was done due to the fact that the number of results returned in the guessed queries turned out to be larger than the numbers found by the translated RNA queries, which was not expected. Results were given by hyperlinked accession numbers that led to links within the UCSC Genome Browser.



C. elegans BLAT Search

BLAT Search Genome

Genome: C. elegans
Assembly: May 2008
Query type: translated RNA
Sort output: query,score
Output type: hyperlink

```
>BX005154/63678-63599
GCAGGCUGAGGUAGUUGGUUGUAUGGUUUUGCAUCAUAAUCAGCCUGGAGUUAACUGUAC
AACCUUCUAGCUUUCCCUGC
>AC094021/47340-47253
CCAGGCUGAGGUAGUAGUUUGUACAGUUUGAGGGUCUAUGAUACCACCCGGUACAGGAGA
UAACUGUACAGGCCACUGCCUUGCCAGG
>AL158152/37901-37981
UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAU
ACAAUCUACUGUCUUUCCUAA
>AL158152/40779-40863
CUAGGAAGAGGUAGUAGGUUGCAUAGUUUUAGGGCAGGGAUUUUGCCCACAAGGAGGUAA
CUAUACGACCUGCUGCCUUUCUUAG
>AP001359/114553-114478
CCAGGUUGAGGUAGUAGGUUGUAUAGUUUAGAAUUACAUCAAGGGAGAUAACUGUACAGC
CUCCUAGCUUUCCUUG
```

Figure 5.5 BLAT search screen at the UCSC Genome Browser.
http://genome.ucsc.edu/cgi-bin/hgBlat

CHAPTER 6

RESULTS AND CONCLUSIONS

6.1 Introduction

This chapter contains the results of the various comparisons and tests run in the previous chapter. Additional analysis is provided to the extent possible. A summary is provided as well, though perhaps no grand conclusion can be drawn from these results.

6.2 Infernal and MiRNAminer

As noted above, miRNAminer and Infernal's cmsearch were used on a similar set of sequence data. Cmsearch is too slow for full-scale genomic searches so it was restricted to search on the areas in which miRNAminer had found potential miRNAs. The results are noted in Table 6.1, and are expressed in numbers of miRNAs found. Potential miRNAs for miRNAminer are simply those found with the default settings (altered slightly as noted in the Methods chapter.) For cmsearch the standard is the one recommended in the Infernal user's guide: anything with an E-value of ten or lower is significant enough to merit further investigation [36, p.40]. As noted above, four sets of cmsearch results were produced: a local set and a glocal set for each of the two covariance models (each representing the full and seed alignments of let-7.)

Table 6.1 MiRNAminer and Cmsearch Hits

Species	miRNAminer	Cmsearch seed	Cmsearch seed (glocal)	Cmsearch full	Cmsearch full (glocal)
C. elegans	1	3	3	3	3
Chicken	4	23	25	25	26
Chimp	3	15	10	15	10
Cow	3	5	6	3	5
Dog	3	15	12	14	8
Human	3	15	10	14	11
Mouse	3	16	15	14	14
Platypus	1	4	6	4	7
Possum	2	9	7	8	7
Rat	3	12	16	13	14
Rhesus	2	10	9	11	8
Total	28	127	119	124	113

Cmsearch did find a greater number of potential miRNAs than miRNAminer, at least four times as many versus the lowest cmsearch value (using the glocal setting on the full let-7 CM.) The glocal searches produced slightly fewer hits than the local searches, which ; the glocal search requires more subsequence to match to produce a hit than the local. Interestingly, the searches done using the CM generated by the full alignment of let-7 produced slightly fewer hits than the seed alignment let-7 CM. Perhaps the larger number of sequences produces a slightly more specialized CM. It should be noted that miRNAminer passed a very basic test of its competency immediately: it found let-7 in *C. elegans*, which of course was the specific let-7 used to search across the 11 genomes to which it has access.

There was some attempt to verify the results via miRBase sequence search (detailed in the Methods chapter.) Organisms selected were *C. elegans*, chicken, human, rat, and platypus. Results are below in Table 6.2. Note that a maximum hit means there were a large number of hits, more than could be efficiently counted. (The miRBase sequence search by default returns a hundred maximum hits, and this was unchanged.

Counting to a hundred did not seem to be useful, so when a large number of hits than could not be quickly eyeballed was produced, that search was recorded as a maximum hit.) For miRNAminer all the results given were used; for cmsearch just the results generated by the seed let-7 CM on local search settings were used. (This was done partially out of convenience, and also because the seed let-7 CM searches done on the local setting were the most sensitive, producing the greatest number of hits.)

Table 6.2 Summary of Cmsearch and MiRNAminer Results Verified on MiRBase

Species	MiRNAminer maximum hits	Cmsearch maximum hits on let-7	Cmsearch hits on let-7, less than maximum	Cmsearch hits on different miRNA families	Cmsearch zero hits
C. elegans	1	2	0	0	0
Chicken	4	14	0	5	4
Cow	3	0	0	3	0
Dog	3	8	1	2	4
Human	3	8	0	2	3
Mouse	3	8	0	4	4
Platypus	1	2	1	1	1
Possum	2	4	0	1	4
Rat	3	8	0	3	1

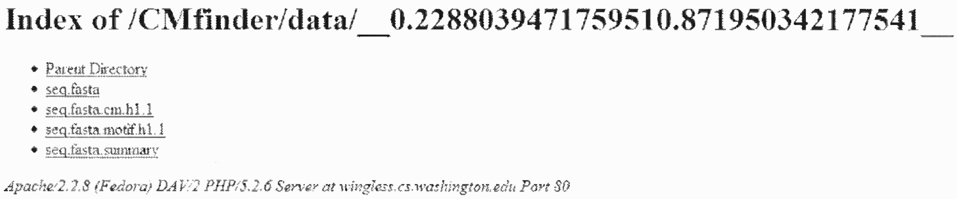
Table 6.2 records the number of incidences of a type of hit. Maximum hits were explained above. The "less than maximum" column indicates times where there was a let-7 hit that was less than maximum—a rare occurrence. The "hits on different miRNA families" indicates miRBase returned matches, just not ones from the let-7 family. The final column indicates a search that provided no hits in miRBase for a potential miRNA identified by cmsearch.

Cmsearch does appear to be more sensitive than miRNAminer, though the results are somewhat deceptive. Cmsearch searches the plus and minus strands of sequence data automatically (that is to say, it searches forwards and backwards) and the minus strand and plus strand results produced identical maximum hits. So the numbers in the maximum hits column for cmsearch are roughly double what they should be, considered fairly. But it still did manage to outperform miRNAminer with confirmed let-7 hits, though one wonders if miRNAminer could be slightly tweaked to produce similar results by changing scoring thresholds. More interesting, perhaps, was cmsearch's ability to find faint homologues of other families in the sequence data. This could be evidence of the strength of covariance models, their ability to detect distantly related sequences.

One curious note was in the cow data, where miRNAminer outperformed cmsearch. Indeed, cmsearch could not find any of the miRBase-confirmed let-7 families that miRNAminer identified. Perhaps there are odd flanking regions that somehow confuse cmsearch in the cow sequence data.

6.3 CMfinder, CARNAC, Pfold, and Mfold

CMfinder and CARNAC initially seemed like a good point of comparison, as their input parameters were similar and they seemed to work towards similar ends (as evidenced by the authors of CMfinder using CARNAC as one of their points of comparison [37].) However, comparing the outputs of the web versions of these two tools proved difficult. Simply, it is difficult to find a common ground between a tool that produces a human-interpretable output (an RNA molecule folded and represented by a jpg image) and a nonhuman-interpretable output (covariance models and—to a lesser degree—Stockholm formatted alignments.) A simple comparison is shown in Figure 6.1.



Index of /CMfinder/data/__0.2288039471759510.871950342177541__

- Parent Directory
- seq.fasta
- seq.fasta.cm.h1.1
- seq.fasta.motif.h1.1
- seq.fasta.summary

Apache/2.2.8 (Fedora) DAV/2 PHP/5.2.6 Server at wingless.cs.washington.edu Port 80

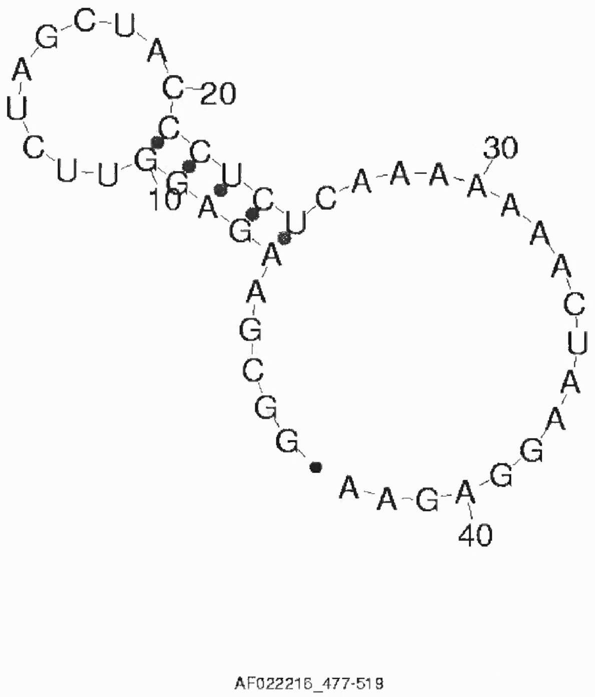


Figure 6.1 CMfinder and CARNAC output on an identical input (the seed sequences for the Hairpin ribozyme from Rfam.) CMfinder returned a CM and a Stockholm alignment. CARNAC returned a jpg image.

The CARNAC output is obvious. CMfinder returns results organized by number of stem loops. In Figure 6.1 the “1.1” in the two lines near the top indicates the number of stem loops in the input configuration and the number of stem loops in the CM: one in the input configuration and one in the output. The corresponding Stockholm alignment is numbered the same way. The Stockholm alignment contains the word motif in the output; the CM contains the letters “cm.”

CMfinder’s output was difficult to analyze due to the lack of human-interpretable results. There is a lack of tools capable of visually representing a Stockholm alignment. Scoring is not obvious either, as it is contained as extra lines within the Stockholm file, and there is little documentation on how to interpret the scores. A brief sample of the scoring lines is shown in Figure 6.2.

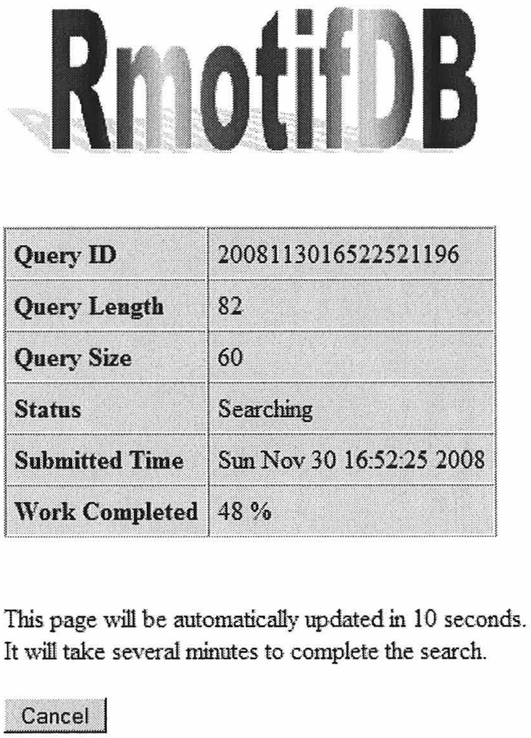
```
#=GS AF022216/477-519           WT    1.00
#=GS AF022216/537-579           WT    1.00
#=GS AP006627/2249570-2249527   WT    1.00
#=GS Z99111/27640-27684         WT    1.00
#=GS AAOX01000006/21305-21347   WT    1.00

#=GS           AF022216/477-519 DE      2.. 37  35.168510
#=GS           AF022216/537-579 DE      3.. 37  33.020340
#=GS    AP006627/2249570-2249527 DE      2.. 38  38.994225
#=GS         Z99111/27640-27684 DE      1.. 39  43.559181
#=GS    AAOX01000006/21305-21347 DE      1.. 37  38.745190
```

Figure 6.2 Sample CMfinder scoring.

The WT lines are the weight, and the DE lines are the start and end of the sequence, followed by the score. The meaning of the scores is not immediately obvious.

CARNAC for its own part also had issues, with some seed alignments producing no results at all. CMfinder always produced results but their meaning was unclear. It was thought that running the Stockholm alignments it generated through RmotifDB would at least allow some kind of verification of results.



Query ID	2008113016522521196
Query Length	82
Query Size	60
Status	Searching
Submitted Time	Sun Nov 30 16:52:25 2008
Work Completed	48 %

This page will be automatically updated in 10 seconds.
It will take several minutes to complete the search.

Figure 6.3 RmotifDB processing a Stockholm alignment.

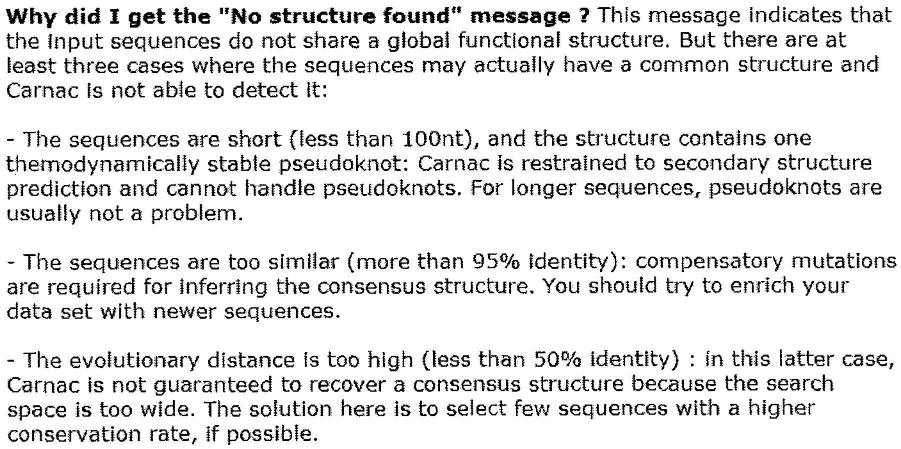
As it turned out, the CMfinder output was not able to be tested in any way via RmotifDB, as RmotifDB did not return any results for CMfinder output. The results are summarized in Table 6.3 along with the CARNAC results.

Table 6.3 Summary of CMfinder and CARNAC Results

Rfam family	CMfinder output (verified on RmotifDB)	CARNAC
5s rRNA	No	Yes
5.8s rRNA	No	No
Small subunit rRNA 5 domain	No	No
Antizyme FSE	No	Yes
PreQ1	No	Yes
Hairpin	No	Yes
Pyrococcus C/D	No	No
tRNA	No	No
Ciliate telomerase	No	Yes

Note: A Yes indicates output returned that was verified by RmotifDB in CMfinder. A no indicates no output, verified or otherwise.

Obviously both programs had problems with the Rfam seed sequences. When no results are returned for CARNAC the error message shown in Figure 6.4 is generated.



Why did I get the "No structure found" message ? This message indicates that the input sequences do not share a global functional structure. But there are at least three cases where the sequences may actually have a common structure and Carnac is not able to detect it:

- The sequences are short (less than 100nt), and the structure contains one themodynamically stable pseudoknot: Carnac is restrained to secondary structure prediction and cannot handle pseudoknots. For longer sequences, pseudoknots are usually not a problem.

- The sequences are too similar (more than 95% identity): compensatory mutations are required for inferring the consensus structure. You should try to enrich your data set with newer sequences.

- The evolutionary distance is too high (less than 50% identity) : in this latter case, Carnac is not guaranteed to recover a consensus structure because the search space is too wide. The solution here is to select few sequences with a higher conservation rate, if possible.

Figure 6.4 CARNAC "No structure found" message.

It is unclear which case from Figure 6.4 applies to the Rfam seed sequences used. The rRNAs, for example, have average shared identities of 78%, 61%, and 43% respectively [69], and are not pseudoknotted structures (though they are fairly complicated.) Antizyme FSE has an 87% identity, and is relatively short (57.6 bp for the average member) but, again, is not pseudoknotted. The lack of hits from RmotifDB from the CMfinder output is also puzzling, and those results will have to be considered solely in reference to Infernal in the section that follows this one. It should be noted that the CMfinder authors mention that the tool works best on unaligned input with "unrelated sequences, long flanking regions and/or low sequence similarity" [37, p.445] so perhaps the Rfam seed sequences were a bit too closely related to produce Stockholm alignments that RmotifDB could locate. It is also possible that the lack of flanking regions was an issue.

In any case, at this point it was decided to compare the CARNAC output with the output produced by Pfold and Mfold. Unfortunately, Pfold also does not allow for the easy visualization of results, as it returns structure in dot-bracket format that is split over multiple lines. There was insufficient time to locate a visualization program that would process both the dot-bracket structure and remove the gap characters (dashes) that Pfold inserts during the processing of results. A sample of Pfold output is shown in Figure 6.5.

```
           1                                                   50
Common     ....((((.( (.(....... ((....((.( .(.(.(.... ..........
X07545/505 --A--CCC-G G-C-C-AU-A GU----GG-C -C-G-G---G C---------
X07545/505 .....(((.( (.(....... ((....((.( .(.(.(.... ..........
M21086/8-1 --A--CCC-G G-C-C-AU-A GC----GG-C -C-G-G---G C---------
M21086/8-1 .....(((.( (.(....... ((....((.( .(.(.(.... ..........
U05019/544 --A--CCC-G G-U-C-AU-A GU----GA-G -C-G-G---G U---------
U05019/544 .....(((.( (.(....... ((....((.( .(.(.(.... ..........
X05870/304 --A--CCC-G G-C-C-AC-A GU----GA-G -C-G-G---G C---------
X05870/304 .....(((.( (.(....... ((....((.( .(.(.(.... ..........
X01588/5-1 --A--CCC-G G-U-C-AC-A GU----GA-G -C-G-G---G C---------
X01588/5-1 .....(((.( (.(....... ((....((.( .(.(.(.... ..........
M16530/8-1 ----ACCC-G G-C-A-AU-A GGC---GC-C -G-G-U---G C---------
M16530/8-1 .....(((.( (.(....... (((...((.( .(.(.(.... ..........
L27163/1-1 ---GUAGC-G G-C-C-AC-A GC----GG-U -G-G-G---G U---------
L27163/1-1 ...(((((.( (.(....... ((....((.( .(.(.(.... ..........
AF034619/5 -----GGC-G G-C-C-AC-A GC----GG-U -G-G-G---G U---------
AF034619/5 .....(((.( (.(....... ((....((.( .(.(.(.... ..........
L27170/1-1 --AGUGGU-G G-C-C-AU-A UC----GG-C -G-G-G---G U---------
L27170/1-1 ...(((((.( (.(....... .(....((.( .(.(.(.... ..........
L27168/1-1 ---UUGGC-G A-C-C-AU-A GC----GG-C -G-A-G---U G---------
L27168/1-1 ....((((.( ..(....... ((....((.( .(.(.(...( ..........
X03407/592 --UAAGGC-G G-C-C-AU-A GC----GG-U -G-G-G---G U---------
X03407/592 ....((((.( (.(....... ((....((.( .(.(.(.... ..........
L27343/3-1 ------GC-G G-C-C-AG-G GC----GG-A -G-G-G---G A---------
L27343/3-1 ......((.( (.(....... ((....((.( .(.(.(.... ..........
X72588/699 ------GC-G G-C-C-AC-A GC----GG-C -G-G-G---G C---------
X72588/699 .......... ..(....... ((....((.( .(.(.(.... ..........
X02128/24- -----GGC-G G-C-C-AG-A GC----GG-U -G-A-G---G U---------
X02128/24- .....(((.( (.(....... ((....((.( .(.(.(.... ..........
X14441/5-1 ----GGGC-G G-C-C-AG-A GC----GG-U -G-A-G---G U---------
X14441/5-1 ....((((.( (.(....... ((....((.( .(.(.(.... ..........
L27169/1-1 --GUAGGC-G G-C-C-AG-A GC----GG-U -A-G-G---G A---------
L27169/1-1 ..((((((.( (.(....... ((....((.( .(.(.(.... ..........
L27166/1-1 --GUAGGC-G G-C-C-AG-A GC----GG-U -A-G-G---G A---------
L27166/1-1 ..((((((.( (.(....... ((....((.( .(.(.(.... ..........
L27236/1-1 --GUAGGC-G G-C-C-AG-A GC----GG-U -A-G-G---G A---------
L27236/1-1 ..((((((.( (.(....... ((....((.( .(.(.(.... ..........
L27167/1-1 --GAAGGC-G G-C-C-AG-A GC----GG-U -G-G-G---G A---------
L27167/1-1 ....((((.( (.(....... ((....((.( .(.(.(.... ..........
L27162/2-1 --GCA-GC-G G-C-C-AU-A GC----GG-U -G-G-G---G C---------
L27162/2-1 ..(((.((.( (.(....... ((....((.( .(.(.(.... ..........
Rel.       ....****.* *.*.*.**.* **....**.* .*.*.+...- *.........
```

Figure 6.5 Partial Pfold output. This was obtained by an input of the seed sequences of 5s ribosomal RNA. Secondary structure is given in dot-parentheses format. A drawback is that the output is spread over several lines; the lines above are picked up again later in the output.

Mfold was also used, though at this point there was no reasonable point of comparison with a CM-using program, which was the general intent of this thesis. It did prove itself to be quite easy to use, as noted in the Methods chapter. It did a decent job folding RNA sequences relative to their known secondary structure. Compare the output shown in Figure 6.6 with the diagram of 5s rRNA in Figure 2.5.

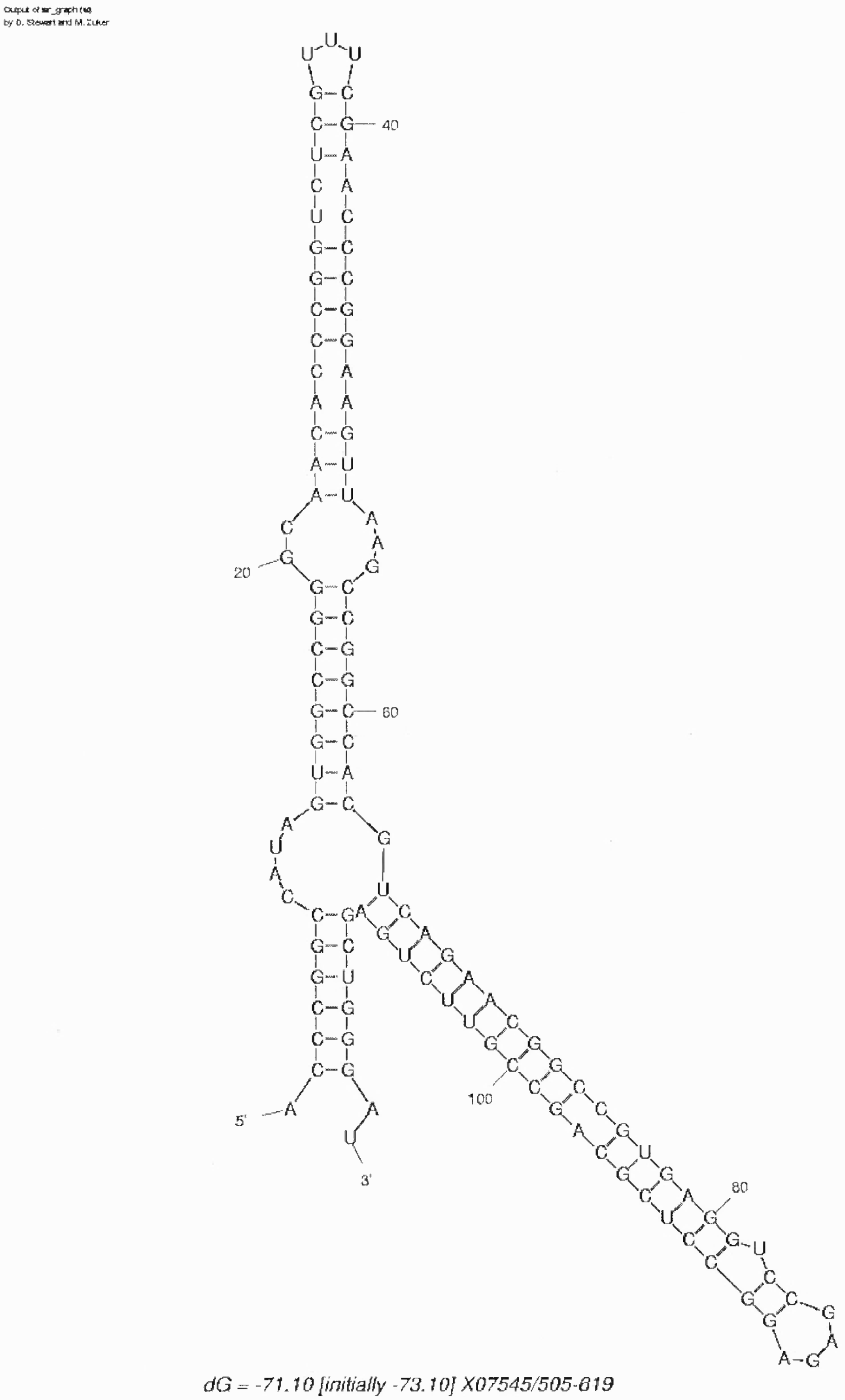

Output of sir_graph (©)
by D. Stewart and M. Zuker
40
20
60
100
5'
3'
80
dG = -71.10 [initially -73.10] X07545/505-619

Figure 6.6 Mfold output. This is of 5s rRNA. Mfold managed to predict three of the loops in the actual structure, though not the correct locations.

6.4 Infernal and CMfinder

As noted above, these results were done on a different machine from the rest of the thesis. The sequence data from the miRNAminer and cmsearch section was edited into a single file. (As the only thing being compared here was number of hits returned by Infernal and CMfinder CMs, this was thought to be a fair measuring stick. It would also not require overly long search time, as the resulting file was about 60,000 bp long.) Infernal 0.72 was installed, due to the aforementioned compatibility problems with 1.0 Infernal releases and pre-1.0 CMs. For the Infernal side of things CMs were downloaded from Rfam. The 8.1 release has the seed CMs available for download, and as Rfam was assembled before Infernal 1.0 the CMs are all in a pre-1.0 format (as it happens, they were created by Infernal 0.72.) CMs downloaded represented the same dataset introduced in Section 5.3.

However, CMfinder produces multiple CMs, and, as detailed in previous section, gives little guidance on which CM, if any, is a good result. The only information to distinguish one CM from another is the suffix. So CMs were chosen that seemed to be the closest to the known structure of the RNA family that generated them. For instance, 5s rRNA has a number of stem loops, so the CM whose suffix seemed to indicate the largest number of stem loops was chosen. Hairpin only has one stem loop, so the CM that ended in a one was chosen. This is obviously not the most rigorous method to determine the best choice of CM, but in the absence of explanatory information it seemed better than simply picking the first CM CMfinder had produced. Results are summarized in Table 6.4.

Table 6.4 Comparison of Infernal and Cmfinder Covariance Models

Rfam family	Hits using Infernal CMs	Hits using CMfinder CMs
5s rRNA	12	0
5.8s rRNA	4	0
Small subunit rRNA 5 domain	0	0
Antizyme FSE	71	0
PreQ1	221	1
Hairpin	148	9
Pyrococcus C/D	149	28
tRNA	100	23
Ciliate telomerase	10	0

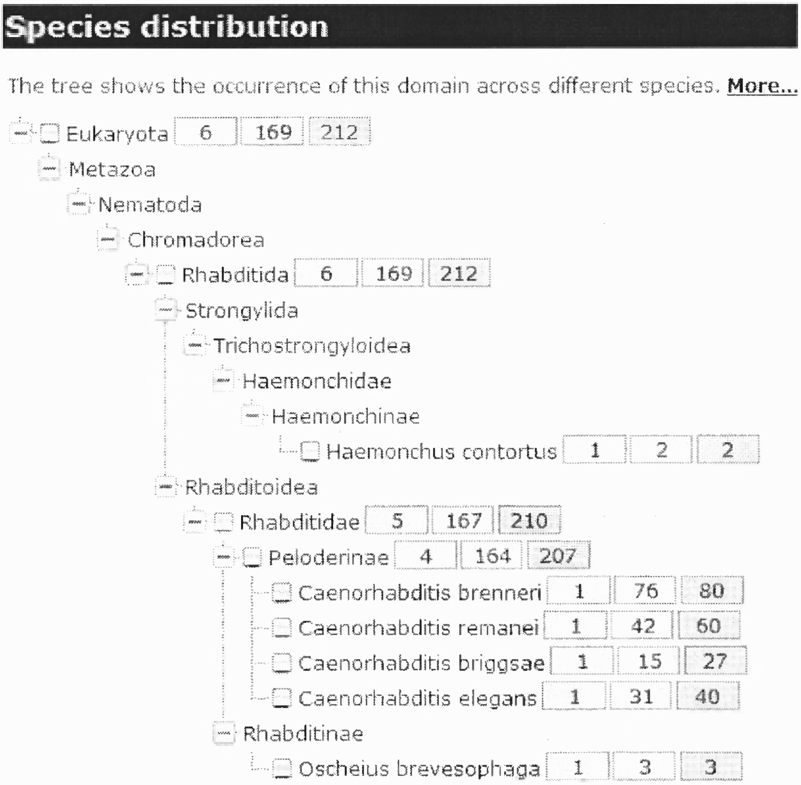
These results are offered more in the spirit of completeness than anything else, as it is difficult to be sure if they represent a fair judgment of CMfinder or not. Certainly the CMfinder CMs located many fewer hits than the Infernal CMs. And at least they generated some results, unlike the CMfinder Stockholm alignments that were used with RmotifDB. Perhaps CMfinder produces CMs of a more stringent type than Infernal.

An attempt was made to compare the two sets of CMs on a full bacterial genome, *C. psychrerythraea*. However there did not appear to be any way to write results to a file in Infernal 0.72. Thus hits were dumped into standard output, which froze the machine where Infernal 0.72 was installed. A full genome search would have been interesting but was not possible with the limited resources afforded this project (though perhaps this is good evidence of the computing requirements needed to search with CMs efficiently.)

Both the Infernal and CMfinder CMs did appear to be producing a good number of hits before the machine froze, though.

6.5 Infernal (Via Rfam) and BLAT

The intention of this section was to produce a clear comparison between a traditional primary alignment tool (like BLAT) with CM-generated data (like that found in Rfam.) But in retrospect it is unclear whether anything was accomplished here. The dataset mentioned in Section 5.6 was assembled. On the Rfam side of things this meant finding the number of times the families in question appeared in *C. elegans* data. Rfam has species trees with this information already recorded, as shown in Figure 6.7.



Species distribution

The tree shows the occurrence of this domain across different species. **More...**

- □ Eukaryota 6 169 212
  - Metazoa
    - Nematoda
      - Chromadorea
        - □ Rhabditida 6 169 212
          - Strongylida
            - Trichostrongyloidea
              - Haemonchidae
                - Haemonchinae
                  - □ Haemonchus contortus 1 2 2
          - Rhabditoidea
            - □ Rhabditidae 5 167 210
              - □ Peloderinae 4 164 207
                - □ Caenorhabditis brenneri 1 76 80
                - □ Caenorhabditis remanei 1 42 60
                - □ Caenorhabditis briggsae 1 15 27
                - □ Caenorhabditis elegans 1 31 40
              - Rhabditinae
                - □ Oscheius brevesophaga 1 3 3

Figure 6.7 An Rfam species tree. This one is for SL2 RNA.
http://rfam.sanger.ac.uk/family?acc=RF00199

The species trees were used to find the number of Rfam family members that came from *C. elegans*. The seed sequences for the dataset were also found at this time and ran through BLAT, using the "Blat's guess" setting and the translated RNA setting. The results, given in number of hits, are in Table 6.5.

Table 6.5 Blat Searches Versus Number of Members Per Family in Rfam

Rfam Family	Number of *C. elegans* members	Blat hits (using Blat's guess)	Blat hits (using translated RNA)
SL2 RNA	40	>100	71
let-7	2	5	1
lin-4	2	1	3
mir-10	1	2	2
mir-9	1	1	1
mir-124	1	16	2
mir-146	2	9	2

It is unclear what, if anything, was accomplished here. The degree of BLAT hits tends to be somewhat proportional to the number of Rfam family members. But this was probably not a valid way to check an Infernal alignment versus a primary alignment tool. The most interesting element here is the difference between BLAT's guesses and the translated RNA settings. One would think BLAT would always guess RNA, due to the presence of uridine in the sequence data, but that was not the case.

6.6 Summary and Epilogue

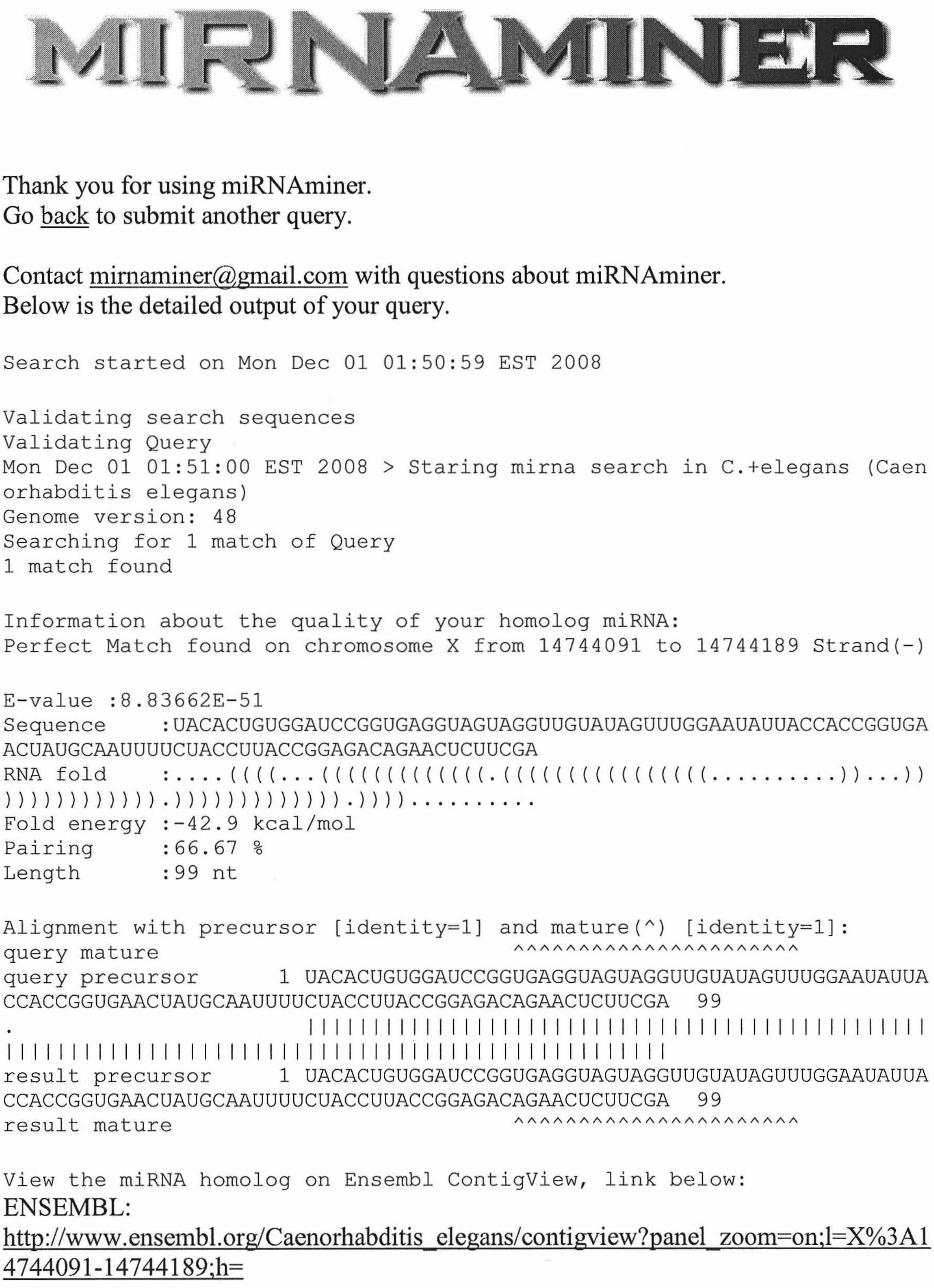
There was no single question answered in these results. It was more of a demonstration of the use of the software involved, with some possible points of comparison between tools. Probably the most successful comparison was that done between miRNAminer and cmsearch, which produced good results that were easy to compare and contrast. CMfinder and Infernal were also compared, though this was not done without issues. Comparing CMfinder with a non-CM-based tool proved very difficult, and a true test of Infernal versus a primary alignment tool was not found. If this thesis had to be rewritten, it would have focused more on miRNAminer and cmsearch, and found more room to explore the strengths and weaknesses of CMfinder and the Infernal suite. The authors of CMfinder have certainly had some success with it [70]; perhaps it performs better in their more structured pipeline setting.

More broadly, it is hoped that this thesis gave a relatively comprehensive introduction to the problem of RNA secondary structure detection and analysis, from both a theoretical and a practical standpoint.

APPENDIX A

SAMPLE MIRNAMINER OUTPUT

This appendix contains sample miRNAminer output. This particular example is the result of searching for the miRNA let-7 on the *C. elegans* genome.



Thank you for using miRNAminer.
Go back to submit another query.

Contact mirnaminer@gmail.com with questions about miRNAminer.
Below is the detailed output of your query.

```
Search started on Mon Dec 01 01:50:59 EST 2008

Validating search sequences
Validating Query
Mon Dec 01 01:51:00 EST 2008 > Staring mirna search in C.+elegans (Caen
orhabditis elegans)
Genome version: 48
Searching for 1 match of Query
1 match found

Information about the quality of your homolog miRNA:
Perfect Match found on chromosome X from 14744091 to 14744189 Strand(-)

E-value :8.83662E-51
Sequence     :UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGA
ACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA
RNA fold     :....((((...(((((((((((((.((((((((((((((((..........))...))
)))))))))))).)))))))))))))).)))).........
Fold energy :-42.9 kcal/mol
Pairing      :66.67 %
Length       :99 nt

Alignment with precursor [identity=1] and mature(^) [identity=1]:
query mature                           ^^^^^^^^^^^^^^^^^^^^^^
query precursor      1 UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUA
CCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA  99
.                      ||||||||||||||||||||||||||||||||||||||||||||||||||
|||||||||||||||||||||||||||||||||||||||||||||||||||
result precursor     1 UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUA
CCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA  99
result mature                          ^^^^^^^^^^^^^^^^^^^^^^

View the miRNA homolog on Ensembl ContigView, link below:
```

ENSEMBL:
http://www.ensembl.org/Caenorhabditis_elegans/contigview?panel_zoom=on;l=X%3A14744091-14744189;h=

```
View the miRNA homolog on UCSC browser, link below:
UCSC: http://genome.ucsc.edu/cgi-
bin/hgTracks?org=C.+elegans&position=chrX:14744091-14744189&miRNA=pack

-----------------------------------------------------------------------
----------------------------------------------------

Search finished at Mon Dec 01 01:51:02 EST 2008

Your search input:

Minimum Precursor Base Pairing=55
Flanking Length=50
Minimum Precursor Length=70
Minimum Alignment Identity With Mature=0.8
Maximum Precursor Length=180
Species searched in=cel,
Maximum Mismatches With Mature=3
MinPrecursorIden=56
CheckSeedConservation=true
debug=false
Your miRNA query=Query,UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUA
CCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA,UGAGGUAGUAGGUUGUAUA
GUU,
MaxAllowJointMatureAndLoop=4
Email sent to =
MaxPrecursorGaps=10
Maximum Blast E-value=0.05
Minimum Blast Alignment Length=18
Maximum Precursor Energy=-21.0
numOfResults=1
-----------------------------------------------------------------------
----------------------------------------------------
```

Figure A.1 Sample miRNAminer output.

APPENDIX B

SAMPLE CMSEARCH OUTPUT

This appendix contains sample cmsearch output. This was produced by the covariance model generated from the seed alignment of let-7, and ran on a stretch of *C. elegans* genome on default settings.

```
CM: let_7_seed.sto-1
>X

  Plus strand results:

 Query = 1 - 83, Target = 1019 - 1091
 Score = 34.88, E = 3.232e-09, P = 4.882e-12, GC =  42

           <<<<<<-<<<<<<<<<<<<<<<<<<<<<~~~~~~>>>>>>>>>>>>>>>>>>>>>>>>>>
         1 ccaGGaUgAGGuAGuAGguuGuauaGUu*[28]*aACuauaCaacCUaCUaCCUuuCCug 82
           :C:GG:  AGGUAG A  UUG AUAGUU      AACUAU CAA  U CUACCU :CC:G
      1019 UCCGGU-AAGGUAGAAAAUUGCAUAGUU*[19]*AACUAUACAACCUACUACCUCACCGG 1090

           >
        83 g 83
           :
      1091 A 1091


  Minus strand results:

 Query = 1 - 83, Target = 1091 - 1019
 Score = 55.59, E = 3.415e-15, P = 5.159e-18, GC =  42

           <<<<<<-<<<<<<<<<<<<<<<<<<<<<---<<<<_______>>>>---------->>>>
         1 ccaGGaUgAGGuAGuAGguuGuauaGUuuuagGGcuaaaauagCCcauuaGGAGAUaACu 60
           :C:GG: GAGGUAG A +UUG:AUAGUU   :GG +A A+UA CC:    GG G  AACU
      1091 UCCGGU-GAGGUAGUAGGUUGUAUAGUU---UGG-AAUAUUA-CCA--CCGGUG--AACU 1042

           >>>>>>>>>>>>>>>>>>>>>>>
        61 auaCaacCUaCUaCCUuuCCugg 83
           AU:CAA+ U CUACCUU:CC:G:
      1041 AUGCAAUUUUCUACCUUACCGGA 1019


 Query = 1 - 83, Target = 216 - 198
 Score = 5.85, E = 0.7728, P = 0.001167, GC =  58

           <<<<<<-<<<~~~~~~>>>>>>>>>
         1 ccaGGaUgAG*[64]*CUuuCCugg 83
            CAGG: GA:      :UU:CCUG
       216 CCAGGCAGAA*[ 0]*UUUGCCUGC 198


//
```

Figure B.1 Sample cmsearch output.

APPENDIX C

SAMPLE STOCKHOLM ALIGNMENT

This appendix contains a sample Stockholm alignment. This is one of seven produced by CMfinder after an input of the seed sequences of ciliate telomerase.

```
# STOCKHOLM 1.0
#=GF AU    Infernal 0.1

#=GS AF417612/228-392   WT    1.00
#=GS U22353/53-206      WT    1.00
#=GS AF399707/2181-2345 WT    1.00
#=GS U22354/216-379     WT    1.00
#=GS U22351/80-240      WT    1.00
#=GS U22352/508-657     WT    1.00
#=GS U22350/52-211      WT    1.00
#=GS U22349/199-360     WT    1.00
#=GS AF417610/218-380   WT    1.00
#=GS AF417611/281-444   WT    1.00
#=GS AF417609/192-355   WT    1.00

#=GS   AF417612/228-392 DE    115..159    54.588234
#=GS      U22353/53-206 DE    108..148    64.246918
#=GS AF399707/2181-2345 DE    114..156    69.430634
#=GS     U22354/216-379 DE    113..155    67.333237
#=GS      U22351/80-240 DE    110..152    55.426781
#=GS     U22352/508-657 DE    100..144    53.997509
#=GS      U22350/52-211 DE    112..154    67.994370
#=GS     U22349/199-360 DE    114..156    60.003937
#=GS   AF417610/218-380 DE    109..154    55.617577
#=GS   AF417611/281-444 DE    111..155    57.132687
#=GS   AF417609/192-355 DE    111..155    57.132687

AF417612/228-392
CAGACAUUC.GACA.UAAGAUACA.CUAUUUAUCuUAUG.GAaG.GUCUA
#=GR AF417612/228-392   SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
U22353/53-206                  AAGAC--
UC.GACA.UUUGAUACA.CUAUUUAUC.AAUG.GA.U.GUCUU
#=GR U22353/53-206      SS
.<<<<..<<...<<.<<.<<<<.......>>>>.>>>>.>>...>>>>.
AF399707/2181-2345
AAGACUAUC.GACA.UUUGAUACA.CUAUUUAUC.AAUG.GA.U.GUCUU
#=GR AF399707/2181-2345 SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
U22354/216-379
AAGACUAUC.GACA.UUUGAUACA.UUAUUUAUC.AAUG.GA.U.GUCUU
#=GR U22354/216-379     SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
U22351/80-240                  AAGGC-
AUC.GACA.UUUGAUACAaAUAUUGAUC.AAUG.GA.U.AUCUU
#=GR U22351/80-240      SS     .<<<-.<<<...<<.<<.<<<-........-
>>>.>>>>.>>.>.->>>.
U22352/508-657
UAGAAUUUC.GACA.UGUGGUACA.CUAUUUAUCuCAUG.GA.GaUUCUA
#=GR U22352/508-657     SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
U22350/52-211
AAGACUAUC.GACA.UUUGGUACA.CUAUUUAUC.AAUG.GA.U.GUCUU
```

```
#=GR U22350/52-211       SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
U22349/199-360                   AAGGC-
AUC.GACA.UUUGAUACAaAUAUUGAUC.AAUG.GA.U.GUCUU
#=GR U22349/199-360      SS      .<<<<.<<<...<<.<<.<<<-........-
>>>.>>>>.>>.>.>>>>.
AF417610/218-380
CAGAUACUCcGACU.UGUGAUACA.CUAUUUAUCaCAUGgGA.G.AUCUA
#=GR AF417610/218-380    SS      .<<<<.<<<...<-.<<.<<<<.......>>>>.>>-
>.>>.>.>>>>.
AF417611/281-444                 AAGAUACUCcGACGaUU-
GAUACA.AUAUUUAUC.AACGgGA.G.GUCUU
#=GR AF417611/281-444    SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
AF417609/192-355                 AAGAUACUCcGACGaUU-
GAUACA.AUAUUUAUC.AACGgGA.G.GUCUU
#=GR AF417609/192-355    SS
.<<<<.<<<...<<.<<.<<<<.......>>>>.>>>>.>>.>.>>>>.
#=GC SS_cons                     :<<<<-<<<.--<<.<<-
<<<<__.______>>>>.>>>>.>>.>.>>>>:
#=GC RF
AAGaCaCUC.GACa.UuUGaUaCA.CUAUUuAuC.aAuG.GA.G.GuCUU
//
```

Figure C.1 A Stockholm formatted alignment.

REFERENCES

1. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**:2079-2088.

2. Eddy SR: **Computational Analysis on RNAs.** *Cold Spring Harb Sym* 2006, **71**:117-128.

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

4. Durbin R, Eddy SR, Krogh A, Mitchison G: *Biological Sequence Analysis.* 11th edition. Cambridge: Cambridge University Press; 2006.

5. Gilbert W: **The RNA World.** *Nature* 1986: **319**:618.

6. Cech T. **Exploring the New RNA World.** [http://nobelprize.org/nobel_prizes/chemistry/articles/cech/index.html] Accessed on 12/3/08.

7. Crick F: **On Protein Synthesis.** *Sym Soc Exp Biol* 1958, **12**:139-163.

8. Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227**:561-563.

9. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell.* 4th edition. New York: Garland Science; 2002. [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.TOC&depth=2]

10. Bird A. **Perceptions of epigenetics.** *Nature* 2007, **447**:396-398.

11. Matzke MA, Birchler JA: **RNAi-mediated paths in the nucleus.** *Nat Rev Genetics* 2005, **6**:24-35.

12. Wassenegger M: **RNA-directed DNA methylation.** *Plant Mol Biol* 2000, **43**:203-220.

13. Aufsatz W, Mette MF, van der Winden J, Matzke AJ, Matzke MA: **RNA-directed DNA methylation in Arabidopsis.** *P Natl Acad Sci USA* 2002, **99**(Suppl 4):16499-16506.

14. Tang W, Luo XY, Sanmuels V: **Gene silencing: Double-stranded RNA mediated mRNA degradation and gene inactivation.** *Cell Research 2001,* **11**:181-186.

15. Coffin JM, Hughes SH, Varmus HE (Eds): **Reverse Transcriptase and the Generation of Retroviral DNA.** In *Retroviruses.* Cold Spring: Cold Spring Harbor Press; 1997. [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=rv.TOC] Accessed on 12/3/08.

16. Becker WM, Kleinsmith LJ, Hardin J: *The World of the Cell.* 5th edition. San Francisco: Benjamin Cummings; 2003.

17. Baxevanis AD, Ouellette BFF: *Bioinformatics: a practical guide to the analysis of genes and proteins.* 3rd edition. Hoboken: John Wiley & Sons; 2002.

18. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R: **Control of translation and mRNA degradation by miRNAs and siRNAs.** *Gene Dev* 2006, **20**:515-524.

19. Berg JM, Tymoczko JL, Stryer L, Clarke ND: *Biochemistry.* 5th edition. New York: W.H. Freeman & Company; 2002. [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=stryer.TOC&depth=2] Accessed on 12/3/08.

20. Brown, TA: *Genomes.* 2nd edition. Oxford: BIOS Scientific Publishers Ltd.; 2002. [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=genomes.TOC&depth=2] Accessed on 12/3/08.

21. Prasanth KV, Spector DL: **Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum.** *Gene Dev* 2007, **21**:11-42.

22. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S: **The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.** *Cell* 1983, **21**:849-857.

23. Tang J, Breaker RR: **Structural diversity of self-cleaving ribozymes.** *P Natl Acad Sci USA* 2000, **97**:5784-5789.

24. Miranda-Rios J: **The THI-box Riboswitch, or How RNA Binds Thiamin Pyrophosphate.** *Structure* 2007, **15**:259-265.

25. Gilbert SF: *Developmental Biology.* 6th edition. Sunderland: Sinauer Associates, Inc.; 2000. [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=dbio] Accessed on 12/3/08.

26. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75**:843-854.

27. Yekta S, Shih I, Bartel DP: **MicroRNA-Directed Cleavage of HOXB8 mRNA.** *Science* 2004, **304**:594-596.

28. Ronemus M, Martienssen R: **RNA interference: Methylation mystery.** *Nature* 2005, **433**:472-473.

29. Paddison PJ, Caudy AA, Bernstein E, Hannon GJ, Conklin DS: **Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells.** *Gene Dev* 2002, **16**:948-958.

30. Egloff S, Van Herreweghe E, Kiss T: **Regulation of Polymerase II Transcription by 7SK snRNA: Two Distinct RNA Elements Direct P-TEFb and HEXIM1 Binding.** *Mol Cell Biol* 2006, **26**:630-642.

31. Jady BE, Kiss T: **A small nucleolar guide RNA functions both in 2'-*O*-ribose methylation and pseudouridylation of the U5 spliceosomal RNA.** *Embo J* 2001, **20**:541-551.

32. Lukowiak AA, Narayanan A, Li ZH, Terns RM, Terns MP: **The snoRNA domain of vertebrate telomerase RNA functions to localize the RNA within the nucleus.** *RNA* 2001, **7**:1833-1844.

33. Rivas E, Eddy SR: **A dynamic programming algorithm for RNA structure prediction including pseudoknots.** *J Mol Biol* 1999, **285**:2053-2068.

34. Jones, NC, Pevzner PA: *An Introduction to Bioinformatics Algorithms.* Cambridge: MIT Press; 2004.

35. Hopcraft JE, Ullman JD: *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley; 1979.

36. The Eddy Lab: **INFERNAL User's Guide.** Version 1.0rc1. 2008. [http://infernal.janelia.org] Accessed on 8/1/08.

37. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder—a covariance model based RNA motif finding algorithm.** *Bioinformatics* 2006, 22:445-452.

38. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder 1.0 Manual.** 2005. [http://wingless.cs.washington.edu/CMfinder/manual.htm] Accessed on 12/3/08.

39. Knudsen B, Hein J: **RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.** *Bioinformatics* 1999, **15**:446-454.

40. Knudsen B, Hein J. **Pfold: RNA secondary structure prediction using stochastic context-free grammars.** *Nucleic Acids Res* 2003, 31:3423-3428.

41. Artzi A, Kiezun A, Shomron S: **MiRNAminer: a tool for homologous microRNA gene search.** *BMC Bioinformatics* 2008, **9**:39.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **285**:2053-2068.

43. Kent WJ: **BLAT—The BLAST-Like Alignment Tool.** *Genome Res* 2002, **12**:656-664.

44. Touzet H, Perriquet O: **CARNAC: folding families of related RNAs.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W142-W145.

45. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res*, 2003, 31:3406-3415.

46. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure.** *J Mol Biol* 1999, **288**:911-940.

47. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33**(Database Issue):D121-D124.

48. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32**(Database Issue):D109-D111.

49. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **MiRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, 34(Database Issue):D140-D144.

50. Griffiths-Jones S, Saini HK, Bateman A, Enright AJ: **MiRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36**(Database Issue):D154-D158.

51. The MIT/ICBP siRNA Database. [http://web.mit.edu/sirna/index.html] Accessed on 12/3/08.

52. Fire A, Xu SQ, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans.*** Nature 1998, **391**:806-811.

53. Andronescu M, Bereg V, Hoos HH, Condon A: **RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database.** BMC Bioinformatics 2008, **9**:340.

54. UCSC Genome Browser Home. [http://genome.ucsc.edu/] Accessed on 12/3/08.

55. Ensembl Genome Browser. [http://www.Ensembl.org/index.html] Accessed on 12/3/08.

56. NCBI Genome. [http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome] Accessed on 12/3/08.

57. The Pfold server. [http://www.daimi.au.dk/~compbio/pfold/] Accessed on 12/3/08.

58. The Mfold server. [http://www.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi] Accessed on 12/3/08.

59. The DINAMelt Server. [http://dinamelt.bioinfo.rpi.edu/quikfold.php] Accessed on 12/3/08.

60. CMfinder 1.0 Web Server. [http://wingless.cs.washington.edu/htbin-post/unrestricted/CMfinderWeb/CMfinderInput.pl] Accessed on 12/3/08.

61. CARNAC web server. [http://bioinfo.lifl.fr/RNA/carnac/carnac.php] Accessed on 12/3/08.

62. Rfam 9.0. [http://rfam.janelia.org/] and [http://rfam.sanger.ac.uk/] Accessed on 12/3/08.

63. Rfam 8.1. [http://www.sanger.ac.uk/Software/Rfam/] Accessed on 12/3/08.

64. BLAT search at the UCSC Genome Browser. [http://genome.ucsc.edu/cgi-bin/hgBlat] Accessed on 12/3/08.

65. Wang JTL, Wen D, Shapiro BA, Herbert KG, Li J, Ghosh K: **Toward an Integrated RNA Motif Database.** In *Proceedings of the 4th International Workshop on Data Integration in the Life Sciences (DILS 2007)*: 27-29 June 2007; Philadelphia. Edited by Istrail S, Pevzner P, Waterman M. Heidelberg: Springer-Berlin; 2007:27-36.

66. RmotifDB: A Database of RNA Structural Motifs. [http://datalab.njit.edu/bioinfo/index.html] Accessed on 12/3/08.

67. RmotifDB help page. [http://datalab.njit.edu/bioinfo/help.html] Accessed on 12/3/08.

68. MiRBase: the home of microRNA data. [http://microrna.sanger.ac.uk/sequences/index.shtml] Accessed on 12/3/08.

69. Browse by family name at Rfam 8.1. [http://www.sanger.ac.uk/Software/Rfam/browse/index.shtml] Accessed on 12/3/08.

70. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL, Breaker RR: **Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.** *Nucleic Acids Res*, 2007, **Advance Access**:1-11.