Use of RNA Secondary Structure for Evolutionary Relationships: Investigating RNase P and RNase MRP

A thesis presented in partial fulfilment of the requirements

For the degree of

Master of Science in Genetics

At Massey University

New Zealand

Lesley Joan Collins

1998

"Science teaches us about the deepest issues of origins, natures, and fates – of our species, of life, of our planet, of the Universe. For the first time in human history, we are able to secure a real understanding of some of these matters. Every culture on Earth has addressed such issues and valued their importance. All of us feel goosebumps when we approach these grand questions. In the long run, the greatest gift of science may be in teaching us, in ways no other human endeavour has been able, something about our cosmic context, about where, when, and who we are."

Carl Sagan - The Demon-Haunted World.

Amendments:

Page 81: Top lines of page should read - The RNAstructure tree (Figure 5.2C) groups the chloroplast sequences together but does not group the cyanobacterial species (*Synechocystis*, *Anabaena* and *Anacystis*) together unless the E. coli outgroup is removed.

Page 126: The following reference should be included:

Pascual, A. and Vioque, A. (1996) Cloning, purification and characterisation of the protein subunit of ribonuclease P from the cyanobacterium Synechocystis sp. PCC 6803. Eur J Biochem 241: 17-24

Abstract

Bioinformatics is applied here to examine whether RNA secondary structure data can reflect distant evolutionary relationships. This is important when there is little confidence in sequence data such as when looking at the evolution of RNase MRP (MRP).

RNase P (P) and RNase MRP (MRP) are ribonucleoproteins (RNPs) that are involved in RNA processing and due to functional and secondary structure similarities, are thought to be evolutionary related. P activity is found in all cells, and fits the criteria for inclusion in the RNA world (Jeffares et al. 1998). MRP is found only in eukaryotes with essential functions in both the nucleus and mitochondria. The RNA components of P and MRP (pRNA and mrpRNA) cannot be aligned with any certainty, which leads to a lack of confidence in any phylogenetic trees constructed from them.

If MRP evolved from P only in eukaryotes then it is an exception to the general process of the transfer of catalytic activity from RNA, to ribonucleoproteins, to proteins (Jeffares et al. 1998). An alternative possibility that MRP evolved with P in the RNA world (and has since been lost from all but the eukaryotes) is raised and examined. Quantitative comparisons of the pRNA and mrpRNA biological secondary structures have found that the third possibility of an organellar origin of MRP is unlikely.

Results show that biological secondary structure can be used in the evaluation of an evolutionary relatedness between MRP and P and may be extended to other catalytic RNA molecules. Although there are many protein families, this may be the first evidence of the existence of a family of RNA molecules, although it would be a very small family.

Secondary structures derived with folding programs from pRNA and mrpRNA sequences are examined for use in the characterisation of catalytic RNA sequences. The high AT content in organellar genomes may hinder the identification of their catalytic RNA sequences. A search strategy is developed here to address this problem and is used to identify putative pRNA sequences in the chloroplast genomes of four green plants. A maize chloroplast pRNA-like sequence is examined in more detail and shows many characteristics seen in known pRNA sequences. Folding programs show some potential for the characterisation of possible catalytic RNA sequences with only a small bias in the results due to sequence length and AT content.

Acknowledgments

Many, many thanks must go to David Penny for his patience and long hours in the air reading this thesis. I appreciate all the work that has gone into getting this bench jockey to wonder into the (RNA) world of theoretical science.

Special thanks to Vince Moulton (the ever travelling mathematician) who was game to team up with the crazy biologist and put some 'real' data into DCA. Thanks also to Soeren Perrey for teaching me the basics of DCA and Unix (a language even stranger than Klingon). Thanks also goes to Robert Pointon for writing all of those wonderful little programs that made life so much easier.

Thanks must also go to the inhabitants of the Boffin lounge for support and occasional coffee. Thanks also to the inhabitants of the BN lab at the NZDRI, for the coffee and time, especially during the writing of this thesis. Many thanks to all my family for their support over the years.

A great, great many thanks must go to my husband Maurice whose unwavering support over the last few years has gotten me through this. I could not have done this without you.

Finally, to everybody who is adventurous enough to work on the fringe... **Qaplah!**

(Klingon for Success)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bioinformatics, a new and exciting field in the biological sciences, is a powerful tool in the investigation of evolutionary relationships. Bioinformatics is applied here to examine two themes. Firstly, RNA secondary structure data is shown to reflect evolutionary relationships where the times of divergence are so old that there is little confidence in sequence data. Secondly, this secondary structure data is combined with sequence and functional data to examine the evolution of RNase MRP (MRP), especially the possibility of it being part of the RNA world.

RNase P (P) is already thought to be part of the RNA world, an early stage in the evolution of life, where RNA was both catalytic and the holder of the genetic information (Jeffares et al. 1998). MRP is thought to be evolutionary related to P due to functional and secondary structure similarities, but due to its presence only in eukaryotes, has not previously been considered to be part of the RNA world. These ribonucleoproteins (consisting of a catalytic RNA and at least one protein subunit) have RNA components (pRNA and mrpRNA) with little sequence homology, resulting in sequence alignments that have not enough reliability to confidently examine their evolutionary relatedness (Sbisà et al. 1996).

P cleaves tRNA precursors to form the mature 5' ends of tRNA molecules with activity being found all cells tested (i.e. universally) including prokaryotes, eukaryotes and also in organelles. Prokaryotic P consists of an RNA strand, and a single protein subunit, whereas the P encoded in the nucleus of eukaryotes has several protein subunits (Pace and Smith 1990). Fungi such as *Saccharomyces cerevisiae* and *Aspergillus nidulans* have retained their mitochondrial -encoded pRNA whereas vertebrate mitochondria and the fission yeast *Schizosaccharomyces pombe* have lost their pRNA gene and use a nuclear-encoded product. In plants, mitochondrial pRNA activity has been shown (Marchfelder and Brennicke 1993), but to date no genes have been characterised.

The secondary structure of prokaryotic pRNA has been seen in the past to show characteristic features for different phylogenetic groups of pRNA (Pace and Brown 1995) and consensus structures have been drawn for these groups of eubacteria and archaebacteria (Haas et al. 1996, Pace and Brown 1995). This is an indication that some features in the pRNA secondary structure are fixed and others variable. For the

purposes of this study, prokaryotic pRNA includes that from eubacteria mitochondria, and plastids (chloroplast and cyanelle). The pRNA from archaebacteria is not covered at this time due to processing power and time considerations.

MRP (Mitochondrial Ribosomal Processing) has been found only in eukaryotes initially as an endoribonuclease that cleaves RNA primers for the initiation of mitochondrial DNA replication (Morrissey and Tollervey 1995). Subsequently a nuclear function in rRNA processing was identified, consistent with its predominant localisation to the nucleolus (Lygerou et al. 1996). MRP consists of an RNA moiety and multiple protein subunits with at least 7 of these, Pop1p (Morrissey and Tollervey 1995), Pop3p (Dichtl and Tollervey 1997) Pop4p (Chu et al. 1997), Pop5p, Pop6p, Pop7p and Pop8p (Chamberlain et al. 1998) proteins being shared with P in the yeast *Saccharomyces cerevisiae*. It is possible that these proteins have structural characteristics that allow them to interact with both mrpRNA and pRNA. mrpRNA secondary structures (Schmitt et al. 1993) have only been characterised for eight species and show great similarity with each other despite being from plant, yeast and vertebrate species. The nucleotide sequences of these mrpRNAs vary greatly in length and nucleotide composition, making alignment of all eight sequences difficult.

Characteristics of MRP, eubacterial, eukaryotic and organellar P are summarised in Table 1.1. Cartoon representations and biological secondary structures of pRNA and mrpRNA show the sharing of some proteins between mrpRNA and the eukaryotic pRNA and the conserved presence of the pseudoknot pairing regions (Figure 1.1).

Comparisons of the RNA secondary structures between mrpRNA and pRNA have shown similarity in shape, especially in the 'cage region' of the RNA molecule in which there is the characteristic pseudoknot formation (Forster and Altman 1990). (Pseudoknots are structural elements that may act as a recognition site for proteins involved in replication initiation or translational regulation. The NMR structure of the classical pseudoknot has been determined (Kolk et al. 1998).) However, to date, there has been no published quantitative comparison of pRNA and mrpRNA secondary structure. When pRNA and mrpRNA secondary structures are broken down into simplified structures it can be seen that a large proportion of the secondary structure is shared between these two RNA molecules (Figure 1.2).

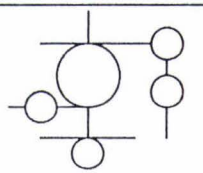
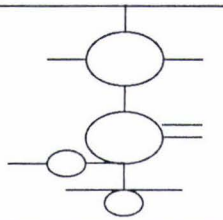
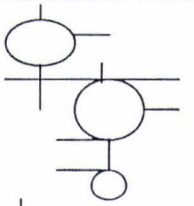
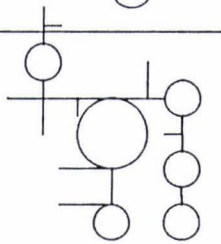
	Activity	Complex	Reaction Catalysed	Encoded	RNA transport	RNA structure (Simplified)	Comments
RNase MRP	Nucleolus, Mitochondria	RNA + Protein More than one protein subunit. In *S. cerevisiae* POP-1 and SMN1 identified.	rRNA processing in nucleolus, Cleaves RNA primers in mitochondria.	Nucleus	To the nucleolus. To the mitochondria.		Pop-1, Pop3 and Pop4 proteins shared with *S. cerevisiae* P. SMN1 unique to MRP.
Eukaryotic RNase P	Nucleus (Can also be found in mitochondria and chloroplasts)	RNA + Protein More than one protein subunit involved.	Cleaves pre-tRNAs to form mature tRNAs	Nucleus	To the mitochondria. Stays within the nucleus.		Many mammalian pRNA sequences in the databases, but none from the 'lower' eukaryotes and the amitochondrial eukaryotes as yet.
Eubacterial RNase P	Cell	RNA + Protein (RNA can be catalytic on its own). One protein in the complex.	Cleaves pre-tRNAs to form mature tRNAs	Chromosome	Within the cell.		pRNA's from many eubacterial species isolated but only *E. coli* RNase P studied in detail.
Mitochondrial RNase P	Mitochondria	RNA + Protein Protein is nuclear encoded. Unsure of how many subunits involved	Cleaves pre-tRNAs to form mature tRNAs	Vertebrates and *S. pombe* use nuclear encoded gene. Other yeasts, plants and *Reclinomonas americana* encode a mitochondrial gene.	To Mitochondria. Within the mitochondria.		RNA structure is much like that of the bacterium *Rhodospirillum*. Is highly A-U rich and very variable in size.

Table 1.1: Summary of characteristics and simplified secondary structure diagrams of MRP, eubacterial, eukaryotic and organellar P.

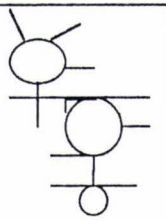
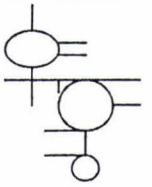
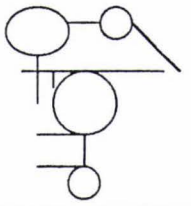
Activity	Where the Activity is found	Complex	Reaction Catalysed	RNA encoded	RNA localisation	RNA structure (Simplified)	Comments
Chloroplast RNase P	Chloroplast	RNA + protein Chloroplast RNA from *Porphyra purpurea* has been sequenced.	Cleaves pre-tRNAs to form mature tRNAs	Chloroplast	Chloroplast		Only sequence found so far is in the *Porphyra purpurea* chloroplast.
Cyanelle RNase P (Cyanophora paradoxa)	Cyanelle	RNA + Protein Thought to be one protein in complex. Eubacterial-like.	Cleaves pre-tRNAs to form mature tRNAs	Cyanelle genome.	Within the cyanelle.		RNA structure is very similar to that of the cyanobacteria.
Archaeal RNase P	Cell	RNA + Protein Thought to be one protein in complex. Eubacterial-like.	Cleaves pre-tRNAs to form mature tRNAs	Cell	Within the cell.		RNA is considered eubacterial-like.

Table 1.1 continued: Summary of characteristics and simplified secondary structure diagrams of MRP, eubacterial, eukaryotic and organellar P.

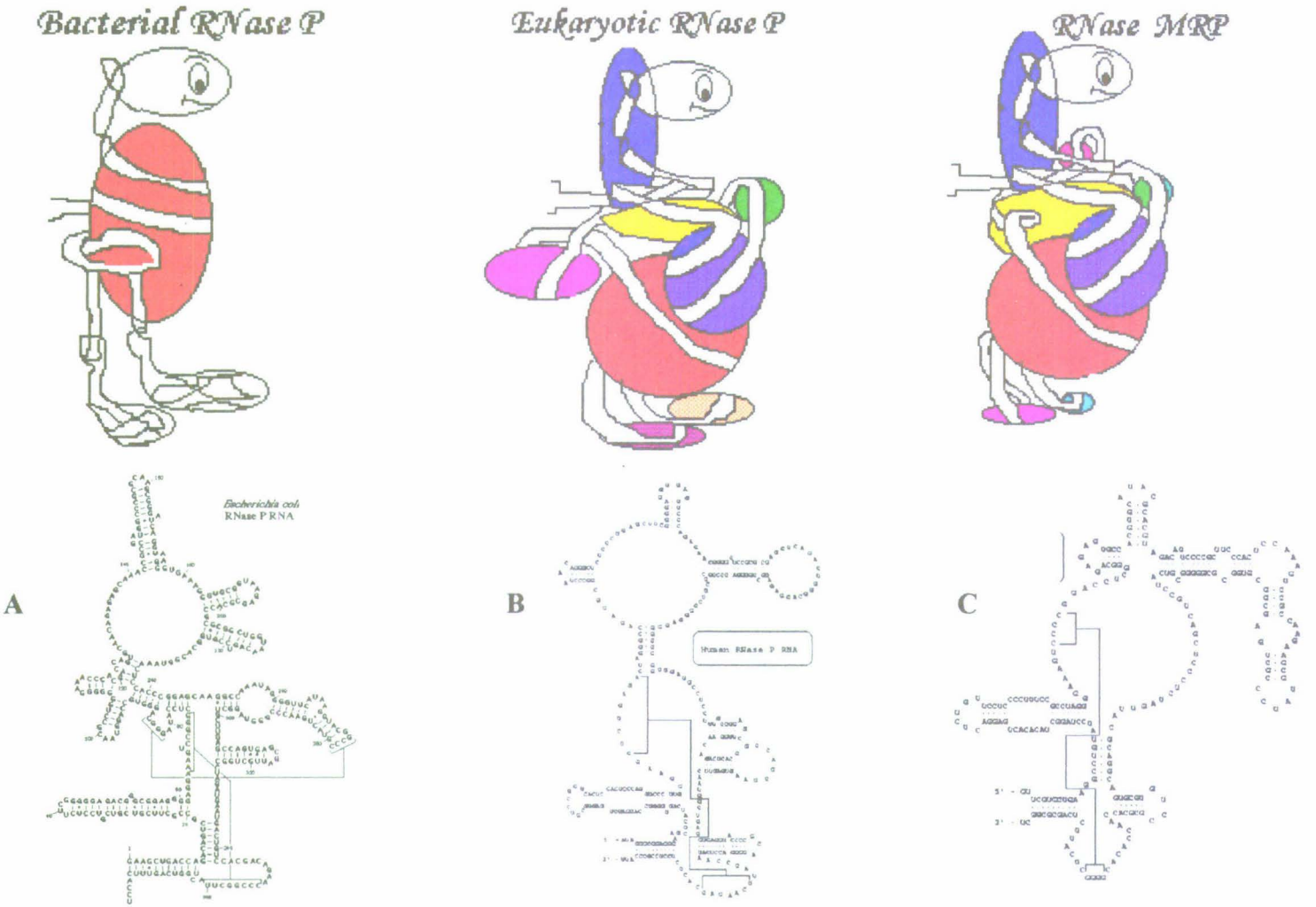


Figure 1.1. Cartoon Representation and RNA biological secondary structure diagrams.
A: *E. coli* pRNA (From the RNase P Database, Brown 1998) associating with one protein, **B:** Human nuclear pRNA (Redrawn from Altman et al. 1993) associating with multiple proteins, **C:** Human RNase mrpRNA (Redrawn from Schmitt et al. 1993) also associating with multiple proteins, some of which are shared with P. The solid lines indicate the long range pairing important in tertiary structure formation (including the Pseudoknot formation).

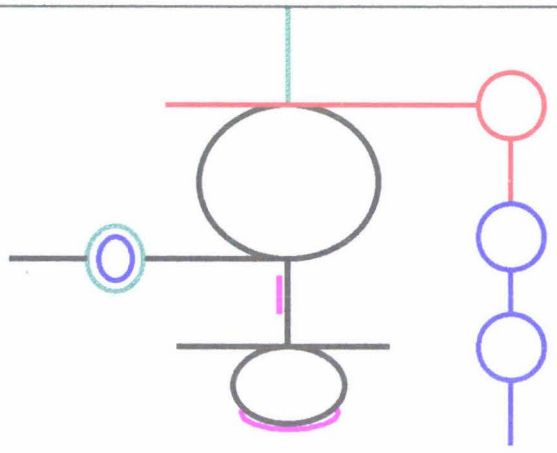


Figure 1.2: Simplified secondary structure of mrpRNA showing features similar to that of eukaryotic, eubacterial and mitochondrial pRNA.
Key:

Features common to mrpRNA and pRNA

Features common to mrpRNA and eubacterial pRNA.

Features common to mrpRNA and eukaryotic pRNA.

Features common to mrpRNA and organellar pRNA.

Features forming the pseudoknot region.

The secondary structure and functional similarities between MRP and P have led to the conclusion that these two ribonucleoproteins (RNP's) are evolutionary related (Morrissey and Tollervey 1995). Both the P and MRP ribozymes cleave RNA's to generate 5' phosphate and 3' hydroxyl termini in a reaction requiring divalent cations (Forster and Altman 1990). They are both sensitive to puromycin, an antibiotic which inhibits pre-tRNA processing (Potuschak et al. 1993), and enzymatic activities from P and MRP isolated from several organisms cofractionate through multiple stages of biochemical purification (Paluh and Clayton 1995). It has been reported that MRP and P may be involved together in a macromolecular complex within the nucleolus (Lee et al. 1996). A contrary theory, however, is that the relationship between MRP and P may be of a functional nature based on their sharing of many protein subunits (Sbisà et al. 1996).

This study investigated three general hypotheses, based on functional characteristics, of the relatedness of P and MRP. pRNA, mrpRNA and 16S rRNA sequences and secondary structures used in this study, are shown in Table 1.2.

The three groups of hypotheses are as follows:

I MRP evolved from an eukaryotic nuclear P in the nucleus of the eukaryotic cell. This could occur by gene duplication followed by divergence of function of the two homologues. This is the theory most commonly suggested in previous studies (Morrissey and Tollervey 1995, Reddy and Shimba 1996, Chamberlain et al. 1996). MRP would have been incorporated into multiple eukaryotic functions and has also gained an essential function in mitochondria. Under this hypothesis MRP is found only in eukaryotes because it was never in any of the other lineages! MRP is present in animals, yeasts, and plants indicating an early divergence from P; however, MRP need not have been present in all early eukaryotes. We would expect under this hypothesis the secondary structures of the mrpRNA to be more similar to eukaryotic pRNA than to prokaryotic pRNA.

Under this hypothesis MRP is an exception to the transfer process of catalysis (RNA to RNP to protein) (Jeffares et al. 1998) with a ribonucleoprotein taking on a new catalytic function after the widespread availability of protein catalysts.

II MRP evolved from an endosymbiont P. MRP could have evolved from the hypothetical endosymbiotic fusion that formed the first eukaryote (Gupta and Golding 1996) or by some later endosymbiosis that led to the mitochondrion. The endosymbiotic origin theory accounts for the essential mitochondrial function of MRP. It has been shown that organellar DNA can be transferred to the nucleus and yet retain a function in the organelle (Brennicke et al. 1993, Wischmann and Schuster 1995, Blanchard and Schmidt 1995). This theory proposes that MRP picked up the additional rRNA processing functions in the nucleus. We might expect here that mrpRNA would retain some organellar characteristics such as a higher A + T content in nucleotide sequence and be more closely related in secondary structure to that of the organellar or prokaryotic pRNA.

III MRP and P evolved in the RNA world. The RNA world hypothesis suggests that DNA and proteins evolved from a world in which RNA was the both the catalytic and information storage molecule, and that today's catalytic RNA species are molecular relics from this time. There are three main criteria used to evaluate the

antiquity of an RNA molecule (Jeffares et al. 1998) and pRNA fits all three of these criteria by being ubiquitous, catalytic and central to metabolism. MRP on the other hand fits only the last two criteria, being present only in the eukaryotic lineage. A central concept to the RNA world is that proteins with superior catalytic properties have gradually replaced RNA as the catalytic molecule (and that no novel catalytic RNAs would be formed after the advent of efficient protein synthesis, Jeffares et al. 1998).

However, it is difficult to see how a molecule such as MRP could have evolved only in the eukaryotic lineage and then integrate itself so intimately into rRNA processing, mitochondrial genome replication, and perhaps other functions central to eukaryotic metabolism. It has been found that eukaryotes carry more proposed 'relics' of the RNA world than prokaryotes. These 'relics' include small nucleolar RNAs, spliceosomes, telomerase, and self-splicing introns, which are all absent from prokaryotes (Jeffares et al 1998). MRP was the only widely occurring catalytic RNA not suggested to be a relic from the RNA world in Jeffares et al. 1998.

Again there are several variants of this hypothesis; MRP could have evolved from P, P evolving from MRP, and MRP and P evolving independently in the RNA world.

With such an early divergence expected between pRNA and mrpRNA (at least back to the divergence of eukaryotes), nucleotide sequence alignments may not be reliable enough to determine with confidence any evolutionary relationship. It is expected, however, that examination of the RNA secondary structure may yield the required information when the sequence data cannot.

It has been shown that many sequences can fit the same secondary structure (Fontana et al. 1993) which allows the catalytic RNA sequence to vary even if the function of the molecule remains unchanged. The secondary structure of the catalytic RNA molecule has both fixed 'motifs' that represent areas that are critical to maintaining the function, and other regions that are free to vary in presence or size. It is expected that these fixed and variable regions of the catalytic RNA secondary structure will change according to the evolution of the function of the molecule, and thus may be used to determine evolutionary relationships when the sequence data may not. Quantitative comparisons of pRNA and mrpRNA secondary structures are used here to calculate distances between these molecules in order to assess their relatedness.

	Accession Number	Length of Sequence	A + T %	Secondary Structure Reference
pRNA Sequences				
Eubacterial pRNA				
Synechocystis sp. PCC6803	X65707	437	48	P
Anabaena sp. PCC 7120	X65648	465	47	P
Anacystis nidulans PCC6301	X63566	385	43	P
Pseudoanabaena sp. PCC 6903	X73135	450	52	P
Escherichia coli	M17569	377	38	P
Bacillus subtilis	M13175	401	51	P
Rhodospirillum rubrum	M59355	429	29	P
Agrobacterium tumefaciens	M59354	402	36	P
Mitochondrial pRNA				
Reclinomonas americana mitochondria	AF007261	312	75	P
Saccharomyces cerevisiae mitochondria	U46121	448	87	No structure
Aspergillus nidulans mitochondria	X93307	300	81	No structure
Plastid pRNA				
Porphyra purpurea chloroplast	U38804	383	63	P
Cyanophora paradoxa Cyanelle	X89853	350	67	P
Eukaryotic pRNA				
Human (nuclear)	X15624	340	36	Altman et al. 1993
Mouse (nuclear)	L08802	288	33	Altman et al. 1993
Danio rerio (nuclear) Zebrafish	U50408	308	43	No structure
Saccharomyces cerevisiae (nuclear)	M27035	368	48	Tranguch and Engelke 1993
Schizosaccharomyces pombe (nuclear)	X04013	373	48	Tranguch and Engelke 1993
mrpRNA Sequences				
Human	X51867	264	36	Schmitt et al. 1993
Bovine	Z25280	277	39	Schmitt et al. 1993
Mouse	J03151	275	36	Schmitt et al. 1993
Rat	J05014	273	35	Schmitt et al. 1993
Xenopus (frog)	Z11844	277	45	Schmitt et al. 1993
Arabidopsis thaliana	X65942	260	49	Kiss et al. 1992
Saccharomyces cerevisiae	Z14231	339	60	Kiss et al. 1992
Schizosaccharomyces pombe	X04013	399	57	Paluh and Clayton 1995
16S rRNA structures	**RDP sequence**			
				RDP
Escherichia coli	E.coli	-	-	RDP
Clostridium innocuum	C.innocuum	-	-	RDP
Methanococcus vannielli	Mc.vanniel	-	-	RDP
Frankia sp.	Fra.spORS	-	-	RDP
Streptomyces coelicolor	Stm.coelic	-	-	RDP
Thermus thermophilus	T.thermoph	-	-	RDP
Bacillus subtilis	B.subtilis	-	-	RDP
Agrobacterium tumefaciens	Ag.tumefac	-	-	RDP
Spirochaeta aurantia	Spi.aurant	-	-	RDP
Thermoplasma acidophilum	Tpl.acidop	-	-	RDP
Mycoplasma capricolum	M.capricol	-	-	RDP
Methanobacterium formicicum	Mb.formici	-	-	RDP
Pseudomonas testosteroni	Ps.testost	-	-	RDP

Table 1.2: pRNA , mrpRNA and 16S rRNA sequences and secondary structures used in this study showing length, accession details, A+T % and from where the secondary structures were obtained
Key: *P* Obtained from the RNase P Database (Brown 1997).
RDP Obtained from the Ribosomal Database Project (Maidak et al. 1997).

This study examined two types of RNA secondary structure. The first is the secondary structure that the RNA forms in nature and is referred to here as the "biological secondary structure". The biological secondary structures of eubacterial pRNA have been studied extensively (Haas et al. 1994, Haas et al. 1996a, Haas et al. 1996b, Green et al. 1996) and consensus structures calculated. Eukaryotic and organellar pRNA biological secondary structures are not as well defined with published hypothetical structures being used here. Some organellar sequences used in this study do not have any published secondary structure and are only used when sequence data alone is required.

The second type of secondary structure is calculated from the nucleotide sequence data using folding programs. Such structures are determined only from the nucleotide sequence data and need not have any relationship to the function of the molecule. Thus, the calculated secondary structures may not have the same fixed and varied regions that are shown in the biological structures (Zuker 1989).

Within the fixed regions of the biological secondary structure it is expected that nucleotide changes in one part of a helix will be met by a corresponding change in another part of the sequence to allow the helix to remain unchanged. Thus it is still expected that sequences of similar functions will form similar secondary structures with the folding programs allowing the formation of a recognisable structural 'motif'. These motifs are possible identification features that could be used in the characterisation of putative catalytic RNA sequences. Secondary structures folded from pRNA and mrpRNA sequences with folding programs are examined for use in the characterisation of putative catalytic RNA sequences.

Organellar genomes (mitochondria and chloroplast) offer a unique opportunity for the testing of searching, gene identification, and characterisation techniques. These genomes are small and many have been completely sequenced, and are available in databases such as GenBank. However the high AT content of organellar genomes often makes them hard to search with standard searching algorithms. Searching databases with a sequence of high AT content gives a high background of non-relevant matches often obscuring meaningful results. The distribution of pRNA and mrpRNA (Figure 1.3) shows that although pRNA is found encoded in the mitochondrial DNA of plants, there is to date, no published green plant chloroplast-encoded pRNA sequences. To test the feasibility of using RNA secondary structure to

characterise potential pRNA sequences, green plant chloroplast genomes were searched for putative pRNA sequences.

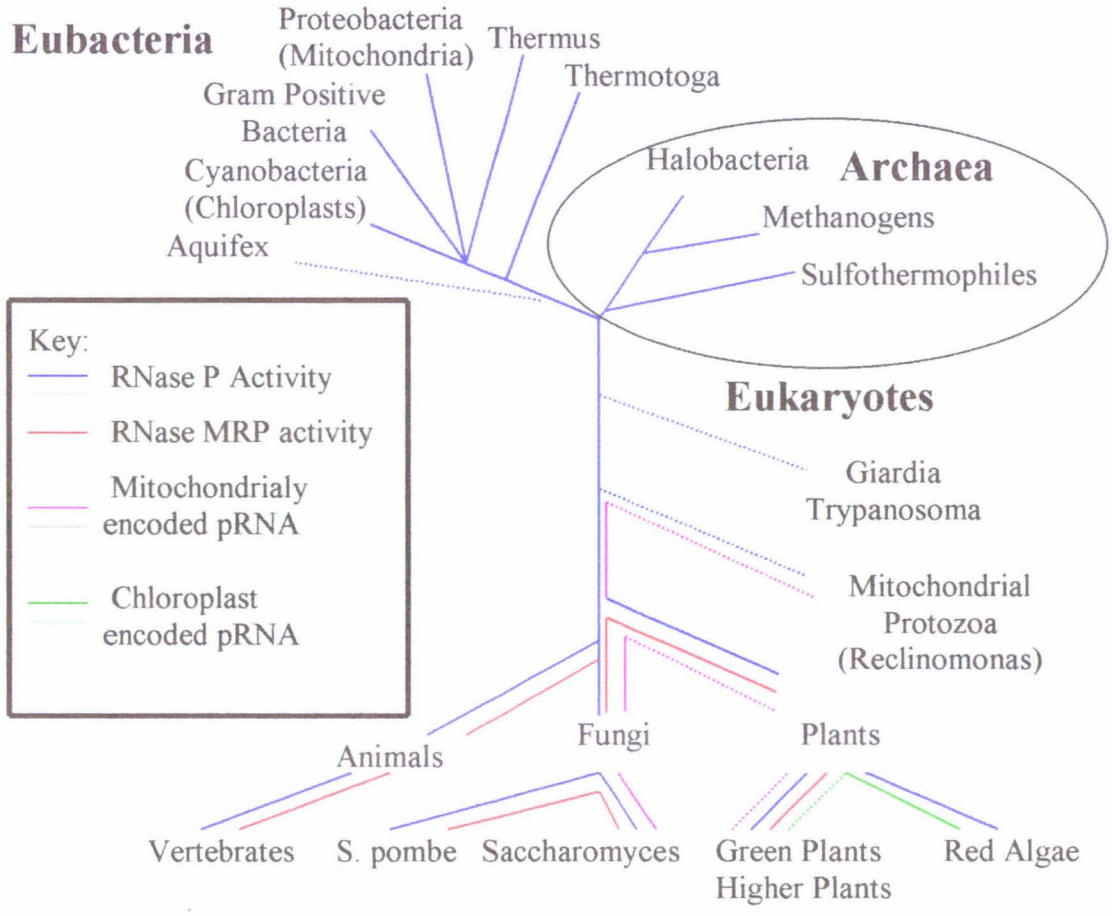


Figure 1.3: Phylogenetic distribution of MRP and P. Dotted lines represent where activity has been shown but the RNA has not yet been characterised. Solid lines represent where the activity has been shown and has been characterised.

It is only recently that pRNA was characterised from the chloroplast of the red alga *Porphyra purpurea* (Reith and Munholland 1995), and from the cyanelle (a chloroplast-like plastid that still retains a cell wall) of *Cyanophora paradoxa* (Baum et al. 1996). Although it is expected that sequence homology between known pRNA sequences and putative green plant chloroplast pRNA sequences would be low, it is still expected that secondary structure (both a theoretical biological structure based on other pRNA structures and a folded structure), would show identifying secondary structure characteristics. One of the putative green plant chloroplast pRNA sequences (from the *Zea mays* – maize chloroplast) is examined more fully with other pRNA and mrpRNA sequences in this study.

There is a possibility that folded structures could be used in the same way as the biological structures, for determining evolutionary relationships. The biological and folded structures from two folding programs are shown for Human pRNA (Figure 1.4), *Escherichia coli* pRNA (Figure 1.5), and Human mrpRNA (Figure 1.6). These figures highlight how different the calculated structures are from the biological structures but also the similarities between the structures formed by the two different folding programs.

Problems with the use of folding programs in the analysis of catalytic RNA may include how much influence characteristics such as the AT content and sequence length, have on the estimated structure. These factors are examined here using random sequences derived by shuffling pRNA and mrpRNA sequences of varying length and AT content. Protein-coding RNA sequences are also used as controls in order to evaluate any trends that may be used in identifying putative catalytic RNA sequences. The amount of pairing that is present in a folded structure could also be another tool in the identification of catalytic RNA sequences.

In summary, this thesis looked at four main issues. The first was the evolution of MRP and its relationship to P. The second was the use of RNA secondary structure in the characterisation of putative pRNA sequences from chloroplasts. The third was the use of biological secondary structure in determining evolutionary relationships, and the fourth was the evaluation of the structural output from folding programs. The techniques developed here may, in future, be applied to other RNA molecules especially those associated with the RNA world as well as the analysis of newly discovered potential RNA molecules.

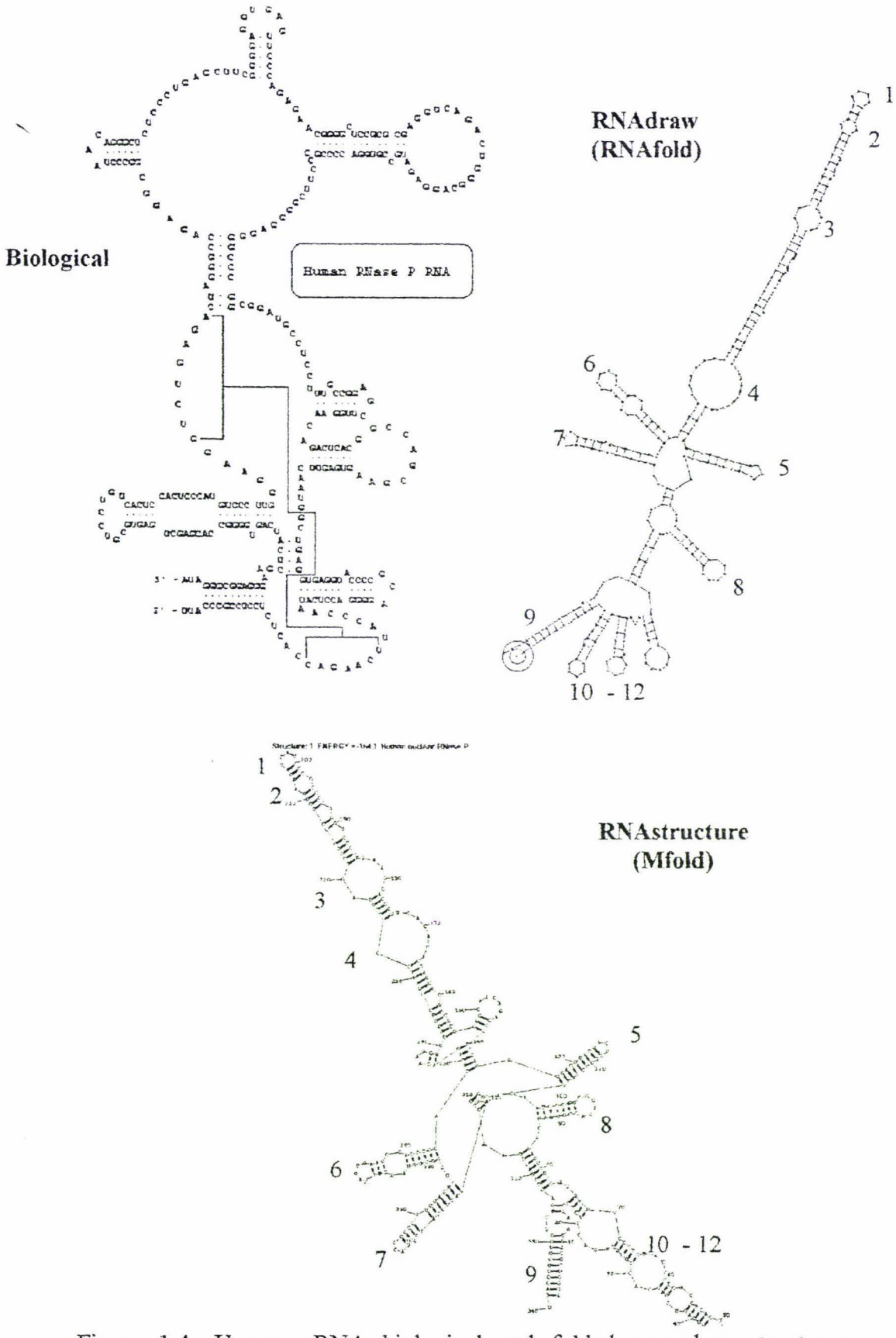


Figure 1.4: Human pRNA biological and folded secondary structures. Numbers 1 to 12 represent features that may be common to both the RNAstructure and the RNAdraw structures.

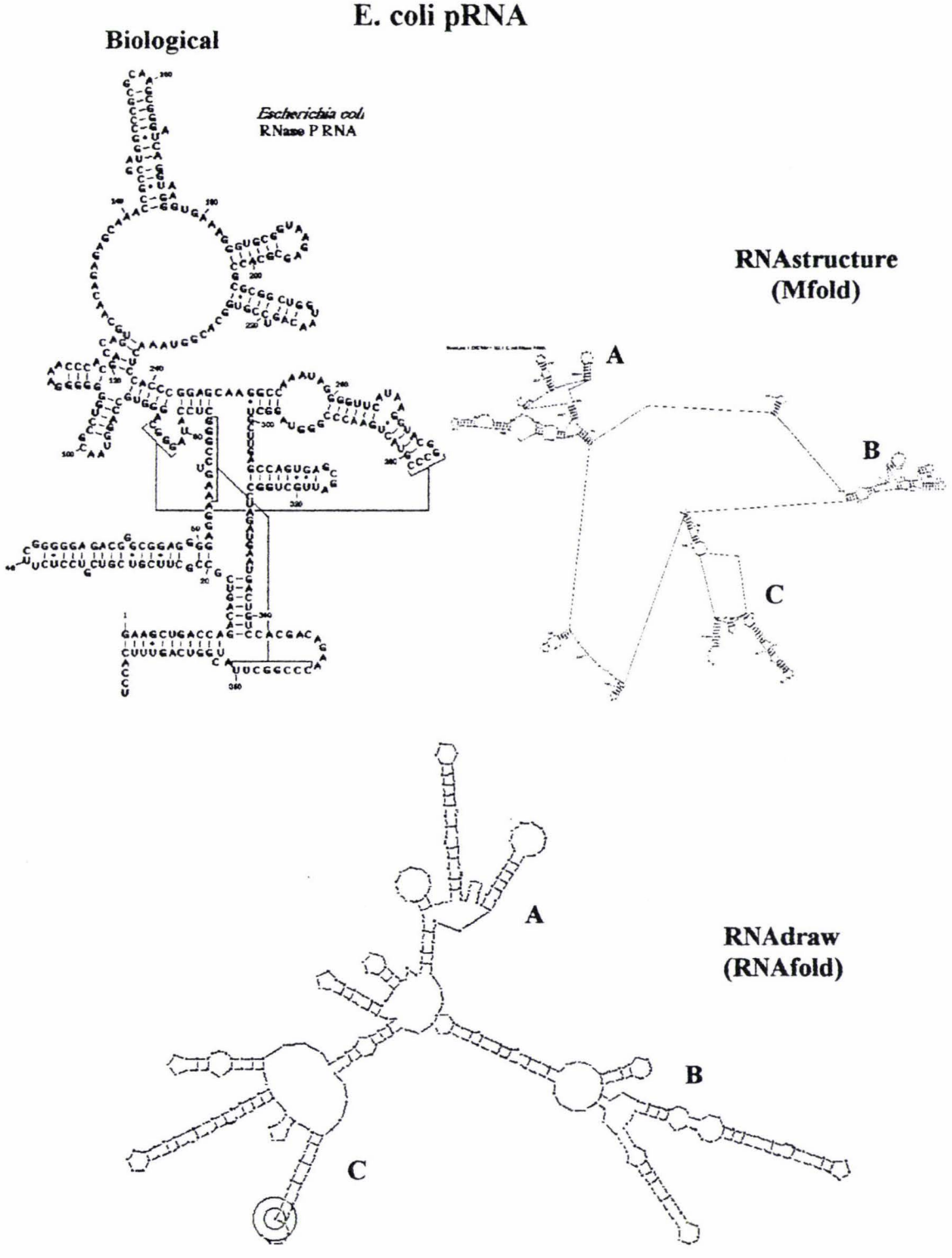


Figure 1.5: *E. coli* pRNA biological and folded secondary structures. A, B, and C represent features that may be common to both the RNAstructure and the RNAdraw structures.

Human mrpRNA

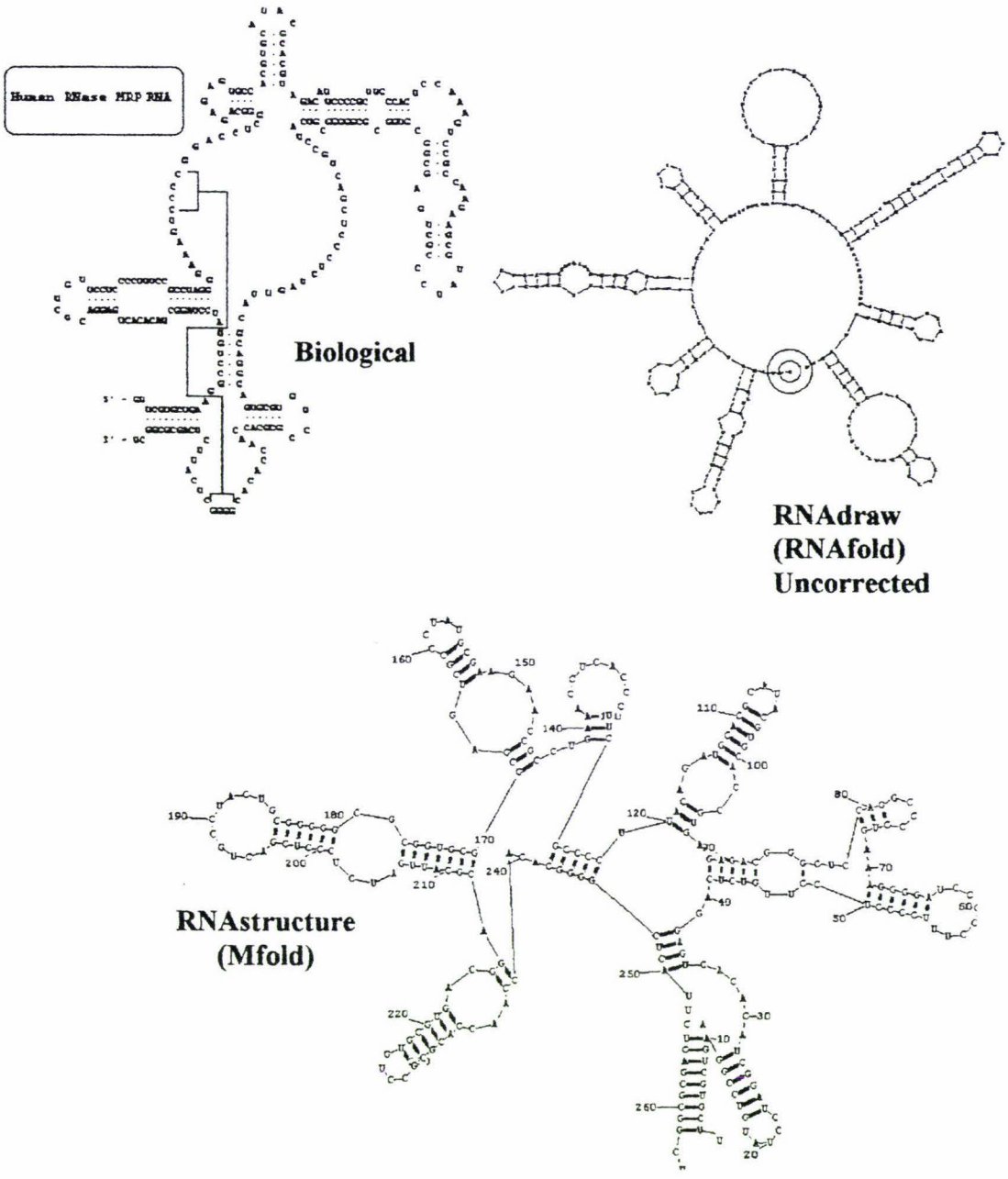


Figure 1.6: Human mrpRNA biological and folded secondary structures. The RNAdraw structure in this case has formed short range pairing in preference to the long range pairing required to pair the 5' and 3' ends. This structure is corrected when required to form 5' - 3' pairing.

Chapter 2

Review of Literature

(To date: July 1998)

This chapter summarises studies on the localisation, transport and activity of MRP and P. As this study integrates nucleotide sequence, secondary structure and published functional data, it is necessary to summarise previous findings. This will necessitate some repetition from the introduction, however, there is considerably more detail available.

RNase MRP

RNase MRP, a site-specific endoribonuclease ribonucleoprotein was first identified in 1987 as an RNA-processing activity in mammalian cells (Clayton 1994). MRP, also known as 7-2 RNA, is encoded in the nucleus but has been found to have activity in both the mitochondria and the nucleolus of the eukaryotic cell. MRP activity has been identified in eight eukaryotes to date including vertebrates (human, bovine, mouse, rat and the frog *Xenopus*), yeasts (*Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*), and plants (*Arabidopsis*).

The MRP complex consists of a single RNA strand of approximately 300 nucleotides and a protein moiety of several protein subunits. The exact composition of the MRP complex is still unclear. The nucleoprotein complex that is immunoprecipitated from nuclear extracts consists of mrpRNA and at least 10 protein components (Li et al. 1994). MRP is transcribed by RNA polymerase III and the 5' flanking sequences of the RNA are required and sufficient for transcription (Yuan and Reddy 1991). Schmitt et al. (1993) observed that the RNA strand folds into a conserved secondary structure that has similarities to the secondary structure of pRNA. However, the nucleotide sequences of the mrpRNA's show little similarity. This indicates that the structure is more important to the function of the mrpRNA than the actual RNA nucleotide sequence.

Protein moiety composition

The protein components of MRP are in the process of being characterised and it has been suggested that MRP processing in the nucleolus and the mitochondria may depend on a different array of protein components (Karwan et al. 1991). Nine proteins

have been characterised so far for MRP in yeasts. In *Saccharomyces cerevisiae* there is yPop-1, Pop3, and Pop4, Pop5p, Pop6p, Pop7p and Pop8p that are common to both MRP and P (Chamberlain et al. 1998). In the fission yeast *Schizosaccharomyces pombe* a protein that is only found in MRP, SNM1 has been identified.

The yPop-1 protein consists of 876 amino acids and has a predicted molecular mass of 100.5 kD (Lygerou et al. 1994). Immunoprecipitation results show that yPop-1 is common to both the P and the MRP in *S. cerevisiae,* and mutational analysis shows that it is essential for viability (Lygerou et al. 1994). The corresponding genes in humans, hPop1 (115 kD) and *Caenorhabditis elegans,* cPop1 (86.2 kD) have also been characterised (Lygerou et al. 1996b). The overall conservation of the three sequences is low. No sequence similarity could be detected between any of the three eukaryotic proteins and the eubacterial P protein, C5, or the mitochondrial P protein from *S cerevisiae* (Lygerou et al. 1996b) indicating that these proteins are quite different.

Pop3 encodes a basic protein of 22.6 kD (predicted molecular weight) (Dichtl and Tollervey 1997) and is essential for both MRP and P. It is not required for the stability or the maturation of these ribonucleoproteins, as depletion of the protein does not affect the cellular levels of either RNA component (Dichtl and Tollervey 1997).

Pop4 is the third protein so far to be found to be essential for both MRP and P and appears to be required for either synthesis or stability of pRNA and to a lesser extent, of mrpRNA (Chu et al. 1997). Pop5p, Pop6p, Pop7p and Pop8 have only been very recently discovered and are in the process of being characterised (Chamberlain et al 1998).

The protein SNM1, isolated from *Schizosaccharomyces pombe*, is the first protein identified that is unique to MRP (Schmitt and Clayton 1994). This protein, of 198 amino acids, showed no striking similarities to any known proteins but contains a leucine zipper motif, a zinc-cluster motif, and a serine/lysine-rich tail. This last region has been found in many RNA-binding proteins involved in splicing (Schmitt and Clayton 1994). This protein exists in three phosphorylated forms and the mature unphosphorylated form. This may be a control mechanism for MRP, in order to regulate it during periods of high and low ribosome biogenesis (Schmitt and Clayton 1994).

Sera from patients suffering from autoimmune diseases often contain antibodies to various nuclear and cytoplasmic ribonucleoprotein complexes. A group of sera

referred to as Th (or To) immunoprecipiate to human MRP and P. The human MRP RNA was also called Th RNA while the human pRNA is also called the H1 RNA for this reason.

The Th/To antisera seem to recognise a 40 kD protein subunit common to the human MRP and P and is referred to as the Th/To antigen, or the Th40 antigen. However immunoblotting experiments with different Th/To sera have failed to reveal a single immunoreactive polypeptide, and cloning of the Th antigen has also not yet been accomplished (Lygerou et al. 1996b). It is considered unlikely that the observation of the Th40 antigen is due to protein degradation or *in vivo* processing of hPop1 (Lygerou et al. 1996b). At this point the exact nature of the Th40 antigen and its possible relationship to hPop1 are yet to be determined. There have been other newly characterised autoantibodies, which do not immunoprecipitate with the Th40 antigen but still with MRP and P (Karwan 1998).

Mitochondrial Activity

In early studies MRP had been shown to process, *in vitro,* the RNA primers required for mitochondrial genome replication. (Jacobson et al. 1995). Controversy arose, however, with little or no detectable mrpRNA being found in the mitochondria of animal and plant cells using standard techniques. Using a combination of biochemical and ultrastructural *in situ* hybridisation, it was determined that MRP was transported to the mitochondria, resolving the question of the existence of pathways for nucleo-mitochondrial transport of nucleic acids in animal cells (Li et al. 1994). Deletions within the 5' or 3' region of the Mouse RNA gene produced transcripts that was still transported to the mitochondria, however deletion of the midportion produced transcripts that failed to transport to the mitochondria. It was concluded that specific structures of the mrpRNA permitted it to be transported to the mitochondria (Li et al. 1994). When MRP is directly injected into the cytoplasm, it is unable to enter the nucleus indicating that mrpRNA lacks signals for cytoplasm to nucleus traffic (Jacobson et al. 1995) and that once MRP has been transported from the nucleus to the cytoplasm it is unable to enter it again.

Dairaghi and Clayton (1993) showed that *in vitro* MRP cleaves at three sites the RNA (from the leading-strand DNA) contained within the mammalian mitochondrial displacement loop (D-loop). Bovine and mouse MRP will process both bovine and

mouse substrates indicating that structural features of the substrate have been conserved despite significant overall nucleotide sequence divergence (Dairaghi and Clayton 1993). Human MRP will cleave a substrate derived from the D-loop region of the mouse mitochondria, and the *S cerevisiae* MRP can process both mouse and human substrates (Li et al. 1994). This shows that these MRPs have retained enough similarity throughout their evolution to function in other organisms despite a large nucleotide sequence divergence.

The transport of MRP from the nucleus to both the nucleolus and the mitochondria has raised questions about the mechanisms by which such intracellular movement is accomplished (Li et al. 1994). A protein has been found that shuttles between the nucleus and the cytoplasm acting as a mediator of RNA export (Lee, Henry and Silver 1996). This study has also shown a link between the nuclear export of poly(A)-RNA and that of a shuttling RNA-binding protein. Another recent study has shown that ribonucleoprotein export is mediated by a structural reorganisation of the nuclear pore basket. The functional role of this basket is to anchor the ribonucleoprotein to the nuclear pore complex and position it in the correct position before it moves through the pore complex (Kiseleva et al. 1996). Future investigations may reveal the mechanism by which MRP can be transported to two very different locations.

Nuclear function

The great majority of mrpRNA in the cell is transported into nucleoli and a nuclear function for MRP has been investigated. *In vitro*, MRP directly and accurately cuts the pre-rRNA at site A3 (Figure 2.1) showing that MRP is directly implicated in rRNA processing. This is consistent with its predominant accumulation in the nucleolus (Lygerou et al. 1996).

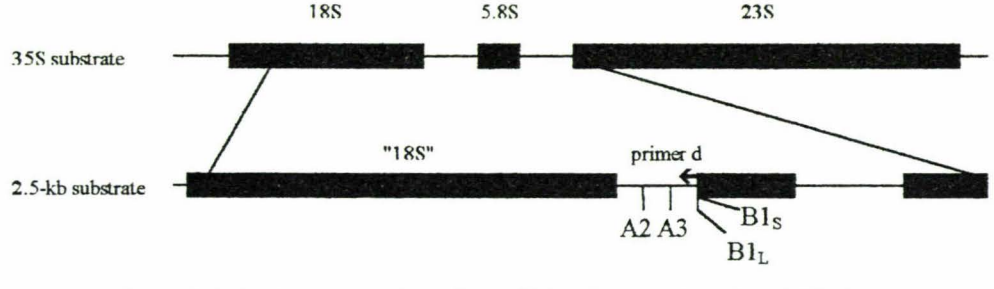


Figure 2.1: *In vitro* processing of pre-rRNA showing the A2 and A3 cleavage sites. Redrawn from Lygerou et al. 1996

It has been shown that the loss of mrpRNA does not affect P activity on precursor tRNA. mrpRNA rapidly accumulates in the dense fibrillar component of the nucleolus as early as thirty seconds after nuclear microinjection (Jacobson et al. 1995). A fragment consisting of nucleotides 1 - 79 is still transported to the nucleus but disappears quickly. The Th40 antigen is required for transport and binds nucleotides 21 - 64 of the human mrpRNA. Deletion or alteration of this binding site resulted in a diffuse distribution of the RNA in the nucleoplasm. Either the Th40 antigen binds and directly targets the mrpRNA to the nucleolus, or the nucleolar localised Th40 antigen is a high affinity site for mrpRNA accumulation (Jacobson et al. 1995). The site of MRP assembly remains unknown. However the rapidity of the transport suggests that the dense fibrillar component of the nucleolus could be the site of assembly as well as the site of the catalytic activity (Jacobson et al. 1995).

Mutational analysis of the *Schizosaccharomyces pombe* mrpRNA has also indicated that there may be unknown nuclear roles for MRP in tRNA maturation and nuclear control of septation (Paluh and Clayton 1996).

RNase P

RNase P is also a ribonucleoprotein and cleaves tRNA precursors giving mature 5' ends of all tRNA molecules. P activity has been found in all cells studied to date. It consists of an RNA component, in prokaryotes a single protein and in eukaryotes multiple proteins (Pace and Smith 1990). Cations such as K^+ and Mg^{2+} are critical in the P reaction (Pace and Brown 1995). Cleavage by P is product (tRNA) inhibited, and may be involved in the regulation of rRNA expression under certain growth conditions *in vivo* (Kirsebom 1995).

Prokaryotic P

The eubacterial pRNA is catalytic without the protein component *in vitro* but only in the presence of a high concentration of salt, such as 100 mM Mg^{2+}. The addition of the protein increases the rate of cleavage by enzyme *in vitro*, but *in vivo* the protein is required for P activity and cell viability (Liu and Altman 1996). It is possible that the lack of 'RNA-alone' catalysis shown in pRNA's other than eubacterial pRNA's, may simply signify the inability for the RNA to fold correctly in the absence of the protein

component (Tranguch and Engelke 1993).

From the analysis of eubacterial pRNA secondary structure mutants, there has not been any single feature of the pRNA secondary structure identified that is absolutely required for catalysis (Pace and Brown 1995). The catalytic core of the eubacterial pRNA is formed by less than one third of the nucleotides in the RNA. Some regions of this core are not required for the actual catalysis but are required for proper folding of the RNA (Green et al. 1996).

The majority of pRNA study has been done using the Gram-negative bacterium *Escherichia coli*. The sequences and structures of the pRNA from Gram-positive eubacteria are different from those of other eubacteria in that it has unusual tertiary interactions (Haas et al. 1996). Work with the *Bacillus subtilis* pRNA has indicated the essential role that magnesium ions play in P function and stability. At least three magnesium binding sites stabilise the folded RNA tertiary structure, at least two sites enhance the formation of complexes of pRNA with pre-tRNA or tRNA and at least one site stabilises the transition state for pre-tRNA cleavage (Beebe et al. 1996).

The C5 P protein subunit of *E coli* contains 119 amino acid residues. The *E. coli* pRNA is protected from enzymatic cleavage and chemical modification, in the presence of the protein C5 at nucleotides 266 to 287 suggesting that the protein binds to the RNA in this region (Hardt and Hartmann 1996). Mutagenic analysis of this protein has shown that certain hydrophobic and basic residues are important for P catalysis *in vivo* and *in vitro*. It was found that some of the amino acid changes could alter the substrate specificity of the P holoenzyme (Gopalan et al. 1997). The crystal structure of the Bacillus subtilis protein subunit has revealed unusual topology that is shared with the ribosomal protein s5 and the ribosomal translocase elongation factor G (Stams et al. 1998). This report suggests that these proteins may have had a common RNA binding ancestor in the primordial translational apparatus.

Protein subunits from other species (e.g. the cyanobacterium *Synechocystis)* have been compared to those above, and show low sequence homology. However, it has been shown that they could adopt a similar three-dimensional structure, and are considered functionally equivalent (Pascual and Vioque 1996).

Archaeal pRNA's are similar in both their primary and secondary structure to the eubacterial pRNA's. However, the archaeal pRNA's are not catalytic without their protein components *in vitro* under any conditions studied to date (Haas et al. 1996, Pannucci et al. 1997). Also there seems to be more diversity in the composition of the archaeal P holoenzyme. Halophile P is composed largely of RNA and resembles the eubacterial enzyme while the *Sulfolobus acidocaldarius* P is predominantly protein and resembles the eukaryotic enzyme (Haas et al. 1996).

Mitochondrial P

The pRNA encoded in the mitochondria of some yeasts and plants is, (as expected in mitochondrial and chloroplast genomes), very A/U rich and can vary in size from 140 to at least 490 nucleotides (Wise and Martin 1991, Martin and Lang 1997). The pRNA is encoded in the mitochondrial genome whereas the protein component is nuclear encoded (Lee et al. 1996b). RNA fragmentation studies on the mitochondrial pRNA from *Aspergillus nidulans* show that the base pair structure formed between the 5' and 3' ends acts as a barrier to nuclease digestion and may be important for maintaining the functional structure of the A/U rich RNA (Lee et al. 1996b).

In *Saccharomyces cerevisiae*, the RNA was thought to be transcribed from the same promoter as two of its substrate tRNA's, However it has been recently shown to be transcribed from its own promoter, SP1, which lies between the $tRNA^{met}$ and the mitochondrial pRNA gene (Biswas 1996).

The mitochondrial P from *A. nidulans* consists of the RNA component and seven smaller polypeptides. The protein components have been shown to be essential *in vivo* (Lee et al. 1996a). The only protein subunits identified to date are Rpm2p from *S. cerevisiae* and the homologous protein from *S. douglaii* (Martin and Lang 1997). It has been suggested that the mitochondrial RNA's as well as other eukaryotic RNA's have lost some elements essential to catalysis and that proteins have taken over some function provided by eubacterial RNA's (Wise and Martin 1991). The smallest known pRNA is that of the *Saccharomycopsis fibuligera* mitochondria. This 140 nucleotide long RNA is not catalytic without its protein component but can be folded into a structure resembling the proposed eubacterial P catalytic core (Green et al. 1996).

In the yeast *S. cerevisiae*, no mutations that alter the structure of the mitochondrial -encoded pRNA have been isolated, the only mutants that are deficient in mitochondrial P activity lack the gene altogether (Sulo et al. 1995). It was also found that if the mitochondrial pRNA gene was introduced into yeast nuclei that lacked the gene in the mitochondria, then P function was restored in the organelle. This showed that pre-existing P activity is not necessary for the biosynthesis of P (Sulo et al. 1995).

The *S. pombe* mitochondrial genome resembles the vertebrate mitochondrial genome in having no mitochondrialy-encoded pRNA (Pulah and Clayton 1995). The P found in human mitochondria also has its components encoded in the nucleus. Rossmanith and Karwan (1998) have found that the human P in the mitochondria is resistant to nucleases and exhibits a protein-like density in Cs_2SO_4 gradients suggesting that,unlike the P found in the nucleus, it is not a ribonucleoprotein but a protein enzyme.

Recently, what is regarded as an "ancient-type" mitochondrial genome was sequenced from the freshwater protozoon *Reclinomonas americana* and a mitochondrial pRNA characterised (Lang et al. 1997). Unlike other mitochondrial pRNA's, the pRNA from *R. americana* is not AU rich and contains almost all of the evolutionarily conserved nucleotide sequence and secondary structure motifs of the eubacterial consensus pRNA model.

The mitochondria in higher plants retain more prokaryotic features than fungal or animal (Gray 1989). The plant mitochondrial P has to recognise and process tRNA precursors imported from the nucleus, those coded by the mitochondrial genome and precursors from tRNA genes imported from the plastid genome (which have since been incorporated into the mitochondrial genome) (Marchfelder and Brennicke 1993). The fact that tRNA's are imported into some mitochondrial genomes from the nucleus again shows that highly charged RNA molecules can pass through the hydrophobic mitochondrial membrane (Marchfelder and Brennicke 1994).

Chloroplast P

Chloroplast P had been thought,at one stage, to lack an RNA component entirely (Baum and Schon 1996b). However, a pRNA from the *Cyanophora paradoxa* cyanelle (Baum and Schön 1996a) and, recently, a pRNA gene from the chloroplast of *Porphyra purpurea* (Reith and Munholland 1995) have been characterised. The cyanelle genome is

more eubacterial than the chloroplast genome. Several genes that have been shown to be transferred from the chloroplast to the nuclear genome early in plant evolution are retained in the cyanelle genome (Baum and Schön 1996b, Martin and Muller 1988). The pRNA from the *C. paradoxa* cyanelle is similar to that of cyanobacterial pRNA's and the secondary structure closely resembles the eubacterial consensus. The RNA is not catalytic by itself but it is essential to P activity (Baum and Schön 1996a).

Eukaryotic (nuclear) P

Eukaryotic (nuclear-encoded) P seems to differ greatly from the eubacterial type of P and has more than one associated protein polypeptide. The *S. cerevisiae* nuclear pRNA has nine associated protein units, 20-fold more protein than in the bacterial enzyme (Chamberlain et al. 1998). Genetic depletion and immunoprecipitation studies have shown that only one polypeptide Rpp1p is unique to the P holoenzyme. These enzymes are large and, like all eukaryotic P, require the RNA component for catalysis (Pace and Smith 1990). Although the proposed secondary structure for the eukaryotic P is different from that of the eubacterial model, parts of the secondary structure of the eukaryotic pRNA can fit into the eubacterial consensus model (Altman et al. 1993). A minimum consensus secondary structure for the vertebrate pRNA has recently been proposed (Pitulle et al. 1998) based on thirteen partial pRNA genes. Mutations in the P10/11-12 domain of yeast nuclear pRNA affect the catalytic rate of the enzyme and magnesium interactions (Ziehler et al. 1997). In vertebrates, the nuclear-encoded P shows activity in both the nucleus and the mitochondria but there have been no pRNA homologs found in the compact vertebrate mitochondrial genome (Li and Williams 1995).

Higher vertebrates may express multiple isoforms of pRNA. In mouse three additional genes homologous to the published mouse pRNA gene were found, but were predicted to produce shorter RNA transcripts in which a conserved secondary structure could still be formed (Li and Williams 1995). A comparison of the putative mouse pRNA homologs and the published sequence suggest that there may have been sequential gene duplication events leading to the present diversity (Li and Williams 1995).

pRNA was characterised from the Zebrafish (*Danio rerio*), and shows 63% sequence homology to the *Xenopus* pRNA and 69% sequence homology to the human pRNA sequence. This gives some idea of the homology of this RNA sequence over a

wide range of vertebrates (Eder et al. 1996).

Recently the P from wheat nuclei (Arends and Schön 1997) and *Tetrahymena thermophila*, a ciliate protozoan (True and Celander 1996), have been purified and characterised enzymatically. As yet there is no RNA component sequenced from these enzymes to compare with the yeast and vertebrate examples. The pRNA from *Drosophila* has also been used recently in kinetic studies (Levinger et al. 1997).

As well as its role in tRNA processing, there may be a nuclear role for P in pre-ribosomal RNA processing. Mutational analysis of the *S. cerevisiae* nuclear pRNA has shown the accumulation of an aberrant rRNA, and a possible function for P after cleavage of the A2 site by MRP (Chamberlain et al. 1996). It is possible that P does not cleave pre-rRNAs directly but may be part of a rRNA processing structure (together with snoRNP's and other associated proteins). It has been found that P exists in the nucleolus where it can be involved in pre-rRNA processing and throughout the nucleoplasm where it can be involved in pre-tRNA processing (Jacobson et al. 1997). Like the mrpRNA it also localises to the dense fibrillar component of the nucleolus requiring the To antigen binding domain (nucleotides 25 - 75). A large macromolecular complex consisting of P, MRP, and possibly preribosomes has been suggested, indicating that P and MRP may function in a co-ordinate fashion in ribosome biogenesis (Lee et al. 1996b).

Evidence for the evolutionary relatedness between MRP and P

Similarities between MRP and P have suggested that these two ribonucleoproteins are evolutionary related. Both cleave RNA's to generate 5' phosphate and 3' hydroxyl termini in a reaction requiring divalent cations (Forster and Altman 1990). Comparison of the eubacteria *E. coli* and the yeast *S. cerevisiae* rRNA's showsconservation of the arrangement of the pre-RNA cleavage sites for P and/or MRP. Morrissey and Tollervey (1995) have shown that a tRNA molecule is found in the pre-rRNA spacer of both eubacteria and archaea and believe that this could represent an ancestral state (or removes the need for MRP in eubacteria, if MRP is in the RNA world).

P and MRP activities from several organisms cofractionate through multiple stages of biochemical purification (Paluh and Clayton 1995). Another similarity is that both are sensitive to the action of puromycin, an antibiotic that inhibits pre-tRNA

processing (a specific inhibitor of P activity) (Potuschak et al. 1993).

The only conserved sequence motif in the pRNA is an 11-nucleotide GAGGAAAGUCC motif. This is also found in human MRP RNA (and other mrpRNAs slightly modified) and corresponds to the long-range pairing helix which is an essential part of the pseudoknot formation for both molecules (Reddy and Shimba 1996).

It has been suggested that the close relationship between MRP and P might be due to the homologous protein components, and the nucleotide and structural similarities in the RNA that are required to bind to them (Sbisà et al. 1996). Another suggestion is that the two enzymes may have maintained closely related catalytic mechanisms that could have resulted in identical evolutionary pressures on the proteins associated with each enzyme (Chu et al. 1997). There is an indication that the interaction between the common proteins and the RNA subunits is not identical. A distortion of the *S. cerevisiae* mrpRNA in the region nt150 results in an unusual conformation of the RNA-protein complex. However there is no obvious equivalent structure in the pRNA based on some basic qualitative comparison studies pRNA (Chu et al. 1997).

Part of the mitochondrial function of MRP can be performed by the *E. coli* P, cleaving the mitochondrial replication primer RNA only at site 2 but not at sites 1 and 3. This cleavage is approximately 250 times less efficient than the normal cleavage reaction of P and requires high $MgCl_2$ concentrations (Potuschak et al. 1993). In the absence of P, however, MRP does not cleave a tRNA precursor (Potuschak et al. 1993). Mutational studies in *S. pombe* did show however an effect on P function, but it was unclear as to whether this was due to a titrating out of a shared protein component (Paluh and Clayton 1996).

It has been proposed that MRP is a variant of P evolved for processing the pre-rRNA (Morrissey and Tollervey 1995), a theory which other papers support (Reddy and Shimba 1996). In this scenario, P originally cleaved a tRNA with a conserved position in the equivalent 16S/23S-spacer region of the rRNA operons from both eubacteria and archaebacteria to ensure separation of the rRNAs. It is suggested here that MRP evolved from P in eukaryotes to cleave this region at the same time this tRNA was transferred elsewhere (Chamberlain et al. 1996). It is unclear as to whether this theory means that MRP does a new function or has specialised to its present functions.

Chapter 3

Finding distantly related pRNA-like sequences in the chloroplast DNA of four green plant species.

Introduction

There have been to date no chloroplast-encoded pRNA genes characterised from green plants. Eubacterial and red algae chloroplast pRNA sequences are used here to search the chloroplast genomes of green plants for any possible pRNA sequences, which are then characterised using RNA secondary structure. It had been thought there was no chloroplast-encoded pRNA, but recently the pRNA was characterised from the chloroplast of the red alga *Porphyra purpurea* (Reith and Munholland 1995), and from the cyanelle (a chloroplast-like plastid that still retains a cell wall) of *Cyanophora paradoxa* (Baum et al. 1996).

Eubacterial pRNAs retain conserved secondary structure even though their nucleotide sequences are highly divergent. The pRNA sequences from two eubacterial species, *E. coli* and *Bacillus subtilis*, are so divergent that sequence homology is too low to be readily detected (Pace and Brown 1995). However, both of these sequences fold into secondary structures that match consensus structures for pRNA (consensus structures available in the RNase P Database, Brown 1998). The conservation of the secondary structure, even with highly divergent sequences, indicates that it is possible that RNA secondary structure can contain identifiable characteristics, and could possibly be used in the identification of other putative pRNA genes.

It is possible that the green plant chloroplast pRNA gene has been lost, and (as in vertebrate mitochondria) instead uses a nuclear-encoded product. There is also the possibility that the pRNA gene sequence in green plant chloroplasts has diverged to the extent that the sequences no longer have any significant homology to the pRNA genes of other species. In this second case, it is expected that the pRNA secondary structure would remain conserved similar to the situation between *E. coli and B. subtilis* pRNA (Pace and Brown 1995).

However, if the secondary structure of a green plant chloroplast-encoded pRNA has changed only slightly, then this may allow a large change to the nucleotide sequence. It has

been shown that many sequences can fit the same secondary structure (Fontana et al.1993). Therefore a slight change in secondary structure may have an even broader set of sequences that will fit this new structure. If the function of the catalytic RNA has changed, then, after perhaps a number of structural changes, the sequence of the new structure may only have limited homology to the old sequence.

Organellar genomes (mitochondria and chloroplast) offer a unique opportunity for testing searching and gene identification techniques, being small and readily obtainable.

However, experience has shown that the high AT content of organellar genomes often makes them hard to search using standard algorithms. Searching databases with a sequence of high AT content (or genomes with a high AT content) gives a high background of non-relevant matches often obscuring meaningful results. The searching strategy developed here has been optimised for working with small genomes and sequences with a high AT content, adapting programs that are not normally for this specific purpose.

The characterisation of any putative pRNA sequences found in green plant chloroplasts involves both the analyses of sequence and secondary structure homology to known pRNA sequences. Structures folded with folding programs are also examined for any structural motifs that are also shown in other eubacterial and chloroplast pRNA folded structures.

The finding of any putative pRNA sequences in green plant chloroplast genomes gives a starting point for future cloning and assay experiments. This will be required to confirm whether these putative sequences are functional. Any effects of postranscriptional RNA editing (a feature found in chloroplast genes (Freyer et al. 1997)) can only be determined by sequencing the active product.

This chapter has been written up as a short communication to form the basis of a manuscript.

Materials and Methods

pRNA sequences and secondary structures were obtained from the RNase P Database (Brown 1998) and the chloroplast genomes were obtained from the Genbank sequence database (Benson et al. 1998) (Table 3.1).

Chloroplast Genome	Accession number/ NID	Length of Sequence	A+T% Content
Zea mays (Maize)	X86563 / g902200	140387	61.6
Porphyra purpurea (Red Algae)	U38804 / g1276652	191028	67.0
Oryza sativa (Rice)	X15901 / g11957	134525	61.1
Nicotiana tabacum (Tobacco)	Z00044 / g11807	155844	62.2
Pinus thunbetgii (Black Pine)	D17510 / g529643	119707	61.5
Marchantia polymorpha (Liverwort)	X04465 / Y00686	121024	71.2
Cyanophora paradoxa cyanelle	U30821 / g1016083	135599	69.5
Spinacia oleracea (Spinach) RNA Pol genes etc.	X06871 / g295119	12360	63.6
pRNA Sequences			
Porphyra Purpurea Chloroplast	U38804	383	63
Cyanophora paradoxa Cyanelle	X89853	350	67
Synechocystis sp. PCC6803	X65707	437	48
Anabaena sp. PCC 7120	X65648	465	47
Anacystis nidulans PCC6301	X63566	385	43
Pseudoanabaena sp. PCC 6903	X73135	450	52
Escherichia coli	M17569	377	38
Saccharomyces cerevisiae Mitochondria	U46121	448	87
Putative pRNA Sequences			
Z. mays chloroplast	from X86563: 19091 - 19419	329	62
Rice chloroplast	from X15901: 17566 - 17910	344	64
Tobacco chloroplast	from Z00044: 29288 - 29603	313	64
Spinach chloroplast	from X06871: 12 - 376	364	63

Table 3.1: Chloroplast genomes, eubacterial , cyanelle and chloroplast pRNA sequences and the putative green plant chloroplast pRNA sequences isolated in this chapter.

BLAST searches (Larlin and Altschul 1990) were done through the NCBI net site (www.ncbi.nlm.nih.gov) using the default values. A BLAST search involves the input of a query sequence, which is then compared to all sequences in a range of databases including the GenBank sequence database. The Ssearch program from the FASTA package (Smith and Waterman 1981) was used to search each chloroplast genome with one cyanobacterial sequence at a time. This program compares a test sequence to all entries in a designated library (a chloroplast genome) using the rigorous Smith-Waterman algorithm, and then calculates a local similarity score. An advantage with this program is that the score allows for deletions and insertions within the sequences, meaning that the length of the test

sequence is not a limiting factor in the search. For this study, a single chloroplast genome (the library) was searched carefully with a single pRNA sequence (the test sequence) giving a one-on-one approach. The top six alignments were then recovered and analysed for any previously known sequence identities.

The putative pRNA sequences were aligned using the ClustalX (Thompson et al. 1997) alignment program. Consensus, cyanobacterial, and the *Porphyra purpurea* chloroplast pRNA secondary structures were obtained from the RNase P Database (Brown 1998).

Sequences were folded with RNAstructure (Mathews et al. 1997) which uses the Mfold (Zuker 1989) folding algorithm and RNAdraw (RNAdraw (Matzura and Wennborg 1996), which uses the RNAfold (Hofacker et al 1994) folding algorithm. Parameters for RNAstructure and RNAdraw are shown in Appendix 4. All secondary structures folded with RNAstructure and RNAdraw are shown in Appendix 1. Only the optimal structures for each sequence were analysed here.

Results

Conventional BLAST searches of the sequence databases with cyanobacterial (*Synechocystis*, *Anabaena* and *Anacystis*) and the *Porphyra purpurea* chloroplast pRNA (designated from here on as the *Porphyra* chloroplast pRNA) sequences failed to recover any sequences from any green plant chloroplasts. This was likely due to the high background of nonrelevent 'matches' often seen when working with sequences of a high AT content.

Using Ssearch, a test search of the *Porphyra purpurea* chloroplast genome with its own pRNA and with the *Synechocystis* pRNA gave maximum scores of 1915 and 249, respectively. This was an indication that any putative pRNA sequences found would not be likely to have a high score (a score of 249 was considered a 'good' score). The *Zea Mays* (maize) chloroplast was then designated as the library and tested with a variety of pRNA sequences. The *Synechocystis* pRNA sequence identified a region with a score of 105 (Figure 3.1). Although this was not a high score, it was the highest score found between any of the cyanobacterial pRNA sequences and the green plant chloroplasts searched.

```
 A:synpdna.txt, 437 nt vs A:zmays.txt library

 140387 residues in     1 sequences
The best scores are:                                    s-w
Z.mays complete chloroplast genome.                      (59550) 105
Z.mays complete chloroplast genome.                      (59550)  92
Z.mays complete chloroplast genome.                      (22161)  83


>>Z.mays complete chloroplast genome.                      (59550 nt)
Smith-Waterman score: 105;    63.3% identity in 90 nt overlap

      250       260       270       280       290       300
Synech TCGAGAGGTACTGGCTCGGTAAACCCCGGTTGGAAGCAAGGTCGGAGGGGCAAAGGTTGG
                                    ::::: :::::    :: : :::: :   :
Z.mays AATTCTTCTACTATTGATCAATAATCATAGTGGAATCAAGGGTACAGAGTCAAAAG---G
   19200     19210     19220     19230     19240     19250

      310       320          330       340       350          360
Synech UCTTTTTCCUGCCCCATGATT---GGTGGAACCGCTTGAGGAATTTGGT---AACAAATT
        : ::   :::::: : :: :   : :: : ::: :  :::::::::  :   ::::::::
Z.mays CCATT---CTGCCCTAAGACTATGGATGAATCCGTTAAAGGAATTTACTCCTAACAAATT
      19260         19270     19280     19290     19300     19310

           370       380       390       400       410       420
Synech TCCCAGATAGATAACTCCCCAAGGGTGCCTCGCATCCTGGAACAGAACCCGGCTTACGAC

Z.mays CTTATATGATTTCTGGTAGAATTGGAGAGCATTAAGTAGAAATATGATACATCGCTCTTT
         19320     19330     19340     19350     19360     19370
```

Figure 3.1: Ssearch output from search of the maize chloroplast genome with the *Synechocystis* pRNA sequence.

This putative maize pRNA sequence was then aligned with the *Porphyra* chloroplast pRNA (Figure 3.2a) and with the *Synechocystis* pRNA (Figure 3.2b). There was a large amount of scattered homology between the putative maize chloroplast pRNA and the *Synechocystis* pRNA and the *Porphyra* chloroplast pRNA even though the maize sequence is much shorter. There are, however, no long stretches of homology between these sequences. The putative maize chloroplast pRNA sequence is hereafter referred to as the maize chloroplast pRNA sequence or the maize chloroplast P-like sequence.

Searching with a high AT rich pRNA sequence in a high AT rich genome can cause many matches between the A's and T's. It is necessary to also observe the amount of G's and C's that are homologous between the sequences in order to determine if a search alignment shows a potential organelle pRNA sequence. The percentage of GC matches between the maize chloroplast pRNA sequence and the *Synechocystis* and *Porphyra* chloroplast pRNA sequences (shown on Figure 3.2) is 38% and 28%, numbers higher than were seen between these pRNA sequences and rRNA sequences also indicated by Ssearch.

A:

```
Synechocystis-PCC6803  GAGAGTTAGGGAGGGAGTTGCGGATTCCTGTCACAGGGAATCTGAGGAAAGTCCGGGCTT
Putative maize pRNA    ------------------------------------------------------TATT
                                                                              **

Synechocystis-PCC6803  CCCAAAGGCCAAACTTGCTGGGTAACGCCCAGTGCGCGCGAGCGTGAGGACAGTGCCACA
Putative maize pRNA    CCCGC--GTCTTCTATTCAGAGCATCCCCTACTCTTTTTTAGGGAAAGAGCGATGTATCA
                       ***    * *     * * * * * * ** * *       ** *  **  * ***   **


Synechocystis-PCC6803  GAAAAATACCGCCCTTTTTAGAAAACAGCAACCAGTAAACAGTTAACAGGGTTTTTCTGC
Putative maize pRNA    TATTTCTACTTAATGCTCTCCAATTC---TACCAGAAATCATATAAGAATTTGTTAGGAG
                             ***       * *  **  *    ***** ** **  *** *   * **

Synechocystis-PCC6803  TGACTAACTGGTAACTGACCACTGAAAAGG-TAAGGGTGCAAAGGTGCGGTAAGAGCGCA
Putative maize pRNA    TAAATTCCTT-TAACGGATTCATCCATAGTCTTAGGGCAGAATGGCCTTTTGACTCTGTA
                       * * *  **  **** **    *  * **  * ****   ** **     * *    * *

Synechocystis-PCC6803  CCAGCAGTATCGAGAGG-TACTGGCTCGGTAAACCCCGGTTGGAAGCAAGGTCGGAGGGG
Putative maize pRNA    CCCTTGATTCCACTATGATTATTGATCAATAG--------TAGAAGAATTCTTTCATAGA
                       **     *  *   * * *  * * **  **         * **** *   *   *  *


Synechocystis-PCC6803  CAAAGGTTGG-TCTTTTTCCTGCCCCATGATTGGTGGAACCGCTTGAGGAATTTGGTAAC
Putative maize pRNA    AATAGGAGACATAATTTACATGG---AT-ATAGTAAGTCTCACTTGGGCTGCTT--TAAT
                        * ***     *  *** * **    ** ** *   *   * **** *    **  ***

Synechocystis-PCC6803  AAATTTCCCAGATAGATAACTCCCCAAGGGTGCCTCGCATCCTGGAACAGAACCCGGCTT
Putative maize pRNA    GGTAGTC-----TTTACATTTTCTCTTTCGCTAGTAGTATGGGGGAGGA-----------
                             **     *  * *  * * *    *    * * **   ***  *

Synechocystis-PCC6803  ACGACTAACTCTCTTTTTTT
Putative maize pRNA    --------------------
```

Total nucleotide matches = 147 (44% of maize pRNA)
Total G-C matches = 57 (38% of total matches)

B:

```
Porphyra purpurea chloroplast    -AAGTAAACGTAGATAGCTGCAGATT-TGATTTAATATAGTCTGAGGAAA
Putative maize pRNA              TATTCCCGCGTCT-TCTATTCAGAGCATCCCCTACTCTTTTTTTAGGGAAA
                                  *      ***   *   * ****   *    ** * *  * *  *****

Porphyra purpurea chloroplast    GTCCGGGCTC-CACAAATACAATTTGTGCTGGAAAAACCCCAGTGTAGGA
Putative maize pRNA              GAGCGATGTATCATATTTCTACTTAATGCTCTCCAATTCTAC---CAGAA
                                 *  **   *  ** * *  * ** ****    **  *       ** *

Porphyra purpurea chloroplast    AACTATGAGGAACGTGCCACAGAAATAAACCGCCAGAACAAATTATGTTA
Putative maize pRNA              ATCATATAAGAATTTGTTA--GGAGTAAATTCCTTTAACGGATTCATCCA
                                 * *    * ***  **  *  * * ****   *   ***  ***     *

Porphyra purpurea chloroplast    GCTGGTAAGGGTGCAAAGGTAAGTTAAAAGCTTACCAAAAATACTGTAAA
Putative maize pRNA              TAGTCTTAGGGCAGAATGGCCTTTTGACTCTGTACCCTTGATTCCACTAT
                                       ****   ** **    ** *     ****    ** *    *

Porphyra purpurea chloroplast    GTATTTGTTAGGTAAACCCCAAAAACGGAG-CAAAGCGACTAAATAAATT
Putative maize pRNA              G-ATTATTGATCAATAGTAGAAGAATTCTTTCATAGAAATAGGAGACATA
                                 * ***  * *   * *    ** **      ** **  *    * * **

Porphyra purpurea chloroplast    TTTTGTATTTTTCAGTAGTTTAGTCTAAAATA--CTGCATGAA-GTTATT
Putative maize pRNA              ATTTACATGGAT--ATAGTA-AGTCTCACTTGGGCTGCTTTAATGGTAGT
                                  ***  **   *   ****  ***** * *  *   **** * ** * ** *

Porphyra purpurea chloroplast    GGCAACAATAACTCTAGAGGAATAGCTATCCTTCGAAATACTTGTGCAAT
Putative maize pRNA              CTTTACATTTTCTCTTTCGCTAGTAGTATGGGGGAGGA-----------
                                     *** *  ****   *  *    ***        *

Porphyra purpurea chloroplast    GTATTCGGAGAACAGAACCCGGCTTATGTAGTACTTTTTA
Putative maize pRNA              ----------------------------------------
```

Total nucleotide matches = 148 (45% of maize pRNA)
Total G-C matches = 41 (28% of total matches)

Figure 3.2: ClustalX sequence alignments of the putative maize pRNA with A: *Synechocystis* pRNA and B: P*orphyra purpurea* chloroplast pRNA. * indicates common nucleotides between the two sequences.

This maize chloroplast pRNA was then used in a BLAST search of Genbank and related databases. There was significant homology found in the rice (Hiratsuka et al. 1989), tobacco (Shinozaki et al. 1996), and spinach (Hudson et al. 1988) chloroplast genomes. Neither the *Porphyra* chloroplast pRNA nor any eubacterial pRNA sequences were identified with this search. Sequences from the Black Pine (*Pinus thunbergii*) (Wakasugi et al. 1994) and Liverwort (*Marchantia polymorpha*) (Ohyama et al. 1986) chloroplast genomes were identified with only limited homology.

Ssearches of the rice, tobacco and spinach genomes with the maize pRNA sequence identified the same regions as were found in the BLAST search with scores of 1645, 624, and 605 respectively.

ClustalX alignments of all four of these chloroplast pRNA sequences (Figure 3.3) showed that these sequences have limited homology to each other. The putative pRNA sequences from Rice, Tobacco and Spinach are referred to from this point on as the green plant pRNA sequences.

```
Maize     ----------------------------AATGGTAACACGAAAAAAGAAAAAATCTCT--CTATTCCCGCGTCTTCTATTCAGAGCATC
Rice      -------ACCCACCGTTCTTTCCTCAAACTGGGGTTT-GTTTTAGCAGTAGTGGCCCAAACCGATACCAATAGGGATGAG--ATAGCGTT
Tobacco   GTGCTCAAGCATCCAATCAAT--TGAAATGAGGGTCA-ACAAAAAAAAGAATAGGACCTA--TTATTCCTACATGTTCCATTAGTAACATT
Spinach   --------------TTCTTCCCCATACTACAAGTGAAAGAGAAAATGTAAAGACTACCA-TTAAAGCAGCCCAAGCGAGACTTACAATA
                                           **       *  * *      *  *          *   *   *

Maize     CCCTACTCTTTTTTAGGGAA-AGA----GCGATGTATCATATTTCTAC--TTAATGCTCTCCAATTCTACCAGAAATCATATAAGAATTT
Rice      TCCTTATTCTTTTTATTAATA-AGAAAGGGCTATGTATCATATTTATAC--TTAATGCTCCCCAATTCTACCAGAAATACTCTAAGAATTT
Tobacco   CCCTTGAGATGTTACTGC---A--------GATTT----TGCTTGTGT--TTAATCTTTCCCGATT-----AGAAATCCTATAGGAATTT
Spinach   TCCATGTGAATTATGTCCTCTATTTCTATGAATATGAAGGAATTATTCCATTATTGTTTACTAATAATAG-TGAAATCC-ATGGTACAGA
           **       *        *         ** *      ** *   *** * *  *  **       *****    *   *

Maize     GTTAG-----GAGTAAATTCCTTTAACGGATTCATCCATAGTCTTAGGGCAGAATGGCCTTTTGACTCTGTAC-CCTTGATTCCACTATG
Rice      GTTAA-----GAGTAAATTCTTTGAACTGATTCATCCATAGCCTTAGGTCAGAACCCCTTTTTGACTCTGTAC-CCTTGATTCTACTATG
Tobacco   CTTATAAAATGAGCGAATTTATTGGATTGGTTTATTAATAGTCTTCGTTCT--------TTTTGACTCTGCGC-CATTGATTCCACTATT
Spinach   GTCAAAAAATGGGCGAACT-ATTCATTTTATGAACCAATCCCATAAATCCA----------ATACCTGAGTACTCCGAGACTCTTTTGAG
           * *      * *  ** *  **       *  *   **    *     *           *  **  *  * *   ** **   *

Maize     ATTATTGATCAA-TAGTAGAAGAATTCTTTCATA------GAAATAGGAGACATAATTTACATGGATATAGTAAGTCTCACTTGGGCTGC
Rice      ATTATTCATCAA-TAGTAGAATAATTCCTTCATA------GAAATAGGAGACATAATTTACATGGATATAGTAAGTCTCGCTTGGGCTGC
Tobacco   ATTAGTGAGGAA-TAACGGAACAATTCCTTTATATTTATAGAGATAGGGGACATAATTCATATGGATATAGTAAGTCTTGCTTGGGCTGC
Spinach   TTTGTATACAAACTATATGCCGCATTCCGCTTT----------TCTTTTCCACAATTACTAATTAGTTAGATATTCC--CTCCCGCCGA
           **    *  ** **   *    ****     *          *      ** ****   *   *  ***  * **   **   ** *

Maize     TTTAATGGTAGTCTTTACATTTTCTCTTTCGCTAGTAGTATGGGGGAGGA-------
Rice      TTTAATGGTAGTCTTTACATTTTCTCTTTCACTAGTAGTATGGGGGAGGAGTGGAC-
Tobacco   TTTAATGGTAGTCTTTACTTTTTCCCTTTCACTCGTAGTGTGGGGAAGAAGTGGACT
Spinach   ATTCAAGGCCCAGTT---AGTAACCACTCCAATAGGAAT---GGAGAGGTTT-----
           ** * **     **     *  *   * *  * * * *   **  **
```

Figure 3.3: ClustalX sequence alignment of all four putative green plant chloroplast pRNAs. * indicates common nucleotides between all the sequences.

Alignments with the *Synechocystis* pRNA sequence (Figure 3.4) and the *Porphyra* chloroplast pRNA sequences (Figure 3.5) also showed very little homology between these sequences and the green plant chloroplast sequences. However, both the *Synechocystis* and the *Porphyra* chloroplast pRNA sequences are longer than the green plant pRNA chloroplast sequences and this may have an effect on the alignments.

```
Maize       ------------------------------------------------------------------AATGGTAAC
Rice        ---------------------------------------ACCCACCGTTCTTTCCTCAAACTGGGGTTTGTTTT
Tobacco     ----------------------------------------------------GTGCTCAAGCATCCAATCAATTG
Spinach     ------------------------------------------------TTCTTCCCCATACTACAAGTGAAAGA
Synechocys GAGAGTTAGGGAGGGAGTTGCGGATTCCTGTCACAGGGAATCTGAGGAAAGTCCGGGCTTCCCAAAGGCCAAACT

Maize       ACGAAAAAAGAAAAAATCTCTCTATTCCCGCGTCTTCTATTCAGAGCATCCCC--TACTCTTTT-TTAGG--GAA
Rice        AGCAGTAGTGGCCCAAAC---CGATACCAATAGGGATGAGATAGCGTTTCCT---TATTCTTTTATTAATAAGAA
Tobacco     AA-ATGAGGGTCAACAAAAAAAGAATAGGACCTATTATTCCTACATGTTCCAT--TAGTAACATTCCCTTGAGAT
Spinach     GAAAATGTAAAGACTACCATTAAAGCAGCCCAAGCGAGACTTACAATATCCATGTGAATTATGTCCTCTATTTCT
Synechocys TGCTGGGTAACGCCCAGTGCGCGCGAGCGTGAGGACAGTGCCACAGAAAAATACCGCCCTTTTTAGAAAACAGCA
                       *                              *                        *

Maize       AGAGCGATGTATCATATTTCTACTTAATGCTCTCCAATTC-TACCAGAAATCAT-ATAAGAATTTGTTAG-----
Rice        AGGGCTATGTATCATATTTATACTTAATGCTCCCCAATTC-TACCAGAAATACT-CTAAGAATTTGTTAA-----
Tobacco     GTTACTGCAGATTTTGCTTGTGTTTAATCTTTCCCGATT------AGAAATCCT-ATAGGAATTTCTTATAAAAT
Spinach     ATGAATATGAAGGAATTATTCCATTATTGTTTACTAATAA-TAG-TGAAATCC--ATGGTACAGAGTCAAAAAAT
Synechocys ACCAGTAAACAGTTAACAGGGTTTTTCTGCTGACTAACTGGTAACTGACCACTGAAAAGGTAAGGGTGCAAAGGT
                    *              **  *  *  *  *          **                    *

Maize       GAGTAAATTCCTTTAACGGATTCATCCATAG-TCTTAGGGCAGAATGGCCTTTTGACTCTGTACCCTTGATTCCA
Rice        GAGTAAATTCTTTGAACTGATTCATCCATAG-CCTTAGGTCAGAACCCCTTTTTGACTCTGTACCCTTGATTCTA
Tobacco     GAGCGAATTTATTGGATTGGTTTATTAATAG-TCTTCGTTC--------TTTTTGACTCTGCGCCATTGATTCCA
Spinach     GGGCGAACT-ATTCATTTTATGAACCAATCC-CATAAATCCAATA---CCTGAGTACTCCGAGACTCT--TTTGA
Synechocys GCGGTAAGA-GCGCACCAGCAGTATCGAGAGGTACTGGCTCGGTAAACCCCGGTTGGAAGCAAGGTCGGAGGGGC
           * *  **                  *   *              *

Maize       CTATGATTATTGATCAATAGTAGAAGAATTCTTTCATA--------GAAATAGGAGACATAATTTACATGGATAT
Rice        CTATGATTATTCATCAATAGTAGAATAATTCCTTCATA--------GAAATAGGAGACATAATTTACATGGATAT
Tobacco     CTATTATTAGTGAGGAATAACGGAACAATTCCTTTATAT--TTATAGAGATAGGGGACATAATTCATATGGATAT
Spinach     GTTTGTATACAAACTATATGCCGCATTCCGCTTTTCTT-------------TTCCACA--ATTACTAATTAGTT
Synechocys AAAGGTTGGUCTTTTTCCUGCCCCATGATTGGTGGAACCGCTTGAGGAATTTGGTAACAAATTTCCCAGATAGAT
                                 *       *                              ***   **   *   *  *

Maize       AGTAAGTCTCACTTGGGCTGCTTTAATGGTAGTCTTTACATTTTCTCTTTCGCTAGTAGTATGGGGGAGGA----
Rice        AGTAAGTCTCGCTTGGGCTGCTTTAATGGTAGTCTTTACATTTTCTCTTTCACTAGTAGTATGGGGGAGGAGTGG
Tobacco     AGTAAGTCTTGCTTGGGCTGCTTTAATGGTAGTCTTTACTTTTTCCCTTTCACTCGTAGTGTGGGGAAGAAGTGG
Spinach     AGATATTCC--CTCCCGCCGAATTCAAGGCCCAGTT---AGTAACCACTCCAATAGGAAT---GGAGAGGTTT--
Synechocys AACTCCCCAAGGGTGCCTCGCATCC-TGGAACAGAACCCGGCTTACGACTAACTCTCTTTTTTT-----------
           *      *            *  *    **                              *     *

Maize       ---
Rice        AC-
Tobacco     ACT
Spinach     ---
Synechocys ---
```

Figure 3.4: ClustalX multiple sequence alignment of the four putative green plant chloroplast pRNA sequences and the *Synechocystis* pRNA sequence. * indicates common nucleotides between all the sequences.

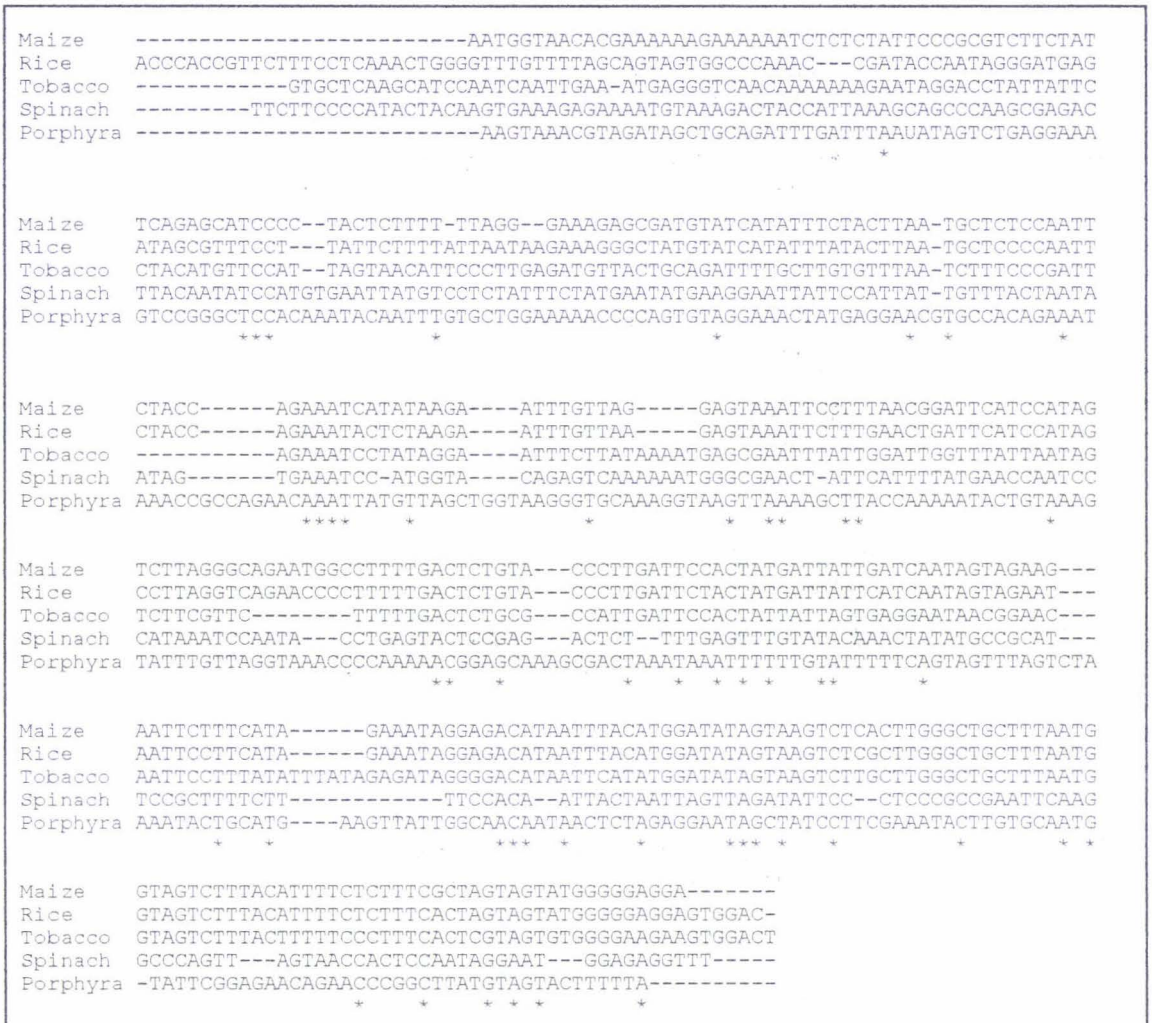


Figure 3.5: ClustalX multiple sequence alignment of the four putative green plant chloroplast pRNA sequences and the *Porphyra purpurea* chloroplast pRNA sequence. * indicates common nucleotides between all the sequences.

Not only are the sequences conserved between chloroplasts, their position in the genome is also conserved (Figure 3.6). They are located downstream from the trnC gene ($tRNA^{cys}$) and overlap an unidentified potential open reading frame, ORF29. Analysis of the upstream regions of these chloroplast pRNA sequences show likely -10 consensus sequences for the start of transcription. In the pRNA gene of the *Cyanophora paradoxa* cyanelle there is a 65-nucleotide separation between the most likely -10 sequence and the proposed start site (Baum and Schön 1996). A 31-nucleotide separation is found in the maize, 121-nucleotide separation in the rice, 10-nucleotide separation in the tobacco and 187-nucleotide separation in the spinach chloroplast pRNA sequences.

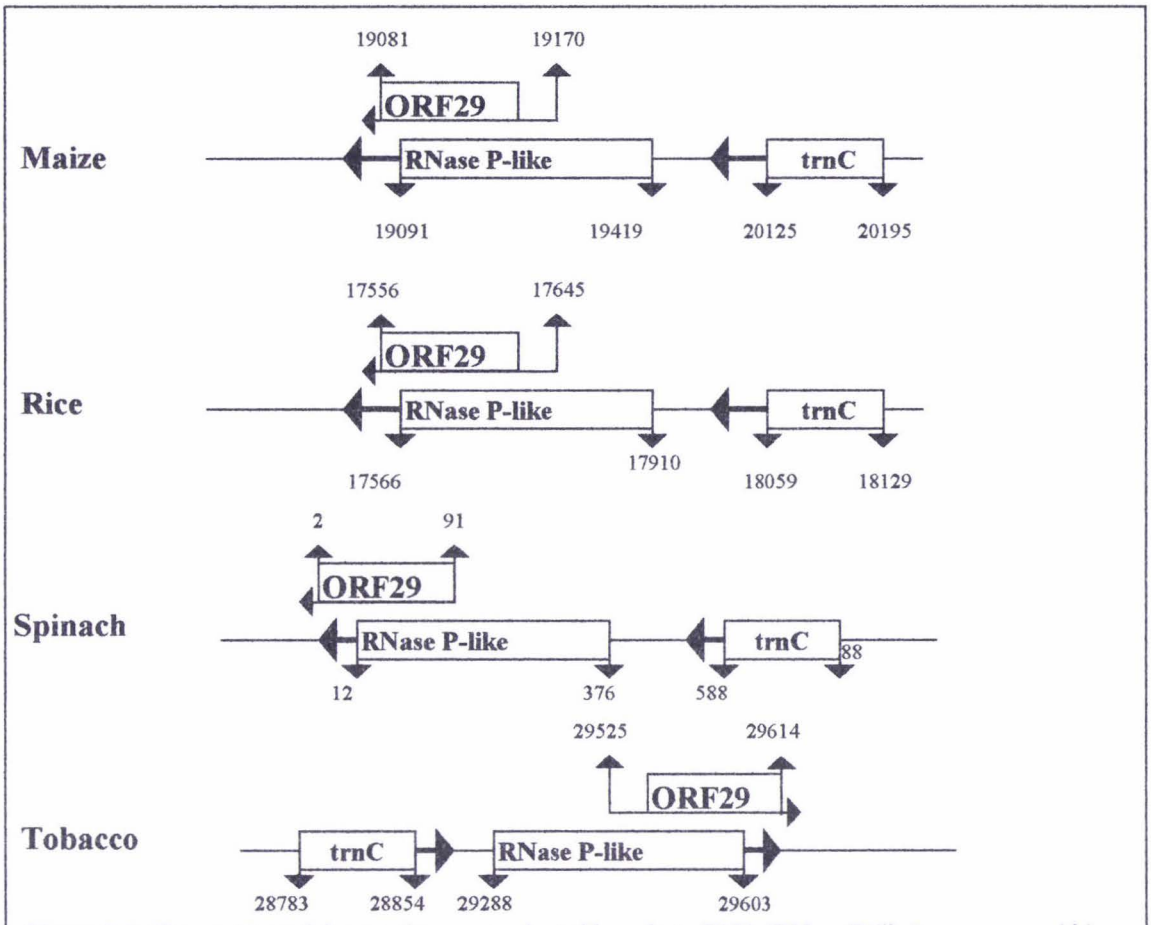


Figure 3.6: The position of the putative green plant chloroplast pRNA (RNase P-like) sequences within the four chloroplast genomes. Numbers refer to the positions of the sequences within the chloroplast genomes according to the GenBank sequences. Note that the Tobacco sequences are in the opposite orientation to the others due to the complementary strand being represented in the databases.

The sequence alignments above were used together with the consensus pRNA secondary structure, and the secondary structures from the cyanobacterial and the *Porphyra* chloroplast pRNAs, to construct hypothetical secondary structures (Figure 3.7). All four chloroplast pRNA sequences can fit a common secondary structure except for the lack of the P19 helix in the rice structure. Comparisons with the eubacterial consensus, cyanobacterial and *Porphyra* chloroplast pRNA secondary structures show that the chloroplast pRNA secondary structures lack the P13 helix. However, the chloroplast pRNA secondary structures have retained the characteristic long-range helices P4 and P6 which are necessary for the pRNA pseudoknot formation (Haas et al. 1996). The short sequences that make up these two helices have been shown to be highly conserved in pRNA sequences (Darr et al. 1992).

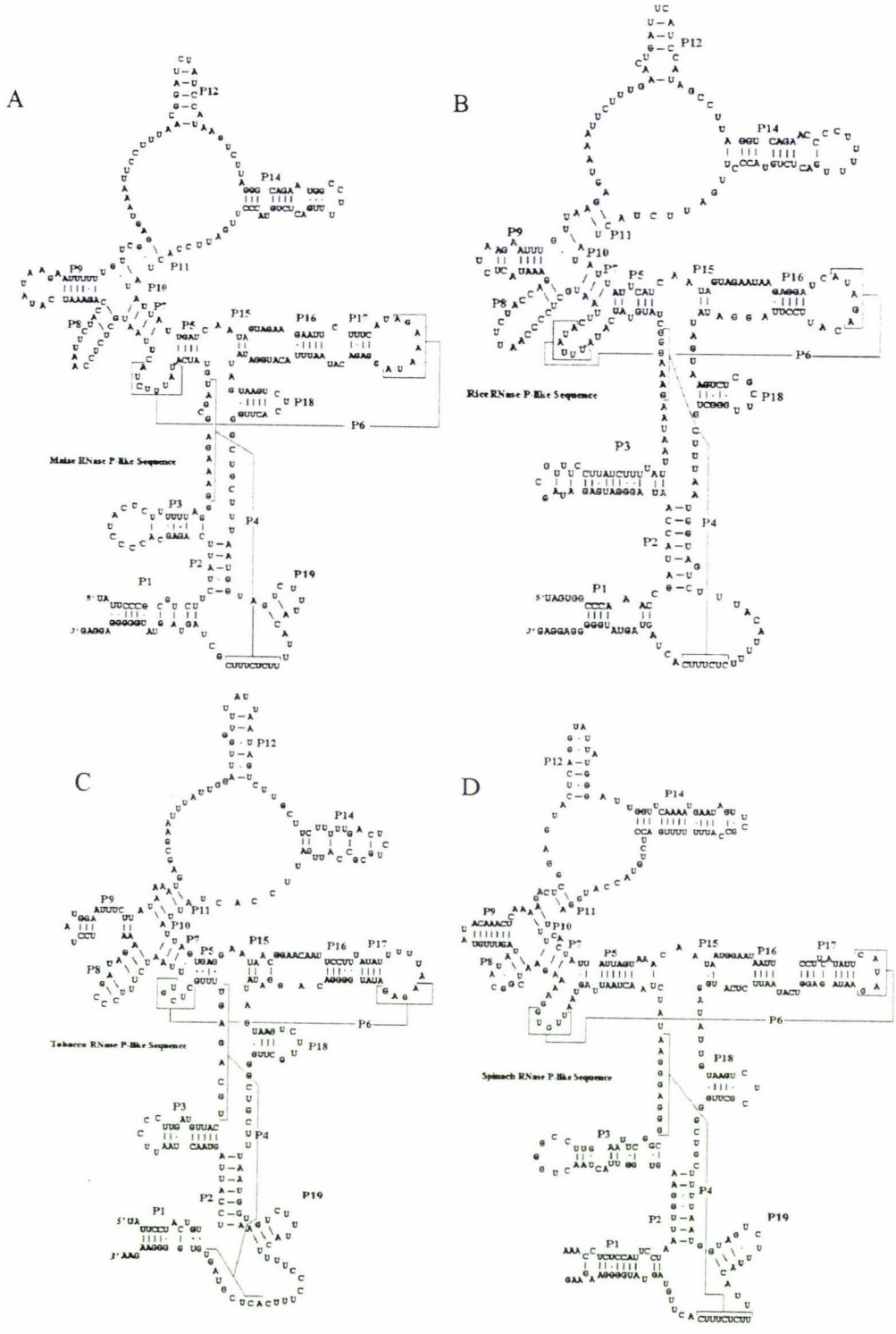


Figure 3.7: Hypothetical secondary structures of the putative green plant chloroplast pRNA sequences from A: Maize, B: Rice, C: Tobacco and D: Spinach. Helices are numbered according to Haas et al. 1994.

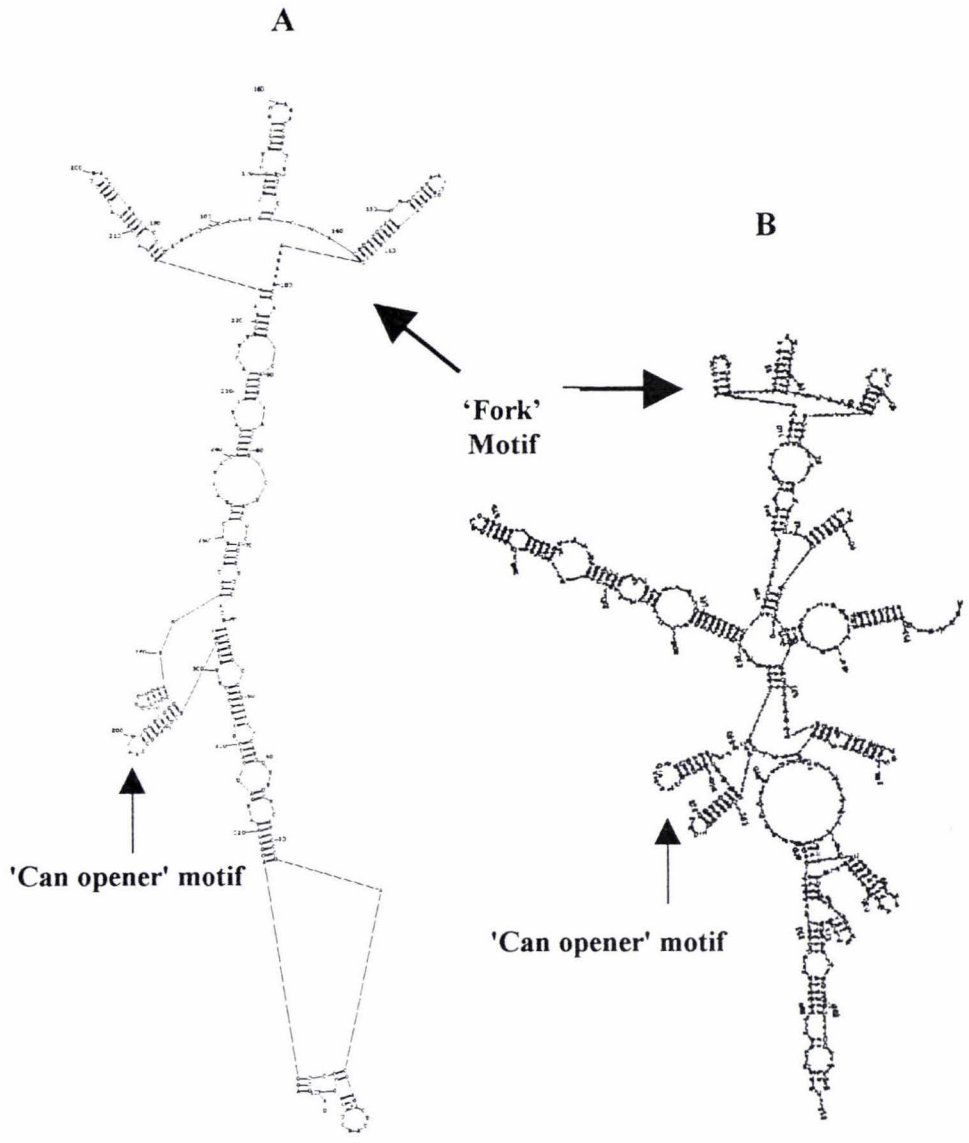


Figure 3.8: Structures folded using RNAstructure (mfold algorithm) of **A**: the maize chloroplast pRNA and **B**: *Synechocystis* pRNA showing both the 'fork' and 'can opener' motifs that have been found in other pRNA folded structures.

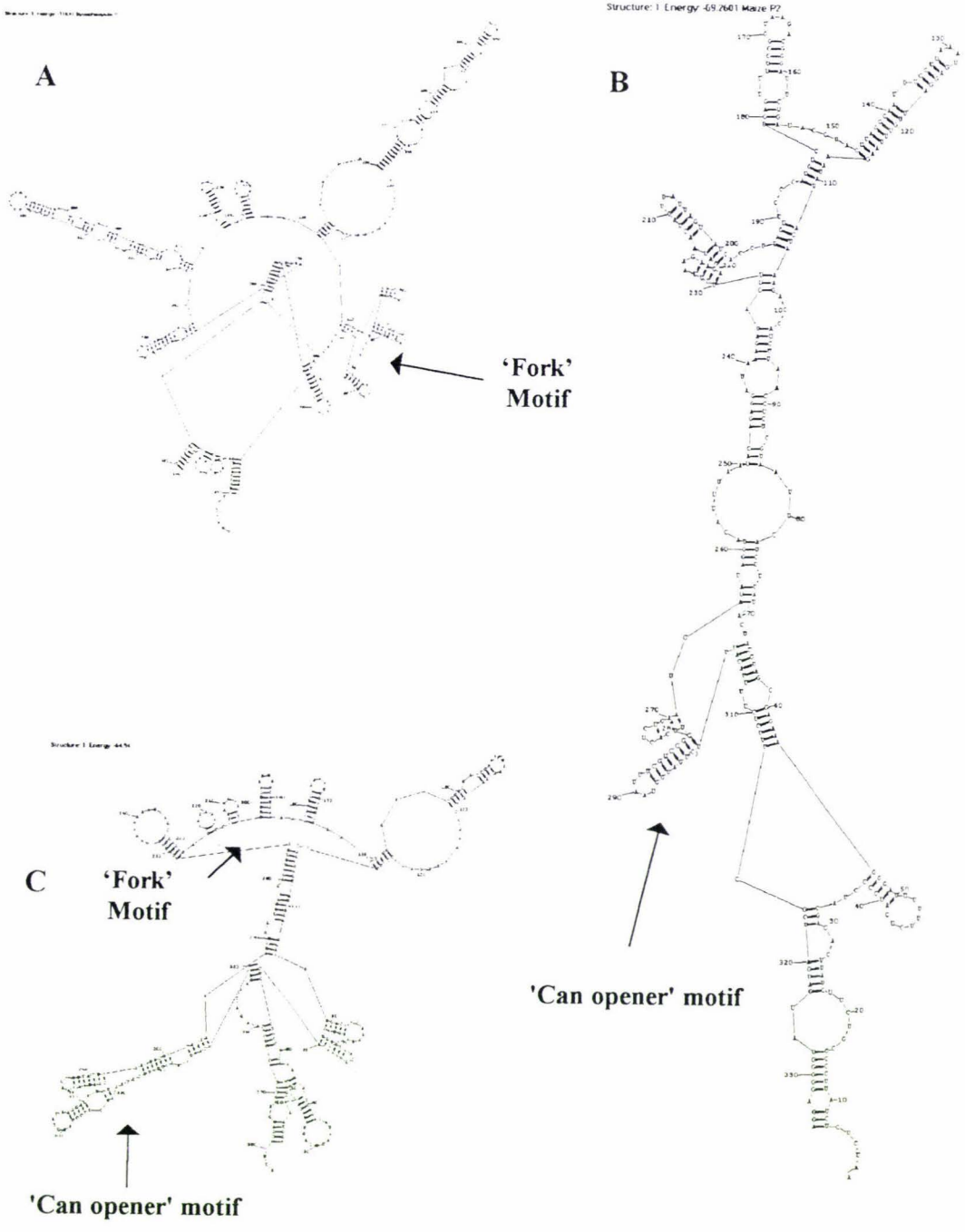


Figure 3.9: Structures folded using the RNAdraw program (RNAfold algorithm) of **A**: *Synechocystis* pRNA, **B**: the maize chloroplast pRNA and **C**: the *Porphyra purpurea* chloroplast pRNA, showing both the 'fork' and 'can opener' motifs found in other pRNA.

The maize chloroplast, cyanobacterial, and the *Porphyra* chloroplast pRNA sequences were folded using the folding programs, RNAstructure, and RNAdraw. The RNAstructure folded structures for *Synechocystis* and the maize chloroplast pRNA sequences show striking similarity and contain two structural 'motifs' which have been arbitrarily labelled the 'fork' and the 'can-opener' (Figure 3.8). These two motifs have also been observed in other pRNA structures folded with RNAstructure and RNAdraw (structures shown in appendix 1). The RNAdraw folded structures (Figure 3.9) do not show similarity between the structures to the same extent as the RNAstructure foldings, but the 'can-opener' motif has been retained in the putative maize chloroplast structure. With this program there appears to be more variability in the structures and both the *Porphyra* chloroplast pRNA and the maize chloroplast pRNA form a circular type structure which is explained more fully in chapter 5. The structures in Figure 3.9 have been 'corrected' for this structure.

The green plant chloroplast sequences identified in this chapter show many features found in other characterised pRNA sequences. However, these identified sequences can only be labelled as 'putative' until biochemical assays can determine if they do encode pRNA activity, or bind to such substrates that would identify them as true chloroplast pRNA sequences.

Discussion

The sequences identified show many features that would be expected in a chloroplast-encoded pRNA. They lie in the same region of the chloroplast genomes and have potential upstream features that have been found in other plastid pRNA sequences. Folding with two folding programs have shown that these sequences do fold into similar secondary structures to those of other pRNA sequences.

The hypothetical biological secondary structures were very hard to construct because the sequences have such a scattered homology to other pRNA sequences. Without any long stretches of homology it was not easy to match possible secondary structure regions to those found in other pRNA secondary structures. It is thus possible then that the biological structures for the green plant chloroplast pRNA sequences are not completely correct. The shortness of these chloroplast pRNA sequences could indicate some helices

are missing, but this could not be determined here. The present way of constructing reasonably accurate secondary structures is to find many sequences and align them. Computer programs such as Covariation (Haas et al. 1996) are used to find nucleotides in the sequence that covary (i.e. change at the same time and maintain their pairing or non-pairing ability). These programs require at least 20 sequences to be effective and limited time for finding more green plant chloroplast sequences ruled out their use. Other methods of constructing secondary structures include biochemical analysis.

The hypothetical open reading frame ORF29 that lies almost completely within the putative pRNA sequences is only 29 amino acids long. Its shortness, and its near total immersion within the putative pRNA sequences, means it is unlikely to code for a protein in maize, rice, tobacco, and spinach chloroplasts. In liverwort and black pine chloroplasts the ORF29 regions are homologous to the maize ORF29 but homology to the putative pRNA sequences or any pRNA sequences was not seen. These genomes are even more AT-rich than those found in other chloroplast pRNA sequences. It is possible that these regions may fold into a pRNA-type secondary structure but this would require much work, possibly involving biochemical analysis. A very high AT content, however, does not seem to be detrimental to a functional pRNA molecule. The mitochondrial -encoded pRNA of the yeast *Saccharomyces cerevisiae* has an AT content of 87% and is still functional.

Several other interesting regions were found throughout the chloroplast genomes having Ssearch scores from 86 to 95. These were identified later as the 16S, 23S rRNA genes, introns and tRNA genes. It is interesting that all these genes have RNA secondary structure and suggests that this searching strategy could be used for finding other catalytic RNA genes in other organelles.

The maize pRNA sequence is examined in subsequent chapters along with other pRNA and mrpRNA sequences. The folded secondary structures do show some striking similarity and this is investigated more fully in later chapters where these structures are compared to other pRNA structures quantitatively.

The finding of potential pRNA sequences in some organellar genomes indicates that it is possible that previous searches of organellar genomes may have missed catalytic RNA sequences hidden in their AT-rich genomes. The Ssearch program used here has been found to be able to eliminate unnecessary background by only allowing a search of the genome in

question, thus allowing a more rigorous search. This program could possibly be used in searching for other catalytic RNA sequences, especially when sequence homology is expected to be low. The use of RNA secondary structure in the characterisation of catalytic RNA has been shown to be an effective way of investigating the potential of a sequence before the expensive process of biochemical analysis is begun. Although not covered in this study, it is hoped that at some stage the pRNA sequences identified here are assayed for P activity, and then confirmed as green plant chloroplast pRNA sequences.

Chapter 4

Evaluation of RNA biological secondary structure for use in determining evolutionary relationships.

Chapter preface

The bulk of this chapter has been written as a paper that is to be submitted to the Journal of Molecular Evolution. This has necessitated some repetition from the introduction of this thesis. Due to the submission format required by this journal there are some formatting differences between this chapter and the rest of this thesis.

The figures for this chapter are not prefixed with the number four, thus Figure 4.1 is simply Figure 1. The figures for this chapter are included, separate from the text, at the end of this chapter. The 16S rRNA Domain I (Figure 4.7) and Domain III (Figure 4.8) data and analysis, which is not shown (but mentioned) in the paper, is included at the end of this chapter. The neighbor-joining trees of the mrpRNA sequences (Figure 4.9), mrpRNA secondary structures (Figure 4.10) and the tree constructed from the secondary structure data of mrpRNA and pRNA (Figure 4.11) are also given at the end of this chapter.

MRP has been thought to be related to P based on secondary structure and functional similarities. Any relationship between P and MRP will go back to at least to an early divergence of eukaryotes. Given the high observed rate of sequence divergence it is expected that the sequence homology would be low; thus sequence alignments may not give clear information. Any lack of confidence in the sequence alignment will cast major doubt on the validity of any phylogenetic tree constructed from them.

The biological secondary structure of catalytic RNA molecules is directly related to the function and therefore may hold information as to the evolution of that molecule. Quantitative secondary structure analysis of the biological secondary structures of pRNA and mrpRNA is used to examine the evolutionary relationship between these molecules.

Preliminary work with 16S rRNA comparing only very coarse structural characteristics (covered in this chapter) is indicative of how the RNA secondary structure of catalytic RNA molecules could indicate evolutionary distances.

Use of RNA Secondary Structure for Evolutionary Relationships: Investigating RNase P and RNase MRP

Lesley Collins, Vincent Moulton* and David Penny

Institute of Molecular Biosciences, Massey University, Palmerston North New Zealand.

*Present address: Department of Physics and Mathematics, Mid Sweden University, Sundsvall, Sweden.

Abstract

RNA secondary structure is evaluated for determining evolutionary relationships between catalytic RNA molecules that are distantly related, and when there is little confidence in sequence alignments. RNase P (P) and RNase MRP (MRP) are used for testing this as they are ribonucleoproteins that due to functional and secondary structure similarities, are thought to be evolutionary related.

P activity is found in all cells, and fits the criteria for inclusion in the RNA world (Jeffares et al. 1998). MRP is found only in eukaryotes with essential functions in both the nucleus and the mitochondria. The RNA components of P and MRP (pRNA and mrpRNA) cannot be aligned with any confidence, which leads to uncertainty in any phylogenetic trees constructed from them.

If MRP evolved from P only in eukaryotes, then it is an apparent exception to the general process of the transfer of catalytic activity from RNA to ribonucleoproteins, and then to proteins (Jeffares et al. 1998). Although there are many protein families, this is evidence for a possible family of RNA molecules. The possibility that MRP evolved with P in the RNA world (and has since been lost from all but the eukaryotes) is considered.

Quantitative comparisons of the pRNA and mrpRNA biological secondary structures have found that the third possibility of an organellar origin of MRP is unlikely. It is concluded that RNA secondary structure can be used in the evaluation of an evolutionary relatedness between MRP and P, and the approach could be extended to other catalytic RNA molecules.

Key Words

RNase MRP, RNase P, RNA secondary structure, RNA world, catalytic RNA.

Introduction

Catalytic RNA is functionally active RNA, which folds into a three-dimensional structure. They are also involved in many essential cellular processes. It has recently been suggested, based on the presumed transfer of catalysis from RNA to ribonucleoproteins, to proteins, that catalytic RNA molecules generally are remnants of the RNA world (Jeffares et al. 1998). The present study examines the quantitative analysis of RNA secondary structure for the determination of evolutionary relationships, especially when there is little confidence in sequence alignments. RNase P (P) and RNase MRP (MRP) are used for testing this as they are ribonucleoproteins (consisting of a catalytic RNA and at least one protein subunit), that due to functional and secondary structure similarities, are thought to be evolutionary related (Karwan 1993). The RNA components (pRNA and mrpRNA) have little sequence homology, resulting in sequence alignments that do not have enough reliability to confidently determine an evolutionary relatedness.

P cleaves tRNA precursors to form the mature 5' ends of tRNA molecules, with activity being found all cells tested including prokaryotes, eukaryotes and also in organelles. Prokaryotic P consists of an RNA strand, and a single protein subunit whereas the P encoded in the nucleus of eukaryotes has several protein subunits (Pace and Smith 1990). In one case it appears that the RNA is lost and P activity is entirely due to proteins (Rossmanith and Karwan 1998). Yeast species such as *Saccharomyces cerevisiae* and *Aspergillus nidulans* have retained their mitochondrialy-encoded pRNA whereas vertebrate mitochondria as well as those from the yeast *Schizosaccharomyces pombe* have lost their pRNA gene, and use nuclear-encoded products. In plants, mitochondrial pRNA activity has been shown (Marchfelder and Brennicke 1993), but to date no genes have been characterised.

The secondary structure of prokaryotic pRNA has been reported to show characteristic features in different phylogenetic groups (Pace and Brown 1995) and consensus structures have been drawn for groups of eubacteria and archaebacteria (Haas et al. 1996, Pace and Brown 1995). This is an indication that some features in the pRNA secondary structure are fixed and others variable. For the purposes of this study, prokaryotic pRNA will include that from eubacteria, mitochondria, and chloroplasts. The pRNA from archaebacteria will not be covered at this time.

MRP (Mitochondrial Ribosomal Processing) has been found only in eukaryotes, initially as an endoribonuclease that cleaves RNA primers for the initiation of mitochondrial DNA replication (Morrissey and Tollervey 1995). Subsequently a nuclear function in rRNA processing was identified, consistent with its predominant localisation to the nucleolus (Lygerou et al. 1996). MRP consists of an RNA moiety and multiple protein subunits with at least 7 of these, Pop1p (Morrissey and Tollervey 1995), Pop3p (Dichtl and Tollervey 1997) Pop4 (Chu et al. 1997) Pop5p, Pop6p, Pop7p and Pop8p (Chamberlain et al. 1998) proteins being shared with P in the yeast *Saccharomyces cerevisiae*. mrpRNA secondary structures (Schmitt et al. 1993) have only been characterised for eight species and show great similarity with each other despite being from plant, yeast and vertebrate species. Although the secondary structures are similar the nucleotide sequences vary greatly in length and nucleotide composition, making alignment difficult.

Comparison of the RNA secondary structures between mrpRNA and pRNA have shown similarity in shape, especially in the 'cage region' of the RNA molecule in which there is the characteristic pseudoknot formation (Forster and Altman 1990). However, to date, there has been no published quantitative comparison of pRNA and mrpRNA secondary structure. The similarities in secondary structure are possibly the direct result of conservation of tertiary and thus functional characteristics.

Functional similarities have led to the conclusion that these two ribonucleoproteins (RNP's) are evolutionary related (Morrissey and Tollervey 1995). Both the P and MRP ribozymes cleave RNA's to generate 5' phosphate and 3' hydroxyl termini in a reaction requiring divalent cations (Forster and Altman 1990). Both P and MRP are sensitive to puromycin, an antibiotic which inhibits pre-tRNA processing (Potuschak et al. 1993) and enzymatic activities from P and MRP isolated from several organisms cofractionate through multiple stages of biochemical purification (Paluh and Clayton 1995). It has been reported that MRP and P may be involved together in a macromolecular complex within the nucleolus (Lee et al. 1996), raising the possibility that the relationship between MRP and P may be of a functional nature, based on their sharing of many protein subunits (Sbisà et al. 1996).

We consider three general hypotheses of how MRP could have evolved from or with P. The three groups of hypotheses are as follows:

I MRP evolved from an eukaryotic nuclear P in the nucleus of the eukaryotic cell. This could occur by gene duplication followed by divergence of function of the two homologues. This is the theory most commonly suggested in previous studies (Morrissey and Tollervey 1995, Reddy and Shimba 1996, Chamberlain et al. 1996). MRP would have been incorporated into multiple eukaryotic functions and has also gained an essential function in the mitochondria. Under this hypothesis MRP is found only in eukaryotes because it was never in any of the other lineages. MRP is present in animals, yeasts, and plants indicating an early divergence from P but would not necessarily have to be in all early eukaryotes. We would expect under this hypothesis the secondary structures of the mrpRNA to be more similar to eukaryotic pRNA than to prokaryotic pRNA.

Under this hypothesis MRP is an exception to the transfer process of catalysis (RNA to RNP to protein) (Jeffares et al. 1998) with a ribonucleoprotein taking on a new catalytic function, after the widespread availability of protein catalysts.

II MRP evolved from an endosymbiont P. There are several variants on this hypothesis. MRP could have evolved from the hypothetical endosymbiotic fusion that formed the first eukaryote (Gupta and Golding 1996, Martin and Muller 1998) or by some later endosymbiosis that let to the mitochondrion. This theory accounts for the essential mitochondrial function of MRP and requires that MRP picked up additional rRNA processing functions in the nucleus. In plants it has been shown that organellar DNA can move to the nucleus and yet retain a function in the organelle (Brennicke et al. 1993, Wischmenn and Schuster 1995, Blanchard and Schmidt 1995). We might expect here that mrpRNA would retain some organellar characteristics such as a higher A + T content in nucleotide sequence and be more closely related in secondary structure to that of the organellar or prokaryotic pRNA.

III MRP and P evolved in the RNA world. The RNA world hypothesis suggests that DNA and proteins evolved from a world in which RNA was the both the catalytic and information storage molecule, and that today's catalytic RNA species are molecular relics from this time. There are three main criteria used to evaluate the antiquity of an RNA molecule (Jeffares et al. 1998) and pRNA fits all three of these criteria by being ubiquitous, catalytic and central to metabolism. MRP on the other hand fits only the last two criteria, being present only in the eukaryotic lineage. A central concept to the RNA world is that proteins with superior catalytic properties have

gradually replaced RNA as the catalytic molecule (and that no novel catalytic RNAs would be formed after the advent of efficient genetically encoded protein synthesis) (Jeffares et al. 1998). It is difficult to see how a molecule such as MRP could have evolved only in the eukaryotic lineage to integrate itself so intimately into rRNA processing, mitochondrial genome replication, and perhaps other functions central to eukaryotic metabolism. It has been found that eukaryotes carry more proposed 'relics' of the RNA world than prokaryotes. These 'relics' include spliceosomes, telomerase and self splicing introns which are all absent from prokaryotes (Jeffares et al.1998). MRP was the only widely occurring catalytic RNA not suggested to be a relic from the RNA world in Jeffares et al. (1998); its status was left unresolved.

There are also several variants of this hypothesis; MRP could have evolved from P, P evolving from MRP, and MRP and P evolving independently in the RNA world. With the possibility that MRP had a function in the RNA world, before the advent of proteins and DNA it is important to know more about the evolutionary relationship of P and MRP.

With such a great divergence expected between pRNA and mrpRNA (at least back to early divergences of eukaryotes) nucleotide sequence alignments may not be reliable enough to determine with confidence any evolutionary relationship. However, examination of the RNA secondary structure may yield the required information when the sequence data cannot. It has been shown that many sequences can fit the same secondary structure (Fontana et al. 1993) which allows the catalytic RNA sequence to vary even if the function of the molecule remains unchanged. The secondary structure of the catalytic RNA molecule has fixed 'motifs' that represent areas that are critical to maintaining the function, and other regions that are free to vary in presence or size. It is expected that these fixed and variable regions of the catalytic RNA secondary structure will change according to the evolution of the function of the molecule, and thus may be used to determine evolutionary relationships when the sequence data cannot. Quantitative comparisons of pRNA and mrpRNA secondary structures are used here to calculate distances between these molecules in order to assess their relatedness.

Materials and Methods

pRNA sequences and prokaryotic pRNA secondary structures were obtained from the RNase P Database (Brown 1998). mrpRNA sequences were obtained from Genbank and the remaining secondary structures from pRNA and mrpRNA were obtained from the literature. All sequences and secondary structure references are given in Table 1.

16S rRNA sequences and secondary structures were obtained from Ribosomal Database Project (RDP) (Maidak et al. 1997). 16S (prokaryotes) and 18S (eukaryotes) rRNA sequence alignments and trees were obtained using the Subalign and Subtree programs available at the RDP (http:\\rdp.life.uiuc.edu). These subtrees, taken from the overall tree of life, were used as standards against which our trees of mrpRNA and pRNA sequences and structures were compared. The advantage in using the subtrees is that they were constructed from many more aligned sequences than were available in our comparisons. A further subtree of 16S rRNA (Figure 2A) sequences from thirteen prokaryotic species was used as a standard tree for the comparison of 16S rRNA secondary structure features. Domain I (as defined by Gutell et al. 1994) of the 16S rRNA secondary structures of 16 prokaryotic species (as shown in Table 1) was divided into areas based on homologous helix formation then the number of nucleotides within that area were determined for each species. Overall differences were calculated and neighbor-joining trees constructed

Prior experience showed that ClustalX (Thompson et al. 1997), Divide and Conquer simultaneous alignment program (Dress et al.1996) and Dialign (Morgenstern et al.1996) were particularly suitable for evaluating distantly related sequences. Genetic distances were obtained from these alignments using DNAdist with the Jukes Cantor multiple substitution correction from the Phylip package (Felsenstein 1989).

RNA secondary structures were converted to bracket notation and then compared with RNAdistance from the Vienna RNA package (Hofacker et al. 1994) using the tree-editing alignment and full structure notation options. The tree-editing algorithm calculates a distance based on the number of steps required to convert one structure to others. By converting secondary structures to the bracket notation we then enabled the RNAdistance program to calculate a distance metric based on the full structural detail available in the RNA secondary structure. The secondary structures were first converted from published structures (noted in table 1) to 'bracket' notation

consisting of a string of '(' and ')' to represent paired nucleotides and '.' to represent unpaired nucleotides (Hofacker et al. 1994).

Phylogenetic trees were built from both the sequence and structural distance data using the neighbor-joining algorithm from the Neighbor program (also from the Phylip package) and the Splitstree package (Huson 1997; Dress et al. 1996). Two of the methods offered by the Splitstree package: Splitstree and refined Buneman produce networks to represent relationships between the taxa. The Splitstree and refined Buneman methods have the advantage that, unlike neighbor-joining, they depend continuously on the distance matrix (i.e. small changes in the distance matrix do not lead to large changes in the resulting tree or network (Moulton et al. 1997; Moulton and Steel 1998)). The neighbor-joining method does not indicate any inconsistencies in the data, unless sometimes when taxa are loaded in a different order (Figure 1). The Splitstree and refined Buneman methods tend to produce networks, rather than simple trees for data that is not very tree-like. It will not accept internal structure unless it is well supported and serves as a good indicator of inconsistencies in the data set (Dress et al. 1996; Bandelt and Dress 1992). However, trees constructed with Splitstree from many taxa can become confusing. The refined Buneman method produces trees that can be easily interpreted but will still indicate any data inconsistency. Where there is the possibility of any inconsistent data only the refined Buneman trees are shown here. All trees were drawn with TreeView (Page 1996).

Results

It was expected that trees from secondary structures would only become useful when sequences were highly diverged, for more recent times sequences would be preferable.

The first question was to determine whether an evolutionary pattern could be detected in an intermediate range, by the quantitative analysis of a simple secondary structure characteristic, when the results could be compared directly with trees constructed from sequences. Only then could the secondary structure be examined in greater detail. rRNA was used for this first stage because the sequences had already been aligned successfully, and well-studied secondary structures were available (The Ribosomal Database Project (RDP) Maidak et al. 1997). Thus we could compare the tree from secondary structure alone with that of the sequence data.

50% of the internal branches were the same when the tree constructed from the Domain I 16S rRNA secondary structure data was compared with the 16S rRNA sequences (Figure 2B). The internal branching of the archaebacterial species, the branching from the archaebacteria to *Thermus thermophilus*, and the internal branching between the archaebacteria and eubacteria were maintained. The internal branching between two pairs of eubacterial species; *Streptomyces coelicolor* and *Frankia sp*, and *Escherichia coli* and *Pseudomonas testosteroni*, was also the same between the two trees. The misplacement of only two other eubacterial species, *Agrobacterium tumefaciens* and *Mycoplasma capricolum*, affected the other half of the internal branches. However, as expected the deep internal structure of the tree was recovered, but there was less resolution at the species level.

These results indicated that simple rRNA secondary structures could be compared in a quantitative manner. However, it was evident that in this case, the internal structure of the tree could be seriously affected by the misplacement of only two taxa. A probable cause of this effect would be that the distances were calculated from only a small number of features. Domain III of the 16S rRNA from the same eubacterial and archaebacterial species was also examined separately and combined with Domain I (data not shown), but this domain had even less variability than Domain I and gave very low distance calculations. The resulting trees had even fewer internal branches the same when compared to the Subtree. Overall this indicated that for RNA secondary structure to be able to recover the correct tree, more secondary structure features would be required.

Prior to looking at the relationship of the secondary structures of pRNA and mrpRNA we compared the evolutionary relationship shown between the mrpRNA sequences and that shown by the mrpRNA secondary structures. This would determine if analysis of the secondary structures could discern the deep divergence of the plant, yeast and vertebrate species in the eukaryotic kingdom.

All eight mrpRNA sequences (five vertebrates, one plant, and two yeast strains, in table 1) were aligned using the Divide and Conquer and Dialign simultaneous alignment methods. These alignments were then maximised for sequence homology rather than for secondary structure as has been done in the past (Sbisà et al. 1996). Alignments tried with the pairwise comparison program, ClustalX (Thompson et al. 1997) did not cope well with the large internal gaps required to align the longer yeast

mrpRNA sequences with the shorter vertebrate and plant mrpRNA sequences. Trees built from the Divide and Conquer and Dialign alignments, using the neighbor-joining, Splitstree and refined Buneman methods, show the same internal branching as is shown in the Subtree of 18S rRNA sequences (Figure 3).

Trees built from the mrpRNA secondary structures (Figure 4) gave very similar results to that given by the aligning the nucleotide sequences. However, the internal branching of the vertebrates has minor differences. In the trees constructed from sequence alignment data, the mouse and rat; and the human and bovine mrpRNA sequences are paired, with the xenopus mrpRNA sequence outside these pairings. In the trees constructed from secondary structure data, we see the human and xenopus mrpRNAs together and the bovine structure grouped with the mouse and rat structures. The internal branch between the mouse and rat mrpRNA structures remains the same as is shown in the alignment trees. Both the alignment and structure trees show that the yeast mrpRNA secondary structures have diverged as much from each other as they have diverged from the vertebrate and plant mrpRNA secondary structures. The internal branching between the vertebrates, the plant species, *Arabidopsis*, and the yeasts (*S. cerevisiae* and *S. pombe*) is the same between the sequence and structural trees (and the subtree constructed from the 18S rRNA sequences). Trees constructed with Splits-tree and neighbor-joining gave very similar results (data not shown).

The consistency between all these trees indicates that the mrpRNA secondary structure data (which does not have to be aligned) is as good as the sequence data for determining reasonably distant relationships, like that between vertebrates, plants, and yeasts. This indicates that the quantitative analysis of catalytic RNA secondary structure can be used to look at evolutionary relationships even if the sequence data cannot be confidently aligned.

The next step was to see if quantitative analysis of the secondary structures of mrpRNA and pRNA could be used to determine an evolutionary relationship between the two. Alignments including both mrpRNA and pRNA sequences show little homology. This lack of homology is also shown between the eukaryotic nuclear and the prokaryotic pRNA sequences, an indication that the sequences have diverged greatly between the eukaryotes and prokaryotes. Neighbor joining, Splitstree, and refined Buneman trees were constructed from this alignment with the only the refined Buneman tree being shown here (Figure 5).

A number of prokaryotic, mitochondrial, chloroplast, and eukaryotic pRNA secondary structures were then quantitatively compared with the mrpRNA structures using RNAdistance. The resulting refined Buneman (Figure 6) tree shows separate groupings of the prokaryotic (including mitochondrial and chloroplast), eukaryotic nuclear pRNA and mrpRNA structures. Within the eubacterial group the cyanobacterial (*Anabaena*, *Anacystis* and *Synechocystis*) structures are grouped together and the proteobacteria (*Agrobacterium* and *Rhodospirillum*) structures are grouped together as is found in the subtree of 16S rRNA sequences.

The *Bacillus* secondary structure has been shown to be a little different from the consensus eubacterial RNase P structure and is shown as such in the RNase P Database (Brown 1998). This different structure is also indicated with *Bacillus* being grouped away from the other eubacteria. The maize P-like structure diverges near the base of the eubacterial tree with the *Bacillus* and the *Reclinomonas* pRNA structures. This structure was modelled on the cyanobacterial and *E.coli* structures but there was no biochemical analysis used in determining this structure. The eukaryotic nuclear pRNA structures show the vertebrate pRNA structures together and the yeast pRNA structures together, which is also shown in the Subtree of 18S rRNA sequences. The mrpRNA structures are grouped the same as is shown previously. Internal branching between the prokaryotic pRNAs, eukaryotic pRNAs, and the mrpRNAs shows that the mrpRNA structures are closer to the eukaryotic pRNA structures than to the prokaryotic structures. The closest species between the mrpRNAs and the eukaryotic pRNAs are the yeast structures from both groups. The eukaryotic pRNAs are shown to be as different from the prokaryotic pRNAs as the mrpRNA structures are. It is not clear from this data where the overall root of the tree is positioned.

Overall the refined Buneman tree shows a closer relationship between the mrpRNA and the eukaryotic pRNA secondary structures indicating that an organellar origin of MRP unlikely. The relationship within the three groups of catalytic RNAs show internal branching of species as would be seen in trees built from nucleotide sequence analysis of 16S or 18S rRNA. This indicates that the mrpRNA and pRNA secondary structure can be compared in a quantitative manner to estimate an evolutionary relationship.

Discussion

Initial analysis of the 16S rRNA Domain I gave good indications that RNA secondary structure could be useful in the analysis of ancient evolutionary relationships, even where sequences cannot be aligned with confidence. Even a very simple characteristic (like the length of helices) behaved in a tree-like fashion. More in depth, quantitative analysis of the whole structure of Domain I (or in fact the whole 16S rRNA) would be expected to result in a very similar tree to what has been determined with sequence analysis. However the nucleotide sequences of 16S rRNA are such that they can be used on their own for determining evolutionary relationships between species. Secondary structure analysis would not expect to add much to the analysis of 16S rRNA although it can be argued that it is the underlying covarion structure that is allowing the 16S rRNA sequences to recover correct trees.

Nucleotide sequence analysis of pRNA and mrpRNA, on the other hand, is not close enough to determine with confidence any evolutionary relationship between the two. However, the trees constructed by neighbor-joining and refined Buneman of the mrpRNA and pRNA secondary structures show that the prokaryotic pRNAs, the eukaryotic pRNAs and mrpRNAs have different structures and that overall mrpRNA is more similar in structure to eukaryotic pRNA than to any eubacterial or organellar pRNA.

Structural analysis has failed to show any close relationship between the any of the prokaryotic pRNA structures and those of the mrpRNA making unlikely the hypothesis that MRP is of an organellar/prokaryotic origin. With an organellar origin we also could have expected the A+T content of the mrpRNA gene to be higher than other genes encoded in the nucleus. A+T contents in mitochondrial genomes are generally high due to a unusually high rate of substitution compared to that of the nuclear genome (Lynch 1996) and that repair systems may show a preference for inserting dATP (Martin 1995). The A+T content of the mrpRNA genes is in fact very similar to those of other nuclear genes in the same species.

MRP is found in vertebrates, plants and yeast and thus if MRP evolved from the nuclear P in the eukaryotic lineage it must have been at an early stage, before the divergence of these three groups. In this context it will be interesting to determine which catalytic RNAs are found in the *Giardia* genome that is currently being sequenced (Sogin, M. personal communication).

Another indication of an early evolution is its essential nature within the eukaryotic lineage both in the nucleus and the mitochondria of plants and yeasts. This is an indication that MRP had to have evolved early enough to have separate but related roles from P in ribosome assembly and tRNA maturation. It had also to have evolved early enough to gain roles in the mitochondria, which in some species (e.g. vertebrates and *S. pombe*) have come to be completely reliant on the nuclear pRNA and mrpRNA products. The question we raise here is how early.

Sequence alignment and structural data alone cannot determine, at this stage, as to which hypothesis of MRP evolution is more likely, evolution early in the eukaryotic lineage or evolution in the RNA world. If the refined Buneman tree (Figure 6) constructed from structural data is rooted on the eubacterial lineage then MRP arose in eukaryotes. However, if the RNA world is used to root the tree of life (Poole et al. 1998), which was the reason for raising the initial question, the answer is not clear. Both hypotheses could be reflected in the highly divergent sequences. Nevertheless, by taking functional data into account we can see that there are some factors that favour the RNA world hypothesis.

Under the RNA world hypothesis MRP could have evolved from P, (or possibly even vice versa), or have arisen independently. It is feasible that some of the similarities seen between the secondary structures of mrpRNA and pRNA have arisen from common protein binding sites or higher structure conformations required for protein binding. The fact that there are multiple proteins shared between MRP and the eukaryotic P (e.g. in S. cerevisiae - Pop1, Pop3 and Pop4) could be an indication of a relationship that stems back to the time when the RNA and protein moieties of MRP and P were first assembled. In the RNA world the first proteins to evolve are expected to be RNA binding proteins with chaperone-like activity which would increase stability and help maintain ribozyme tertiary structure (Poole et al 1998). It is possible that MRP and P picked up chaperone-like proteins that had a function that was required by both (e.g. stability and transport) at a time when there were only a few proteins around.

It has been a misconception, perhaps, that the prokaryotic P is the more ancient form of P, because of the ability of its RNA to be catalytic on its own (with high Mg^{2+} concentrations) and because it contains only one protein subunit. In essence it is a simpler form of P than the eukaryotic version. However, the presence of multiple protein subunits may be an indication of an ancient origin. Ribosomal proteins, for example,

have been found to have a multifunctional nature with most of these proteins having functions additional to their role in the ribosome (Wool 1996). This sort of co-opting of a few ancestral proteins by multiple processes may have occurred with P and MRP. Prokaryotic P is specialised for one function, the maturation of tRNA, whereas eukaryotic P and MRP are involved in many functions in the eukaryotic cell. The eukaryotic pRNA and mrpRNA had likely a very complicated evolutionary pathway. Each interaction between the RNA and protein subunits (and also between the RNA and each substrate), would have been optimised to such a point that even with high Mg^{2+} concentrations, the RNA is no longer stable without any stabilising chaperone proteins. The prokaryotic RNase P, the eukaryotic P and MRP have evolved considerably since the first ancient ribonucleoprotein complex. Under the RNA world hypothesis these ancient P and MRP molecules would have had a stable RNA in the RNA world then gained multiple protein subunits as protein synthesis evolved.

Although many protein families and super-families have been isolated there do not appear to be any other potential families of catalytic RNA molecules. Analysis of RNA secondary structure and functional data have led us to suggest here a family of catalytic RNA consisting of P and RNase MRP. Other RNA families are possible, which also may not be able to be identified on sequence homology data alone.

Tertiary structure characteristics may have a greater potential for the determination of evolutionary relationships, as they are in an even closer relationship to the function of the molecule than is the secondary structure. There are two models at present for the tertiary structure of pRNA and none yet for mrpRNA (Pace and Brown 1995). It has been shown for proteins that analysis of the tertiary structure may be more revealing for distantly related structures and primary sequence analysis can be more useful for closer, less diverged structures (Gutell 1992).

Quantitative analysis of secondary structures on the other hand, produce distances that indicate that an organellar origin of MRP is unlikely. The two remaining hypotheses each show MRP to be an interesting molecule. If MRP evolved in eukaryotes then it seems that a RNP took on a catalytic function in preference to a protein, an exception to the general process of the transfer of catalysis. An RNA world origin of MRP, however, allows a new perspective in the analysis of this molecule.

In summary, although sequence alignments of pRNA and mrpRNA were obtained, there can be little confidence in any phylogenetic trees constructed from them due to the low

homology shown between the sequences. Analysis of secondary structure data offers an additional approach to evaluate ancient divergences.

Acknowledgments

This work was supported by the Marsden Fund. Thanks also to the Department of Mathematics and Statistics (FIMS) at Massey University for UNIX support.

References

Altman S, Wesolowski D and Puranam R (1993) Nucleotide sequences of the RNA subunit of RNase P from several mammals. Genomics 18:418-422

Bandelt H-J and Dress A (1992) A new and useful approach to phylogenetic analysis of distance data. Molecular Phylogenetics and Evolution 1:242-252

Blanchard J and Schmidt G (1995) Pervasive migration of organellar DNA to the nucleus in plants. J Mol Evol 41:397-406

Brennicke A, Grohmann L, Hiesel R, Knoop V and Schuster W (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Letters 325:140-145

Brown J (1998) The Ribonuclease P Database. Nucleic Acids Research 26:351-352

Bryant D and Moulton V (1998) A polynomial time algorithm for constructing the refined Buneman tree. Applied Mathematics Letters (in press)

Chamberlain J, Pagan-Rámos E, Kindelberger D and Engelke D (1996) An RNase P RNA subunit mutation affects ribosomal RNA processing. Nucleic Acids Research 24:3158-3166

Chamberlain J Lee Y, Lane W and Engelke D (1998) Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. Genes Dev 12:1678-1690

Chu S, Zengel J and Lindahl L (1997) A novel protein shared by RNase MRP and RNase P. RNA 3:382-391

Dichtl B and Tollervey D (1997) Pop3p is essential for the activity of the RNase MRP and RNase P ribonucleoproteins *in vivo*. The EMBO Journal 16:417-429

Dress A, Huson D, and Moulton V (1996) Analyzing and visualizing sequence and distance data using Splits-Tree. Discrete Applied Mathematics 71:95-109

Felsenstein J (1989) PHYLIP- Phylogeny inference package (version 3.2). Cladistics 5:164-166

Fontana W, Konings D, Stadler P, and Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389-1404

Forster A and Altman S (1990) Similar cage-shaped structures for the RNA components of all ribonuclease P and ribonuclease MRP enzymes. Cell 62:407-409

Gupta R and Golding G (1996) The origin of the eukaryotic cell. TIBS 21:166-171

Gutell R (1992) Evolutionary characteristics of 16S and 23S rRNA structures. In: Hartman H and Matsumo K (eds.) The origin and evolution of prokaryotic and eukaryotic cells. World Scientific Publishing Co, New York, p243

Gutell R, Larsen N and Woese C (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiological reviews 58:10-26

Haas E, Banta A, Harris J, Pace N and Brown J (1996) Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. Nucleic Acids Research 24:4775-4782

Hofacker I, Fontana W, Stadler P, Bonhoeffer L, Tacker M and Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie 125:167-188

Huson D (1998) Splitstree - a program for analysing and visualizing evolutionary data. Bioinformatics (in press)

Jeffares D, Poole A and Penny D (1998) Relics from the RNA world. J Mol Evol 46:18-36

Kirsebom L (1995) RNase P - a 'scarlet pimpernel'. Molecular Microbiology 17:411-420

Kiss T, Marshallsay C and Fillpowicz W (1992) 7-2/MRP RNAs in plant and mammalian cells: association with higher order structures in the nucleus. The EMBO Journal 11:3737-3746

Lang B, Burger G, O'Kelly C, Cedergren R, Golding G, Lemieux C, Sankoff D, Turmel M and Grey M (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493-497

Lee B, Matera A, Ward D and Craft J (1996) Association of RNase mitochondrial RNA processing enzyme with ribonuclease P in higher ordered structures in the nucleolus: A possible coordinate role in ribosome biogenesis. Proc. Natl. Acad. Sci. USA 93:11471-11476

Lygerou Z., Allmang C, Tollervey D. and Seraphin B (1996a) Accurate processing of a eukaryotic precursor ribosomal RNA by Ribonuclease MRP *in vitro*. Science 272:268-270

Lynch M (1996) Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes. Mol Biol. Evol 13:209-220

Maidak B, Olsen G, Larsen N, Overbeek R, McCaughey M and Woese C (1997) The RDP (Ribosomal Database Project). Nucleic Acids Research 25:109-111

Martin A (1995) Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol Biol Evol 12:1124-1131

Martin W, Muller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392: 37-41

Morgenstern B, Dress A and Werner T (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci USA 93:12098-12103

Morrissey JP and Tollervey D (1995) Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system. TIBS 20:78-82

Moulton V and Steel M (1998) Retractions of finite distance functions onto tree metrics Discrete Applied Mathematics (in press)

Moulton V, Steel M and Tuffley C (1997) Dissimilarity maps and substitution models. Proceedings of the DIMACS workshop on mathematical hierarchies. AMS 37:111-131

Moulton V, Zuker M, Steel M, Pointon M and Penny D (1988) Metrics on RNA secondary structure. Submitted to J Comput Biol

Pace N and Brown J (1995) Evolutionary Perspective on the structure and function of Ribonuclease P, a ribozyme. Journal of Bacteriology 177:1919-1928

Pace N and Smith D (1990) Ribonuclease P: function and variation. The Journal of Biological Chemistry 265:3587-3590

Page R (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12:357-358.

Paluh J and Clayton D (1995) *Schizosaccharomyces pombe* RNase MRP RNA is homologous to metazoan RNase MRP RNAs and may provide clues to interrelationships between RNase MRP and RNase P. Yeast 11:1249-1264

Penny D, McCormish B and Hendy M (1998) Modeling the covarion model by a hidden Markov chain. Submitted to Proc Natl Acad Sci USA

Poole A, Jeffares D and Penny D (1998) The Path from the RNA world. J Mol Evol 46:1-17

Potuschak T, Rossmanith W and Karwan R (1993) RNase MRP and RNase P share a common substrate. Nucleic Acids Research 21:3239-3243

Rossmanith W and Karwan R (1998) Characterization of human mitochondrial RNase P: novel aspects in tRNA processing. Biochem Biophys Res Commun 247: 234-241.

Saitou N and Nei M (1987) The Neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425

Sbisà E, Pesole G, Tullo A and Saccone C (1996) The evolution of the RNase P- and RNase MRP- associated RNAs: Phylogenetic analysis and nucleotide substitution rate. J Mol Evol 43:46-57

Shapiro B and Zhang K (1990) Comparing multiple A secondary structures using tree comparison. CABIOS 6:309-318

Schmitt M, Bennett J, Dairaghi D and Clayton D (1993) Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison. The FASEB Journal 7:208-213

Schuster P, Fontana W, Stadler P, and Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond B 225:279-284

Thompson J, Gibson T, Plewniak F, Jeanmougin F and Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res 25:4876-4882

Tranguch A and Engelke D (1993) Comparative structural analysis of nuclear RNase P RNAs from yeast. The Journal of Biological Chemistry 268:14045-14053

Wischmann C and Schuster W (1995) Transfer of *rps10* from the mitochondrion to the nucleus in *Arabidopsis thaliana*: evidence for RNA-mediated transfer and exon shuffling at the integration site. FEBS Letters 374:152-156

Winker S and Woese CR (1991) A definition of the domains *Archaea, Bacteria* and *Eucarya* in terms of small subunit ribosomal RNA characteristics. System Appl Microbiol 14:305-310

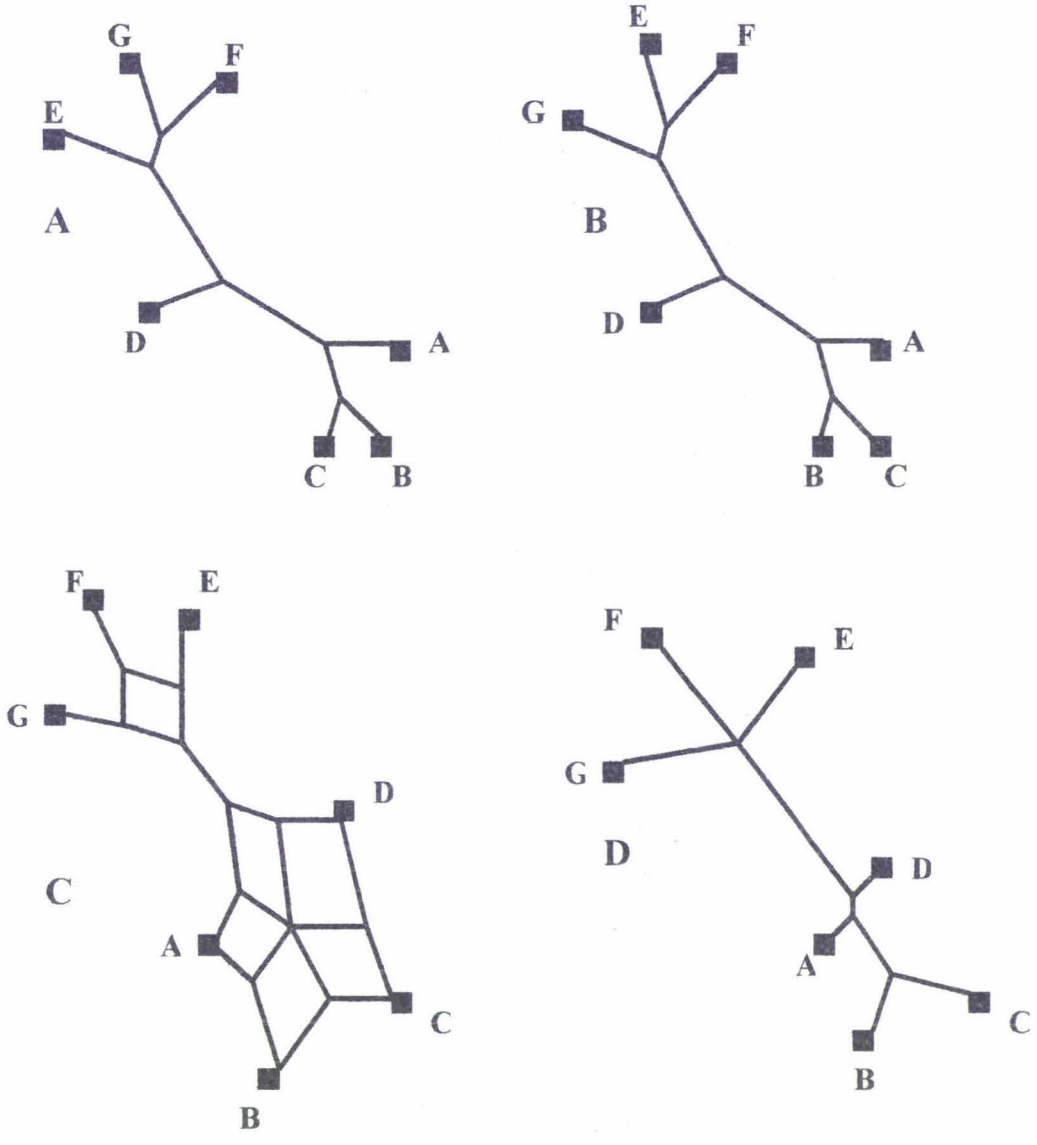
Woese C (1987) Bacterial Evolution. Microbiology Reviews 51:221-271

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

Wool IG (1996) Extraribosomal functions of ribosomal proteins. TIBS 21:164-165

	Accession Number	Length of Sequence	A + T %	Secondary Structure Reference
RNase P Sequences				
Synechocystis sp. PCC6803	X65707	437	48	P
Anabaena sp. PCC 7120	X65648	465	47	P
Anacystis nidulans PCC6301	X63566	385	43	P
Pseudoanabaena sp. PCC 6903	X73135	450	52	P
Escherichia coli	M17569	377	38	P
Bacillus subtilis	M13175	401	51	P
Rhodospirillum rubrum	M59355	429	29	P
Agrobacterium tumefaciens	M59354	402	36	P
Reclinomonas americana mitochondria	AF007261	312	75	P
Porphyra purpurea chloroplast	U38804	383	63	P
Human (nuclear)	X15624	340	36	Altman et al. 1993
Mouse (nuclear)	L08802	288	33	"
Danio rerio (nuclear) Zebrafish	U50408	308	43	No structure
Saccharomyces cerevisiae (nuclear)	M27035	368	48	Tranguch and Engelke 1993
Schizosaccharomyces pombe (nuclear)	X04013	373	48	"
RNase MRP Sequences				
Human	X51867	264	36	Schmitt et al. 1993
Bovine	Z25280	277	39	"
Mouse	J03151	275	36	"
Rat	J05014	273	35	"
Xenopus (frog)	Z11844	277	45	"
Arabidopsis thaliana	X65942	260	49	Kiss et al. 1992
Saccharomyces cerevisiae	Z14231	339	60	"
Schizosaccharomyces pombe		399	57	Paluh and Clayton 1995
RNase P-like Sequence				
Z. mays (maize) chloroplast	from X86563: 19091 - 19419	329	62	Note 1
16S rRNA structures	**RDP sequence**			
				RDP
Escherichia coli	E.coli	-	-	"
Clostridium innocuum	C.innocuum	-	-	"
Methanococcus vannielli	Mc.vanniel	-	-	"
Frankia sp.	Fra.spORS	-	-	"
Streptomyces coelicolor	Stm.coelic	-	-	"
Thermus thermophilus	T.thermoph	-	-	"
Bacillus subtilis	B.subtilis	-	-	"
Agrobacterium tumefaciens	Ag.tumefac	-	-	"
Spirochaeta aurantia	Spi.aurant	-	-	"
Thermoplasma acidophilum	Tpl.acidop	-	-	"
Mycoplasma capricolum	M.capricol	-	-	"
Methanobacterium formicicum	Mb.formici	-	-	"
Pseudomonas testosteroni	Ps.testost	-	-	"

Table 1. RNase P and RNase MRP RNA sequences used in this study showing length, Accession details, A+T % and from where the secondary structures were obtained
Key: *P* Obtained from the RNase P Database (Brown 1998),
RDP Obtained from the Ribosomal Database Project (Maidak et al. 1997).
Note 1: The maize chloroplast P-like sequence is a putative pRNA sequence isolated by sequence homology from the maize chloroplast and has yet to undergo biochemical analysis. It remains unpublished at this point in time.



SpeciesA 0
SpeciesB 2 0
SpeciesC 4 2 0
SpeciesD 5 5 3 0
SpeciesE 6 8 8 5 0
SpeciesF 7 9 9 6 3 0
SpeciesG 6 8 8 5 4 3 0

Figure 1: Comparison of three methods of constructing trees. **A**: Neighbor-joining with taxa *A, B, C, D, E*; **B**: Neighbor-joining with taxa *C, A, B, D, G, F, E*; C: Splitstree and D: refined Buneman.

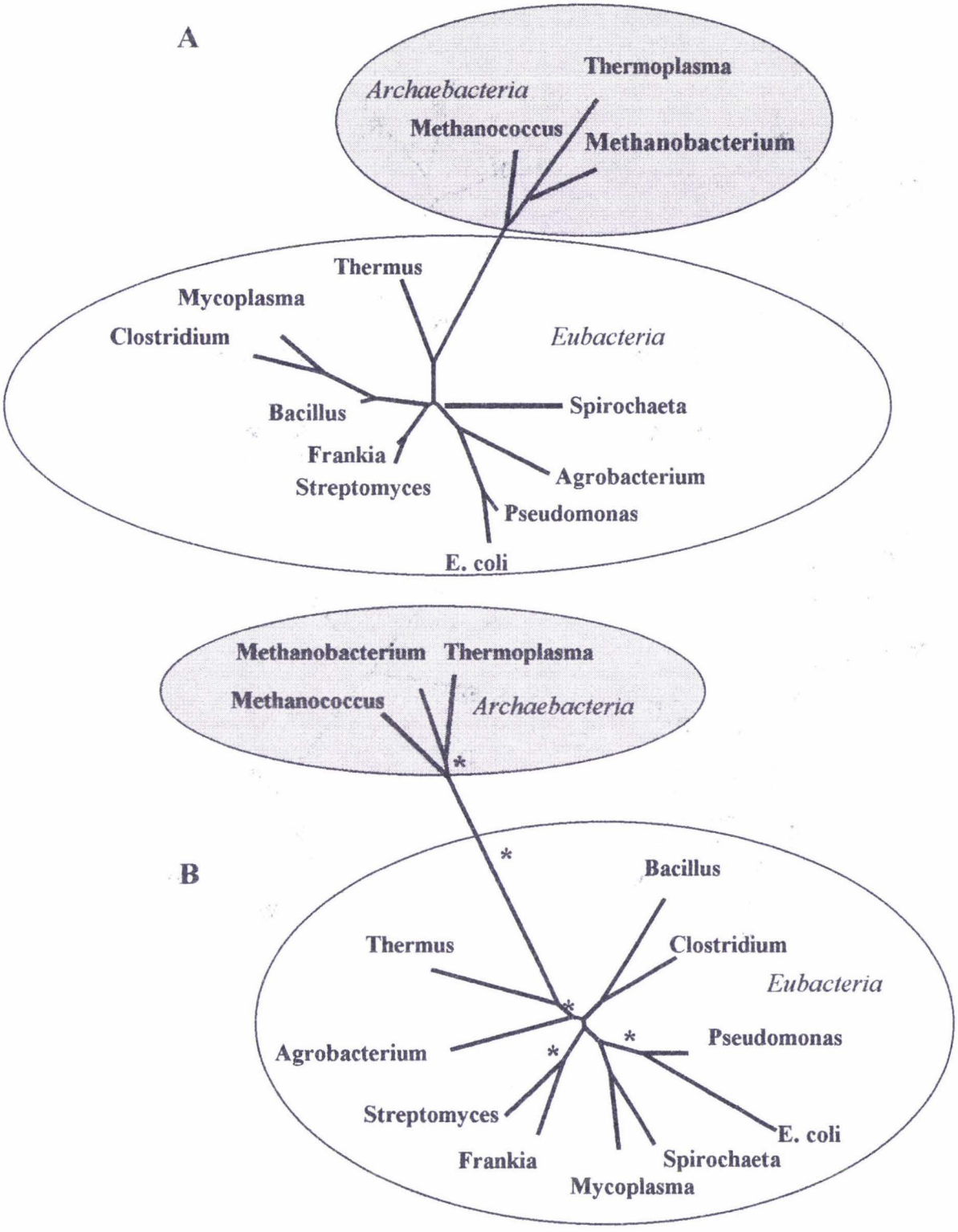


Figure 2: **A**: Subtree of 16S rRNA eubacterial and archaebacterial sequences from the Ribosomal Database Project. **B**: Neighbor-joining tree of 16S rRNA Domain I length data. * indicates the internal branches that are the same as those in the subtree.

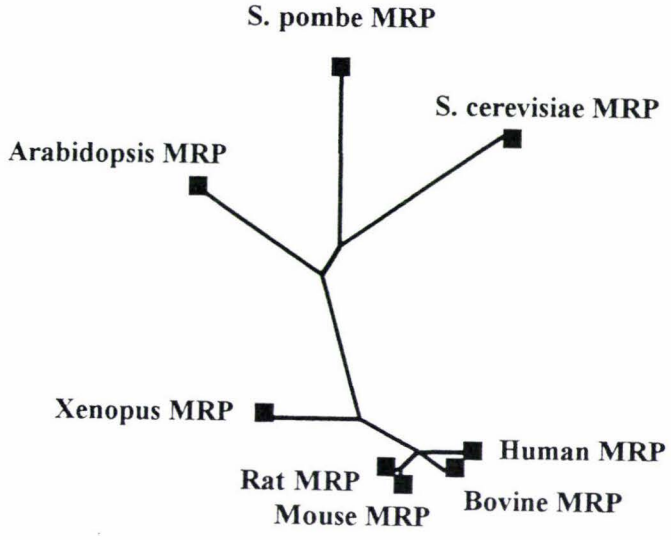


Figure 3: Refined Buneman tree of mrpRNA sequences aligned by Divide and Conquer. The Splitstree and Neighbor-joining trees were identical to this tree.

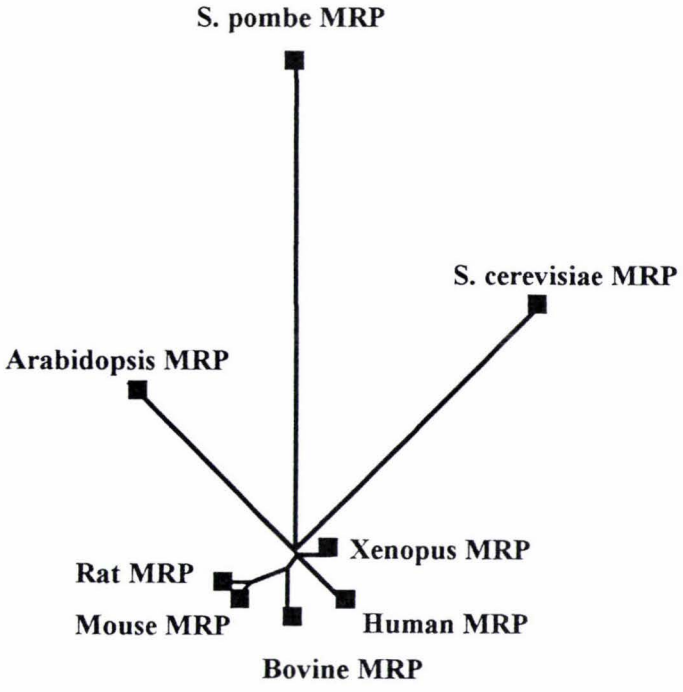


Figure 4: Refined Buneman tree of mrpRNA secondary structures compared by RNAdistance.

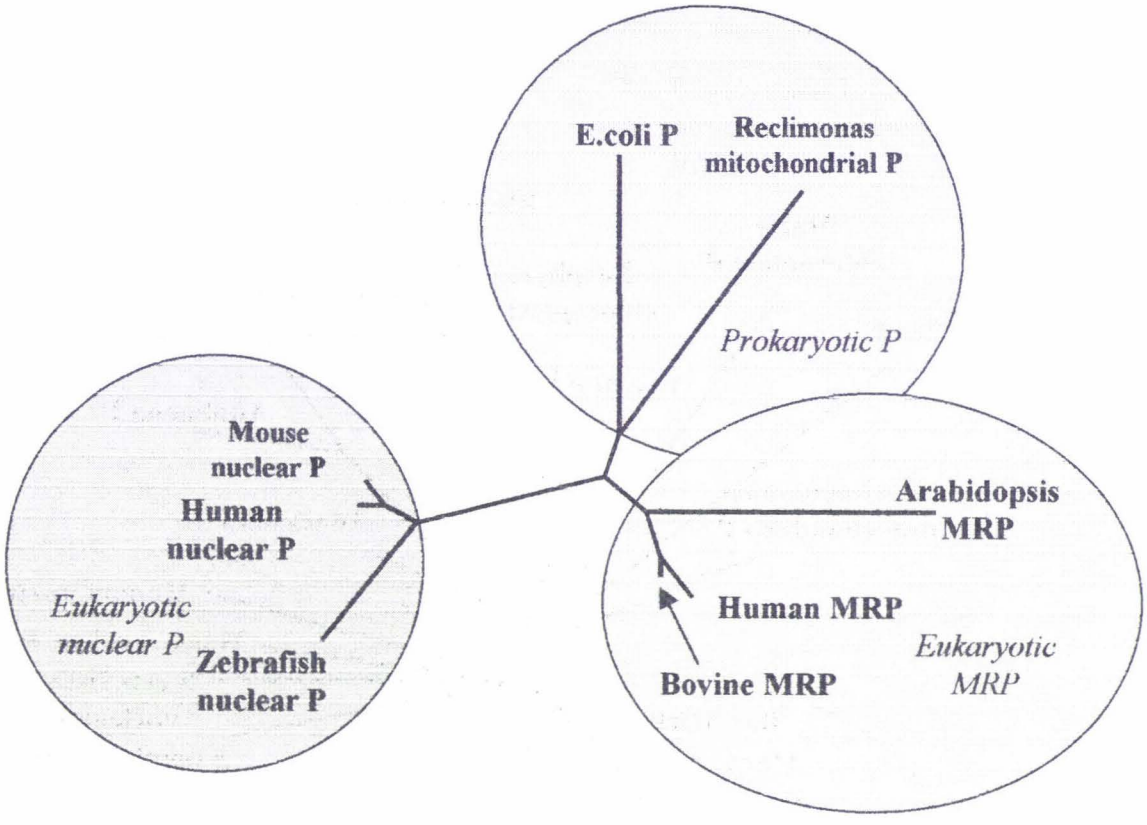


Figure 5: Refined Buneman tree of mrpRNA and pRNA sequences aligned by Divide and Conquer.

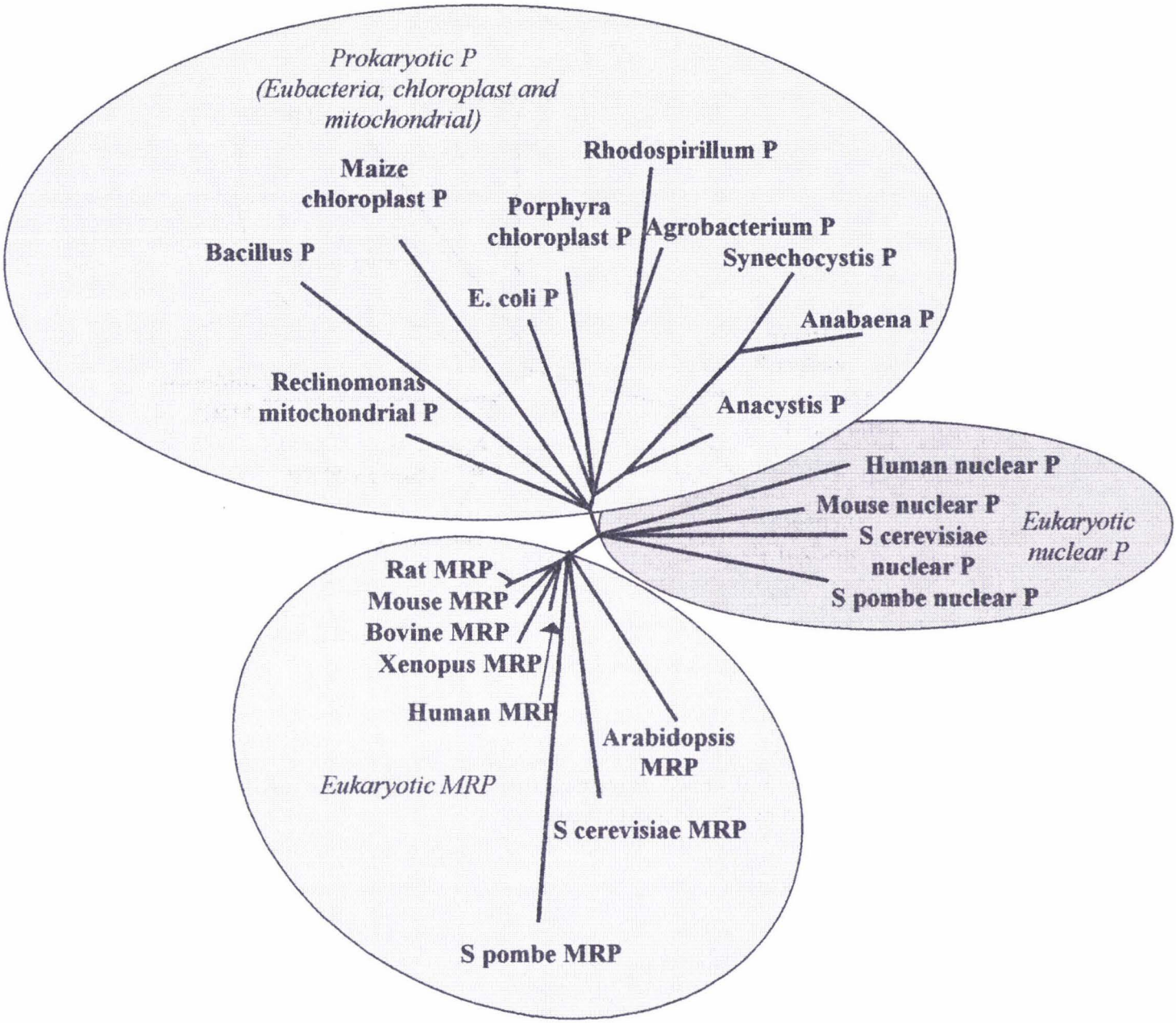


Figure 6: Refined Buneman tree constructed from pRNA and mrpRNA secondary structures.

A

Domain I Areas

Area	Area according to E.coli numbering
1	1 - 12
2	13 - 21
3	26 - 38
4	39 - 46
5	47 - 58
6	59 - 106
7	107 - 121
6	122 - 142
9	143 - 178
10	179 - 197
11	198 - 220
12	240 - 288
13	289 - 311
14	316 - 337
15	338 - 353
16	367 - 394
17	404 - 436
18	437 - 499
19	500 - 546

B

Species	RDP Primary Sequence	Code
Escherichia coli	E.coli	A
Clostridium innocuum	C.innocuum	B
Methanococcus vannielli	Mc.vanniel	C
Frankia sp.	Fra.spORS	D
Streptomyces coelicolor	Stm.coelic	E
Thermus thermophilus	T.thermoph	F
Bacillus subtilis	B.subtilis	G
Agrobacterium tumefaciens	Ag.tumefac	H
Spirochaeta aurantia	Spi.aurant	I
Thermoplasma acidophilum	Tpl.acidop	J
Mycoplasma capricolum	M.capricol	K
Methanobacterium formicicum	Mb.formici	L
Pseudomonas testosteroni	Ps.testost	M

C

```
A
B  6.0000
C  11.000  10.000
D  7.0000  4.0000  9.0000
E  7.0000  5.0000  9.0000  3.0000
F  8.0000  6.0000  9.0000  6.0000  5.0000
G  8.0000  4.0000  10.000  5.0000  6.0000  5.0000
H  7.0000  5.0000  9.0000  5.0000  5.0000  6.0000  5.0000
I  6.0000  4.0000  10.000  5.0000  4.0000  6.0000  6.0000  6.0000
J  11.000  10.000  4.0000  9.0000  9.0000  9.0000  10.000  9.0000  10.000
K  6.0000  4.0000  9.0000  4.0000  4.0000  6.0000  6.0000  6.0000  3.0000  10.000
L  11.000  10.000  3.0000  9.0000  9.0000  9.0000  10.000  8.0000  10.000  3.0000 10.000
M  4.0000  4.0000  10.000  5.0000  5.0000  5.0000  5.0000  5.0000  4.0000  10.000
4.0000 10.000
```

Figure 4.7: Data used in 16S rRNA secondary structure analysis. **A**: Domain I of the 16S rRNA divided into areas. Numbering of the divisions is based on the E.coli 16S rRNA secondary structure. **B**: Species of eubacteria and archaebacteria used in this study and their reference codes in the Ribosomal Database Project (RDP, Maidak et al. 1996) and **C**: Matrix of differences between the areas in Domain I of the 16S rRNA secondary structure.

A

Domain III Areas

	Area according to E. coli numbering
33	938 - 944
34	945 - 955
35	956 - 983
36	984 - 990
37	991 - 1045
38	1046 - 1067
39	1068 - 1073
40	1074 - 1085
41	1086 - 1099
42	1108 - 1117
43	1118 - 1155
44	1156 - 1182
45	1236 - 1252
46	1253 - 1284
47	1297 - 1339
48	1346 - 1372

B

A											
B	2										
C	5	5									
D	3	3	6								
E	4	4	5	3							
F	3	5	5	3	3						
G	1	1	5	3	4	5					
H	4	3	4	6	5	6	2				
I	2	1	5	3	4	5	1	3			
J	5	5	3	6	5	7	5	3	5		
K	4	2	5	3	2	3	3	4	3	6	
L	4	4	3	5	4	6	4	2	4	2	5
M	3	2	5	4	3	4	3	4	3	5	3	4

C

A												
B	8											
C	16	15										
D	10	7	15									
E	11	9	14	8								
F	11	11	14	9	8							
G	9	5	15	8	10	10						
H	11	8	13	11	10	12	7					
I	8	5	15	8	8	11	7	9				
J	16	15	7	15	14	16	15	12	15			
K	10	6	14	7	6	9	9	10	6	16		
L	15	15	6	14	13	15	14	10	14	5	15	
M	7	6	15	9	8	9	8	9	7	15	7	14

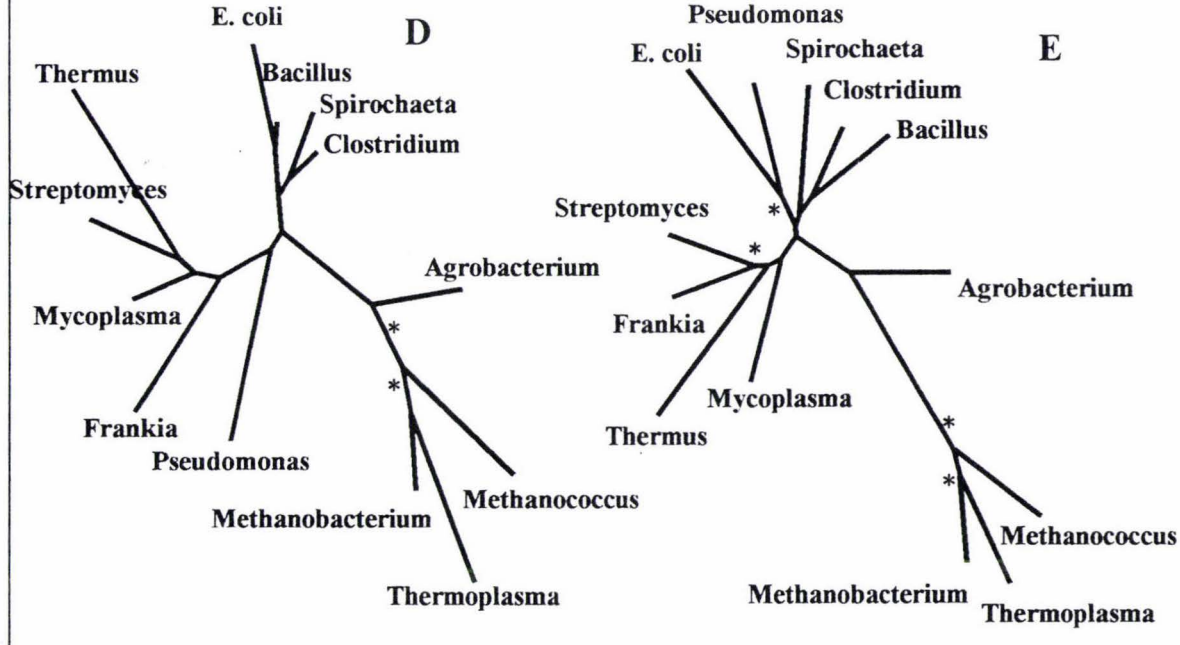


Figure 4.8: Data used in 16S rRNA Domain III and combined Domain I and III secondary structure analysis. **A**: Domain III of the 16S rRNA divided into areas. Numbering of the divisions is based on the E.coli 16S rRNA secondary structure. **B**: Matrix of differences between the areas of Domain III. **C**: Matrix of differences between the combined areas of Domains I and III. **D**: Neighbor-joining tree of Domain III lengths. **E**: Neighbor-joining tree of combined Domains I and III lengths. * indicates the internal branches that are the same as those in the subtree in Figure 4.2.

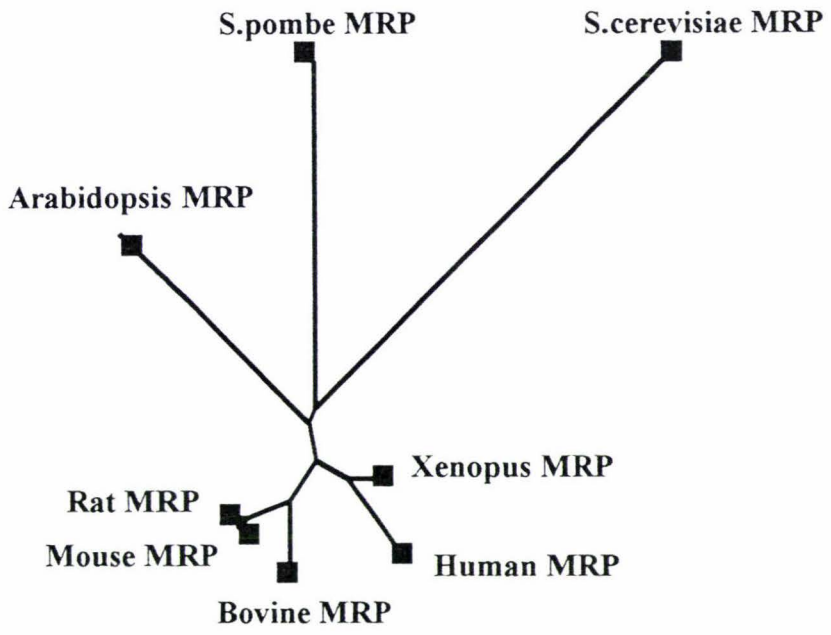


Figure 4.9: Neighbor-joining of mrpRNA secondary structures compared by RNAdistance.

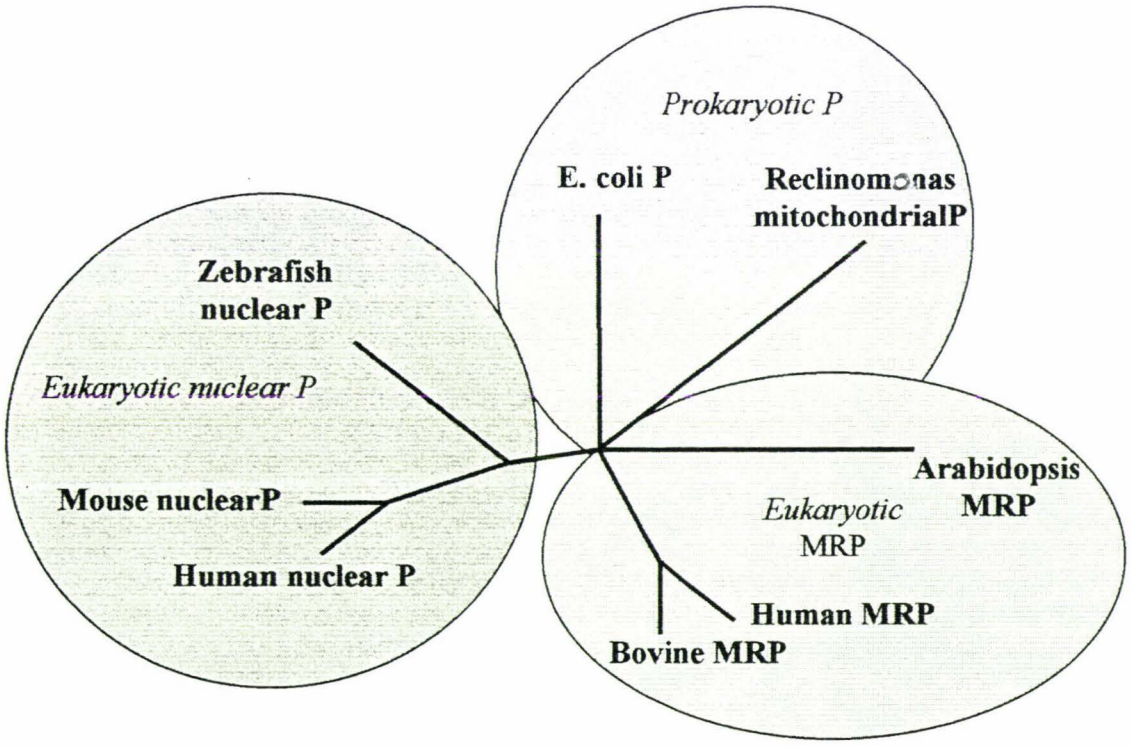


Figure 4.10: Neighbor-joining tree of MRP and pRNA sequences aligned by Divide and Conquer.

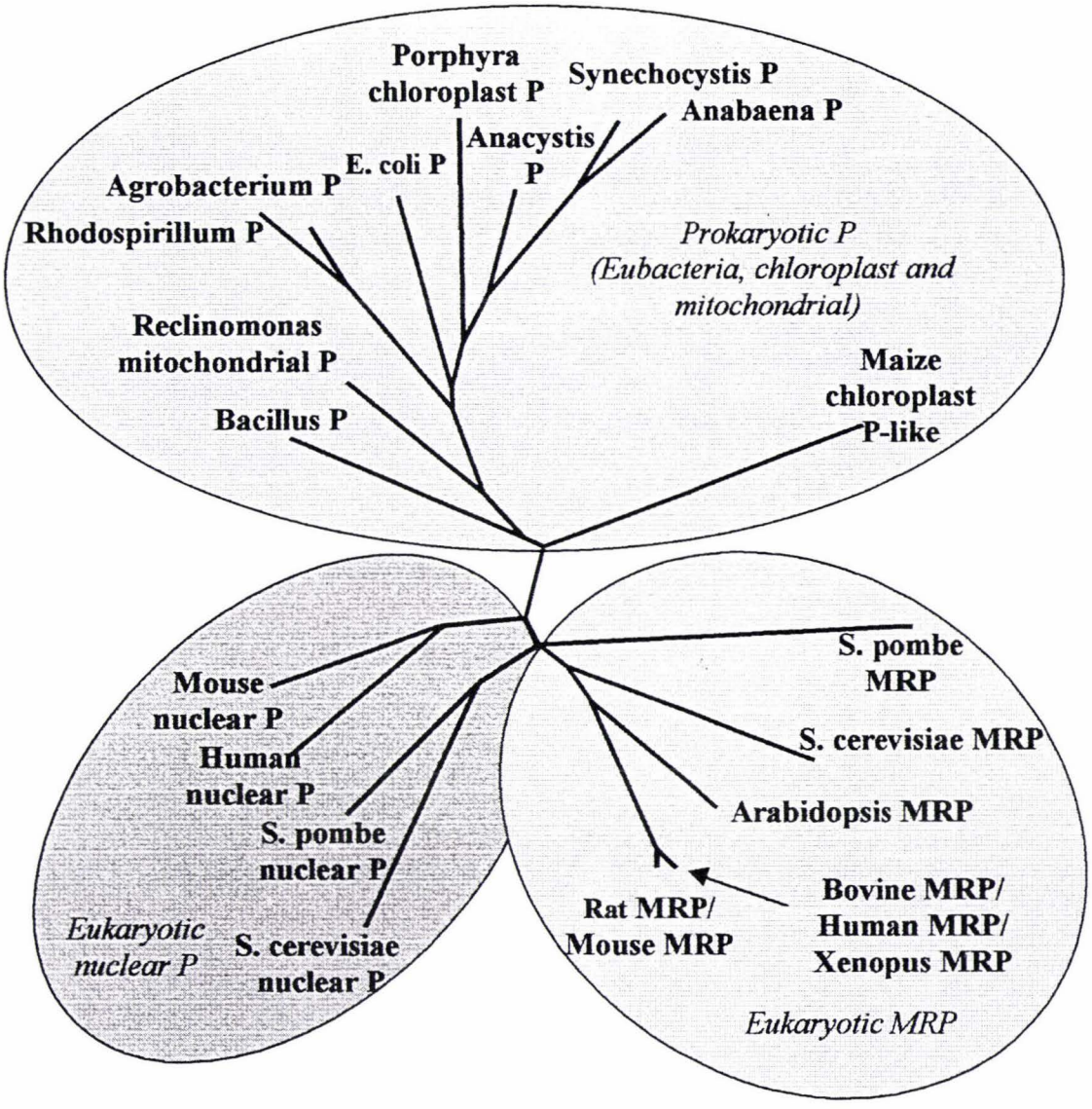


Figure 4.11: Neighbor-joining tree constructed from pRNA and mrpRNA secondary structures.

Chapter 5

Evaluation of folding programs for the analysis of evolutionary relationships of catalytic RNA molecules.

Folding programs have often been used to analyse RNA for possible secondary structure motifs e.g. terminators of protein coding genes (personal experience). Some secondary structure motifs, arbitrarily labelled the 'fork' and the 'can-opener', could be seen in a qualitative comparison of the *Synechocystis* pRNA and the putative maize chloroplast pRNA (Maize RNase P-like) (Figure 3.8). In this chapter, pRNA and mrpRNA calculated secondary structures are analysed in a quantitative manner, to determine if these folding programs can be used in some way to evaluate evolutionary relationships of catalytic RNA.

It has been recognised that folding programs do not produce the same secondary structure of an RNA molecule as is seen biologically (Zuker 1989). These methods minimise the free energy of the folded molecule by assigning energy to the stacking of one hydrogen bond over another, and assigning destabilising energies to various loops. The overall free energy of a folding is the sum of the energies of the stacked base pairs and the loops (Zuker 1989). Base pairing is generally of three types G with C, A with U and G with U. More unusual base pairing can occur with G-G, U-U, C-C, C-A, A-A, A-G and U-C pairs forming but these pairings are considered weaker. For the basis of this project G - U pairing was allowed, but none of the other unusual pairings were considered.

There are, at least three major problems associated with using folding algorithms for the determination of RNA secondary structure (Zuker 1989). The tertiary (three-dimensional) structure of the RNA molecule is ignored; energy rules used to assign free energies to structures are derived from melting data on small oligonucleotides only; and many different foldings can be possible close to the minimum energy. An additional problem specific to ribonucleoproteins is that in nature, the RNA component must include effects of the binding of proteins and metal ions. It is possible to turn this last 'problem' into an advantage if a 'biologically correct' structure is not required. Essentially the folding by these programs is dependent on the sequence data and useful comparisons between catalytic RNA molecules could be made when the protein and metal binding is essentially unknown. For this purpose a structure that is biologically correct is not required as long as the structures formed follow an evolutionary pattern that is similar to that followed by the

biological structures. This chapter examines different groupings of pRNA and mrpRNA thermodynamically folded structures, to see if they follow the same evolutionary pattern as the biological secondary structures have been shown to do in chapter 4.

The RNAdistance program can compare secondary structures in four structure formats. Only the full structure format was used in the previous chapter as it gives the maximum amount of information about the structures being compared. Trees constructed from all four formats are necessary to see whether the level of coarseness of the structure makes any difference on the evolutionary tree produced, and whether any particular level will be useful for analysing evolutionary relationships.

Materials and Methods

pRNA and mrpRNA sequences were obtained from Genbank or from the RNase P Database (Brown 1998) (Table 1.1). A Subtree of 16S RNA sequences was downloaded from the Ribosomal Database Project (Maidak et al. 1997).

Sequences were folded using one of two programs, RNAdraw (Matzura and Wennborg 1996) which uses the RNAfold algorithm from the Vienna RNA package (Hofacker et al. 1994), and RNAstructure (Mathews et al. 1997) which uses the Mfold algorithm (Zuker 1989). All structures were folded at settings equivalent to 37°C with the other folding parameters set at the default settings (listed in the program descriptions in appendix 4). The secondary structure diagrams for all sequences constructed with RNAdraw and RNAstructure are shown in Appendix 1.

The calculated RNA secondary structures for each sequence were converted from the ct files to bracket notation using the program 'ct2bracket' written by R. Pointon for this project (Appendix 4). Representations of the bracket notation for pRNA and mrpRNA secondary structures are shown in Appendix 2. The folding programs used here calculate a number of sub-optimal structures as well as the optimal structure. In most cases the optimal structure from each folding was compared. However, when the *S. cerevisiae* nuclear pRNA sequence was folded with the RNAstructure program, the first three optimal structures bore little resemblance to the other pRNA folded sequences. For this sequence the third suboptimal (fourth structure in order of minimum energy) structure was used as it showed a close resemblance to the other pRNA folded sequences used in this study.

During the folding of the pRNA and mrpRNA sequences it was discovered that both the RNAstructure and RNAdraw programs formed two types of structures for some sequences. The first was the expected type with the 5' and 3' ends of the RNA molecule paired (Figure 5.1B). The other was a 'circular-type' structure that had short range pairing of the ends to sequences near those ends (Figure 5.1A). It was possible to 'correct' these circular structures by forcing the pairing of the 5' and 3' ends.

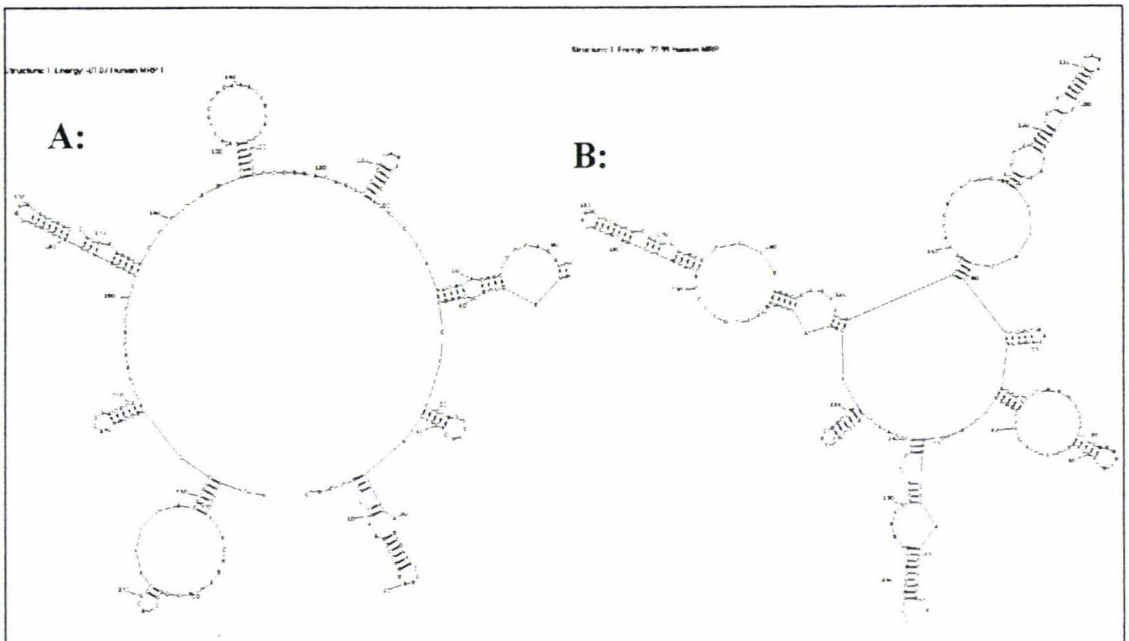


Figure 5.1: Human mrpRNA folded with RNAdraw to show **A**: uncorrected circular structure and **B**: 5' – 3' corrected structure.

However, it was recognised that, in any screening process that uses these programs, such corrections are currently not possible, so trees of both uncorrected and corrected structures are compared here. Table 5.1 shows which structures required corrections for each of the programs.

RNAstructure (Mfold)	RNAdraw (RNAfold)
Arabidopsis mrpRNA	Arabidopsis mrpRNA
Aspergillus mitochondrial pRNA	Human mrpRNA
Rat mrpRNA	Mouse mrpRNA
	Porphyra chloroplast pRNA
	Maize chloroplast pRNA
	Rat mrpRNA
	Reclinomonas mitochondrial pRNA
	Saccharomyces cerevisiae mitochondrial pRNA

Table 5.1: mrpRNA and pRNA secondary structures, which gave a circular structure when, folded with either RNAstructure and/or RNAdraw.

Secondary structures were compared using the RNAdistance program from the Vienna RNA package (Hofacker et al.1994). The tree-editing alignment option was used in this study rather than the string alignment option. The tree editing option in RNAdistance calculates the distance between two trees by calculating the smallest sum of costs required, transforming one tree into another tree by a series of editing operations (Hofacker et al 1994).

This RNAdistance program has the option of comparing the secondary structures in 4 grades of coarseness. The full structure format (f), was that used in the previous chapter for the comparison of the biological secondary structures, and contains the full amount of structural detail available. The next level of coarseness is the HIT (Homeomorphically irreducible tree) structure which represents a stack of pairings by P and a number indicating the number of base pairs in that stack. Unpaired bases are represented as U and a number indicating how many bases are unpaired before the next stack of paired bases. The full tree is converted to the tree by assigning an internal node to each base pair and a leaf note to each unpaired nucleotide. This apparently simpler representation of the secondary structure retains complete information on the structures but may reveal different information about the structure than what is revealed by the full structure.

The third level of coarseness is the weighted coarse (w) structure, which does not retain the full information of the secondary structure. Different structure elements are represented by single matching bracket and are labelled as H (hairpins loop), I (interior loop), B (bulge), M (multiloop), S (stack), and E (external elements). Again a number may be used to indicate the number of unpaired bases or base pairs comprising each structural element. The fourth level is coarse format (c) and compares the structures based only on the presence of helices and loops without any reference to the lengths of either, looking only at the branching structure. A representation of all four levels of structure is given in Appendix 4. The RNAdistance program converts the full structure input into the other three structures when required. Structures folded from the first three sets of sequences were compared in all four formats with the fourth group only being compared in the full structure format.

Phylogenetic trees were created from this distance data using the neighbor-joining algorithm from the Phylip package (Felsenstein 1989), and trees were drawn using

TreeView (Page 1996). The RNAdistance and ct2bracket programs were run on a Unix SunOS release 4.1.3 whereas RNAstructure, RNAdraw, Neighbor, and TreeView were run using the Windows 95 operating system on a PC.

Three data sets were used here to compare trees constructed either directly from sequences, biological structures, or from folded secondary structures. The first group contains five pRNA from photosynthetic organisms (three cyanobacteria and two chloroplast) and *E. coli* as an outgroup looking at the grouping of the cyanobacteria and chloroplast species. The second group examines the eight mrpRNA sequences in a similar manner, but at the placement of the individual species more detail. The third group includes mrpRNA and pRNA from eubacteria, eukaryotes, and organelles examining whether structures calculated from these sequences fall into the distinct groupings that were shown with the biological secondary structures.

Results

In the first group of pRNA sequences, the folded structures from four bacterial (*E. coli*, *Synechocystis*, *Anabaena*, and *Anacystis*) and the two chloroplasts (*Porphyra purpurea* and the putative maize sequence) were compared. This group of pRNA sequences required no correction for circular structure when using the RNAstructure program. However, the structures for the *Porphyra* chloroplast pRNA and the putative maize chloroplast pRNA required correction for the RNAdraw program.

A subtree of 16S rRNA sequences for the same species (Figure 5.2A) was compared the trees of the folded structures together with trees constructed from the biological secondary structures for these species. For the first group of pRNA sequences, there are some differences between the tree constructed from the 16S rRNA sequences (Figure 5.2A) and the tree constructed from the secondary structures (Figure 5.2E) using the full structure format. Only two trees (Figures 5.2D and 5.2E) are identical and the position of the outgroup (*E. coli*) is quite variable. However, (excluding *E. coli*) the two chloroplast sequences (*Porphyra* and Maize) always come together on the tree, while the three cyanobacterial sequences swap as to which one diverges first. Since the biological secondary structure used for the putative maize chloroplast pRNA is only hypothetical it is not too surprising to see that this structure is grouped away from the other chloroplast structure. This indicates (as implied in the previous chapter) that this hypothetical structure may not be entirely correct.

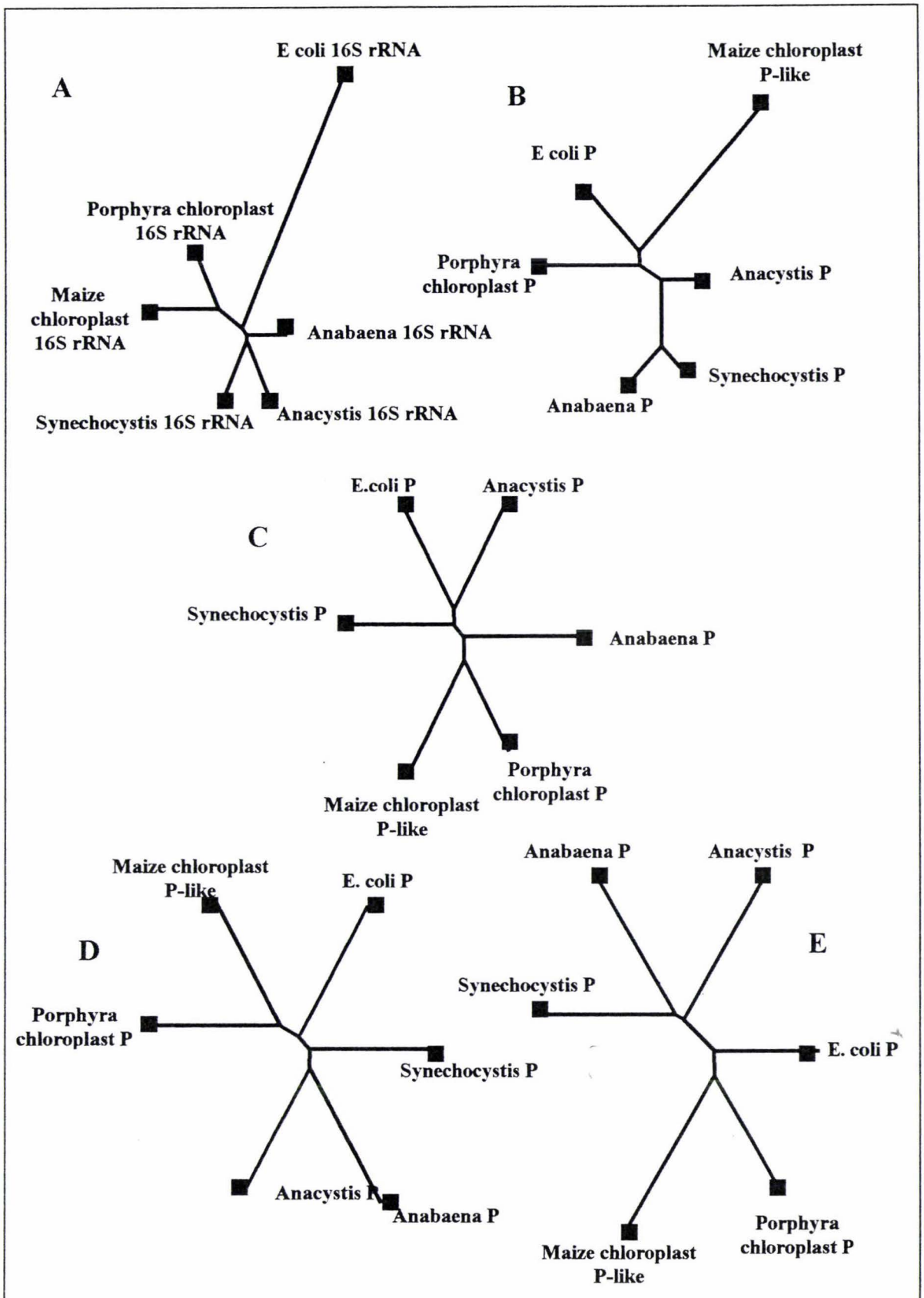


Figure 5.2: pRNA sequences: **A**: Subtree of 16S rRNA sequences. Neighbor-joining trees of full structure format (f) of **B**: biological secondary structures, **C**: RNAstructure and RNAdraw folded secondary structures **D**: uncorrected, **E**: corrected.

does not group the cyanobacterial species (*Synechocystis*, *Anabaena* and *Anacystis*) together unless the *E. coli* outgroup is removed. The RNAdraw trees for both the uncorrected (Figure 5.2D) and the corrected (Figure 5.2E) are very similar, grouping both the chloroplast and the cyanobacterial species together. However the pairing of the cyanobacterial species are different between the two trees. Correcting the RNAfold structures for the chloroplast sequences has changed the order of the cyanobacteria to that shown in the biological structures tree but has grouped the chloroplast structures in the same way as is shown in the 16S rRNA tree.

The HIT structure gives results very similar to those shown for the full format. The biological structures tree (Figure 5.3A), the RNAstructure tree (Figure 5.3B), the RNAdraw uncorrected tree (Figure 5.3C) and corrected tree (Figure 5.3D) are also unchanged from those shown in the full format.

In the weighted coarse format the biological structure tree (Figure 5.4A) groups the chloroplast structures together, and keeps the pairing of the cyanobacteria shown in the full and HIT structures format. The grouping of the chloroplasts together at this level shows that the hypothetical biological structure for the maize chloroplast pRNA is more similar to the other chloroplast pRNA sequence than to the bacterial pRNAs at this coarser level of structure format. The RNAstructure tree (Figure 5.4B) splits the cyanobacterial species by the *E. coli* outgroup, but keeps the chloroplast sequences together (as is shown in the full and HIT trees). However this tree has placed the cyanobacterial species differently. The RNAdraw uncorrected tree (Figure 5.4C) is the same as shown in the full and HIT structures but the corrected tree (Figure 5.4D) has split the chloroplast sequences while keeping the cyanobacteria together.

The tree constructed from the coarse biological structures (Figure 5.5A) has split both the cyanobacterial and chloroplast sequences but has maintained the pairing of the *Anabaena* and *Synechocystis* structures as is shown in the full structure. The RNAstructure tree (Figure 5.5B) has also split the cyanobacterial and chloroplast sequences and has not paired any of the cyanobacterial structures. The RNAdraw uncorrected tree (Figure 5.5C) has split the chloroplast sequences but has maintained the cyanobacteria grouping. The corrected tree (Figure 5.5D) is like the RNAstructure tree with the chloroplast and cyanobacterial structures in no particular groupings.

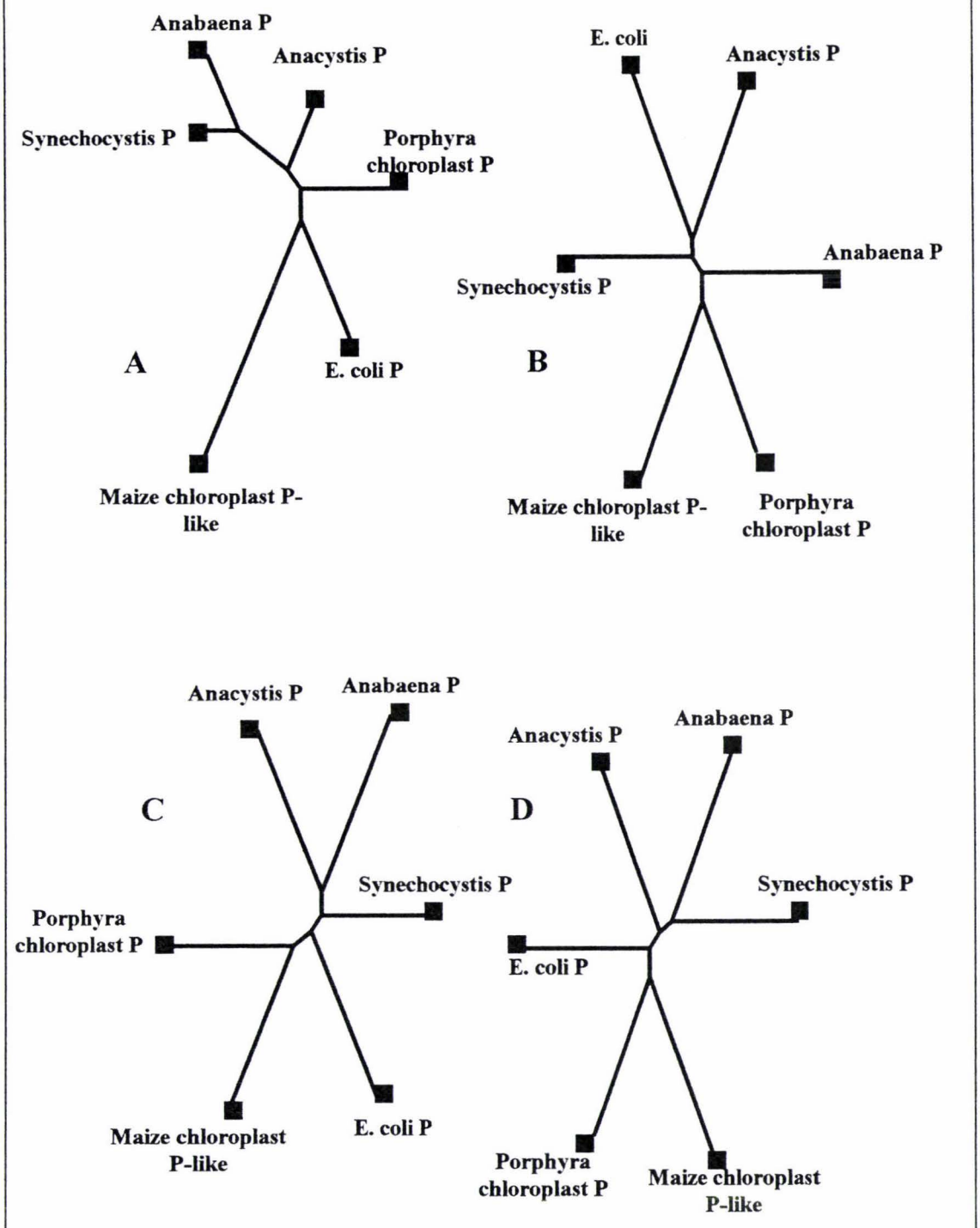


Figure 5.3: Neighbor-joining tree of pRNA structures compared in the HIT structure format (h); **A**: biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C**: uncorrected and **D**: corrected.

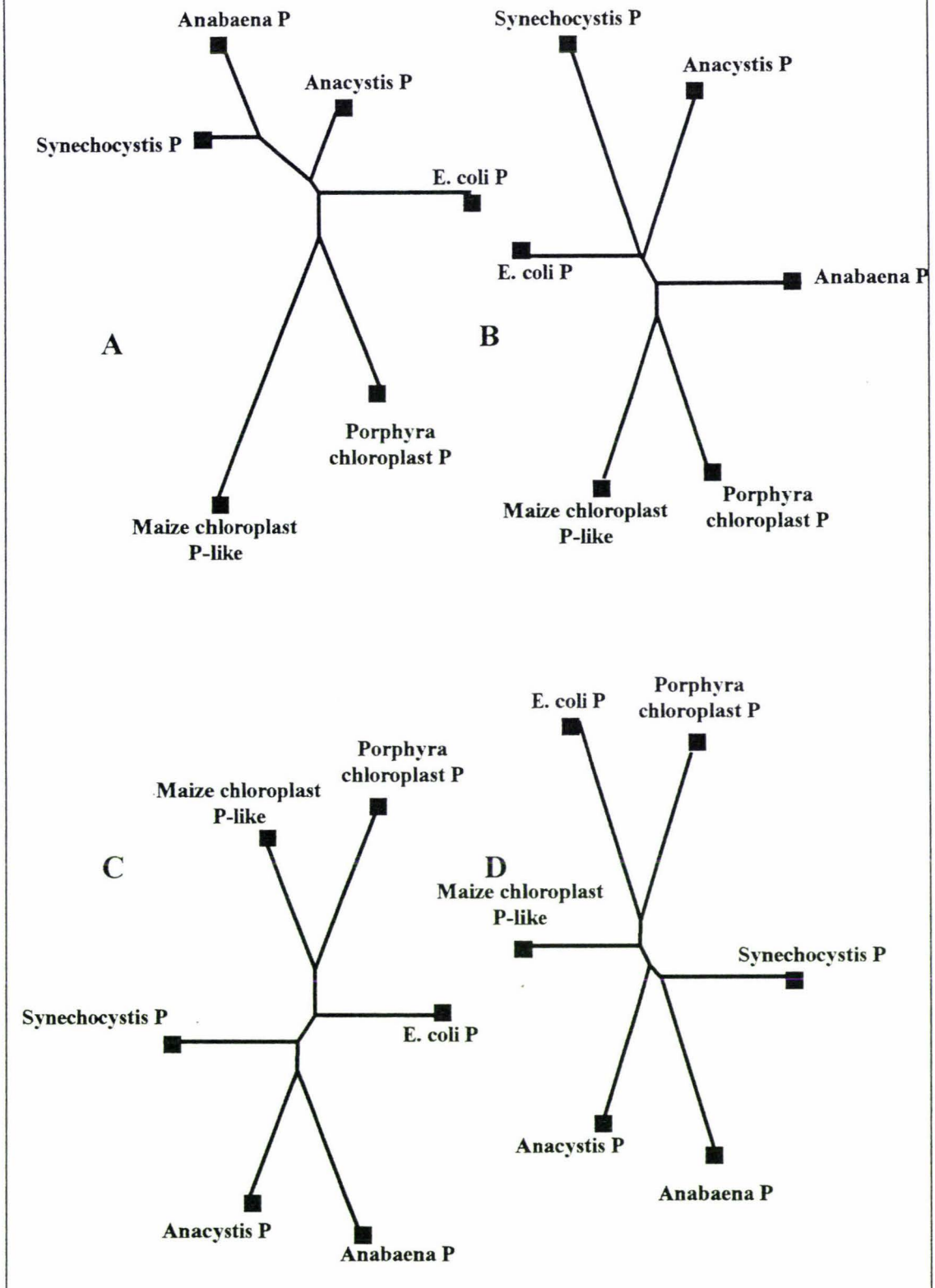


Figure 5.4: Neighbor-joining tree of pRNA structures compared in the Weighted coarse format (w); **A**: biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C**: uncorrected and **D**: corrected.

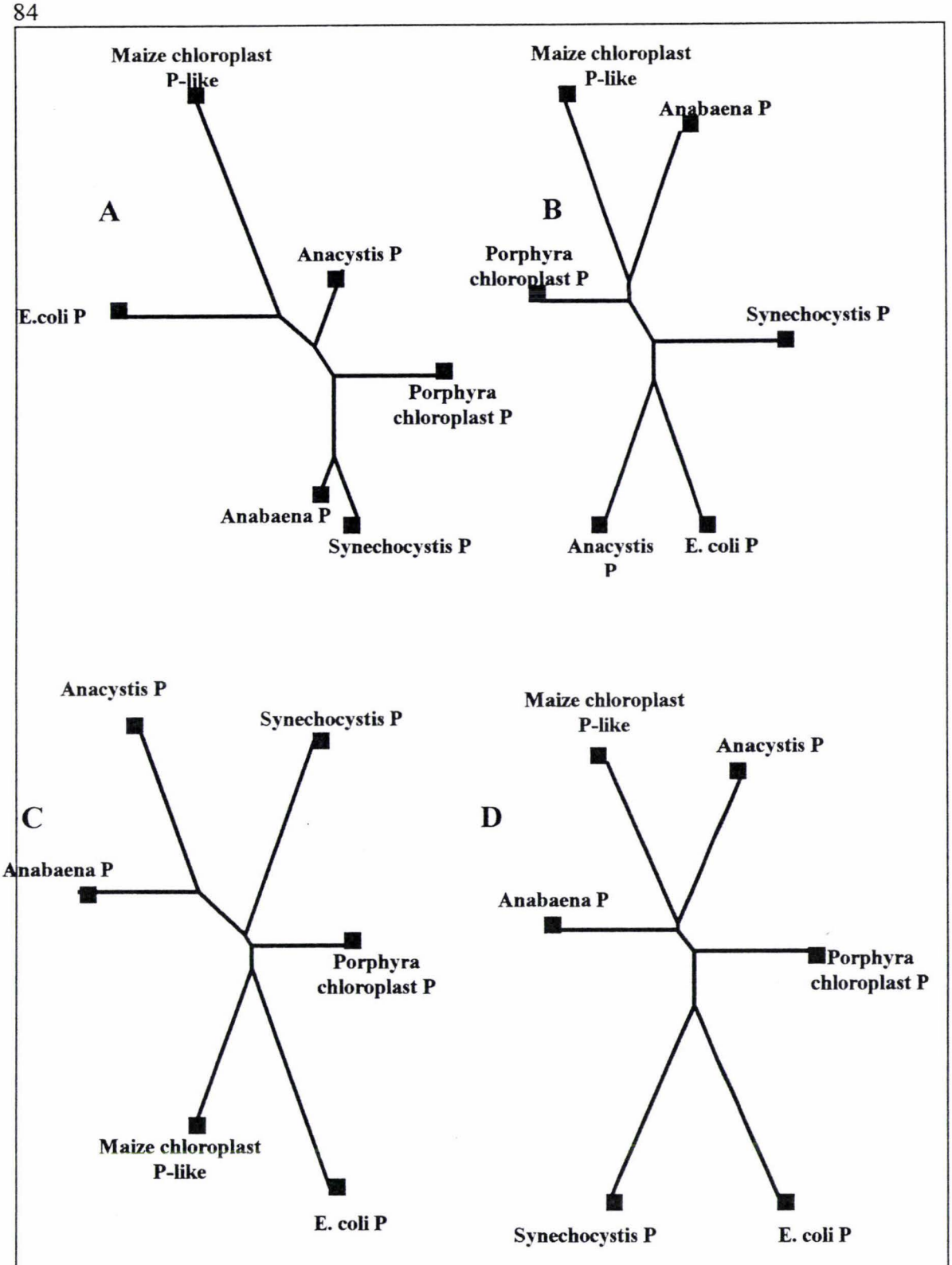


Figure 5.5: Neighbor-joining tree of pRNA structures compared in the Coarse format (c); **A**: biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C**: uncorrected and **D**: corrected.

Overall, for the initial group of pRNA sequences, the RNAdraw folding program gives a tree closer to the 16S rRNA tree and the biological secondary structure trees. For both the biological and the thermodynamically folded structures the coarser the structure format the more unlike the 16S tree the resulting trees become. However, this is only a qualitative observation. None of the biological or thermodynamically folded trees gives a tree exactly like the 16S tree. However, it is interesting that although the order of the cyanobacterial structures is different, the relative groupings of the cyanobacterial and the chloroplast structures are the same as for the 16S tree.

Given that the cyanobacterial and chloroplast groupings were maintained with the folded structures, the next step was to look in more detail at the placement of individual folded structures on the trees. A second group of mrpRNA folded secondary structure trees was compared to the tree constructed from mrpRNA sequence data (Figure 5.6A) and the tree constructed from the mrpRNA biological secondary structures (Figure 5.6B). The only differences between these two trees are the grouping of the vertebrate species.

In contrast, comparisons of the folded mrpRNA structures in the full structure (Figure 5.6C - F) show virtually no similarity with the sequence and biological structure trees. Both RNAstructure and RNAdraw structures, uncorrected and corrected, show little difference. The two fungal species are not grouped together and the vertebrates do not show any grouping similar to that of the biological structure and sequence alignment trees. Similarly the HIT (Figure 5.7), weighted coarse (Figure 5.8) and coarse structure (Figure 5.9) formats give trees less similar to the sequence alignment and biological full structure trees as the structure format gets coarser. There is a much more apparent difference between the trees compared in the full structure format and the HIT format, than that seen between the pRNA structures. The folded structures of this group of mrpRNA sequences do not follow the evolutionary pattern seen either in the biological secondary structures or in sequence alignment data.

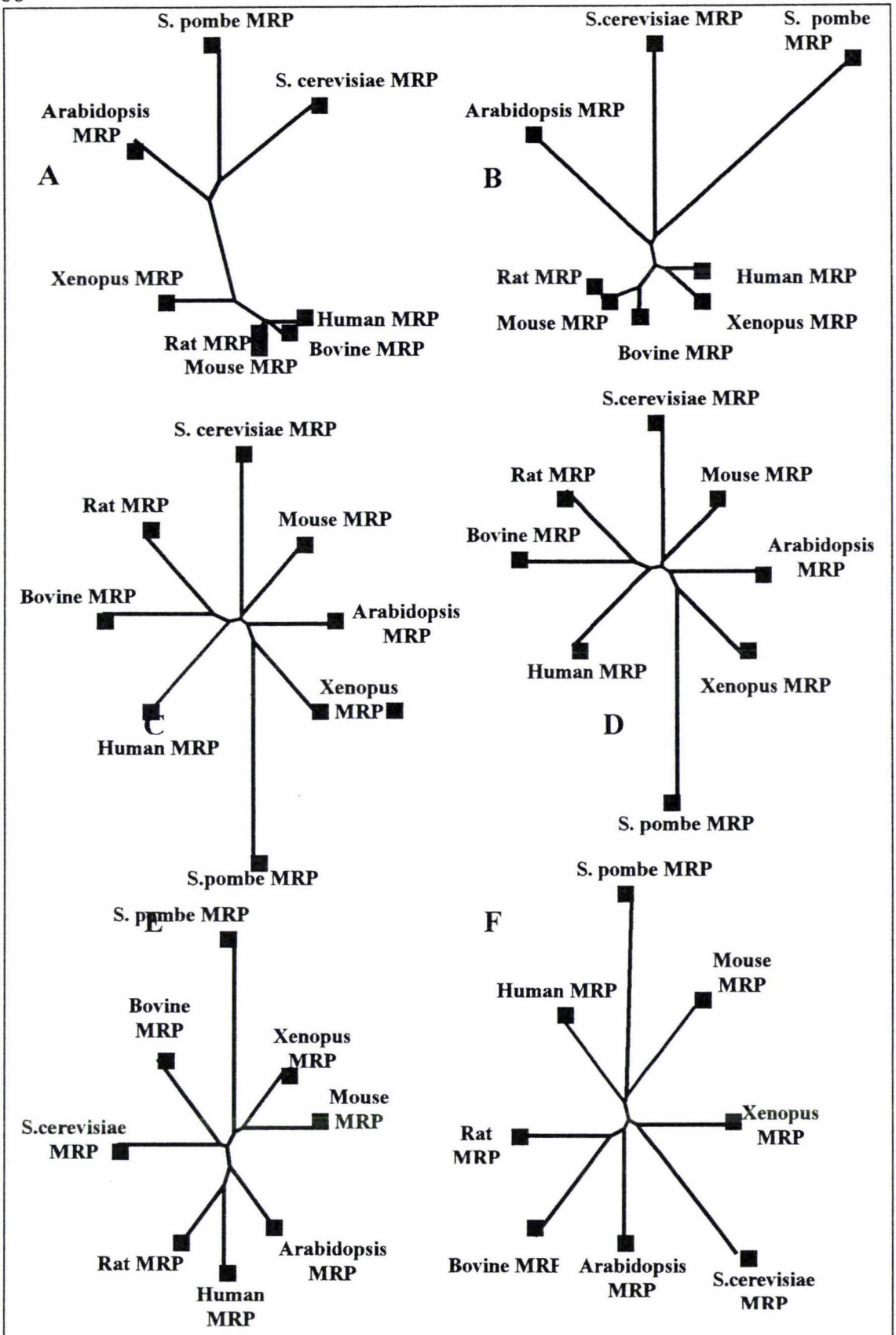


Figure 5.6: Neighbor-joining trees of mrpRNA **A**: aligned sequences, **B**: biological secondary structures; sequences folded by RNAstructure **C**: uncorrected and **D**: corrected; sequences folded by RNAdraw **E**: uncorrected and **F**: corrected. All structures are compared in the full (f) format.

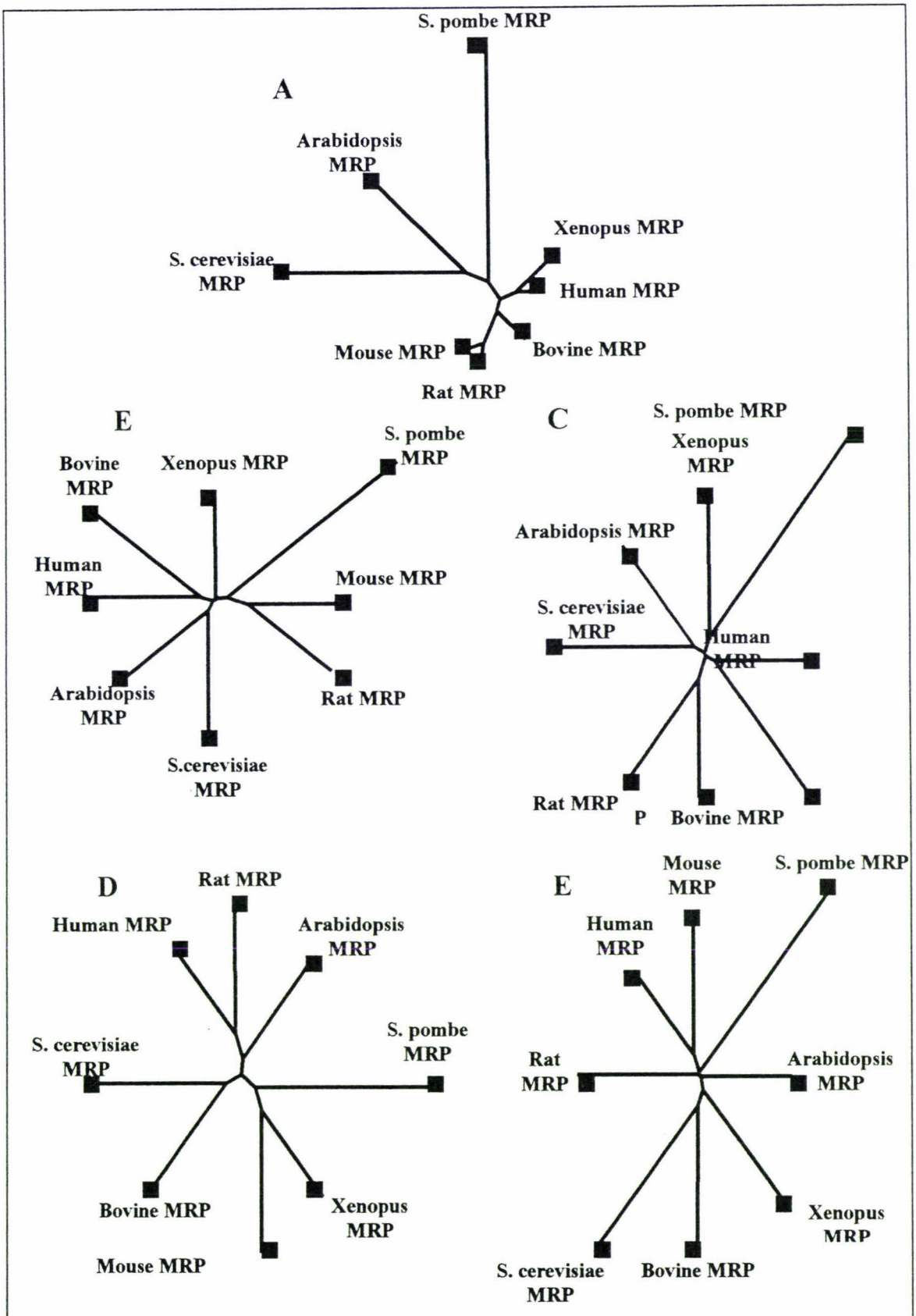


Figure 5.7: Neighbor-joining trees of mrpRNA: **A** biological secondary structures; sequences folded by RNAstructure **B**: uncorrected and **C**: corrected; sequences folded by RNAdraw **D**: uncorrected and **E**: corrected. All structures are compared in the HIT (h) format.

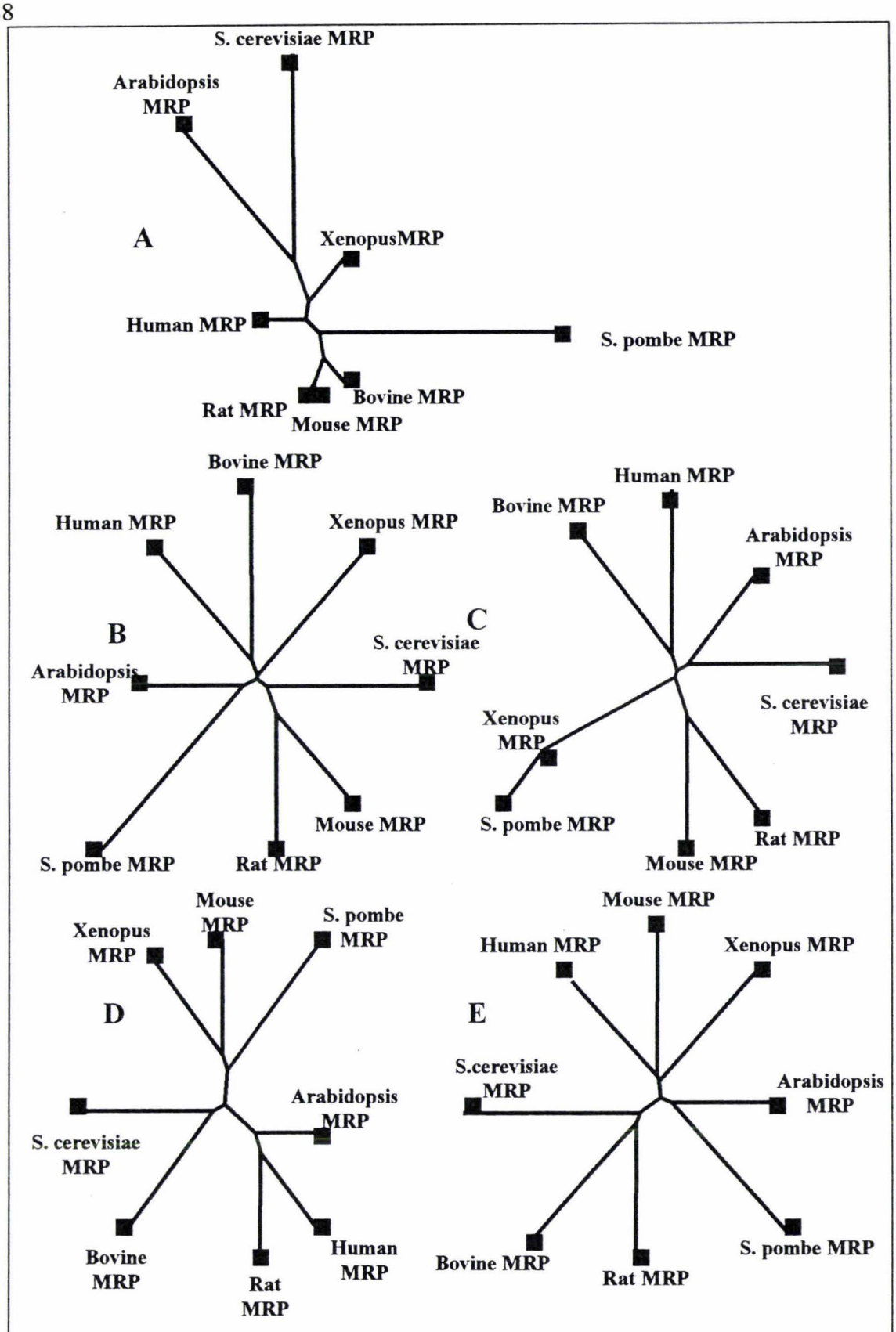


Figure 5.8: Neighbor-joining trees of mrpRNA: **A** biological secondary structures; sequences folded by RNAstructure **B**: uncorrected and **C**: corrected; sequences folded by RNAdraw **D**: uncorrected and **E**: corrected. All structures are compared in the Weighted Coarse (w) format.

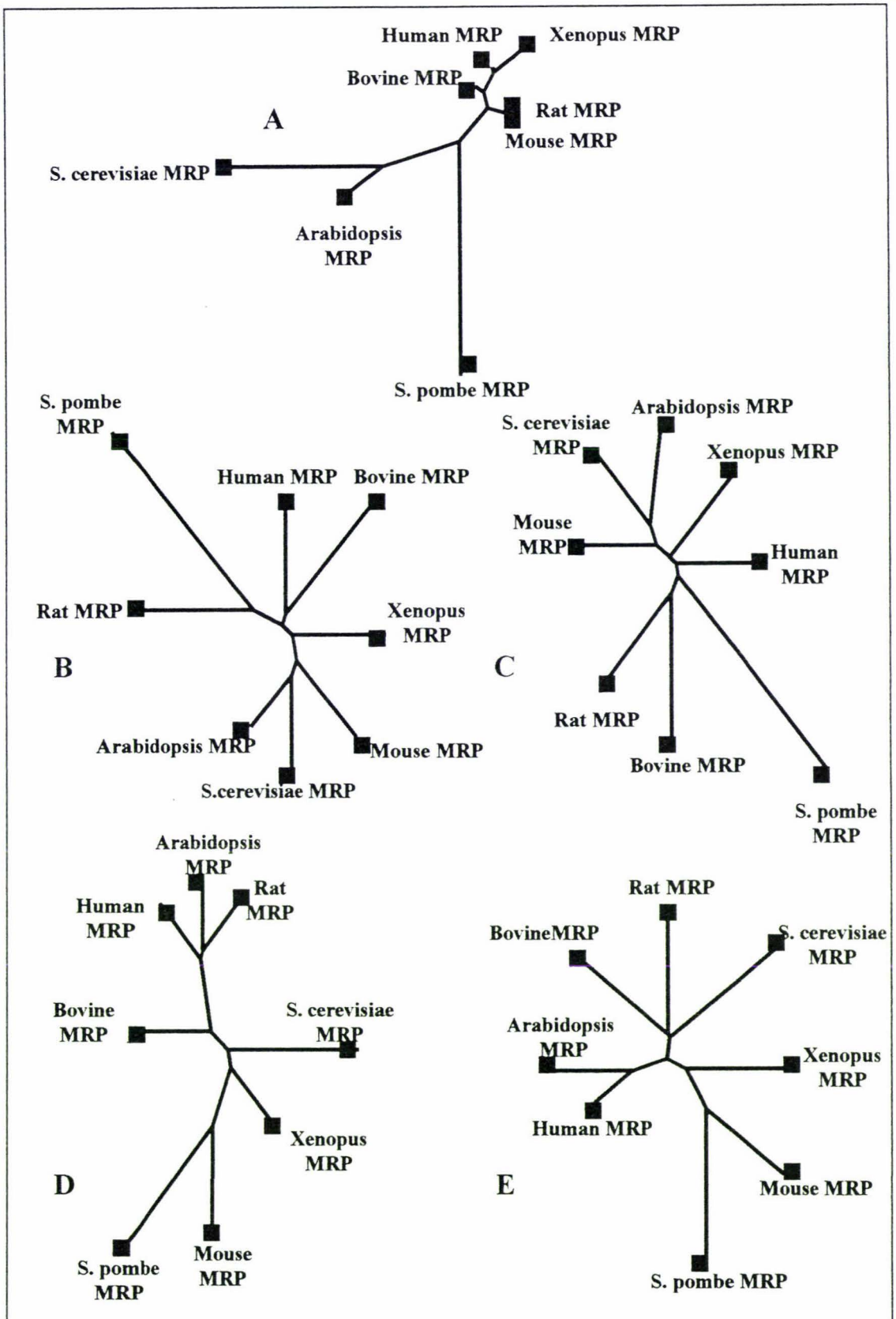


Figure 5.9: Neighbor-joining trees of mrpRNA: **A** biological secondary structures; sequences folded by RNAstructure **B**: uncorrected and **C**: corrected; sequences folded by RNAdraw **D**: uncorrected and **E**: corrected. All structures are compared in the Coarse (c) format.

Folded structures from a third group of mrpRNA and pRNA sequences was then examined. When the biological secondary structures from this group of sequences were compared to each other (See figure 4.6), they fell into three distinct groups; prokaryotic pRNA, eukaryotic nuclear pRNA and mrpRNA. Although the folded structures (Figures 5.10- 5.17 for the RNAstructure program and Figures 5.18 - 5.25 for the RNAdraw program, all of these figures are shown at the end of this chapter) do not fall into the three distinct groupings, there are still some interesting results. The weighted coarse and the coarse structure formats follow the same pattern as in the pRNA (Figures 5.4 and 5.5) and the mrpRNA (Figures 5.8 and 5.9) comparisons and scatter species all over the tree, that in the biological structures tree fall together. This scattering is still seen after structures were corrected to exclude circular structure, or in which program (RNAstructure or RNAdraw) were used.

One observation with this third group of species was that when circular structures are corrected they do not always affect where they themselves lie on the tree, but often affect where other species lie on the tree. For example, with the RNAstructure program there are only three structures in the third group that required correction. When this group is compared in the full structure format, those that required correction do not change position between the uncorrected tree (Figure 5.10) and the corrected tree (Figure 5.11), but the position of the bovine mrpRNA changed considerably. Another example in the HIT format comparison of this group of structures, the positions of the rat mrpRNA (circular) and the mouse mrpRNA change but the position of the bovine mrpRNA remains the same between the uncorrected (Figure 5.12) and the corrected tree (Figure 5.13).

There are some interesting groupings formed between the pRNA and mrpRNA folded structures. The nuclear pRNA and mrpRNA from the yeast *S. cerevisiae* are grouped together by RNAstructure for both the uncorrected and corrected forms, and in the both the full and HIT formats. RNAdraw groups them together in the full uncorrected format only. The *S. pombe* structure is grouped with the *S. cerevisiae* structure in the same way.

Within this third group of mrpRNA and pRNA sequences there are some groupings of structures that are seen in both the biological structures tree and that constructed from the folded structures. The rat mrpRNA and mouse mrpRNA are grouped together, but only

with RNAstructure in the full and HIT formats. The three cyanobacterial pRNAs (*Anabaena, Anacystis* and *Synechocystis*) are grouped together but only with the RNAdraw program in the full uncorrected format. However, the other bacterial pRNAs do not fall into the expected groupings. The *Agrobacterium* and *Rhodospirillum* pRNA structures are at opposite ends of the tree. There is little consistency between the trees constructed from the RNAstructure folded sequences and the RNAdraw folded sequences. However, there are many more sequences that formed circular structures with the RNAdraw program than with the RNAstructure program and it is uncertain as to the effect of this on the overall trees constructed.

Overall there are two trends that are formed by the mrpRNA and pRNA data. The first is that the more coarse the secondary structure format, the less that the mrpRNA and eukaryotic nuclear pRNAs form separate groups (i.e. the more disorganised the tree becomes). The second is that 5' - 3' corrections to the circular structures formed by RNAstructure and RNAfold do not make the trees any closer to those formed from biological secondary structures. The putative maize chloroplast pRNA sequence forms structures that group either with the *S. cerevisiae* mitochondrial pRNA or with the *Aspergillus* mitochondrial pRNA especially when there have been no corrections to circular structures. Thus the current programs do not seem to be useful in predicting secondary structure with sufficient accuracy to obtain evolutionary information.

A possibility is that the high percentage of A and T in the mitochondrial and chloroplast sequences could be a factor in the placement of these structures on the trees. However, it is shown in figures 5.10 and 5.18 that sequences with similar AT contents are not grouping together. The effect of AT content is examined in more detail in the next two chapters.

The next step was to remove the mrpRNA and eukaryotic nuclear pRNA sequences from the third group to see how only the bacterial and organellar folded structures compare to each other. Because of earlier results this was only done using the full structure format. With RNAdraw (Figure 5.26) the organellar structures are largely grouped together whether the structures are uncorrected or corrected. The cyanobacterial sequences are not grouped together and are grouped with other bacterial species such as *Bacillus* and *Agrobacterium*. RNAstructure (Figure 5.27) does not group all the organellar structures or the cyanobacterial structures together. However, the majority of the bacterial and organellar

pRNA group together in the same way as is seen in trees constructed from their biological secondary structures.

When the circular structure is uncorrected, there is little difference between the trees of the bacterial and organellar folded structures and the previous trees including the eukaryotic pRNA and mrpRNA structures. However, in the corrected trees there are some significant improvements. The presence of the eukaryotic pRNA and mrpRNA structures have little effect on the placing of structures on the tree in the uncorrected format but have a greater effect when structures are corrected. This is an indication that the formation of many circular structures in RNAdraw may be detrimental to the determination of the correct tree for evolutionary comparison.

Discussion

For this study of the evolutionary characteristics of folded sequences, only the optimal structures were compared. There was the possibility that structures folded from similar sequences could be related at a coarse level, as the detail inherent in the full structure might obscure underlying structure similarities. This did not occur with the data shown here. The trees constructed from the coarse structure formats showed more apparent randomness in the position of structures on the tree as was shown in the trees constructed from full and HIT formats. This may be an indication that the structures formed using RNAstructure and RNAdraw are in fact quite alike, and it is only at the detailed levels of structure format that meaningful differences can be seen using the RNAdistance comparison program.

Comparisons of structures in the weighted coarse and coarse formats are perhaps more likely to seem less 'organised' due to the differences being slight between the structures. The number of structures that are compared in the coarser formats are much less (approx. 30 - 50), than in the full and HIT formats (approx. 300 - 400), indicating that a slight difference in the coarse structure will have a huge impact on the tree produced. The analysis of structures in all four formats from full and HIT to weighted coarse and coarse, shows that by rendering a structure into its most basic units, evolutionary information carried in the sequence may actually be lost. This could produce trees that are almost random in the distribution of pRNA and mrpRNA structures. The weighted coarse and coarse structure formats may be more useful, however, with the comparison of structures

formed from large sequences (such as the complete 16S rRNA), than with comparatively small structures (such as are formed from pRNA and mrpRNA sequences).

It has been recognised that the use of free energy minimisation (as is used in the two folding programs RNAstructure and RNAdraw) is not enough to determine a secondary structure with any confidence (Zuker 1989). Other data usually obtained through experimental means is used to determine single-stranded and paired regions as well as areas involved in protein binding. This is perhaps the primary reason why the trees formed from sequences folded with RNAstructure and RNAdraw, do not follow the same evolutionary pattern as is seen in the biological secondary structures. The lack of homology between the pRNA and mrpRNA sequences that fails to produce a confident alignment (shown in chapter 4) is probably also be responsible for the differences that are shown by the folded structures.

The formation of circular structure by the folding programs highlights an inconsistency with the folding algorithms. This circular structure is formed through short-range pairing being preferred over long-range pairing,such as the pairing of the 5' and 3' ends of the RNA molecule. However, when these ends were forcibly paired, the resulting structure still did not perform like the biological structure, ruling out the possibility of automatically forcing the 5' and 3' ends. The only way, at present, that one can determine if a structure is circular or not, is by drawing it. This could become extremely tedious and time consuming if many structures are being compared (and should only be attempted by the really keen and/or stupid). The RNAstructure program had fewer circular structures form than what was seen with RNAdraw. However, the ideal program would not produce any circular structures. It is recommended at present, that structures being folded with either of these programs or with corresponding algorithms, be drawn and circular structure determined visibly. In future, it is certainly preferable to develop an algorithm that checked for circular structure.

The programs used in this study can only measure the amount of energy required to form a structure. In nature, there could be several mechanisms involved in RNA folding, depending on intrinsic factors such as the size and sequence of the RNA as well as external factors such as temperature, free Mg^{2+} and pH (Pan et al. 1997). The length of the sequence and its AT content may be important factors in the formation of structures by these folding programs. These issues will be examined in the next chapter.

There is evidence also that, in nature, RNA can fold by multiple pathways, and that

the folding kinetics of large molecules is dominated by non-native or misfolded intermediates (Pan et al. 1997). The Kinetic Partitioning mechanism of Thirumalai and Woodson (summarised in Pan et al. 1997) states that the under-lying free energy landscape is complex, and not only contains the global minimum (corresponding to the biological structure), but also contains other low energy minima (suboptimal structures). The number of these competing minima, in which the RNAs adopt misfolded conformations, increases with the size of the RNA, a result being that the folding process includes direct and indirect pathways to the biological state. The problem, therefore, with using minimal energy programs for RNA folding is that the optimal structure need not correspond to the biological structure. It is not known yet as to whether any suboptimal structures formed from the folding of mrpRNA and pRNA sequences are identical or even very close to the biological structure. Due to constraints on the processing time and power of the computer systems used in the study, only the optimal structures could be examined (except for the *S. cerevisiae* nuclear pRNA). A future investigation could be to look at a large number of suboptimal structures for a sequence such as a pRNA sequence, to see how close any of them come to being like the biological sequence, if any.

There is consistent grouping of the maize chloroplast pRNA with the *Porphyra* chloroplast pRNA (when a small number of sequences are compared), and with either the *Aspergillus* mitochondrial pRNA or the *S. cerevisiae* mitochondrial pRNA (when large numbersof pRNA and mrpRNA sequences are compared). This could be an indication that the hypothetical biological structure of the maize chloroplast pRNA sequence may have some features that indicate that it is of organellar origin. The most likely feature is the elevated AT content of organellar sequences. However, this is unlikely in this case, as folded structures from sequences of the same AT contents do not appear to group together on any of the trees in this study.

The folding programs used here do not consider pseudoknot formations. Pseudoknots are characteristic structures formed by long-distance pairing and found in such RNA structures as pRNA, mrpRNA, and 16S rRNA (Pleij and Bosch 1989). Programs based on genetic algorithms (Gultyaeu et al. 1995) are able to take into account pseudoknots as well as the disruption of temporary structures, the folding of a molecule during its synthesis and kinetically driven transition to more stable structures. These types of programs with the ability to include more information during the folding process may in

the future generate more 'evolutionarily consistent' structures.

This chapter has shown that sequence data alone cannot be reliably used to determine an evolutionary relationship between mrpRNA and pRNA structures. Only the biological structures analysed in the previous chapter seem to have this capability. Tertiary structure analysis of mrpRNA and pRNA may be more revealing but this process is still being developed for proteins (where there are many more tertiary structures available). Some relationships between some pRNA sequences (especially the bacterial pRNA sequences) was possible, indicating that some evolutionary information is maintained in the sequence.

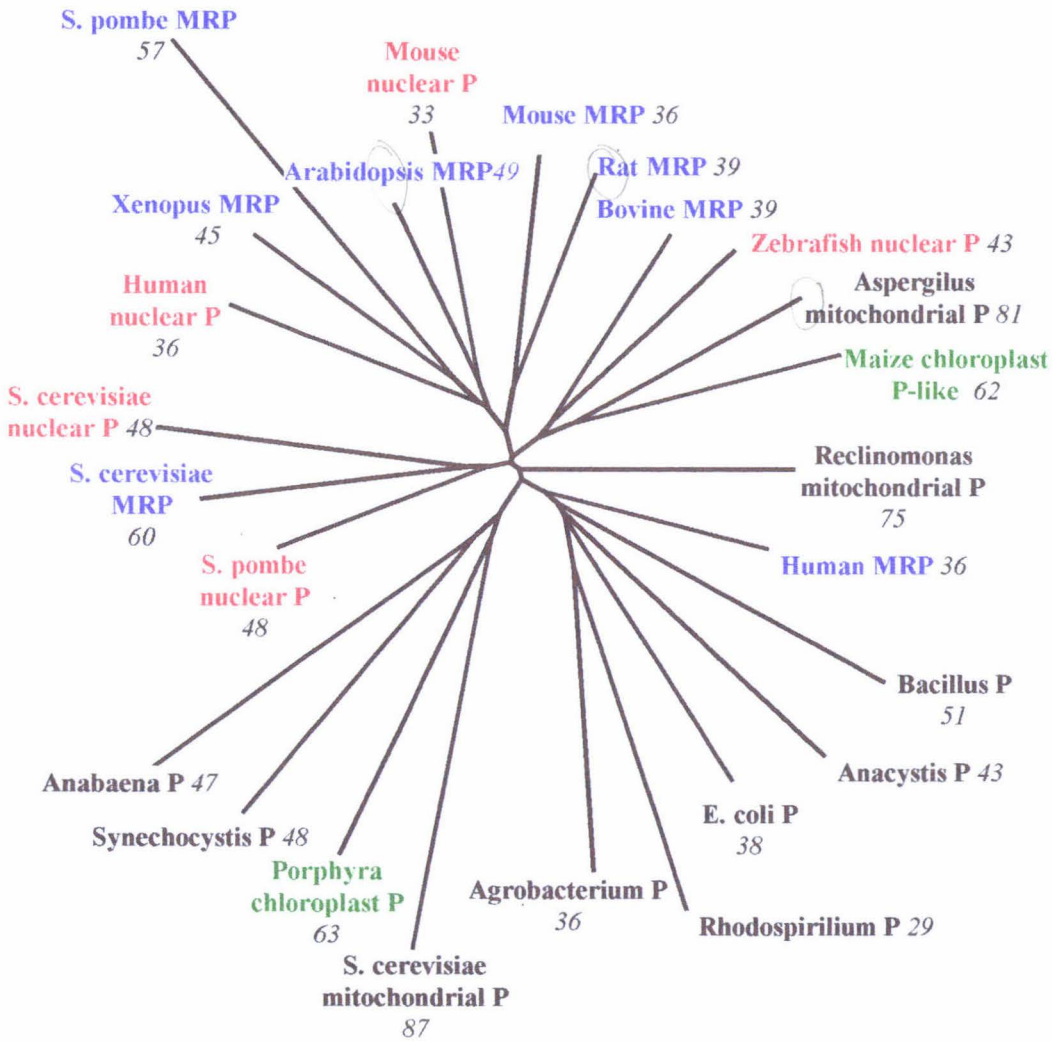


Figure 5.10: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Full format (f) - uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

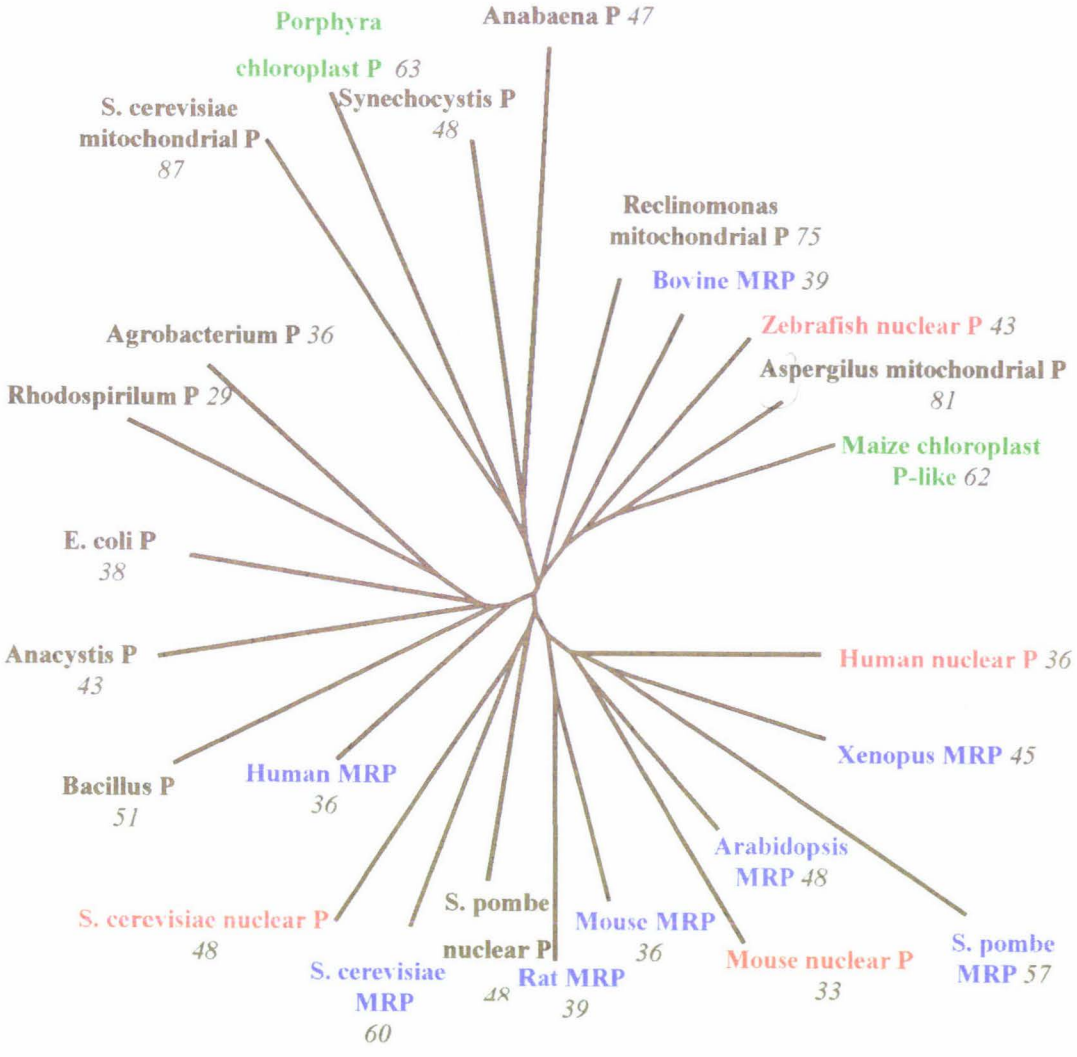


Figure 5.11: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Full format (f) - corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

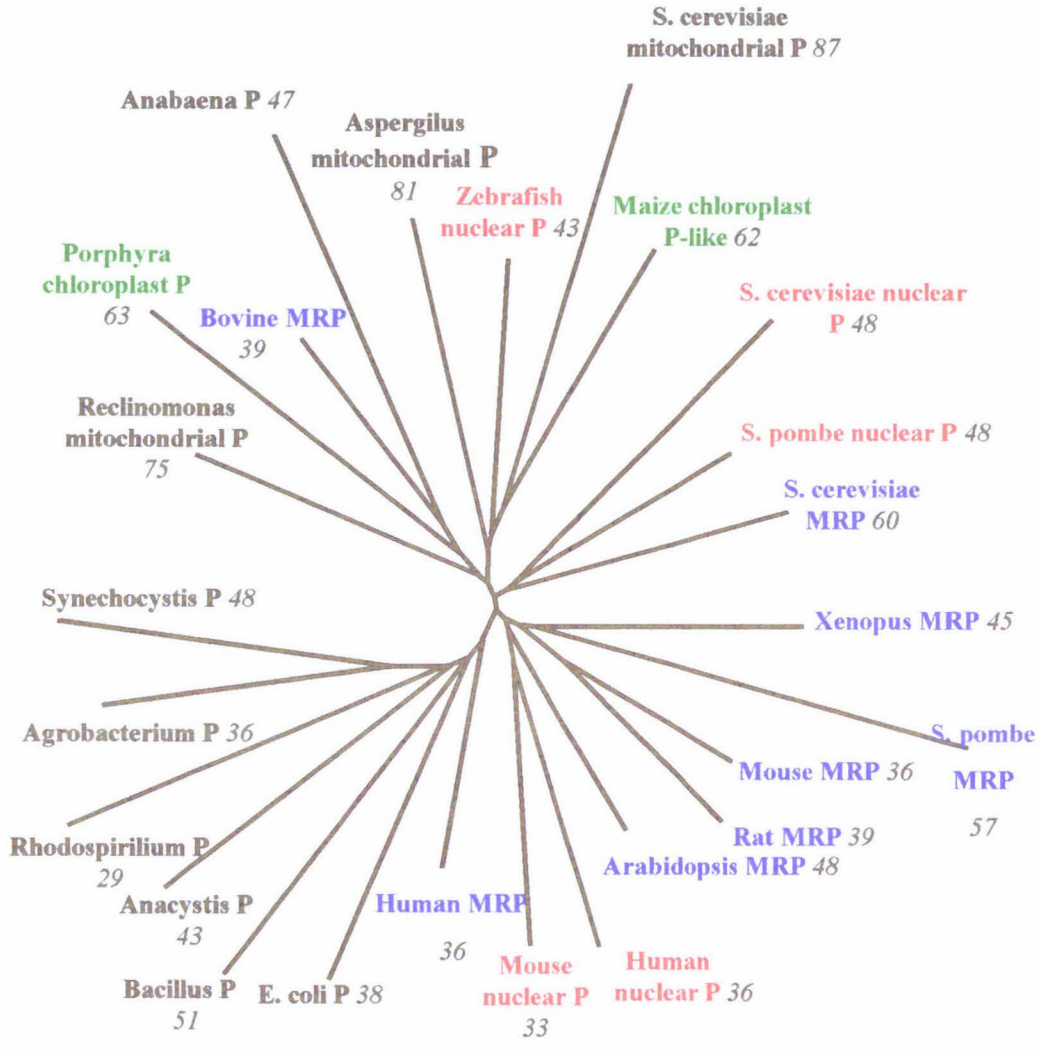


Figure 5.12: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. HIT format (h) – uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

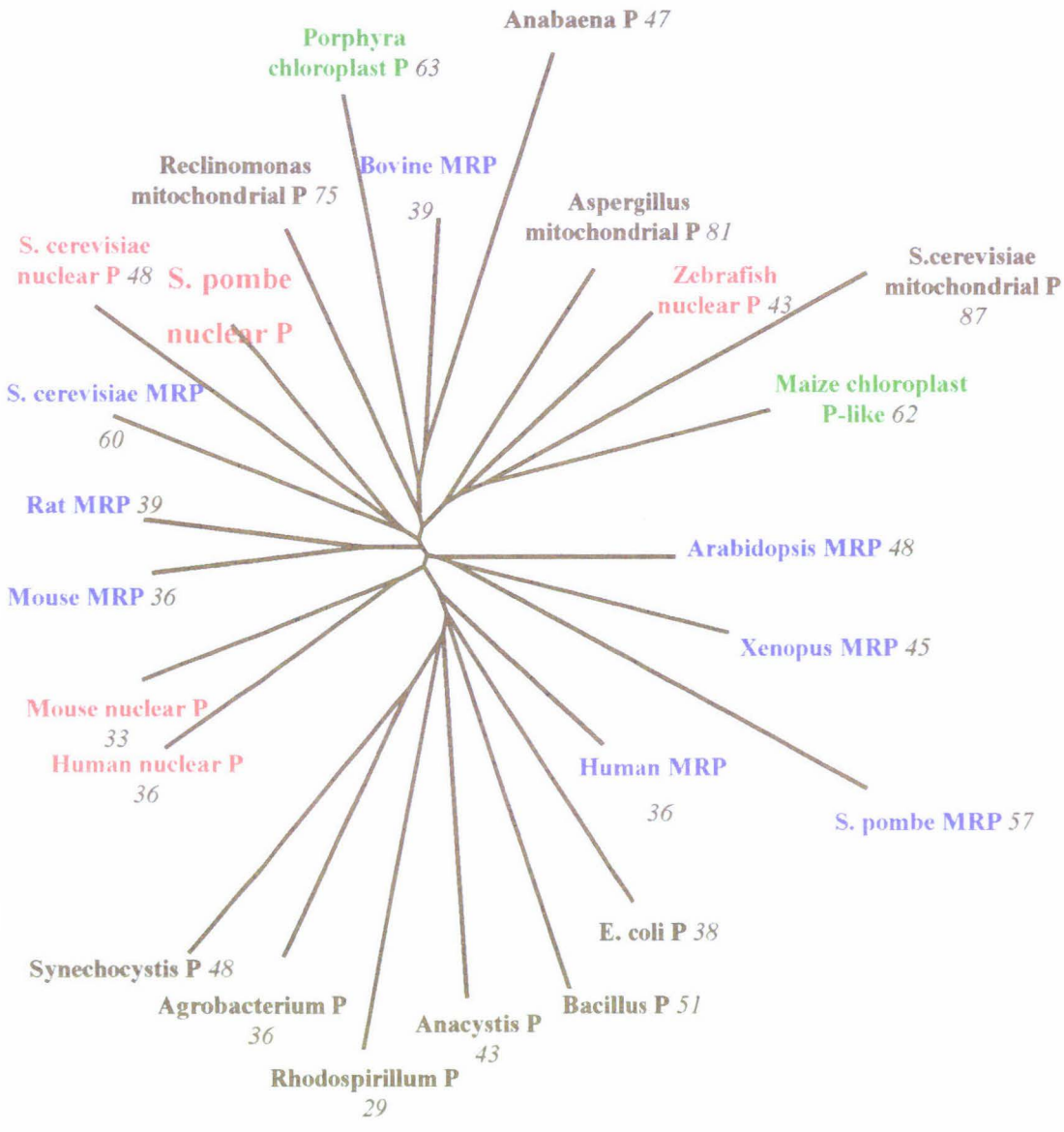


Figure 5.13: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. HIT format (h) – corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

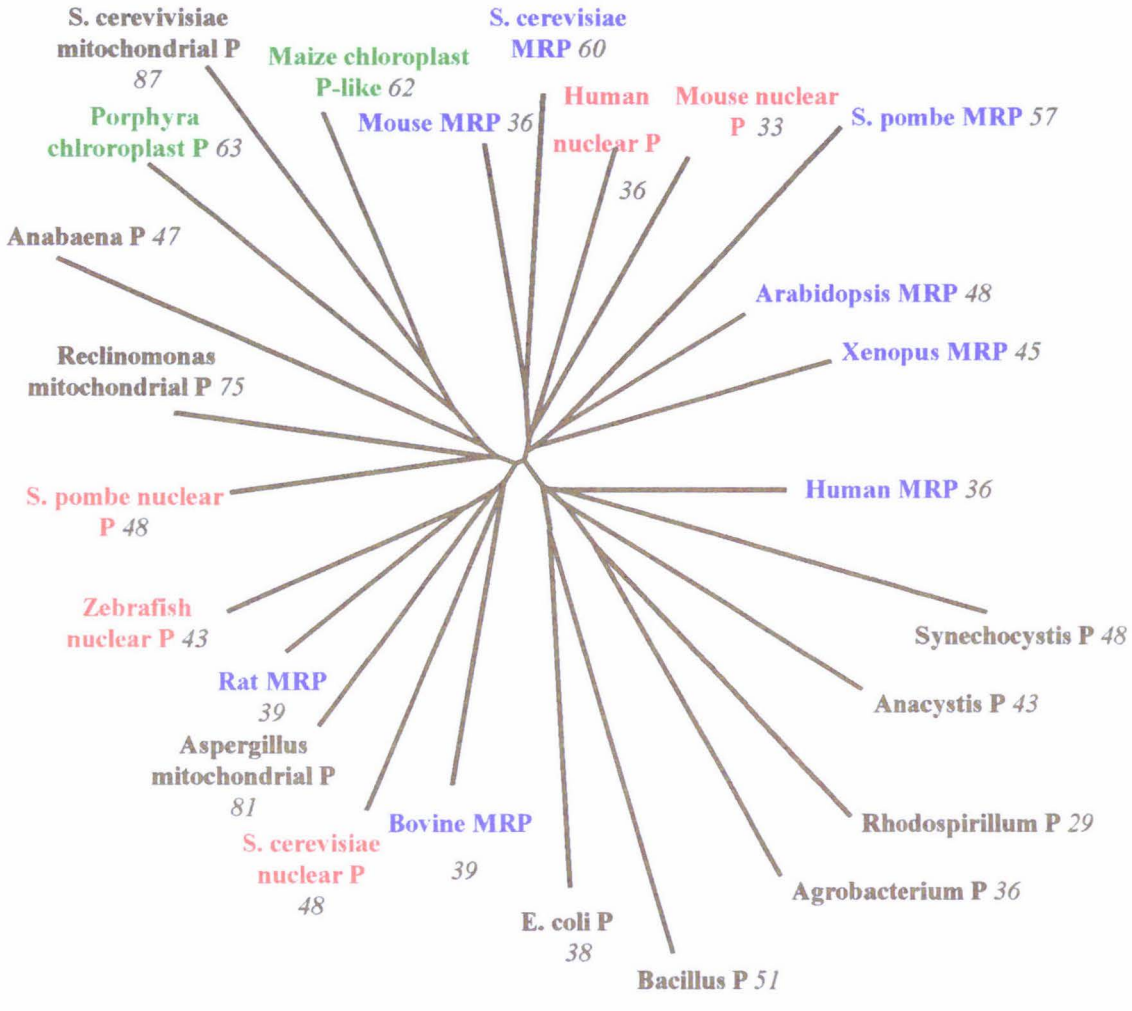


Figure 5.14: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Weighted coarse (w) - uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

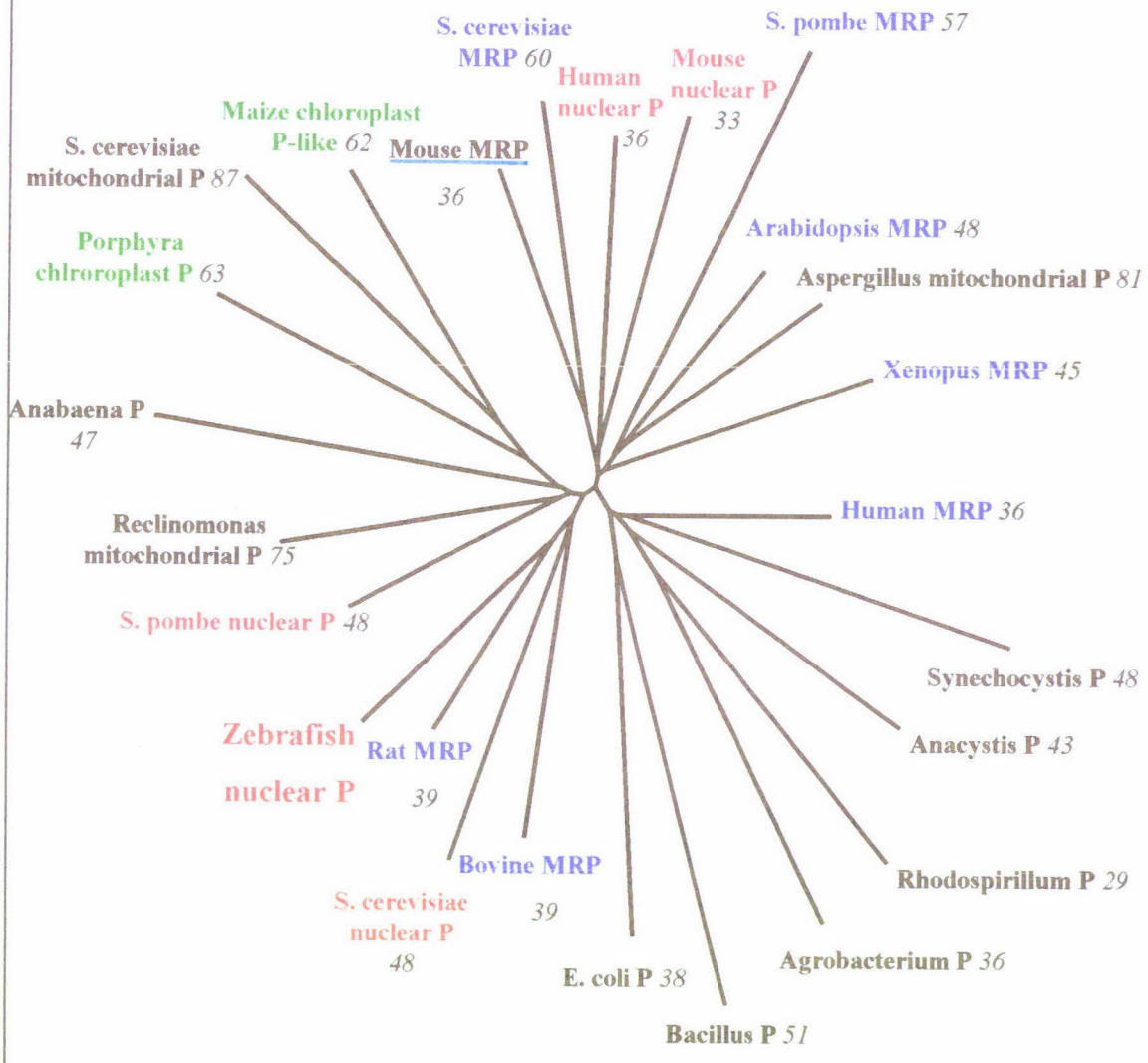


Figure 5.15: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Weighted coarse (w) - corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

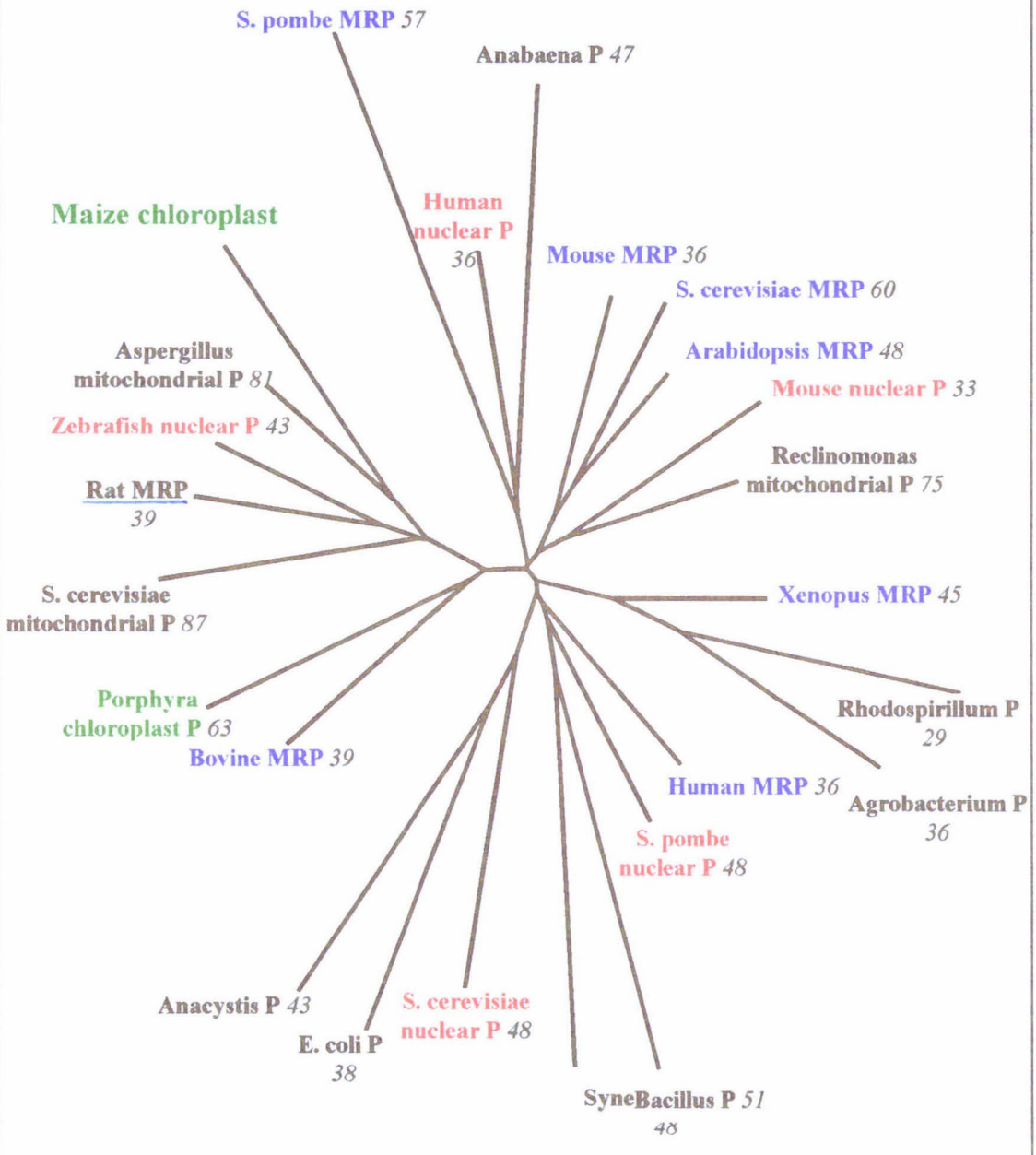


Figure 5.16: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Coarse structure (c) – uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

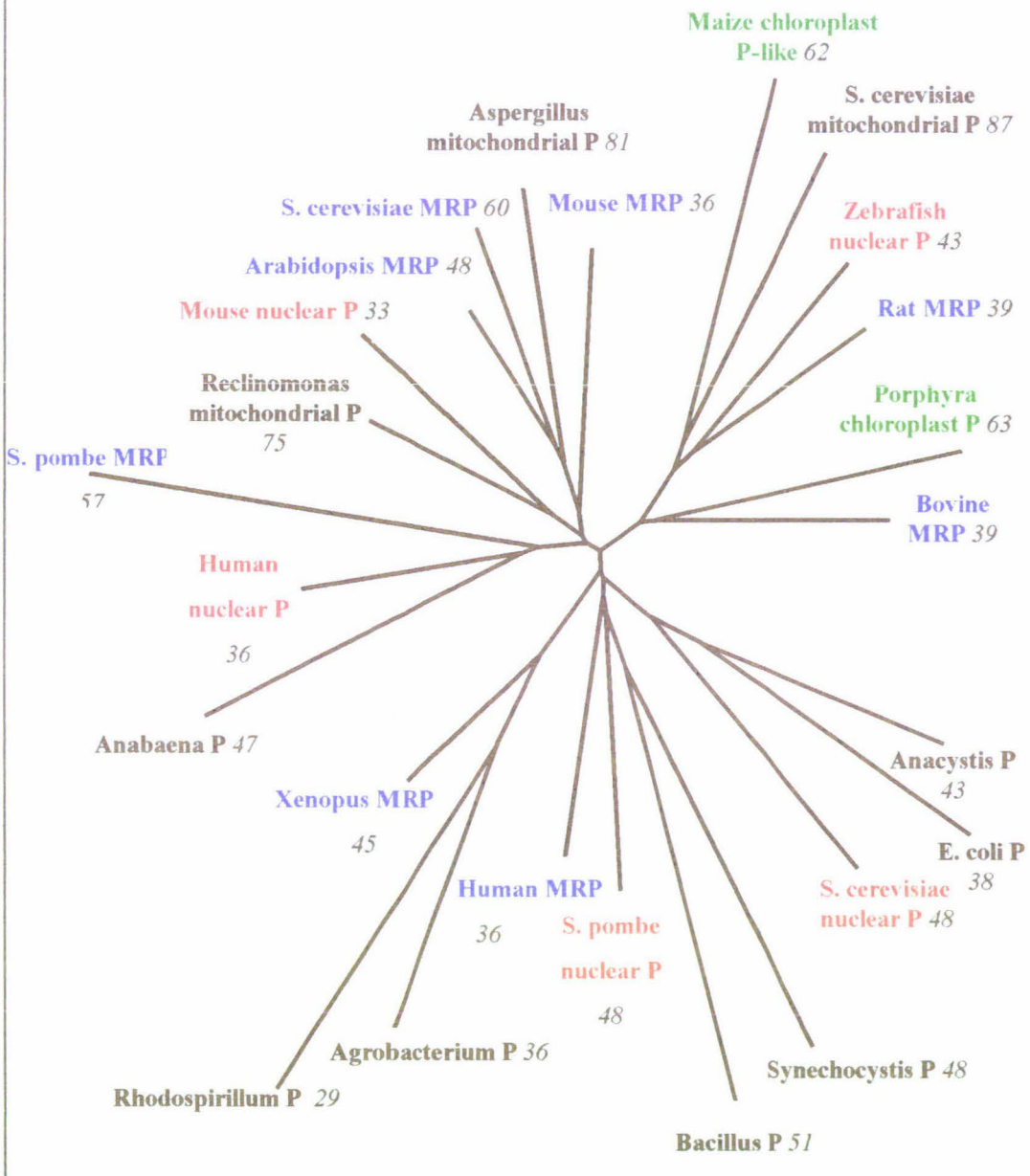


Figure 5.17: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Coarse structure (c) – corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

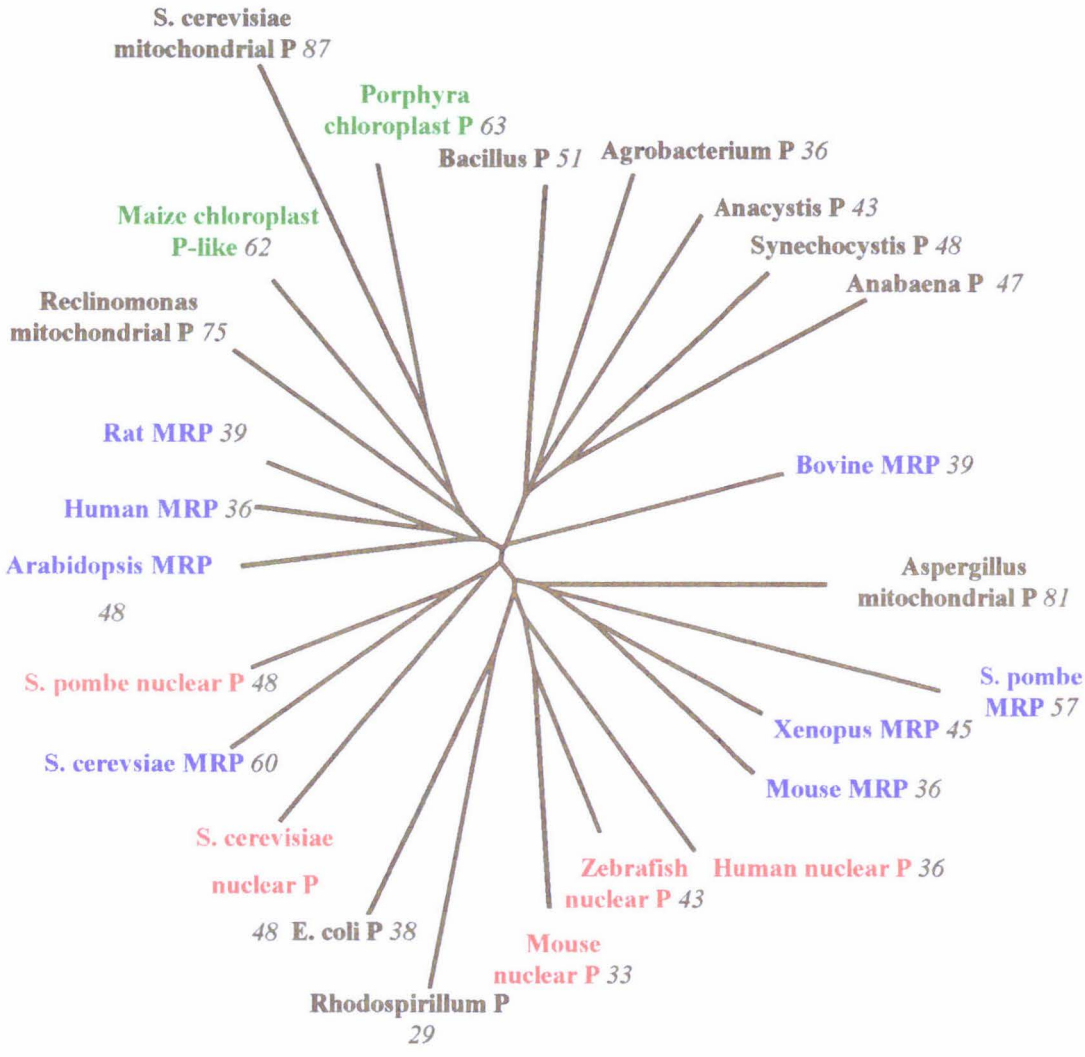


Figure 5.18: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNADraw. Full structure (f) – uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

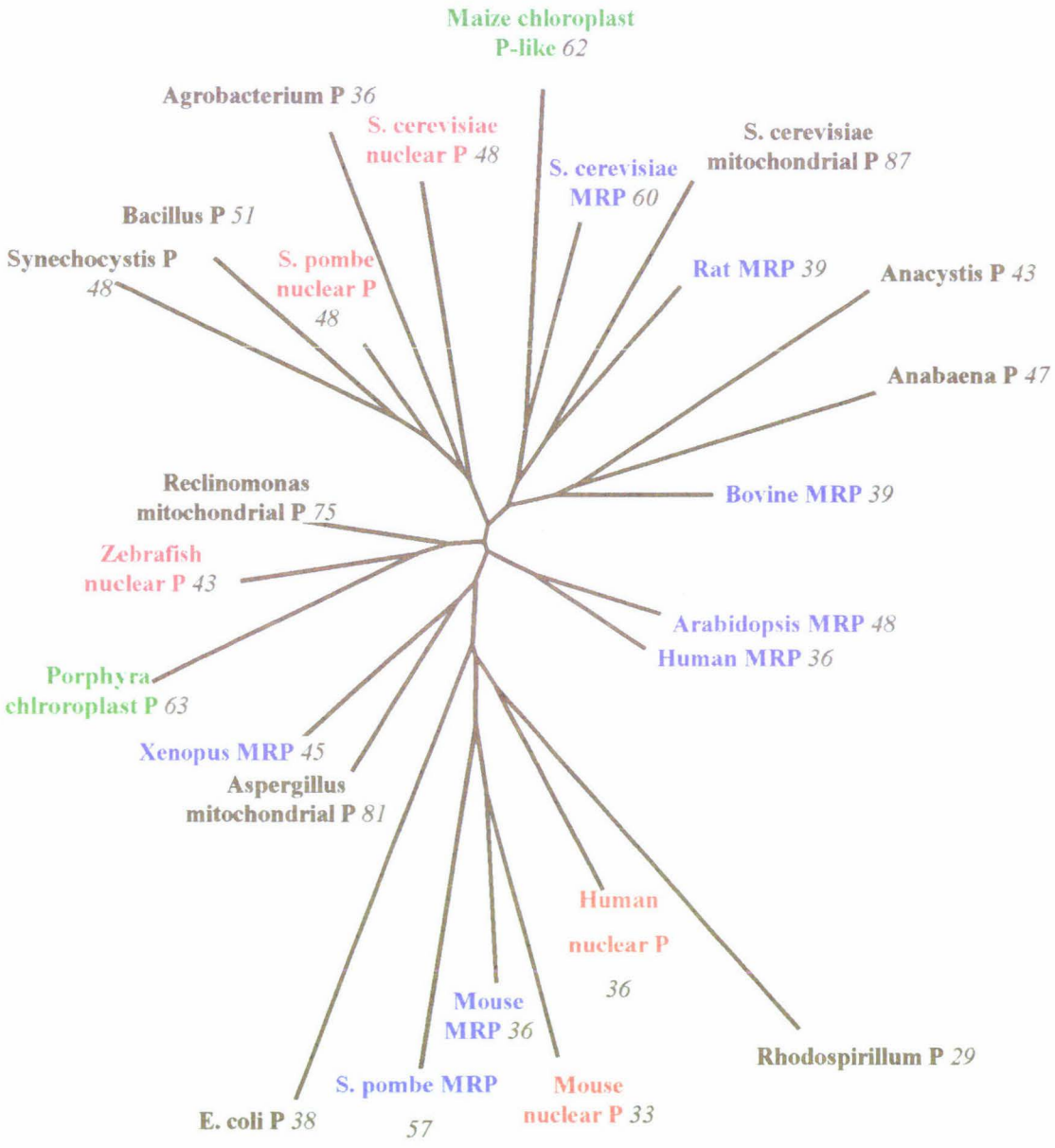


Figure 5.19: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Full structure (f) – corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

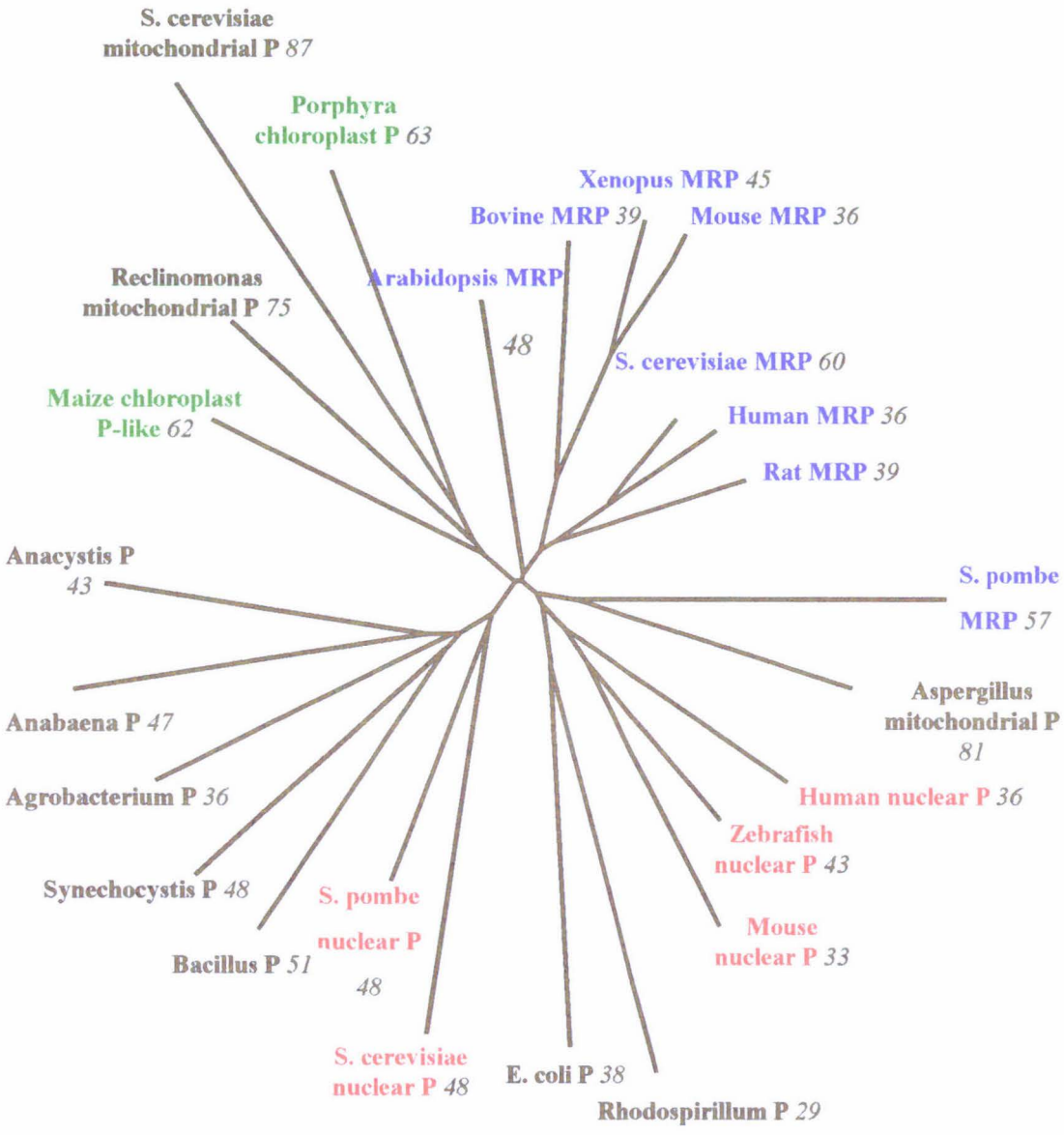


Figure 5.20: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. . HIT structure (h) - uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

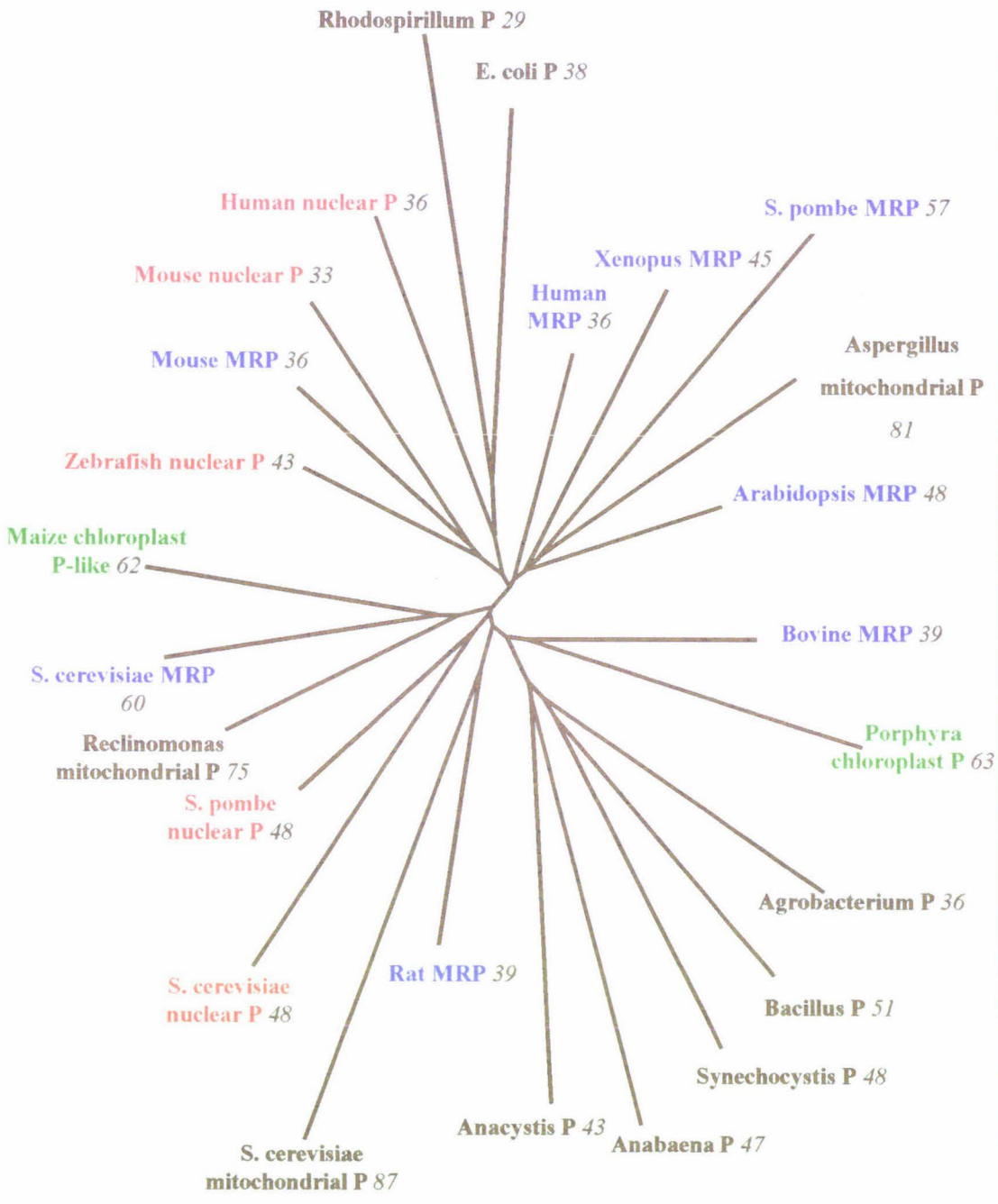


Figure 5.21: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. HIT structure (h) - corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

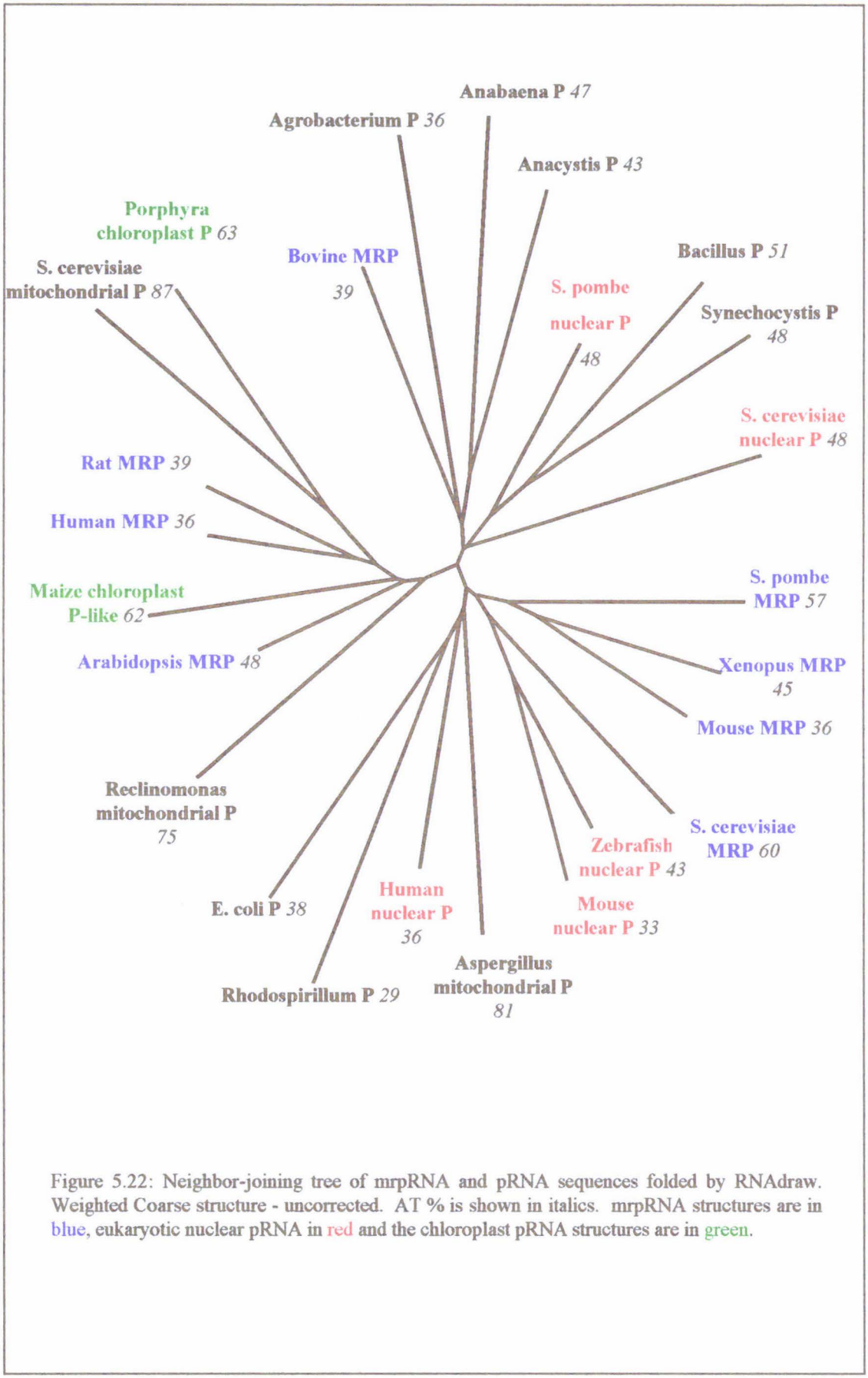


Figure 5.22: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Weighted Coarse structure - uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

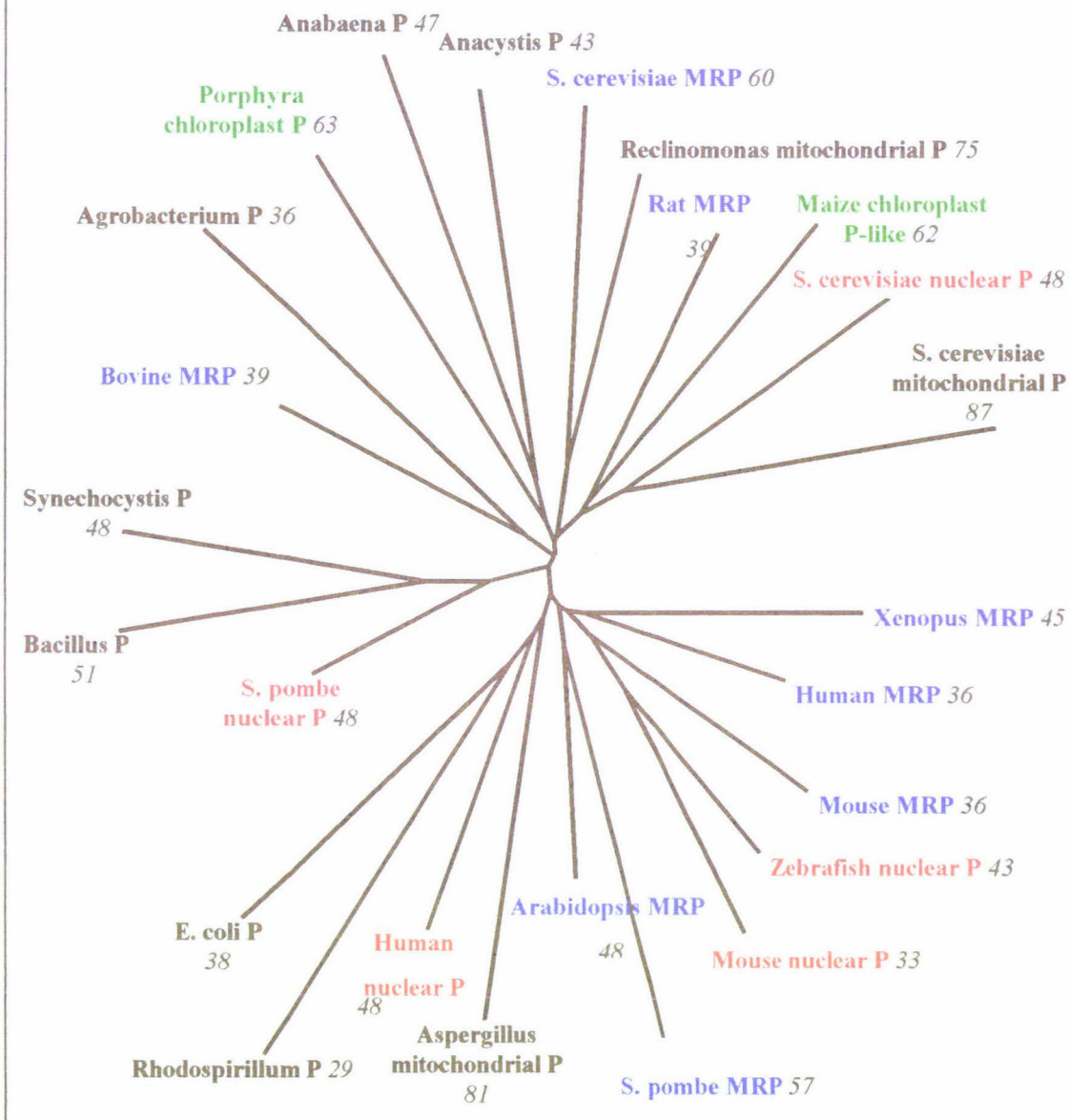


Figure 5.23: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Weighted Coarse structure - corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

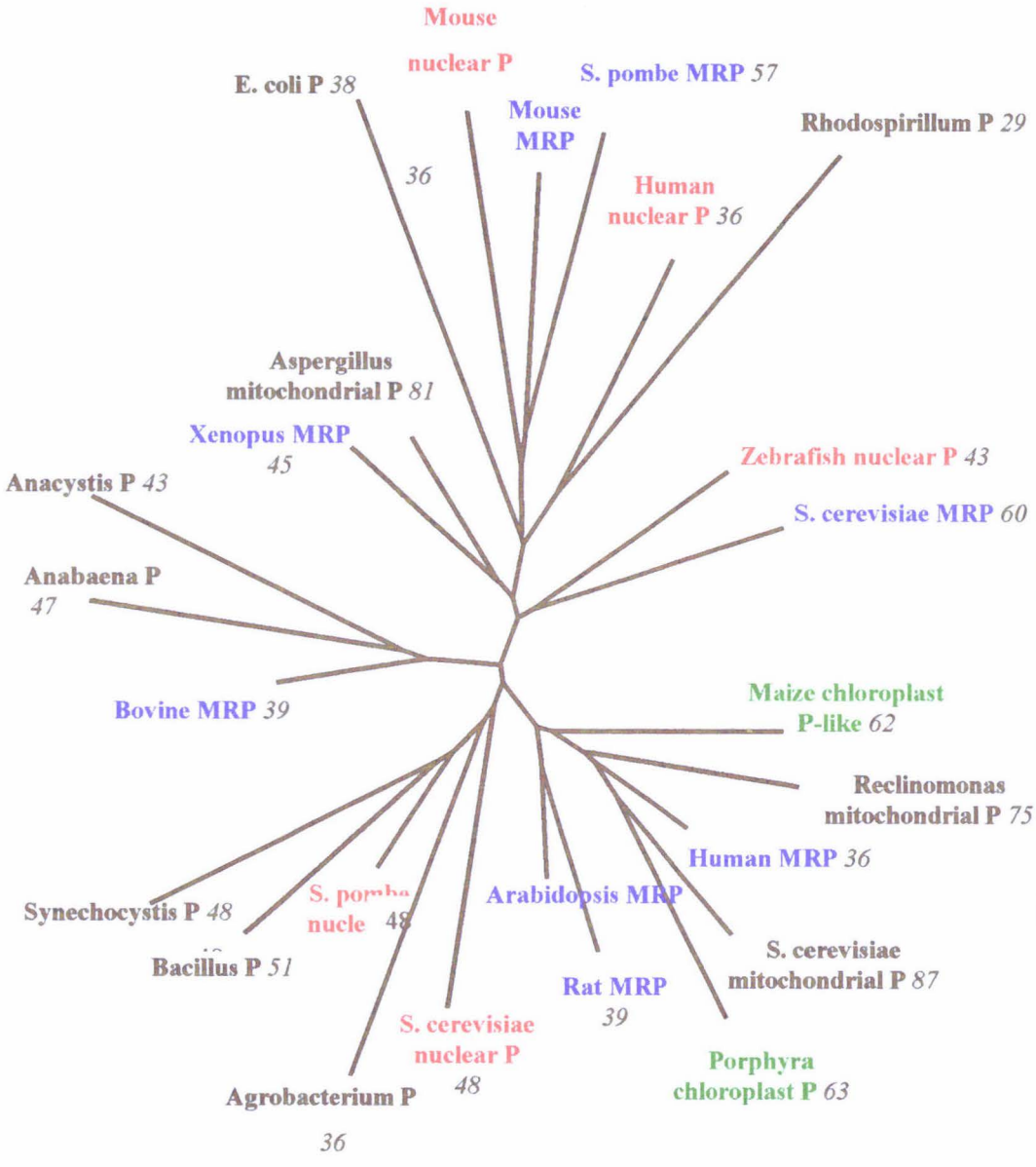


Figure 5.24: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Coarse structure - uncorrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

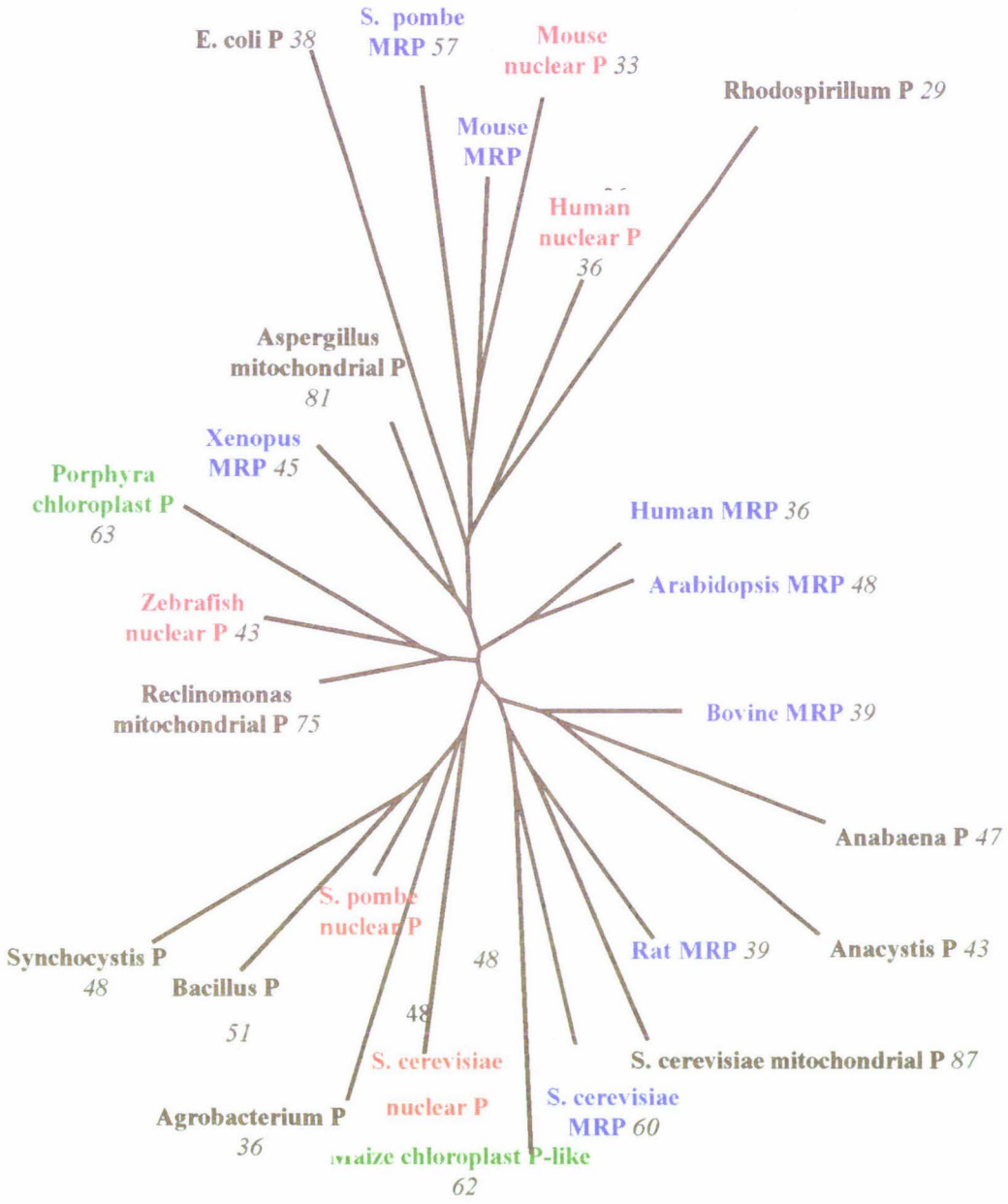


Figure 5.25: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Coarse structure - corrected. AT % is shown in italics. mrpRNA structures are in blue, eukaryotic nuclear pRNA in red and the chloroplast pRNA structures are in green.

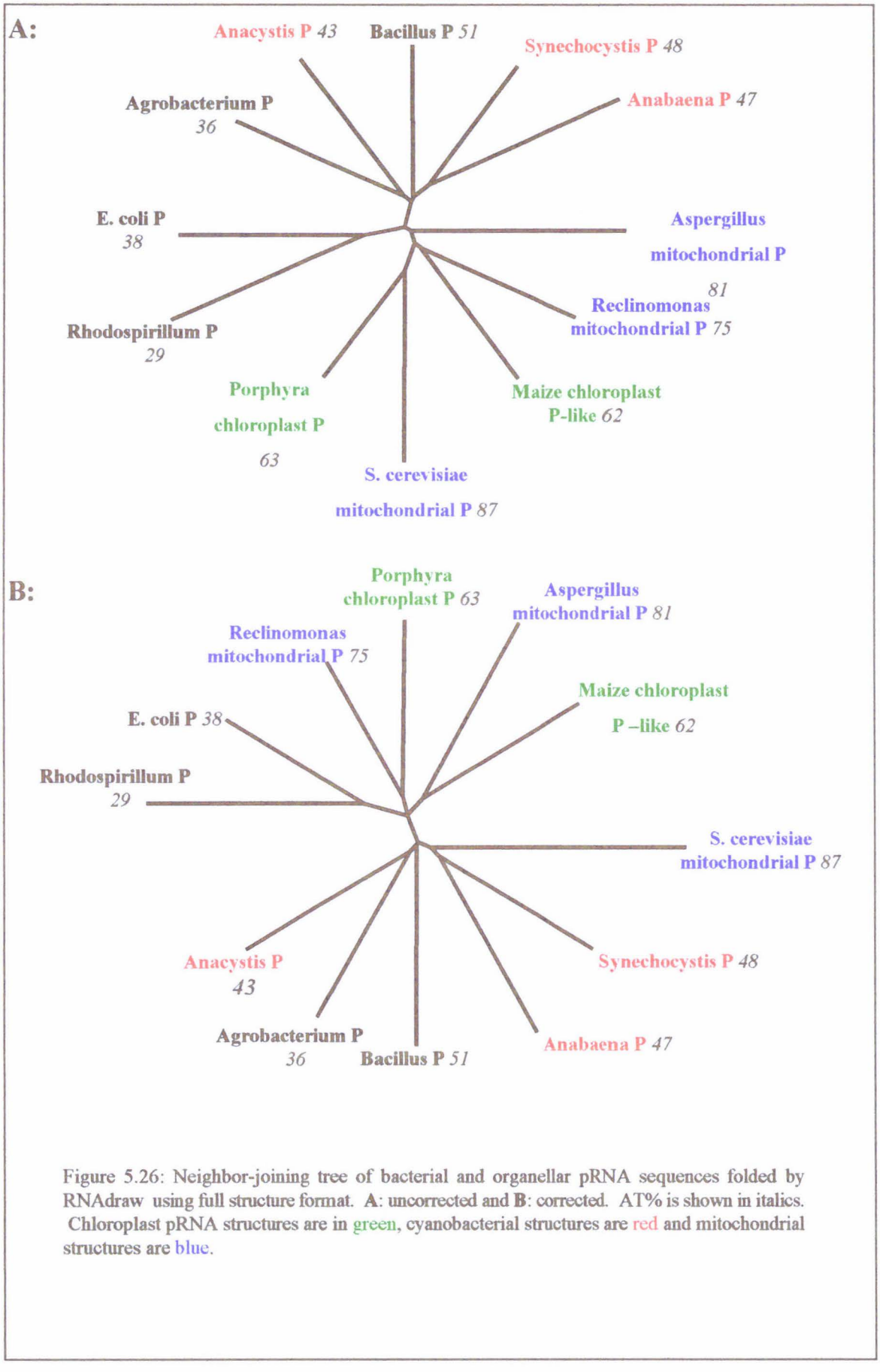


Figure 5.26: Neighbor-joining tree of bacterial and organellar pRNA sequences folded by RNAdraw using full structure format. **A**: uncorrected and **B**: corrected. AT% is shown in italics. Chloroplast pRNA structures are in green, cyanobacterial structures are red and mitochondrial structures are blue.

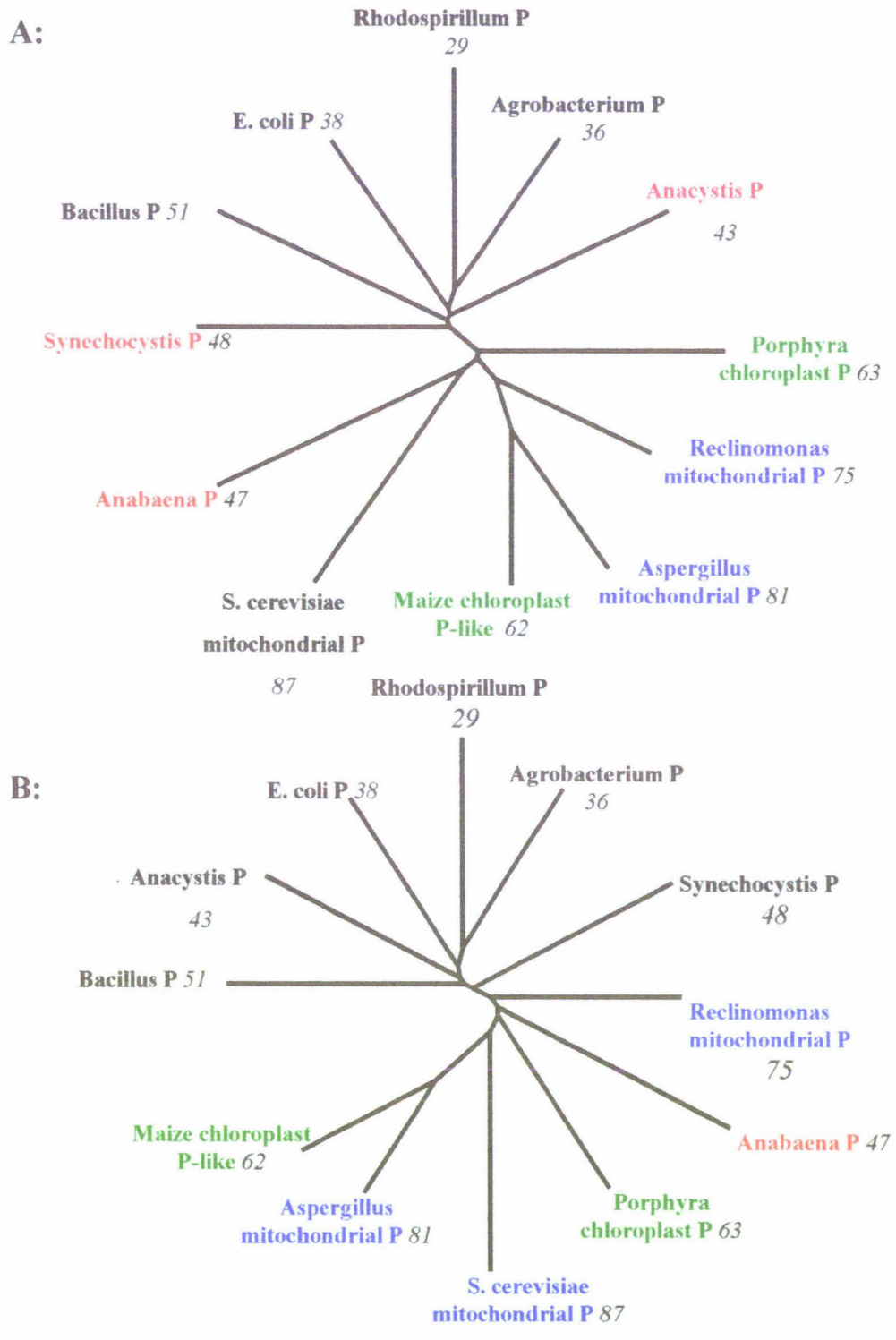


Figure 5.27: Neighbor-joining tree of bacterial and organellar pRNA sequences folded by RNAstructure using full structure format. **A**: uncorrected and **B**: corrected. AT% is shown in italics. Chloroplast pRNA structures are in green, cyanobacterial structures are red and mitochondrial structures are blue.

Chapter 6

Investigation of AT content and length on the comparison of folded pRNA sequences.

Introduction

The question has been raised in previous chapters, as to whether the AT content and length of the sequence have an effect on the trees produced from comparisons of folded secondary structures. This chapter will look at some folded pRNA structures to see if they are more similar to other pRNA structures, or to structures generated from random sequences of the same length and AT content.

Folded structures from the maize chloroplast pRNA (Maize chloroplast P-like) have consistently grouped either with *Porphyra* chloroplast pRNA or with mitochondrial pRNA folded structures. This chapter will look to see if this is due to similar AT content and/or length of sequence, or due to the sequence having other organellar pRNA 'characteristics'.

Six pRNA sequences (the same that were used in the first group of sequences in chapter 5), were folded and compared with structures of sequences shuffled from the *E. coli*, *Porphyra* and maize chloroplast pRNAs. If the length or the AT % of the sequence were having an effect on the placement of the structures on the trees, it is expected that the pRNA structures would group with structures constructed from shuffled sequences of identical length and AT% than with other pRNA structures.

The lengths of protein sequences can have a large effect on tertiary folding. Five or six protein fragments with identical amino acid composition may adopt totally unrelated 3D folding (Abagyan and Batalov 1997). With RNA is also possible that the length of the sequence would have an effect on the amount of folding in the secondary structure. This can be looked at by examining how folded structures from the *Anacystis* pRNA and the *Porphyra* chloroplast pRNA (two sequences which have very similar lengths but different AT contents) are placed on the trees constructed here.

Materials and Methods

Random sequences were calculated by shuffling the nucleotides in the *E. coli*, *Porphyra*, and the maize chloroplast pRNA sequences. This gave shuffled sequences with the same length and AT content that are found in original pRNA sequences. The programs *snfold* (shuffles the sequence then folds these random sequences with the folding program Mfold (Zuker 1989)) and *rsnfold* (shuffles the sequence then folds RNAfold (Hofacker et al 1994)) were written by R. F. Pointon for this project (These programs are summarised in appendix 4).

The six pRNA sequences from *E. coli*, *Anacystis*, *Anabaena*, *Synechocystis*, *Porphyra* chloroplast and the maize chloroplast pRNA sequences were folded using Mfold and RNAfold. Only the optimal secondary structure out of those calculated is used here. The pRNA structures were then compared to the folded shuffled sequences using the RNAdistance program (Hofacker et al 1994) calculating distances in the full structure format only.

Phylogenetic trees were created from distance data using the neighbor-joining algorithm from the Phylip package (Felsenstein 1989), and trees were drawn using TreeView (Page 1996). All programs for shuffling, folding and comparing the sequences were run on a Unix system running SunOS release 4.1.3.

There were no corrections for the circular structures of the *Porphyra* chloroplast pRNA and the maize chloroplast pRNA, which are formed with RNAfold (same algorithm as RNAdraw). This is because it would be expected that some of the random structures generated from these sequences might also form circular structure. None of the pRNA structures generated circular structure when folded with the Mfold program (same algorithm as RNAstructure).

Results

The results here consist of six figures, two each comparing shuffled sequences from the *E. coli*, *Porphyra*, and maize chloroplast pRNA sequences to the six pRNA sequences (three cyanobacterial, two chloroplast and *E. coli* pRNA sequences). Each group of shuffled and pRNA sequences is folded once by RNAfold (Figures 6.1 - 6.3) and then once by Mfold (Figures 6.4 - 6.6).

When the six pRNA sequences are folded with RNAfold and compared to the folded shuffled *E. coli* pRNA sequences (rsnfold) (Figure 6.1) all of the pRNA sequences group together. For thirty-two sequences already on a tree, by random chance alone there is a 1/32 chance that two species will group together. From this it can be calculated that there is a 1/28596 chance (probability of 0.00003496) that all six of the pRNA sequences have grouped together. Looked at in another way, if the six pRNA sequences form a tree then all twenty-six shuffled sequences group onto a single edge on that tree of six pRNA sequences. Thus the *E. coli* pRNA RNAfold structure is more similar to the other pRNA sequences (folded with RNAfold) than to any of the structures shuffled from itself. The maize chloroplast pRNA structure is also grouped here with the *Porphyra* chloroplast pRNA structure, and the three cyanobacterial structures are grouped together. However, the AT contents of the cyanobacterial pRNA sequences are similar to each other and very different from the two chloroplast pRNA sequences and the *E. coli* pRNA sequence. This last result may not be too surprising as it is still unclear whether chloroplasts were formed by a single or multiple endosymbiotic events (Lockhart et al. 1996). The lengths of the pRNA sequences from the cyanobacterial species *Anabaena* and *Synechocystis* are again similar to each other and different from *Anacystis* pRNA sequence. Similarly the length of the *Porphyra* chloroplast pRNA is more similar to the *E. coli* pRNA, than to the maize chloroplast pRNA with which it is grouped. The *Anacystis* pRNA and the maize chloroplast pRNA sequences also have very similar lengths, but are placed away from each other. Thus when the pRNA sequences and shuffled *E. coli* sequences are folded with RNAfold, the length and AT content are not determining their positions on the tree.

When a sequence of a higher AT content such as the *Porphyra* chloroplast pRNA is used to generate shuffled sequences then folded with RNAfold (Figure 6.2), the eubacterial pRNA structures are grouped separately from the chloroplast pRNA structures, though still close to them. The maize and *Porphyra* chloroplast pRNA sequences are still grouped together. However, they are also grouped with some of the random structures (of the same AT content and length as the *Porphyra* chloroplast pRNA sequence). The *Porphyra* chloroplast pRNA sequence has very similar AT content to the maize chloroplast pRNA sequence. As with the *E. coli* shuffled sequences, the lengths of the pRNA sequences, except for the *Anabaena* and *Synechocystis* pRNAs, do not seem to group

together.

The maize chloroplast pRNA structure does not group with the *Porphyra* chloroplast pRNA structure when random structures generated from the maize sequence are used (Figure 6.3). The *E. coli* sequence is also grouped away from the cyanobacterial sequences. This tree does not tell us if the AT content is having an effect or not. The *Porphyra* chloroplast pRNA, and the putative maize chloroplast pRNA sequences give circular structures with the RNAfold program. It is not known whether any of the structures folded from the shuffled sequences are also circular and are having an affect on the placement of the two chloroplast structures on the tree.

With Mfold there are no circular structures formed by these six pRNA sequences. With this program, the eubacterial pRNAs are split into two groups when compared to the structures folded from the shuffled *E. coli* pRNA sequence (constructed by snfold) (Figure 6.4). The *E. coli* and *Anacystis* pRNA structures in one group and the *Anabaena* and *Synechocystis* pRNA structures are in another. As with the RNAfold program, the maize chloroplast pRNA is grouped with the *Porphyra* chloroplast pRNA but away from the eubacterial pRNA structures. The *Anacystis* and *E. coli* pRNA sequences do not have similar AT content or lengths but the other two groupings of the *Anabaena* and *Synechocystis* pRNA structures and the two chloroplast pRNA structures do.

When structures folded from the shuffled *Porphyra* chloroplast pRNA sequence are folded with Mfold and compared to the six pRNA Mfold structures (Figure 6.5), all six pRNA structures fall into a group well separated from the shuffled structures. Again there are three pairs of structures, the *E. coli* and the *Anacystis* pRNA structures, the *Anabaena* and *Synechocystis* pRNA structures, and the two chloroplast structures. This pairing is also shown when structures folded from the shuffled maize chloroplast pRNA sequence are used (Figure 6.6). However, the chloroplast structures are grouped away from the eubacterial pRNA structures in this tree.

The Mfold folding program is less influenced by the formation of circular structure and thus may be more advantageous when looking at the effects of AT content and length on the comparison of folded pRNA sequences. It is not known whether there are any circular structures formed by the random sequences but it appeared that this is less likely with Mfold than with RNAfold.

The consistent pairing of the *Anabaena* and *Synechocystis* pRNA sequences may be due to three factors. They are both cyanobacteria and their pRNA sequences have very similar AT contents and lengths. Overall, there is an indication that AT content is not having a major effect on where structures are placed on the tree. There is some grouping of the pRNA sequences with similar AT contents, but not consistent grouping of the pRNA structures with structures from the shuffled sequences of the same AT content. This indicates that there is some other distinguishing feature in the pRNA sequence, which is not found in the shuffled sequences.

Anacystis pRNA and the *Porphyra* chloroplast pRNA structures (whose sequences have similar length) do not group together in any of the trees. There is no indication in any of the trees that the length of the pRNA sequences is having a major effect on the placement of structures.

Discussion

In the trees constructed from folded structures of mrpRNA and pRNA (chapter 5), the maize chloroplast pRNA sequence consistently grouped with the *Aspergillus* mitochondrial pRNA sequence or the *S. cerevisiae* mitochondrial pRNA sequence. Both these sequences have a much higher AT content than is seen in both the maize and the *Porphyra* chloroplast pRNA sequences. The *Porphyra* chloroplast pRNA and the maize chloroplast pRNA sequences have very similar AT contents (63 and 62). If the two chloroplast pRNA sequences grouped together when both the *Porphyra* chloroplast shuffled structures and the maize chloroplast pRNA shuffled structures were compared, then it would be easy to say that the AT content is having a major effect on the placement of the chloroplast pRNA structures. However, the two chloroplast pRNA structures act differently when compared to the structures folded from the shuffled *Porphyra* and maize chloroplast pRNA sequences. The pairing of the maize chloroplast pRNA Mfold structure with the *Porphyra* chloroplast pRNA Mfold structure, even when sequences shuffled from the maize sequence were used, is an indication that the AT content is not the only factor in the placement of these structures.

Another factor that could be involved in the grouping of the structures on the trees is the amount of base pairing that is occurring in the structure. There is a possibility that the random structures do not contain the same percentage of pairing, as do the pRNA

structures. AT rich sequences could expect to form fewer bases pairing as the destabilisation of the structure by loops is more readily compensated by GC pairs than by AT pairs (Fontana et al. 1993). This factor is examined in the next chapter.

Overall the results from this chapter suggest that there is likely some 'characteristic' present in the pRNA sequences. This produces pRNA-like features in the folded structures which is lost when the sequence is shuffled. These structural features will group the folded structures with other pRNA structures rather than with random structures of the same AT content and length. For these sequences, at least, AT content and length of the RNA sequence is having a lesser effect on the folded structures. Any effects on much longer sequences or sequences with extreme AT contents (i.e. sequences composed of only A's and T's or sequences composed of only G's and C's) will be left for future study.

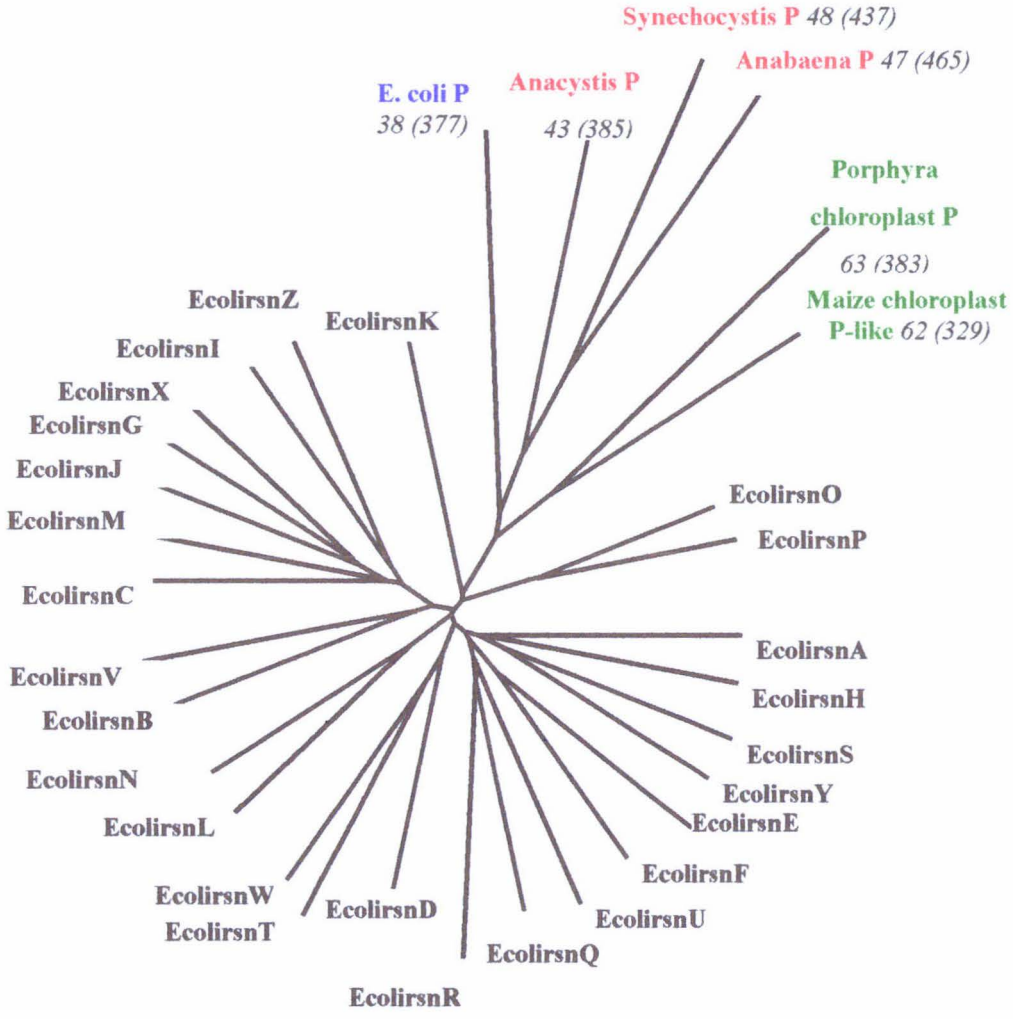


Figure 6.1: Neighbor-joining tree of pRNA sequences compared to 26*E. coli* shuffled sequences (EcolirsnA - Z) folded by RNAfold. The AT% is shown in italics with the sequence length in brackets. Cyanobacterial sequences are shown in red, the chloroplast sequences are in green and the *E. coli* sequence is in blue. The shuffled sequences have the same AT% and length as the *E. coli* sequence but do not group with it.

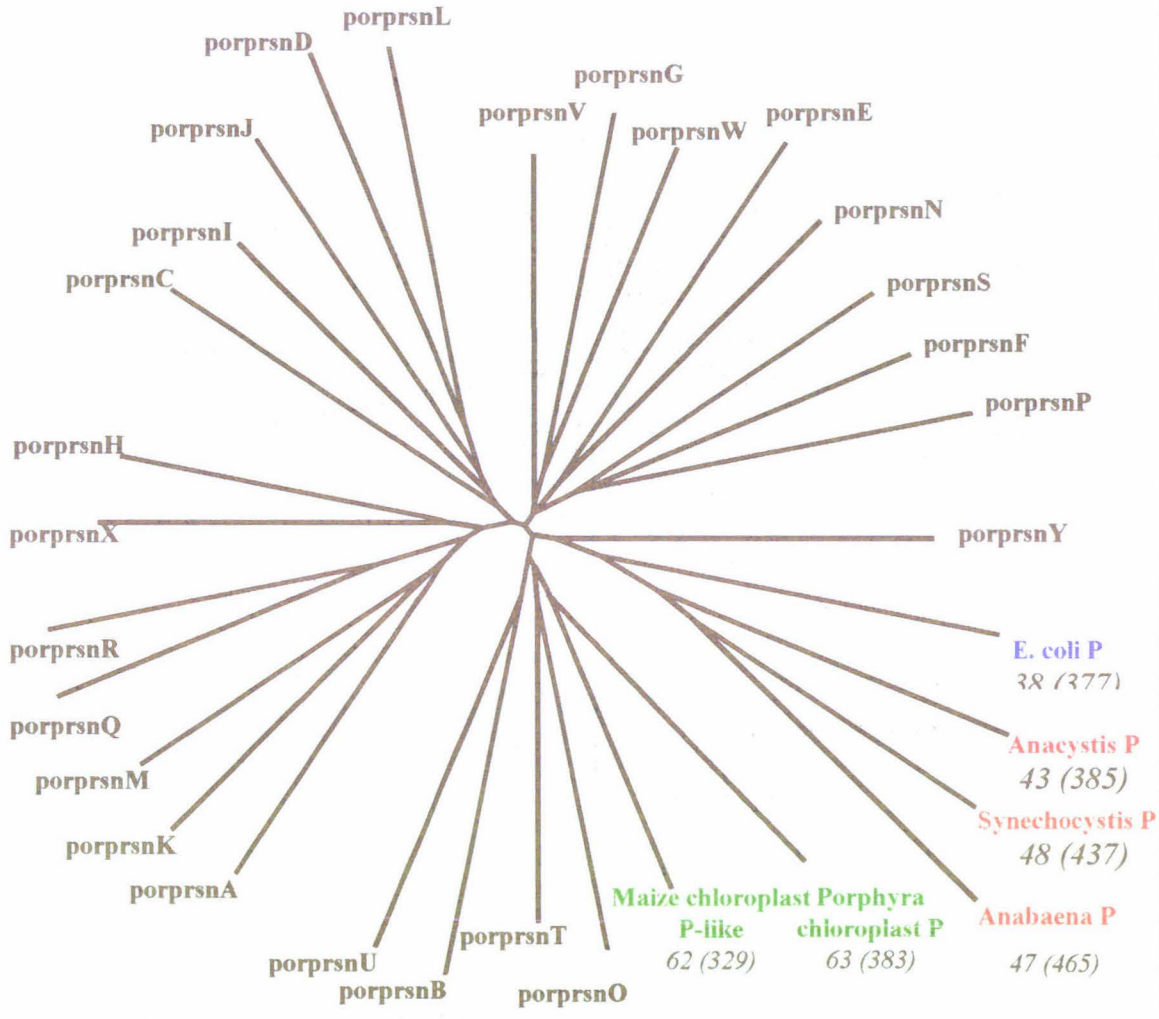


Figure 6.2: Neighbor-joining tree of pRNA sequences compared to 26 *Porphyra* shuffled sequences folded by RNAfold (porprsnA - Z). The AT% is shown in italics with the sequence length in brackets. Cyanobacterial sequences are shown in red, the chloroplast sequences are in green and the *E. coli* sequence is in blue. The shuffled sequences have the same AT% and length as the *Porphyra* sequence with only one shuffled sequence between the chloroplasts and the eubacteria.

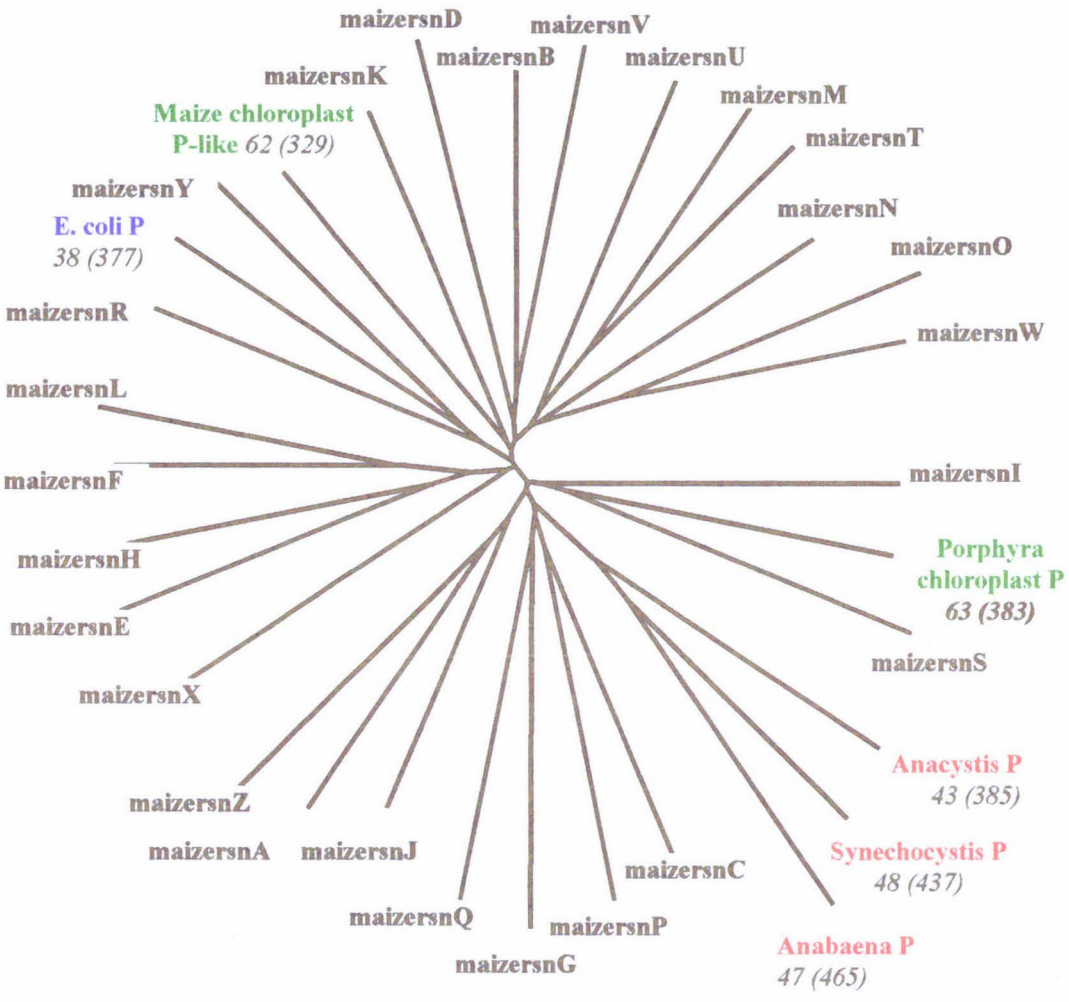


Figure 6.3: Neighbor-joining tree of pRNA sequences compared to 26 maize chloroplast pRNA shuffled sequences (maizersnA - Z) folded by RNAfold. The AT% is shown in italics with the sequence length in brackets. Cyanobacterial sequences are shown in red, the chloroplast sequences are in green and the *E. coli* sequence is in blue. The shuffled sequences have the same AT% and length as the maize sequence. The maize and *E. coli* sequences are grouped well away from the other eubacteria with the *Porphyra* chloroplast sequence only slightly away. The formation of circular structure may be having an effect on this tree.

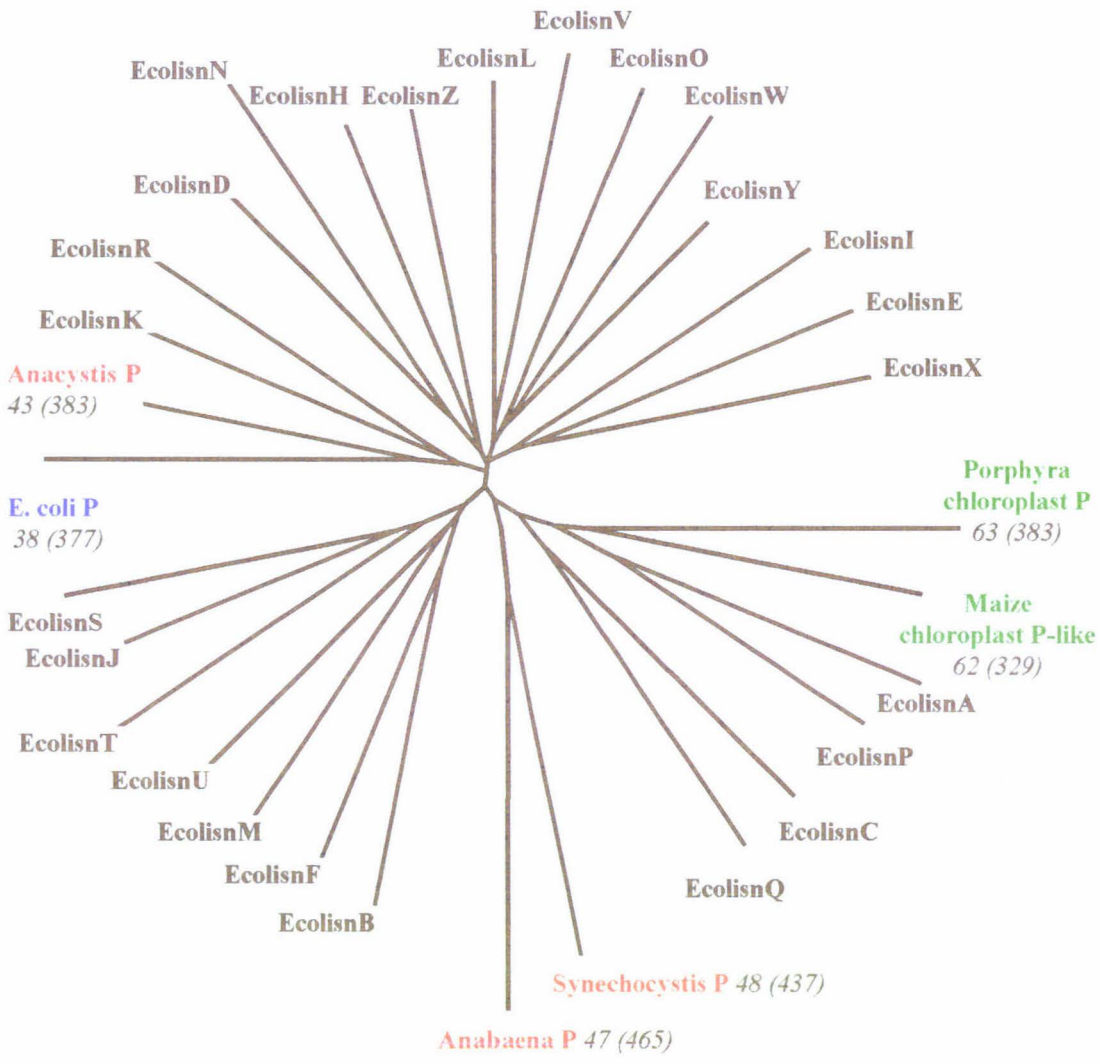


Figure 6.4: Neighbor-joining tree of pRNA sequences compared to 26 *E. coli* shuffled sequences folded by Mfold (EcolisnA - Z). The AT% is shown in italics with the sequence length in brackets. Cyanobacterial sequences are shown in red, the chloroplast sequences are in green and the *E. coli* sequence is in blue. The shuffled sequences have the same AT% and length as the *E. coli* sequence but here the *E. coli* and *Anacystis*, *Anabaena* and *Synechocystis* and the chloroplast sequences have formed separate groups.

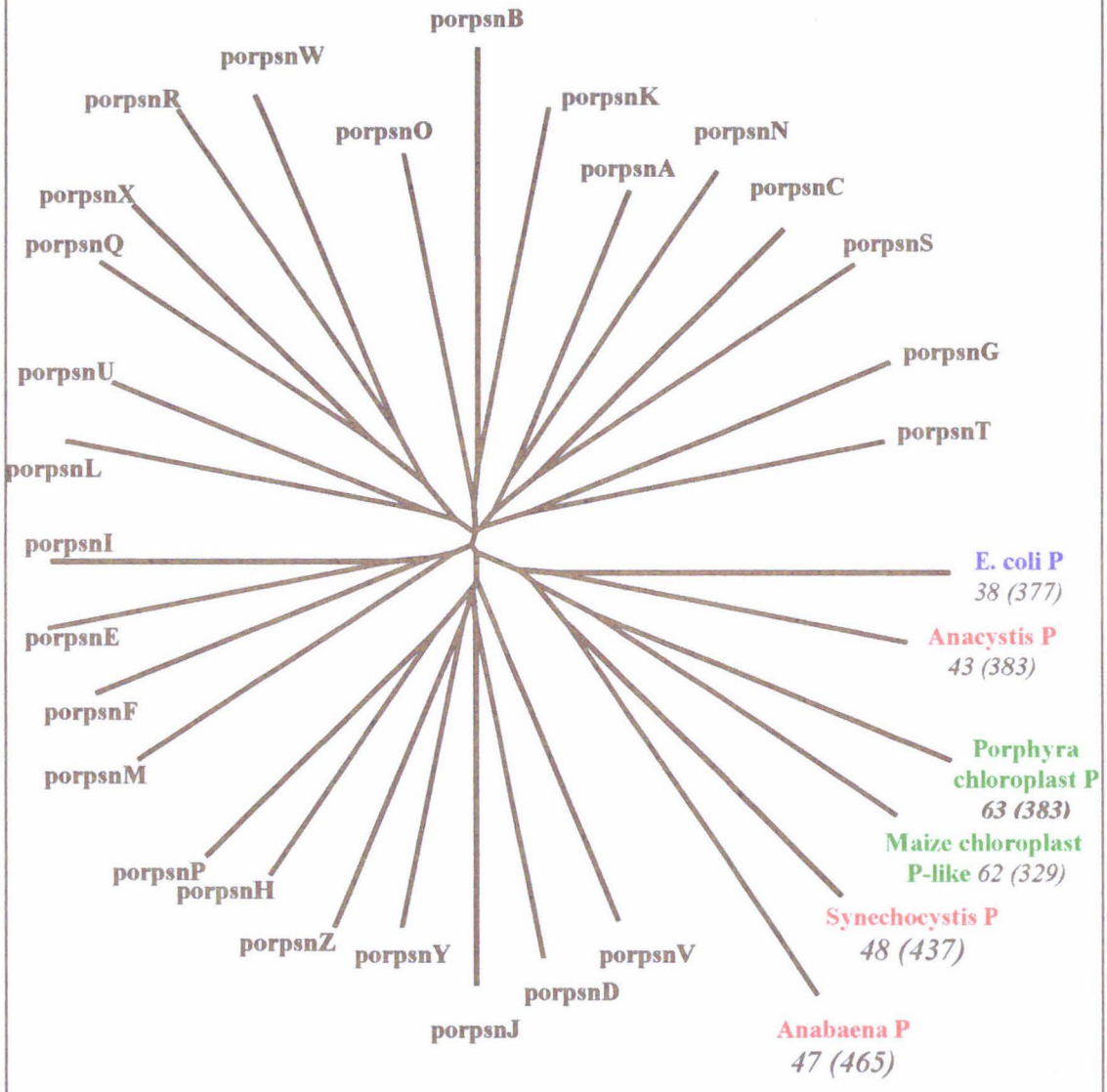


Figure 6.5: Neighbor-joining tree of pRNA sequences compared to 26 *Porphyra* shuffled sequences folded by Mfold (porpsnA - Z). The AT% is shown in italics with the sequence length in brackets. Cyanobacterial sequences are shown in red, the chloroplast sequences are in green and the *E. coli* sequence is in blue. The shuffled sequences have the same AT% and length as the *Porphyra* sequence but all pRNA sequences are grouped together. However, the *Anacystis* sequence has grouped away from the other cyanobacteria.

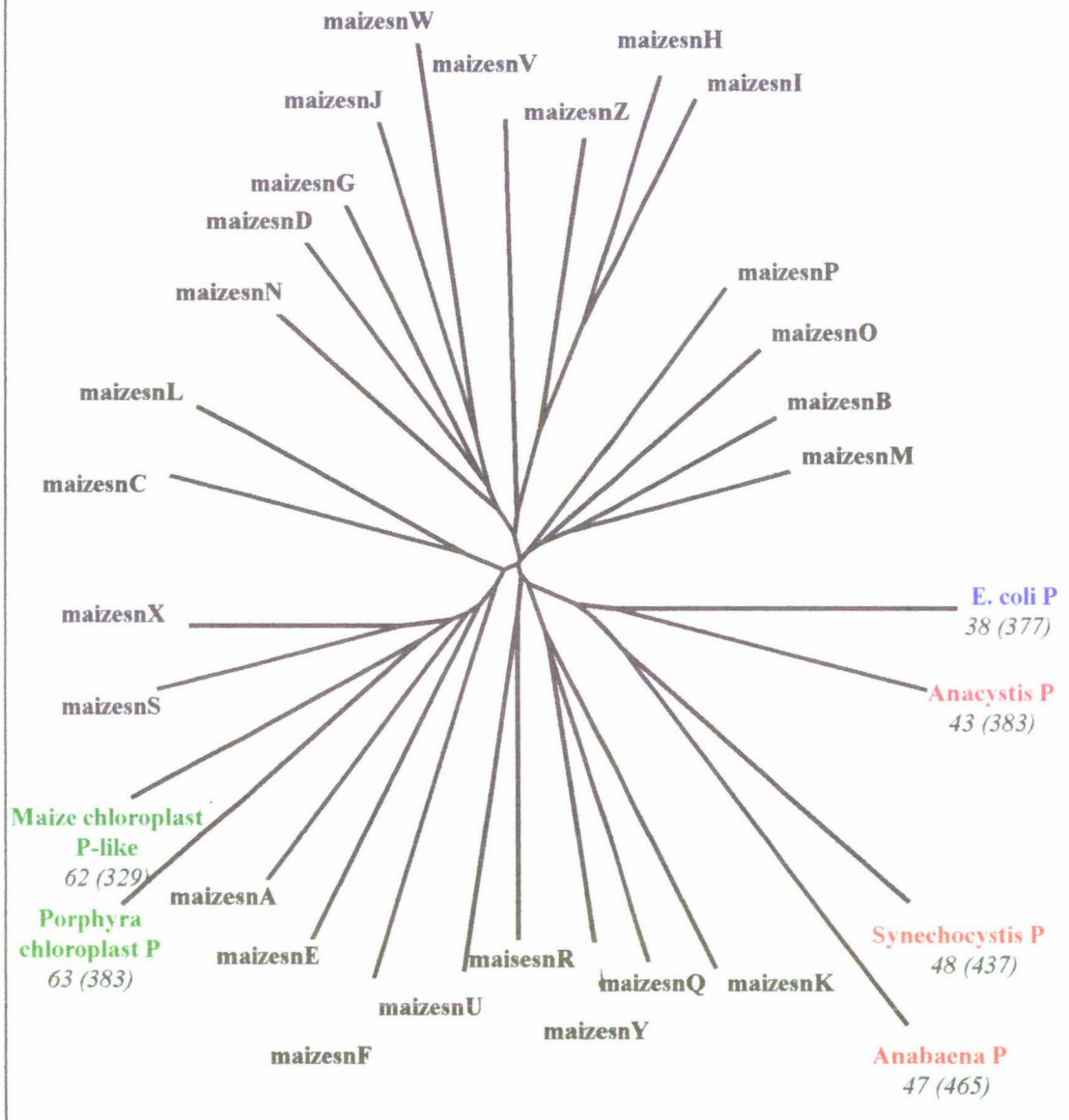


Figure 6.6: Neighbor-joining tree of pRNA sequences compared to 26 maize chloroplast pRNA shuffled sequences folded by Mfold (maizesnA - Z). The AT% is shown in italics with the sequence length in brackets. Cyanobacterial sequences are shown in red, the chloroplast sequences are in green and the *E. coli* sequence is in blue. The shuffled sequences have the same AT% and length as the maize sequence. The chloroplast sequences are grouped away from the eubacterial sequences but again the Anacystis sequence has grouped away from the other cyanobacteria.

Chapter 7

Investigation of the percentage of pairing between nucleotides in folded secondary structures.

Introduction

One aspect of the folding of pRNA and mrpRNA sequences by folding programs is the percentage of pairing present in the folded secondary structure. There was a possibility that catalytic RNAs may have a higher percentage of folding in folded structures than random sequences of the same length and AT content. Similarly, it was suspected that the amount of pairing in sequences of catalytic RNA such as pRNA and mrpRNA might be higher than that of RNA sequences that encode protein sequences. If so, then it is possible that this characteristic could be used in identifying unknown catalytic RNA sequences. This chapter examines the amount of pairing present in pRNA and mrpRNA sequences which then is compared to the amount of pairing present in random sequences. The possible relationships between AT content, length and the amount of pairing in the folded structures are examined here.

There has been little work published on the relationships between the different variables involved in the folding of RNA sequences. It has been reported that for small random sequences (50 to 100 nucleotides), the mean number of base pairs increased linearly with the sequence length, and high GC sequences had the highest amount of base pairing (Fontana et al. 1993). Since GC pairs more readily compensate destabilisation of the structure by loops than AU pairs, it was reasonable to suggest that AU sequences would form fewer base pairs, on average, than GC sequences (Fontana et al. 1993). However, it could also be possible that biological sequences with high AT contents, such as those from chloroplast and mitochondria, could have a high percentage of folding due to the increased chance of pairings of A's and T's.

Fontana et al. (1993) also calculated statistical references for the pairing of sequences. Sequence composed of an AUGC alphabet had on average 29% pairing; sequences composed of only an AU alphabet had on average 35.4% pairing; and sequences composed of only a GC alphabet had on average 40% pairing. Secondary structures folded from ß-globin mRNA show 31% pairing, and those folded from rRNA sequences found in eubacteria, 33% and mitochondria 26%. Studies of the folding of ß-globin mRNA and

rRNA also show less folding in mitochondrial sequences than is shown in the protein-coding RNA and eubacterial rRNAs (Fontana et al. 1993). These results suggest that it is possible that the AT content of a sequence has an effect on the amount of pairing in a structure. This effect could then lead to a bias in how AT rich sequences (e.g. organellar sequences such as the *S. cerevisiae* mitochondrial pRNA which is 87% AT rich) are placed on trees constructed from secondary structure data. Examining the relationship between the AT content, sequence length and the amount of pairing present in folded pRNA and mrpRNA structures, is a start to understanding how these features may be used in the future to characterise potential catalytic RNA sequences.

Materials and Methods

Random sequences were generated by shuffling some mrpRNA and pRNA sequences and eight protein-coding mRNA sequences, which were then folded with RNAfold using the program *rsnfold*.

Another program *Pairs* (also written by R. Pointon) was used to obtain the percentage of pairing contained in the structures (Appendix 4) The percentage of pairing was calculated for the mrpRNA and pRNA (Table 7.1a), and the protein-coding mRNA sequences (Table 7.1b). The amount of pairing in the shuffled sequences is shown in Table 7.2. All programs for shuffling and folding the sequences were run on a Unix system running SunOS (release 4.1.3).

Histograms of the amount of pairing in the structures constructed from shuffled pRNA, mrpRNA, and protein-coding sequences were used to show any differences in pairing between the original and shuffled sequences. Both the uncorrected and corrected folded secondary structures were plotted when necessary. If there was no 5'-3' correction necessary (i.e. no circular structure was formed) then only one indication is shown on the histograms. Graphs of the percentage of pairing in the secondary structures against the AT content and length as well as preliminary regression analysis was carried out by the Excel program of the Microsoft Office 97 package.

A:

	% pairing			
	Uncorrected	Corrected	AT%	Length
Xenopus MRP	59	*	45	277
Human MRP	53	59	36	264
Bovine MRP	59	*	39	277
Mouse MRP	63	60	36	275
Rat MRP	57	62	35	273
S. cerevisiae MRP	69	*	60	339
S. pombe MRP	70	*	57	399
Arabidopsis MRP	61	65	49	260
Zebrafish nuclear P	66	*	43	308
Human nuclear P	70	*	36	340
Mouse nuclear P	72	*	33	288
S. cerevisiae nuclear P	67	*	48	368
S. pombe nuclear P	67	*	48	373
Reclinomonas mitochondrial P	59	64	75	312
Aspergillus mitochondrial P	54	*	81	300
S. cerevisiae mitochondrial P	59	70	87	448
Porphyra chloroplast P	50	63	63	383
Maize chloroplast P-like	59	67	62	329
Rhodospirillum P	64	*	29	429
Bacillus P	61	*	51	401
Agrobacterium P	64	*	36	402
E. coli P	67	*	38	377
Synechocystis P	64	*	48	437
Anabaena P	66	*	47	465
Anacystis P	64	*	43	385

MRP
Uncorr Corr
Min 53 66
Max 79 70

Nuclear P
Uncorr Corr
Min 66 66
Max 72 72

Organellar P
Uncorr Corr
Min 50 54
Max 59 70

Eubacterial P
Uncorr Corr
Min 59 61
Max 67 67

B:

	% pairing		
	uncorrected	AT %	Length
E. coli cr1	58	59	401
Maize chloroplast rps14	52	60	314
Anabaena nifx2	60	56	404
Reclinomonas mitochondrial rsp12	48	66	395
Arabidopsis mitochondrial nad41	48	64	302
Bacillus nasBD	51	54	320
Porphyra chloroplast apcD	59	68	501
Porphyra chloroplast rp121	45	60	314

Protein RNA
Uncorr
Min 45
Max 60

Table 7.1: The percentage of pairing in the structure folded by RNAfold for both the uncorrected and corrected structure (* indicates that there was no correction necessary for this structure), AT contents and lengths of **A**: mrpRNA and pRNA **B**: mRNA for protein sequences used in this chapter.

Random Structures (RNAfold)			% pairing		
	AT%	Length	Min	Max	Mean
pRNA and mrpRNA					
E. coli P	38	377	55	66	60.36
Porphyra chloroplast P	63	383	48	64	56.93
Maize chloroplast P-like	62	329	53	68	59.42
Arabidopsis MRP	49	260	55	65	59.58
Human MRP	36	264	50	64	58.48
S. cerevisiae MRP	60	339	56	70	62.78
S. cerevisiae nuclear P	48	368	53	67	59.09
Human nuclear P	36	340	55	70	63.5
Zebrafish nuclear P	43	308	48	64	56.93
S. cerevisiae mitochondrial P	87	448	52	69	62.17
Protein-coding mRNA					
Maize chloroplast rps14	60	314	48	64	56.93
Porphyra chloroplast apcD	68	501	55	65	59.58
Porphyra chloroplast rp121	60	314	44	67	56.22
Arabidopsis mitochondrial nad41	64	302	53	67	59.09
Bacillus nasBD	54	320	55	66	60.41
Reclinomonas mitochondrial rsp12	66	395	56	70	62.78
E. coli cr1	59	401	52	68	59
Anabaena nifx2	56	404	56	71	64.18

Table 7.2: The percentage of pairing in the structures folded from the shuffled sequences.

Results

The study by Fontana et al. 1993 indicated that a relationship between the amount of pairing in a structure and AT content (and/or sequence length) was possible. This can be tested by comparing the amount of pairing within a structure with the pairing of structures of shuffled sequences (which will have the same length and AT content). If the amount of pairing in a structure is similar to that of the shuffled sequences then it can be concluded that either AT content and/or length is having an effect on the amount of pairing. A consistent higher or lower amount of pairing indicates that the amount of pairing is not greatly affected by AT content and/or length.

The uncorrected RNAfold secondary structures from the eubacterial and organellar pRNA generally showed a low amount of pairing compared to that of the structures generated from the shuffled sequences (Figures 7.1 and 7.2). When the structures were corrected for circular structure, the amount of pairing was in all cases higher than that of the uncorrected structure. The *E. coli* pRNA (which did not require any correction) had pairing that fell in the middle of the range of its random sequences, a feature that was also

shown with the *Reclinomonas* mitochondrial pRNA. However the *S. cerevisiae* mitochondrial pRNA (Figure 7.1), the *Porphyra* chloroplast pRNA and the maize chloroplast pRNA (Figure 7.2) showed a high amount of pairing for their corrected structures when compared to the random sequences. The three eukaryotic nuclear pRNA sequences show a higher percentage of pairing than the shuffled sequences (Figure 7.3). None of the pRNA RNAfold structures show any circular structure and did not require any corrections. This was not so with the mrpRNA sequences (Figure 7.4) as both the human and the *Arabidopsis* mrpRNA had circular structure. The corrected *Arabidopsis* mrpRNA and the *S. cerevisiae* mrpRNA (which did not require any correction), also have a higher percentage of folding than is shown in structures constructed from their shuffled sequences. The corrected human mrpRNA structure showed a higher amount of pairing than the uncorrected structure but was only in the middle of the range of the pairing shown by its random structures.

The amount of pairing in the protein-coding mRNA sequences on the other hand generally showed less pairing than that of the structures formed from random sequences (Figures 7.5 and 7.6). The two exceptions to this were the maize chloroplast ribosomal protein *A14* mRNA (shown in Figure 7.5) and the *E. coli* cr1 protein mRNA (shown in Figure 7.6). None of the protein-coding mRNA RNAfold structures form 5' - 3' paired structures, but they do not form what has been classed here as circular structures (all protein-coding mRNA structures are shown in Appendix 1). In these protein-coding mRNA structures, short-range pairing generally forms the helices, but in the middle of the sequence there is a long stretch of pairing which is different from what is seen in the circular structures. There was nothing noticeably different about the maize chloroplast ribosomal A14 protein mRNA and the *E. coli* cr1protein mRNA, when compared to the other protein RNAfold structures.

Generally there is an indication that for pRNA and mrpRNA RNAfold structures have a higher percentage of folding shown than that calculated from their shuffled sequences (thus had the same length and AT content). However, for the protein coding mRNA sequences examined here, the amount of pairing is generally lower. This is possibly a feature by which putative pRNA and mrpRNA sequences can be characterised. It is somewhat comforting to note that the maize pRNA sequence has the same pattern as is seen with other pRNA sequences.

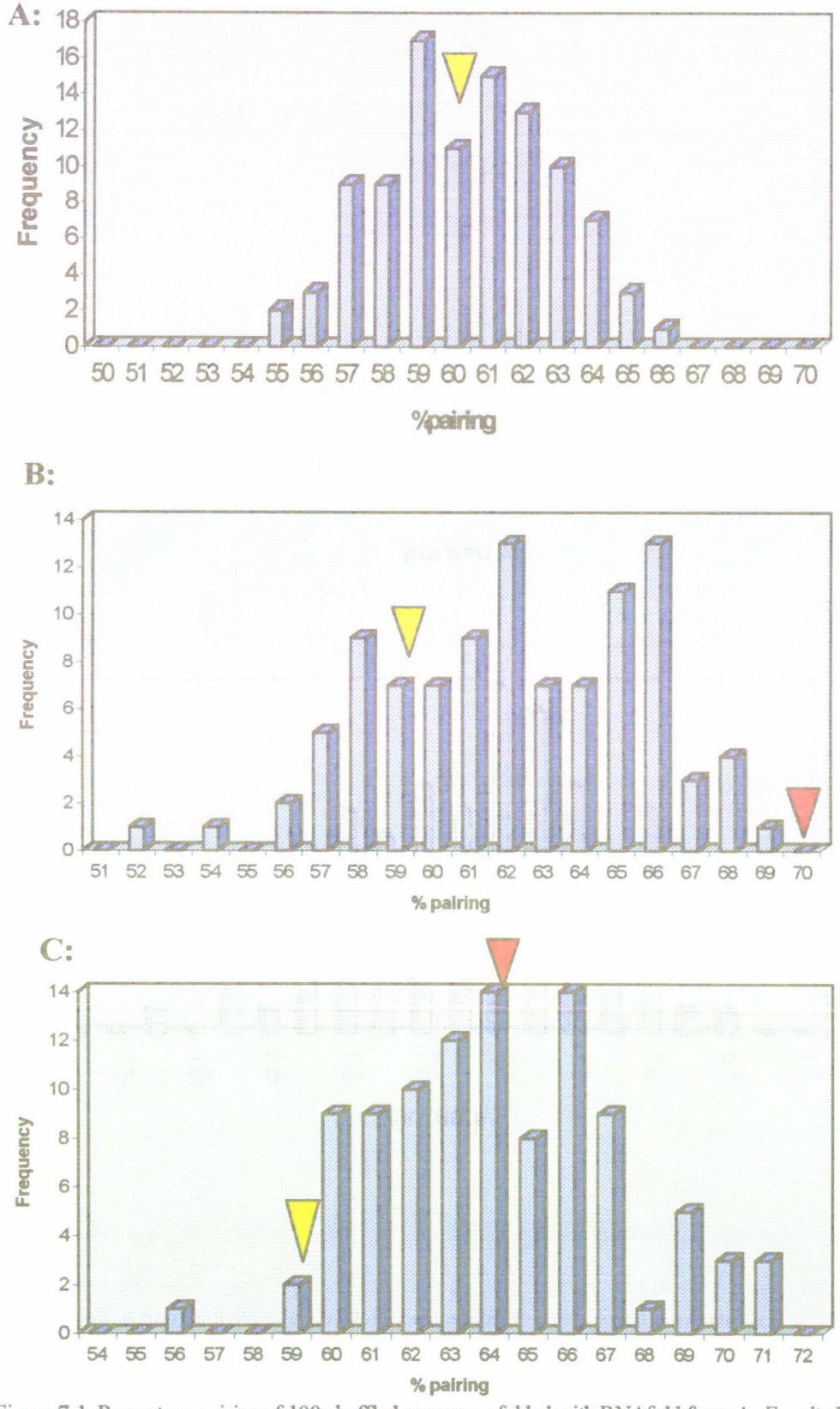


Figure 7.1: Percentage pairing of 100 shuffled sequences folded with RNAfold from **A**: *E. coli* pRNA (no correction was required), **B**: *S. cerevisiae* mitochondrial pRNA and **C**: *Reclinomonas* mitochondrial pRNA. ▽ represents the % pairing of the uncorrected RNAfold structure. ▽ represents the % pairing of the corrected RNAfold structure.

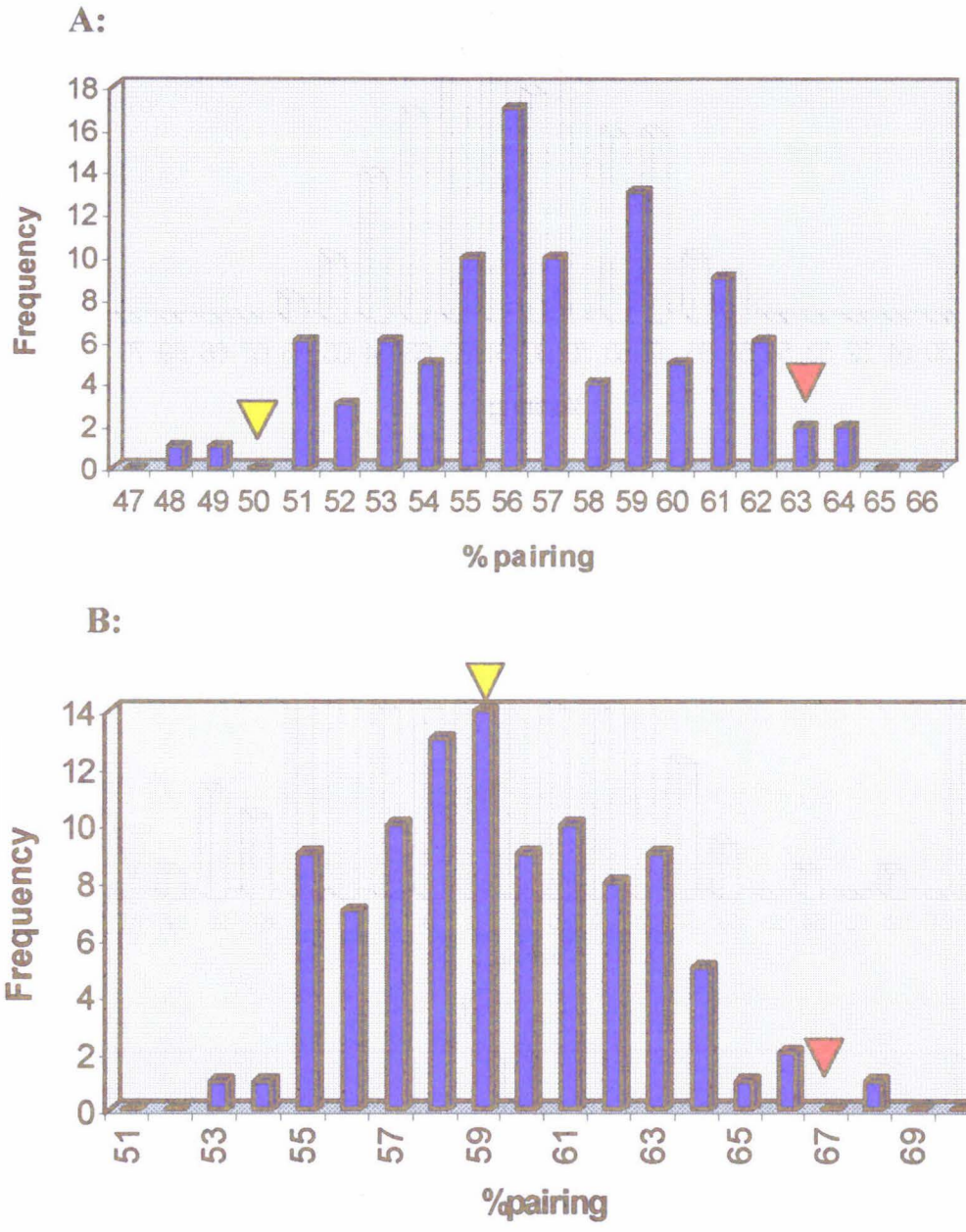


Figure 7.2: Percentage pairing of 100 shuffled sequences folded with RNAfold from **A**: *Porphyra* chloroplast pRNA. **B:** the maize chloroplast pRNA. ▽ represents the % pairing of the uncorrected RNAfold structure. ▼ represents the % pairing of the corrected RNAfold structure

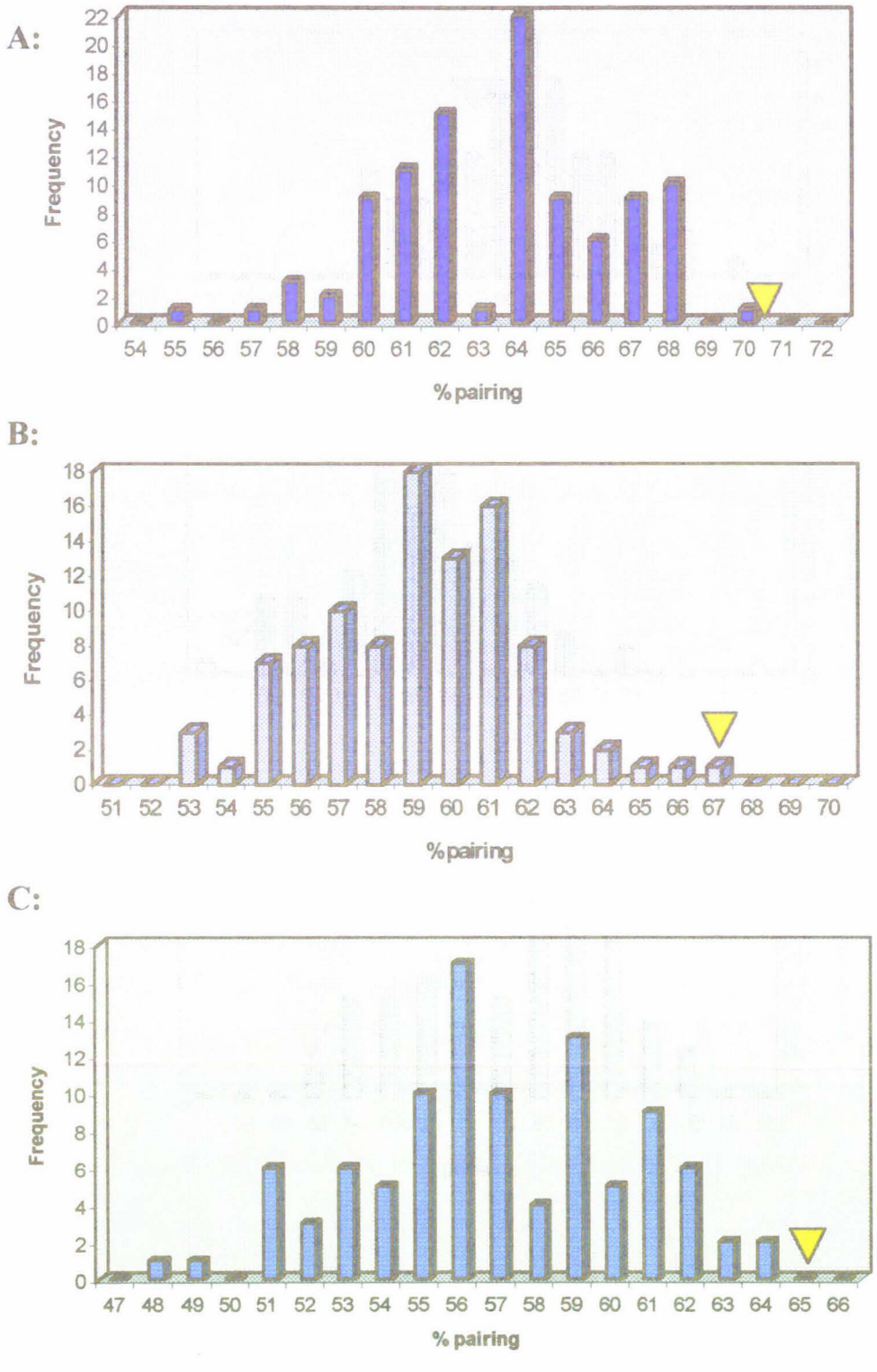
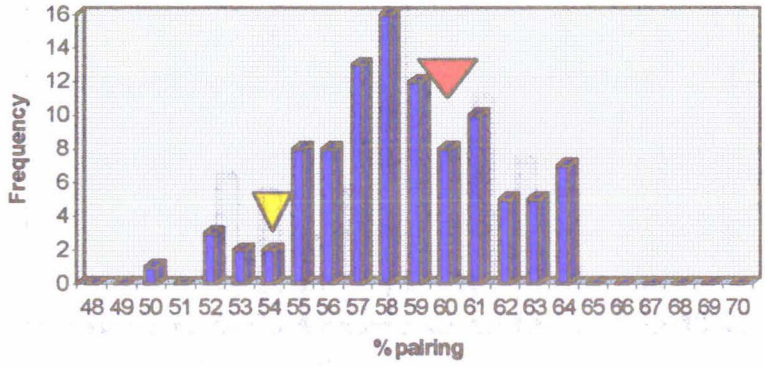
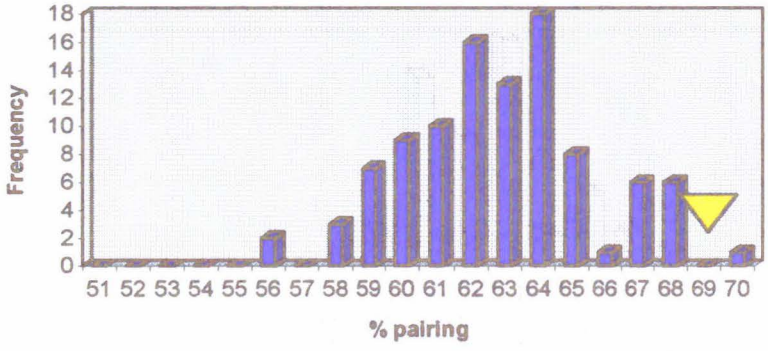


Figure 7.3: Percentage pairing of 100 shuffled sequences folded with RNAfold from **A**: Human nuclear pRNA. **B:** S. cerevisiae nuclear pRNA. **C**: Zebrafish nuclear pRNA. ▼represents the % pairing of the RNAfold structure (no corrections were required for these sequences.

A:



B:



C:

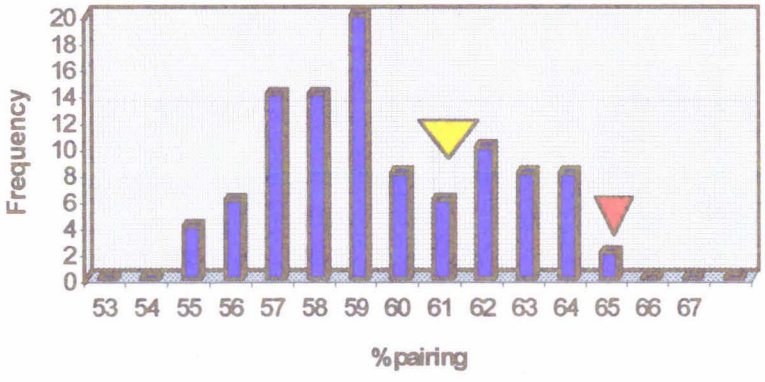


Figure 7.4: Percentage pairing of 100 shuffled sequences folded with RNAfold from **A**: Human mrpRNA. **B**: S. cerevisiae mrpRNA (no correction was required) and **C**: Arabidopsis mrpRNA. ▽ represents the % pairing of the uncorrected RNAfold structure. ▽ represents the % pairing of the corrected RNAfold structure.

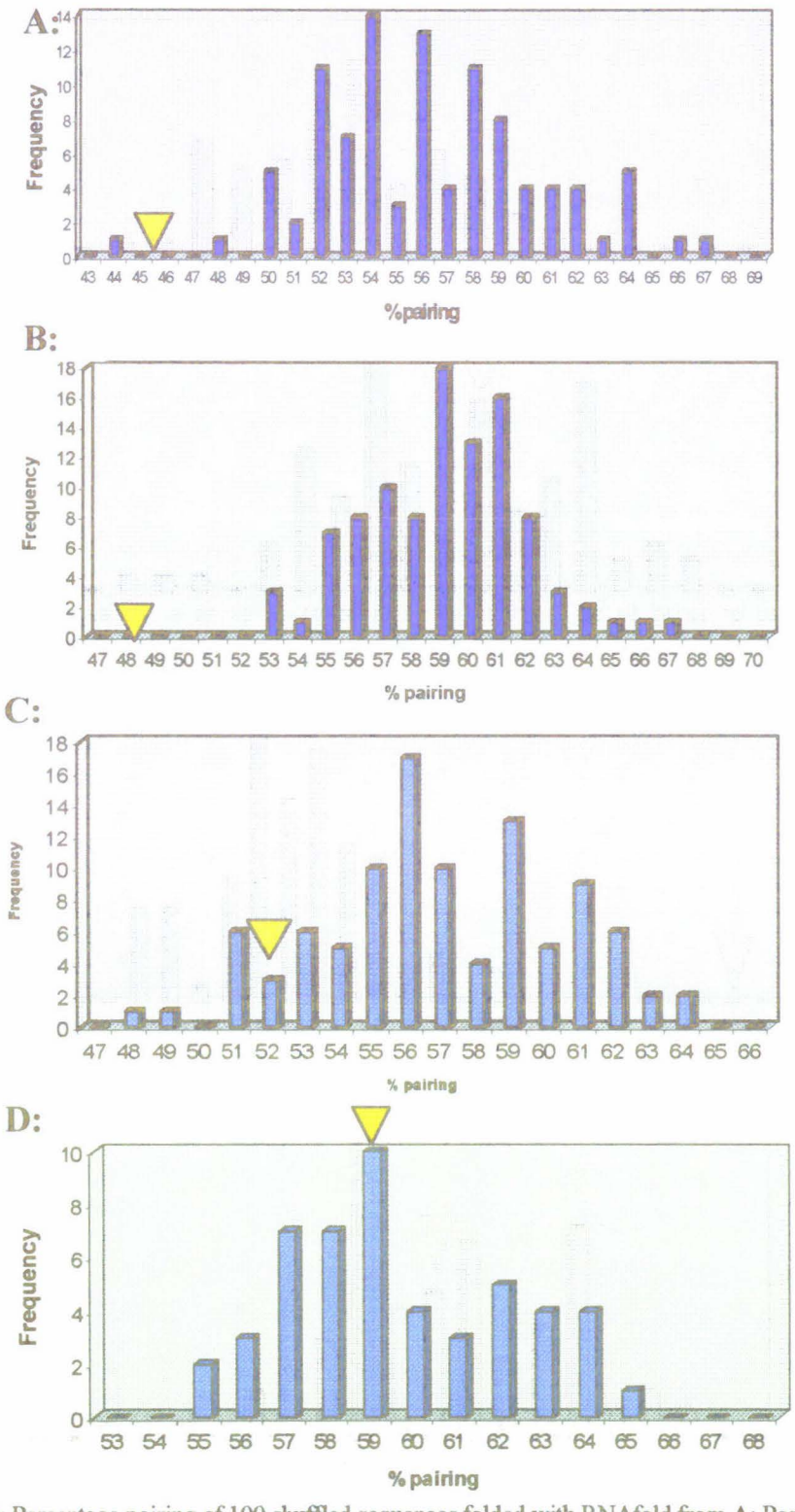


Figure 7.5: Percentage pairing of 100 shuffled sequences folded with RNAfold from **A**: Porphyra chloroplast 50S ribosomal protein *L21*, **B**: Arabidopsis mitochondrial NADH dehydrogenase subunit 4L *nad41*, **C**: maize chloroplast ribosomal protein A14 *rps14* and **D**: *Porphyra* chloroplast allophycocyanin gamma chain protein *apcD*. ▽ represents the % pairing of the RNAfold structure. No corrections were calculated for these structures.

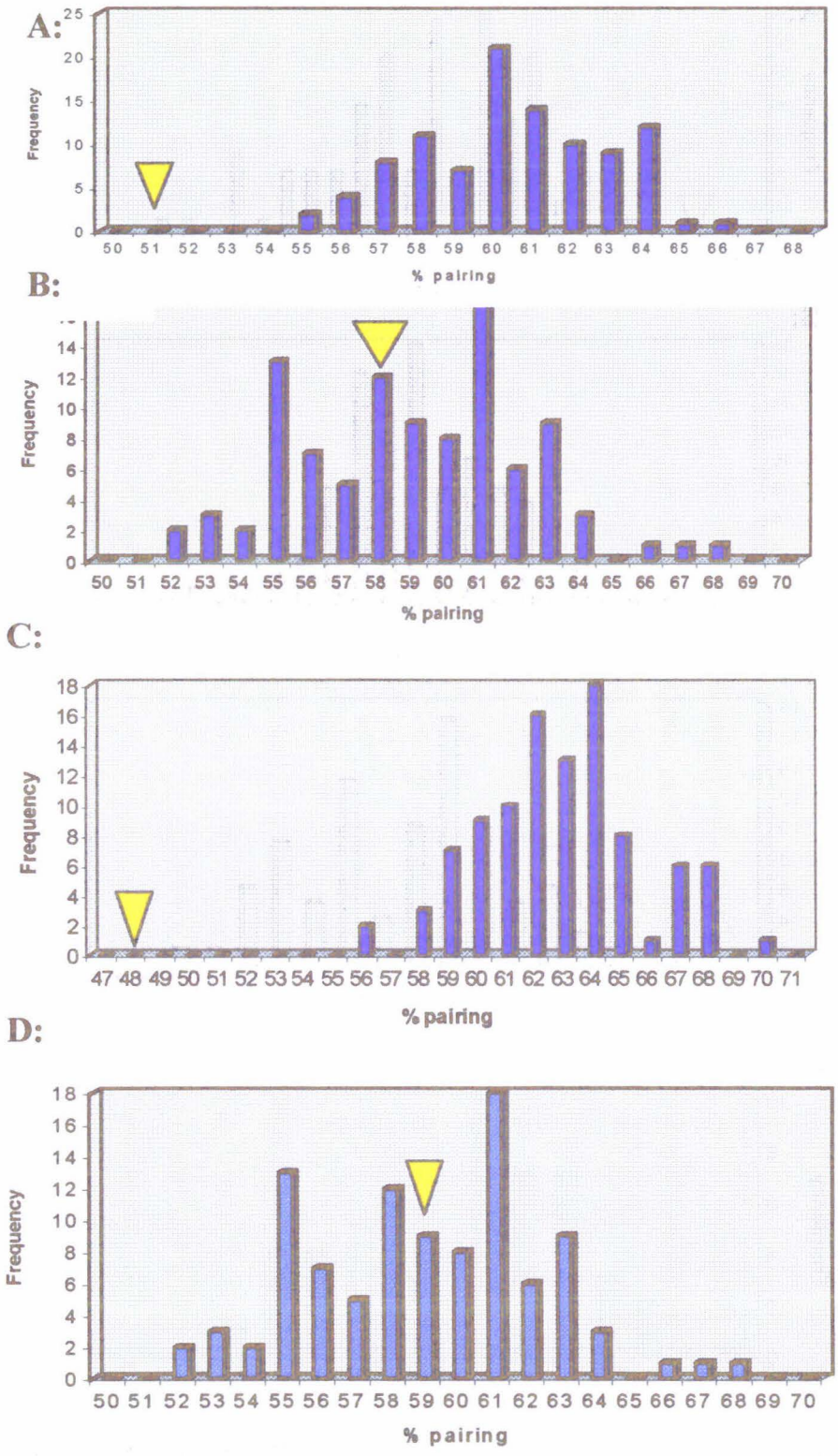


Figure 7.6: Percentage pairing of 100 shuffled sequences folded with RNAfold from **A**: Bacillus nitrite reductase subunit *nasBD*. **B**: E. coli *cr1* protein. **C**: Reclinomonas mitochondrial ribosomal protein S12 *rps12*, and **D**: *Anabaena sp*. Nitrogen fixation protein *nifX2*. ▽ represents the % pairing of the RNAfold structure. No corrections were calculated for these structures.

A scatter-plot of the amount of pairing against the AT content for the pRNA, mrpRNA and the protein-coding RNA sequences (Figure 7.7) shows some distinct grouping of the different RNAs. However, there is no noticeable trend within these groupings supporting an effect of the AT content on the amount of pairing. A scatter-plot of the amount of pairing against the length of the pRNA, mrpRNA and the protein-coding mRNA sequences (Figure 7.8) shows no definitive trend within the RNA groups, supporting a relationship between the length of sequence and the amount of pairing.

Figure 7.7: Scatter plot of the % pairing against % AT for the RNAfold secondary structures for ◆ mrpRNA, □ eukaryotic pRNA, △ organellar and eubacterial pRNA and ● protein coding mRNA sequences.

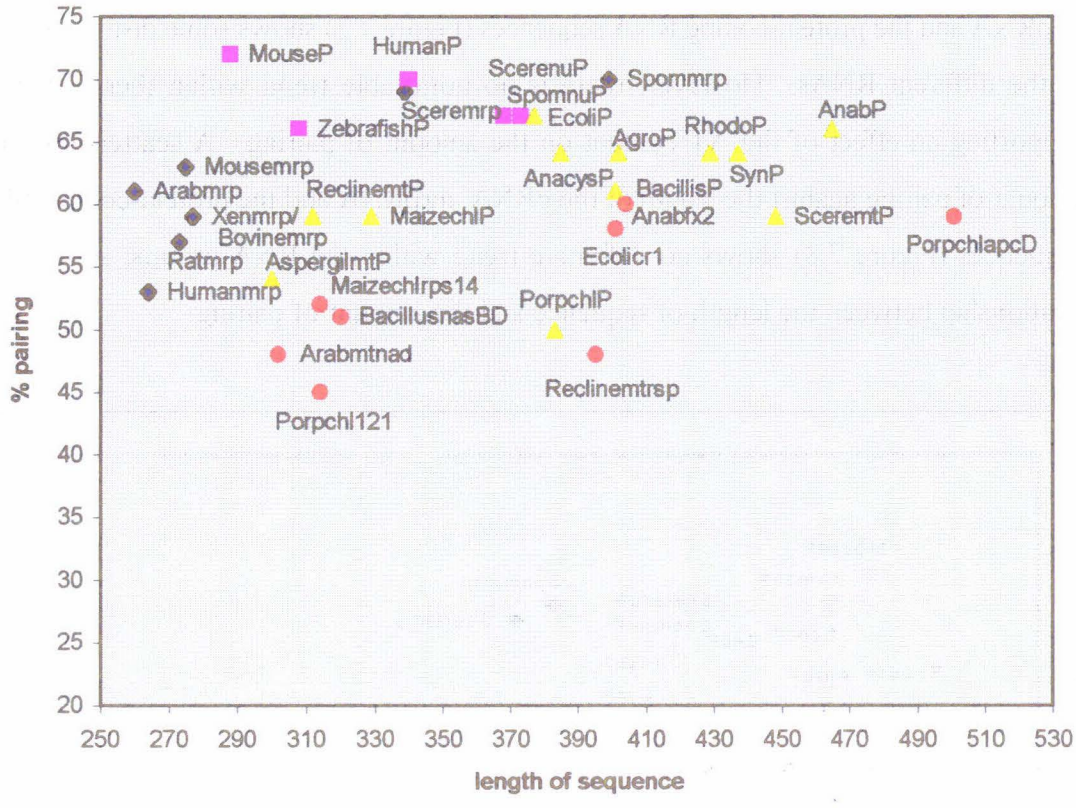


Figure 7.8: Scatter plot of the % pairing against length for the RNAfold secondary structures for ◆ mrpRNA, ■ eukaryotic pRNA, △ organellar and eubacterial pRNA and ● protein coding RNA sequences.

Plots of the minimum, maximum and mean % pairing for the 100 random structures against the AT content (Figure 7.9) and the sequence length (Figure 7.10) for each sequence also show no distinctive trends towards any relationship between these variables. There are also no obvious groupings of the mrpRNA, pRNA and protein-coding mRNA structures folded from shuffled sequences on this graph.

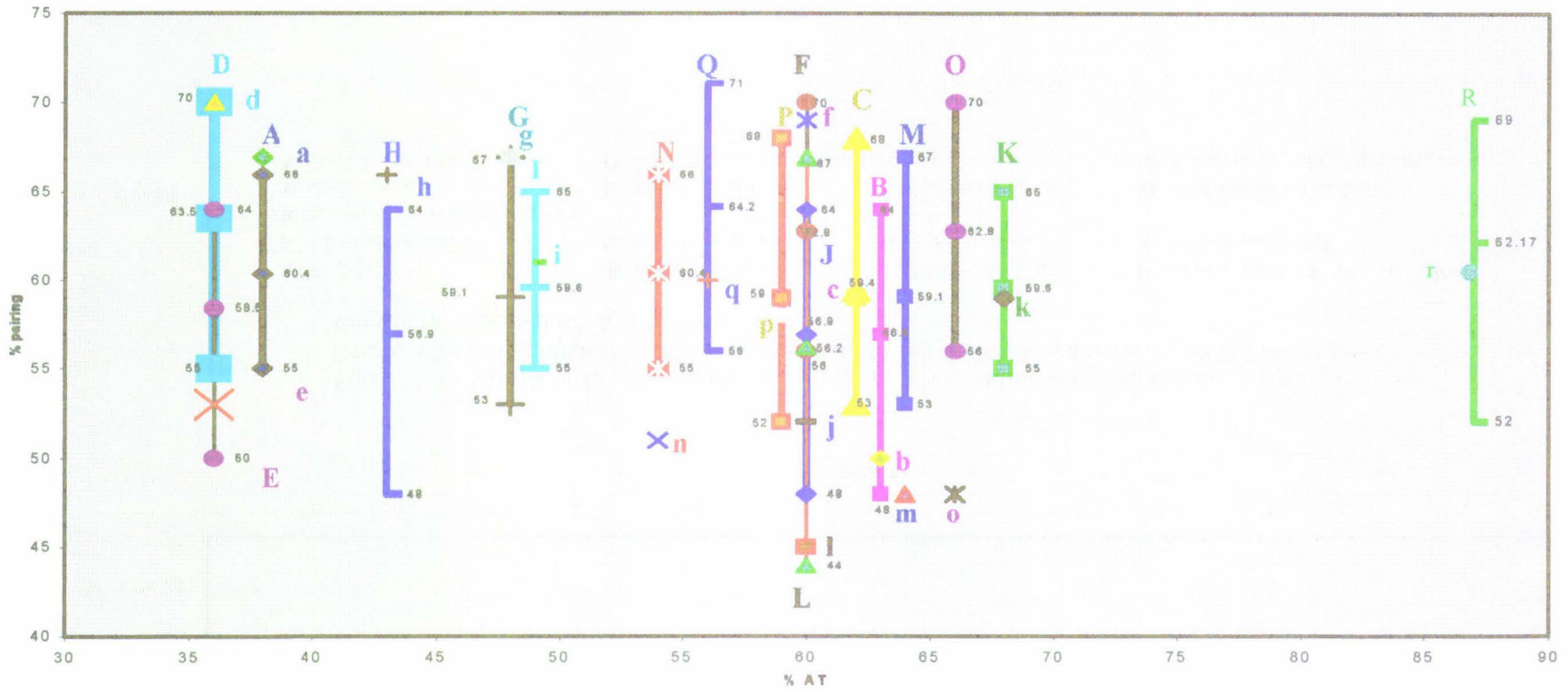


Figure 7.9: Graph of AT% against % pairing for RNAfold random sequences. Upper case letters refer to the random structures constructed from an RNA sequence. Lower case letters refer to the original RNA RNAfold structure.

A: E. coli P **B**: Porphyra chl P **C**: Maize chl P-like **D**: Human nu P **E**: Human MRP
F: S.cerevisiae MRP **G** :S.cerevisiae nu P **H**: Zebrafish nu P **I**: Arabidopsis MRP
R: S.cerevisiae mitochondrial P
J: Maize chl rps14 **K**: Porphyra chl apcD **L**: Porphyra chl rp21 **M**: Arabidopis mt nad41
N: Bacillus nasBD **O**: Reclinomonas mitochondrial rsp12 **P**: E.coli cr1
Q: Anabaena nifx2

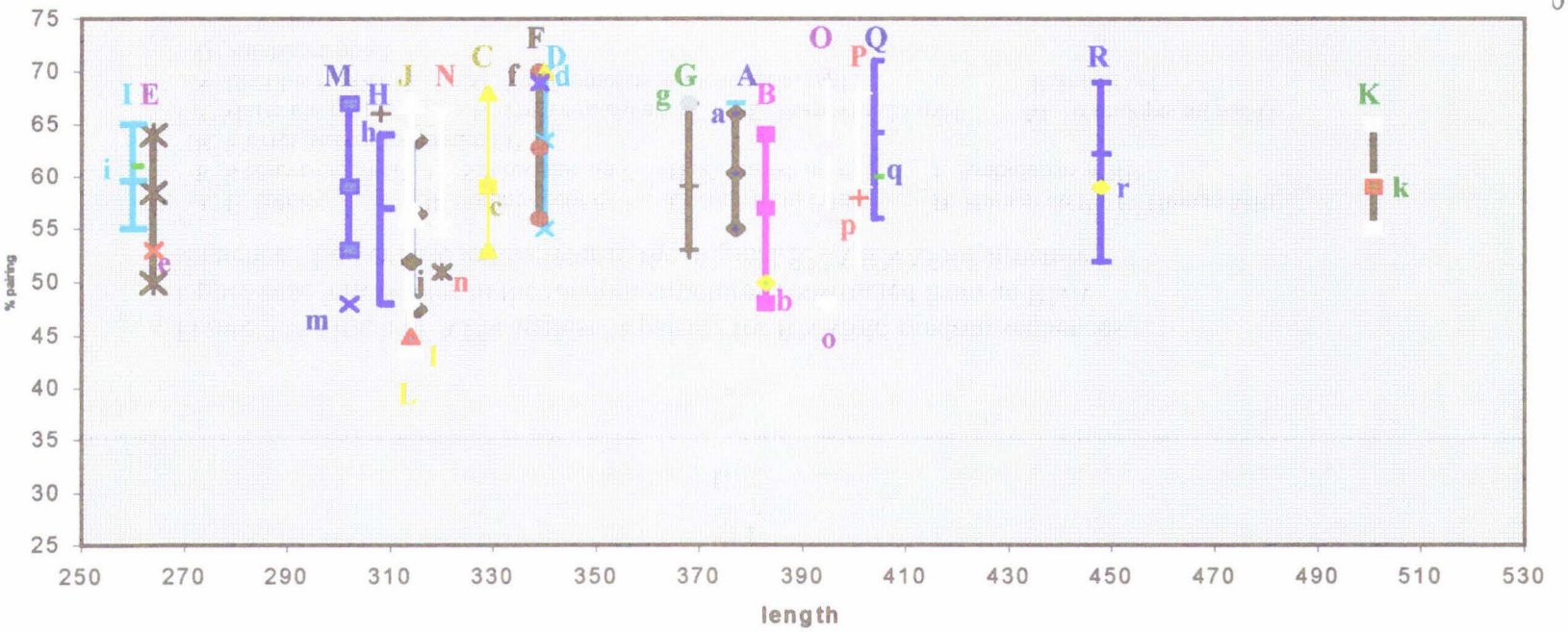


Figure 7.10: Graph of length against % pairing for RNAfold random sequences. Upper case letters refer to the random structures constructed from an RNA sequence. Lower case letters refer to the original RNA RNAfold structure.

A: E. coli P
B: Porphyra chl P
C: Maize chl P-like
D: Human nu P
E: Human MRP
F: S.cerevisiae MRP
G: S.cerevisiae nu P
H: Zebrafish nu P
I: Arabidopsis MRP
R: S.cerevisiae mitochondrial P
J: Maize chl rps14
K: Porphyra chl apcD
L: Porphyra chl rp21
M: Arabidopis mt nad41
N: Bacillus nasBD
O: Reclinomonas mitochondrial rsp12
P: E.coli cr1
Q: Anabaena nifx2

Regression analysis on the AT content against the percentage of pairing (Table 7.3) and on the sequence length against the percentage of pairing (Table 7.4) support the lack of trend shown in the scatter-plots above. In both cases, a good fit with any trend drawn for the data would show an r-value (a measure of the goodness of fit between the data points and the regression trend calculated) between 0.5 and 1.0. Alternatively an R^2 value can be used but this gives less information about any calculated trend. The R^2 value is calculated directly from the trend line fitted to the data points on the graphs above (Figures 7.7 and 7.8), there is no r-value returned using this function of Excel). There is little evidence to support any linear equation relating the amount of pairing to the AT content (Table 7.3). The only significant results ($R^2 > 0.8$) come from fitting a polynomial trend equation with an order of at least five to the protein-coding mRNA random and original sequence. Since there are only eight data points in this data set, it is unlikely that this equation has any real meaning.

When the amount of pairing is compared to the length of the sequence, there are slightly higher R^2 values than was seen when compared to the AT contents. Significant results are found with both the mrpRNA and pRNA sequences and the random sequences, but these again only appear with the polynomial equation with an order of at least four. This suggests that any relationship between the amount of pairing and either AT content or sequence length is a very subtle one with the RNAfold program.

Discussion

The amount of pairing found in this study is much higher than that found by Fontana et al. (1993) (mentioned in the introduction above). However, it is unclear what folding program was used to compute their data. This is an indication that it is important to calculate individual statistical references for different folding programs.

The analysis of mrpRNA and pRNA sequences against random sequences of the same AT content and length is consistent, with the amount of folding in these catalytic RNAs higher than non-catalytic random sequences.

	equation	R^2 (goodness of fit)							
%AT against %pairing		MRP and P			Protein-coding mRNA		All sequences		
		Random sequences	Uncorr	Corr	Random sequences	Uncorr	Random sequences	Uncorr	Corr
Linear	$y = ax + b$	0.0271	0.1030	0.1267	0.2525	0.0112	0.0086	0.1673	0.0428
Logarithmic	$y = a \ln(x) + b$	0.0114	0.0995	0.1138	0.2630	0.0136	0.0180	0.1949	0.0721
Polynomial order = 2	$y = ax^2 + bx + c$	0.1647	0.1039	0.1431	0.3564	0.1312	0.1049	0.2527	0.2662
Polynomial order = 3	$y = ax^3 + bx^2 + cx + d$	0.1915	0.2337	0.1765	0.4977	0.6509	0.1484	0.3295	0.3339
Polynomial order = 4	$y = ax^4 + bx^3 + cx^2 + dx + e$	0.5241	0.2613	0.1858	0.5363	0.7297	0.2826	0.3604	0.3421
Polynomial order = 5	$y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$	0.7251	0.5935	0.3651	0.9307 *	0.7448	0.3338	0.3609	0.3479
Polynomial order = 6	$y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$	0.7307	0.6096[a]	0.3677	0.9967 *	0.8925 *	0.4359	0.3825	0.3504
Power	$y = ax^b$	0.0114	0.091	0.1152	0.2615	0.0147	0.0192	0.1819	0.0755
Exponential	$y = a (e)^b$	0.0272	0.094	0.1260	0.2513	0.0121	0.0095	0.1548	0.0464

Table 7.3: Regression analysis of AT% against % pairing. The R^2 value was calculated from a trend line fitted to the data and indicates how well the trend equation fits the data (an r-value was not returned). A good fit has a value close to 1.0.

[a] indicates that the equation allows for a negative percent of pairing.

* indicates a significant result.

length against %pairing	equation	R^2 (goodness of fit)							
		MRP and P			Protein-coding mRNA		All sequences		
		Random sequences	Uncorr	Corr	Random sequences	Uncorr	Random sequences	Uncorr	Corr
Linear	$y = ax + b$	0.0856	0.0017	0.3124	0.0352	0.5245	0.0251	0.0143	0.0221
Logarithmic	$y = a \ln(x) + b$	0.0841	0.0074	0.3363	0.0339	0.5376	0.0249	0.0168	0.0226
Polynomial order = 2	$y = ax^2 + bx + c$	0.0866	0.2083	0.3593	0.0360	0.5467	0.0252	0.0252	0.0247
Polynomial order = 3	$y = ax^3 + bx^2 + cx + d$	0.1562	0.3030	0.5575	0.1025	0.5583	0.0276	0.0254	0.0610
Polynomial order = 4	$y = ax^4 + bx^3 + cx^2 + dx + e$	0.7836	0.4765	0.8145	0.7029	0.8302	0.0308	0.1055	0.0893
Polynomial order = 5	$y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$	0.7916	0.5271	0.8366*	0.9960*	0.8698*	0.2439	0.3680	0.5215
Polynomial order = 6	$y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$	0.8089*	0.7944[a]	0.9399*	0.9967*	0.8925[a]	0.2594	0.3700	0.5481
Power	$y = ax^b$	0.0827	0.0063	0.3348	0.0300	0.5239	0.0219	0.0201	0.0242
Exponential	$y = a (e)^b$	0.0845	0.0014	0.3100	0.0318	0.5110	0.0225	0.1810	0.0243

Table 7.4: Regression analysis of length against % pairing. The R^2 value indicates how well the trend equation fits the data. A good fit has a value close to 1.0.
[a] indicates that the equation allows for a negative percent of pairing.
indicates a significant result.

The low amount of folding in the protein-coding sequences, when compared to their shuffled sequences is very different from the pRNA and mrpRNA sequences. This does suggest that this is an identifying feature of these catalytic RNA sequences. It is interesting to note that the maize chloroplast pRNA sequence (Figure7.2) follows the same pattern as the pRNA and mrpRNA sequences. This is another piece of evidence for the maize chloroplast pRNA sequence to be an actual pRNAsequence.

Many more catalytic and non-catalytic (protein-coding and random sequences) will need to be tested before this could be considered a standard characteristic of catalytic RNA. If so then it is possible that a simple test for pairing characteristics could be used in searching procedures for new catalytic RNA sequences, especially when highly divergent sequences are being compared.

Overall the data suggests that the AT content and the length of the RNA sequences have little effect on the amount of pairing present in pRNA, mrpRNA and protein-coding mRNA when using the RNAfold program. However, this does not rule out any other effects that AT content or length may have on the formation of RNA secondary structure, or any effects on the amount of pairing in other -folding programs. Initial results with the Mfold program (results not shown) indicated that the Mfold program had the same characteristics as the RNAfold program in regards to the amount of pairing, AT content and sequence length. The lack of any relationship found between the amount of pairing that a folded structure has, and either the AT content or length lends more credibility to the groupings formed by the analysis of folded structures in the previous chapters of this thesis.

Chapter 8

Conclusions and future considerations

This thesis looked at four main issues; the evolution of MRP and its relationship to P, the use of biological secondary structure in determining evolutionary relationships, the evaluation of the structural output from folding programs and the use of secondary structure in the characterisation of putative pRNA sequences from chloroplasts. Conclusions and issues relating to future work in these areas are discussed below.

Evolution of mrpRNA

MRP has become a very interesting molecule. If it was a part of the RNA world then it is interesting that it has only been retained in the eukaryotic lineage. If, on the other hand, MRP evolved only in eukaryotes then it is interesting that it goes against the general rule that proteins took over the catalytic role from ribonucleoproteins as protein synthesis evolved.

A major problem in looking at the evolution of MRP is that there have been only eight species characterised. The relationship between MRP and P is not likely to be completely resolved until more eukaryotic P and MRP, (both the protein and RNA moieties) have been characterised. Especially helpful will be the characterisation of pRNA and possibly mrpRNA from lower eukaryotes such as *Giardia* (presently underway, M. Sogin Personal Communication) and *Entamoeba*. This will indicate whether the evolution of mrpRNA extends back further than the divergence of yeasts, plants and animals.

The low homology between the eubacterial and eukaryotic pRNA sequences is an indication of how a catalytic RNA sequence can change while retaining the common function. It is possible that the eukaryotic P has functions additional to those shown by the prokaryotic P, which may account for some of the structural differences between the two. With such a difference in sequence and structural homology between the eukaryotic and eubacterial pRNA, it is not surprising to see a similar lack of homology between the eukaryotic pRNA and mrpRNA. If these two molecules are related, then differences in their functions are also likely to account for the lack of homology shown here. Future functional studies of MRP may reveal the full extent of any complex associations between MRP and P, and may reveal a greater range of activities within the cell for these two molecules.

Comparison of biological secondary structures.

Comparison of mrpRNA and pRNA biological secondary structures have shown that a quantitative analysis of these structures can be used in determining evolutionary relationships between catalytic RNA. Differences in the biological secondary structure due to the association of a single protein (prokaryotic pRNA) or multiple proteins (eukaryotic pRNA and mrpRNA) made little difference with the ability of this data to produce tree-like data.

There is currently, a logistical problem associated with using RNAdistance to compare many biological secondary structures. The input of the structures into this program is required to be in bracket notation, but at present, the conversion of the biological secondary structures to bracket notation can only be done manually. A sophisticated program would be required to be able to scan in the structure and then convert to the bracket or any other notation. Alternatively, the RNAdistance program could be modified to give the required output format.

Structures of different lengths (e.g. rRNA or introns) may also require different parameter settings from those used here. A quantitative comparison between trees constructed from sequences, and the structural data available for the 16S rRNA, could be very informative. The length of the 16S rRNA genes (approx 1500 nucleotides) precluded this comparison in this study. Using 16S rRNA secondary structure data could enable comparisons of closely related species and subspecies to determine as to what level secondary structure data is useful.

Trees constructed from the biological secondary structure data are not identical to the control subtrees constructed from 16S or 18S rRNA. For the comparison of secondary structure, this study used only the default settings of the RNAdistance program, which may account for some of the minor differences. Different matrices and conditions may be required to fine-tune the use of this program for comparing different groups of secondary structures.

When aligning catalytic RNA sequences it should be noted that a change in the binding of a nucleotide might change the presence or absence of a helix and have a large effect on the secondary structure. For this reason it is advisable to compare trees constructed from secondary structure data to those constructed from sequence data, this will show up any similarity in sequence that is not accompanied by a similarity in structure. Alignment programs (such as 'Divide and Conquer' and 'Dialign') which

allow for individual sections of the sequences to be compared may be very useful in the future for aligning sequences with considering secondary structure.

Folding algorithms.

Folding programs produce sub-optimal structures as well as the optimal structure based on the energy required for folding. This study has used only the optimal structures except with the *S. cerevisiae* nuclear pRNA folded with RNAstructure (the suboptimal structure #4 was used instead). Further study may be required to understand more fully the distances between the sub-optimal structures and the optimal structures, and how these could be related to evolutionary studies.

The largest problem found here with using thermodynamically based folding programs is the inconsistency in the structure within the pRNA and mrpRNA groups of folded structures. The formation of circular structure caused a bias in the results that could not be completely overcome by forcing the pairing of the 5' and 3' ends. In order for folded structural characteristics to be useful in a searching strategy, this consistency problem needs to be overcome. A 'search and fold' searching algorithm could be developed, but may require a great amount of study on the folding parameters used in the folding programs. It may also be possible that current thermodynamic folding programs are not suitable for this purpose and another type of folding program may be required.

The identification of coarse structural motifs (eg the fork and the can opener found in pRNA) may be a very useful characterisation technique. It would be interesting to see if the motifs found in the pRNA occur in any other catalytic RNA molecules.

An interesting feature within the thermodynamically folded structures has been the amount of nucleotide pairing shown for both the catalytic RNA and protein coding mRNA sequences in this study. When the amount of pairing within a structure is compared to that shown in structures constructed from random sequences of the same length and AT content, this feature becomes a characteristic by which prospective catalytic RNA molecules may be identified. Further development is required but use of this characteristic in searching and identification procedures looks promising.

The putative maize chloroplast pRNA.

The identification of the putative green plant chloroplast pRNA sequences shows that organellar genomes may harbour more catalytic RNA genes than originally thought. The high AT content does make finding and characterising putative catalytic RNA genes difficult. Searching individual genomes rather than whole databases seems to be more effective in these cases. The Ssearch program used here was a searching tool developed for searching small libraries of sequences. Faster programs such as BLAST (Altschul et al. 1990) have been used for searching within large genome databases (eg. with the *S. cerevisiae* genome database) and could possibly be adapted for small genomes.

The search for other catalytic RNA genes may also extend to nuclear eukaryotic genomes. It has been shown that higher vertebrates may express multiple isoforms of pRNA with mice having three genes homologous to the mouse mrpRNA gene (Li and Williams 1995). This suggests that other eukaryotes may have multiple genes for mrpRNA and possibly pRNA. The sequencing of more eukaryotic genomes will allow this to be examined and may uncover more information about MRP evolution.

The closer examination of the putative maize chloroplast pRNA gene shows that it does have many characteristics that are also shown in known pRNA genes. Structures folded with folding programs from this sequence tend to group with other organellar pRNA genes, even when compared to sequences of the same length and AT content. The "acid test" for this sequence is whether it has the catalytic properties expected in a pRNA molecule determinable only by biochemical analysis.

Final Remarks

The use of folded secondary structures in evolutionary and characterisation studies is a relatively new field of research. The increasing investigation of the RNA world will require more tools for the analysis of catalytic RNA molecules especially in cases where conventional sequence alignments are unreliable. RNA secondary structure analysis still requires some development to be reliably used in evolutionary studies but results here are encouraging.

References

Abagyan, R. and Batalov, S. (1997) Do aligned sequences share the same fold. J. Mol. Biol. **273**:355-368.

Altman, S., Wesolowski, D., and Puranam, R. S. (1993) Nucleotide sequences of the RNA subunit of RNase P from several mammals. Genomics **18**:418-422.

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

Arends, S. and Schön, A. (1997) Partial purification and characterisation of nuclear ribonuclease P from wheat. Eur. J. Biochem. **244**:635-645.

Bandelt, H-J. and Dress, A. (1992) A new and useful approach to phylogenetic analysis of distance data. Molecular Phylogenetics and Evolution **1**:242-252.

Baum, M. and Schön, A. (1996) Localization and expression of the closely linked cyanelle genes for RNase P RNA and two transfer RNAs. FEBS letters **382**:60-64.

Baum, M,, Cordier, A. and Schön, A. (1996) RNase P from a photosynthetic organelle contains an RNA homologous to the cyanobacterial counterpart. J. Mol. Biol. **257**:43-52.

Beebe, J., Kurz, J. and Fierke, C. (1996) Magnesium Ions are required by *Bacillus subtilis* ribonuclease P RNA for both binding and cleaving precursor $tRNA^{Asp}$. Biochemistry **35**:10493-10505.

Benson, D., Boguski, M., Lipman, D., Ostell, J. and Ouellette, B. (1998) Genbank. Nucleic Acids Research **26**:1-7.

Biswas, T. K. (1996) Expression of the mitochondrial RNase P RNA subunit-encoding gene from a variant promoter sequence in *Saccharomyces cerevisiae*. Gene **170**: 23-30.

Blanchard, J. L. and Schmidt, G. W. (1995) Pervasive migration of organellar DNA to the nucleus in plants. J. Mol. Evol. **41**:397-406.

Brennicke, A., Grohmann, L., Hiesel, R., Knoop, V. and Schuster, W. (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Letters **325**:140-145.

Brown, J. W. (1997) The Ribonuclease P Database. Nucleic Acids Research **25**:263-264.

Brown, J. W. (1998) The Ribonuclease P Database. Nucleic Acids Research **26**: 351-352.

Brown, J. W., Nolan, J. M., Haas, E. S., Rubio M-A. T., Major, F. and Pace, N. R. (1996) Comparative analysis of Ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. Proc. Natl. Acad. Sci. USA **93**: 3001-3006.

Bryant, D. and Moulton, V. (1998) A polynomial time algorithm for constructing the refined Buneman tree. Applied Mathematics Letters (in press).

Carrara, G., Calandra, P., Fruscoloni, P. and Tocchini-Valentini, G. P. (1995) Two helices plus a linker: A small model substrate for eukaryotic RNase P. Proc. Natl. Acad. Sci. USA **92**: 2627-2631.

Chamberlain, J. R., Pagan-Rámos, E., Kindelberger, D. W. and Engelke, D. R. (1996) An RNase P RNA subunit mutation affects ribosomal RNA processing. Nucleic Acids Research **24**:3158-3166.

Chamberlain, J. R., Lee, Y., Lane, W. S. and Engelke, D. R. (1998) Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. Genes Dev. **12**: 1678-1690.

Chu, S., Archer, R., Zengel, J.M. and Lindahl, L. (1994) The RNA of RNase MRP is required for normal processing of ribosomal RNA. Proc. Natl. Acad. Sci. USA **91**: 659-663.

Chu, S., Zengel, J. M. and Lindahl, L. (1997) A novel protein shared by RNase MRP and RNase P. RNA **3**:382-391.

Clark, C. G. and Roger, A. J. (1995) Direct evidence for secondary loss of mitochondria in *Entamoeba histolyica*. Proc. Natl. Acad. Sci. USA **92**: 6518-6521.

Dairaghi, D. J. and Clayton, D. A. (1993) Bovine RNase MRP cleaves the divergent bovine mitochondrial RNA sequence at the displacement-loop region. J. Mol. Evol. **37**: 338-346.

Darr, S.C., Brown, J.W. and Pace, N. R. (1992) The varieties of ribonuclease P. TIBS **17**: 178-182.

Dichtl, B. and Tollervey, D. (1997) Pop3p is essential for the activity of the RNase MRP and RNase P ribonucleoproteins *in vivo*. The EMBO Journal **16**:417-429.

Dress, A,, Huson, D. and Moulton, V. (1996) Analyzing and visualizing sequence and distance data using Splits-Tree. Discrete Applied Mathematics **71**:95-109.

Dress, A. W. M., Perrey, S. W. and Fuellen, G. (1996) Fast approximation to the NP-hard problem of multiple sequence alignment. Information and Mathematical Sciences Reports, Series F: 96/06 Department of Mathematics, Massey University, Palmerston North.

Felsenstein, J. (1989) PHYLIP- Phylogeny inference package (version 3.2). Cladistics **5**:164-166.

Fontana, W., Kinings, D., Stadler, P. and Schuster, P. (1993) Statistics of RNA Secondary Structures. Biopolymers **33**: 1389-1404.

Forster, A. C. and Altman, S. (1990) Similar cage-shaped structures for the RNA components of all ribonuclease P and ribonuclease MRP enzymes. Cell **62**:407-409.

Freyer, R., Kiefer-Meyer, M-C. and Kössel, H. (1997) Occurrence of plastid RNA editing in all major lineages of land plants. Proc. Natl. Acad. Sci. USA **94**: 6285 – 6290.

Glemarec, C., Kufel, J., Földesi, A., Maltseva, T., Sandström, A., Kirsebom, L. A. and Chattopadhyaya, J. (1996) The NMR structure of 31mer RNA domain of Escherichia coli RNase P RNA using its non-uniformly deuterium labelled counterpart [the 'NMR-window' concept. Nucleic Acids Research **24**: 2022-2035.

Green, C. J., Rivera-Leon, R. and Vold, B. S. (1996) The catalytic core of RNase P. Nucleic Acids Research **24**: 1497-1503.

Gultyaev, A., van Batenburg, F. and Pleij, C. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. J. Mol. Biol. **250**: 37-51.

Gupta, R. S. and Golding, G. (1996) The origin of the eukaryotic cell. TIBS **21**:166-171.

Gutell, R. R. (1992) Evolutionary characteristics of 16S and 23S rRNA structures. In: Hartman, H. and Matsumo, K (Eds.) The origin and evolution of prokaryotic and eukaryotic cells. World Scientific Publishing Co, New York, p243

Gutell, R. R., Larsen, N. and Woese, C. R. (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiological reviews **58**:10-26.

Haas, E. S., Brown, J. W., Pitulle, C. and Pace, N. R. (1994). Further perspective on the catalytic core and secondary structure of ribonuclease P RNA. Proc. Natl. Acad. Sci. USA **91**: 2527-2531.

Haas, E..S., Armbruster, D. W., Vucson, B. M., Daniels, C. J. and Brown, J.W. (1996a) Comparative analysis of ribonuclease RNA structure in Archaea. Nucleic Acids Research **24**: 1252-1259.

Haas, E. S., Banta, A. B., Harris, J. K., Pace, N. R. and Brown, J. W. (1996b) Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. Nucleic Acids Research **24:**4775-4782

Hardt, W-D. and Hartmann, R. K. (1996) Mutational analysis of the joining regions flanking Helix P18 in *E. coli* RNase P RNA. J. Mol. Biol. **259**: 422-433.

Hiratsuka J and 16 other authors. (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. **217**: 185-194.

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

Hudson, G. S., Holton, T. A., Whitfield, P. R. and Bottomley, W. (1988) Spinach chloroplast rpoBC genes encode three subunits of the chloroplast RNA polymerase. J. Mol. Biol. **200**: 639-654.

Huson, D. (1998) Splitstree - a program for analysing and visualizing evolutionary data. Bioinformatics (in press)

Jacobson, M. R., Cao, L-G., Wang Y-L. and Pederson, T. (1995) Dynamic localization of RNase MRP RNA in the nucleolus observed by fluorescent RNA cytochemistry in living cells. The Journal of Cell Biology **131**: 1649-1658.

Jeffares, D. C., Poole, A. M. and Penny, D. (1998) Relics from the RNA world. J. Mol. Evol. **46**:18-36.

Karlin, S. and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA **87**:2264-2268.

Karwan, R. (1993) RNase MRP /RNase P: a structure-function relation conserved in evolution? FEBS **319**: 1-4.

Karwan, R. M. (1998) Further characterization of human RNase MRP/RNase P and related autoantibodies. Mol. Biol. Rep. 25: 95-101.

Karwan, R., Bennet, J. L. and Clayton, D. (1991) Nuclear RNase MRP processes RNA at multiple discrete sites: interaction with an upstream G box is required for subsequent downstream cleavages. Genes and Development **5**: 1264-1276.

Kiresbom, L. A. (1995) RNase P - a 'scarlet pimpernel'. Molecular Microbiology **17**: 411-420.

Kiseleva, E., Goldberg, M. W., Daneholt, B. and Allen, T .D. (1996) RNP Export is mediated by structural reorganization of the Nuclear Pore Basket. J. Mol. Biol. **260**: 304-311.

Kiss, T., Marshallsay, C. and Fillpowicz, W. (1992) 7-2/MRP RNAs in plant and mammalian cells: association with higher order structures in the nucleus. The EMBO Journal **11**:3737-3746.

Kolk, H. M., van der Graaf, M., Wijmenga S. S., Pleij, W. A., Heus H. A. and Hilbers C. W. (1998) NMR structure of a classical pseudoknot: Interplay of single- and double-stranded RNA. Science **280**: 434-438.

Labuda, D. and Zietkiewicz, E. (1994) Evolution of secondary structure is the family of 7SL-like RNAs. J. Mol. Evol. **39**: 506-518.

Lang, B. F., Burger, G, O'Kelly C. J., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M. and Grey, M. W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature **387**:493-497.

Lee, B., Matera, G. A., Ward, D. C. and Craft, J. (1996) Association of RNase mitochondrial RNA processing enzyme with ribonuclease P in higher ordered structures in the nucleolus: A possible co-ordinate role in ribosome biogenesis. Proc. Natl. Acad. Sci. USA **93**:11471-11476.

Lee, M. S., Henry, M. and Silver, P.A. (1996) A protein that shuttles between the nucleus and the cytoplasm is an important mediator of RNA export. Genes and Development **10**: 1233-1246.

Lee, Y. C., Lee, B. J. and Kang, H. S. (1996a) The RNA component of mitochondrial ribonuclease P from *Aspergillus nidulans*. European Journal of Biochemistry **235**: 297-303.

Lee, Y. C., Lee, B. J., Hwang, D. S. and Kang, H. S. (1996b) Purification and Characterization of mitochondrial Ribonuclease P from *Aspergillus nidulans*. European Journal of Biochemistry **235**: 289-296

Levinger, L., Bourne, R., Kolla, S., Cylin, E., Russell, K., Wang, X. and Mohan, A. (1997) Processing kinetics and minimal substrates for Drosophila RNase P and 3'-tRNAse. Nucleic Acids Symp. Ser. **36**: 78-82.

Li, K. and Williams, R. S. (1995) Cloning and characterization of three new murine genes encoding short homologues of RNase P RNA. The Journal of Biological Chemistry **270**: 25281-25285.

Li, K., Smagula, C. S., Parsons, W. J., Richardson, J. A., Gonzalez, M., Hagler, H. K. and Williams, R.S. (1994) Subcellular partitioning of MRP RNA assessed by ultrastructural and biochemical analysis. The Journal of Cell Biology **124**: 871-882.

Liu, F. and Altman, S. (1996) Requirements for cleavage by a modified RNase P of a small model substrate. Nucleic Acids Research **24**: 2690-2696.

Lockhart, P. J., Larkum, A. W. D., Steel, M. A., Waddell, P.J. and Penny, D. (1996) Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA **93**:1930-1934.

Lygerou, Z., Allmang, C., Tollervey, D. and Seraphin, B. (1996a) Accurate processing of an eukaryotic precursor ribosomal RNA by Ribonuclease MRP *in vitro*. Science **272**:268-270.

Lygerou, Z., Pluk, H., van Vernroij, W. J. and Séraphin, B. (1996b) hPop1: an autoantigenic protein subunit shared by the human RNase P and RNase MRP ribonucleoproteins. The EMBO Journal **15**: 5936-5948.

Lynch M (1996) Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes. Mol. Biol. Evol **13**:209-220.

Mahapatra, S. and Adhya, S. (1996) Import of RNA into *Leichmania* Mitochondria occurs through direct interaction with membrane-bound receptors. The Journal of Biological Chemistry **271**: 20432-20437

Maidak, B., Olsen, G., Larsen, N., Overbeek, R., McCaughey, M. and Woese, C. (1997) The RDP (Ribosomal Database Project) Nucleic Acids Research **25**:109-111.

Maier, R. M., Neckermann, K., Igloi, G. L., Kössel, H. (1995) Complete sequence of the maize chloroplast genome: gene content, hotshots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. **251**: 614-628.

Marchfelder, A. and Brennicke, A. (1993) Plant mitochondrial RNase P and *E. coli* RNase P have different substrate specificities. Biochemistry and Molecular Biology International **29**: 621-633.

Marchfelder, A. and Brennicke, A. (1994) Characterization and partial purification of tRNA processing activities from potato mitochondria. Plant Physiol. **105**: 1247-1254.

Martin, A. (1995) Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol. Biol. Evol. **12**:1124-1131.

Martin, N, C. and Lang, B. F. (1997) Mitochondrial RNase P: the RNA family grows. Nucleic Acids Symp. Ser. **36**: 42-44.

Martin, W. and Muller, M. (1998) The hydrogen hypothesis for the first eukaryote. Nature 392: 37-41

Matzura, O. and Wennborg, A. (1996) RNAdraw: an integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. Computer Applications in the Biosciences (CABIOS) **12**: 247-249.

Morgenstern, B., Dress, A. and Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. USA **93**:12098-12103.

Morrissey, J. P. and Tollervey, D. (1995) Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system. TIBS **20**:78-82.

Moulton, V., Steel, M. and Tuffley, C. (1997) Dissimilarity maps and substitution models. Proceedings of the DIMACS workshop on mathematical hierarchies. AMS **37**:111-131.

Moulton, V. and Steel, M. (1998) Retractions of finite distance functions onto tree metrics. Discrete Applied Mathematics (in press).

Moulton, V., Zuker, M., Steel, M., Pointon, M. and Penny, D. (1988) Metrics on RNA secondary structure. Submitted to J. Comput. Biol.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota S., Inokuchi, H. and Ozeki H. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature **322**: 572-574.

Ouzounis, C., Casari, G., Sander, C., Tarnames, J. and Valencia, A. (1996) Computational comparisons of model genomes. TIBTECH **14**: 280-285.

Pace, N. and Brown, J. (1995) Evolutionary Perspective on the structure and function of Ribonuclease P, a ribozyme. Journal of Bacteriology **177**:1919-1928.

Pace, N. and Smith, D. (1990) Ribonuclease P: function and variation. The Journal of Biological Chemistry **265**:3587-3590.

Page, R. D. M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences **12**:357-358.

Palmer, J. D, and Delwiche, C. F. (1996) Second-hand chloroplasts and the case of the disappearing nucleus. Proc. Natl. Acad. Sci. USA **93**: 7432-7435.

Paluh, J. L. and Clayton, D. A. (1995) *Schizosaccharomyces pombe* RNase MRP RNA is homologous to metazoan RNase MRP RNAs and may provide clues to interrelationships between RNase MRP and RNase P. Yeast **11**:1249-1264.

Paluh, J. L. and Clayton, D. A. (1996) A functional dominant mutation in *Schizosaccharomyces pombe* RNase MRP RNA affects nuclear RNA processing and requires the mitochondrial-associated mutation ptp1-1 for viability. The EMBO Journal **15**: 4723-4733.

Pan, J., Thirumalai, D. and Woodson, S. (1997) Folding of RNA involves parallel pathways. J. Mol. Biol. **273**: 7-13.

Pan, T. (1995) Higher order folding and domain analysis of the ribozyme from *Bacillus subtilis* Ribonuclease P. Biochemistry **34**: 902-909.

Pannucci, J., Haas, E. S. and Brown, J. W. (1997) RNase P cleavage assays using archaeal cell extracts and in vitro transcribed archaeal RNase P RNA. Nucleic Acids Symp. Ser. **36**: 90-92.

Pearson, W. R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics **11**: 635-650.

Perez-Terzic, C., Pyle, J., Jaconi, M., Stehno-Bittel, L. and Clapham, D. E. (1996) Conformational states of the Nuclear Pore Complex induced by Depletion of Nuclear Ca^{2+} Stores. Science **273**:1875-1877.

Pitulle, C., Garcia-Paris, M., Zamudio, K. R. and Pace N. R. (1998) Comparative structure analysis of vertebrate ribonuclease P RNA. Nucleic Acids Research **26**: 3333-3339.

Pleij, C. and Bosch, L. (1989) RNA Pseudoknots: Structure, Detection, and Prediction. Methods in Enzymology **180**:289-303.

Poole, A. M., Jeffares, D. C. and Penny, D. (1998) The Path from the RNA world. J. Mol. Evol. **46**:1-17.

Potuschak, T., Rossmanith, W. and Karwan, R. (1993) RNase MRP and RNase P share a common substrate. Nucleic Acids Research **21**: 3239-3243.

Preston, B. D. (1996) Error-prone retrotransposition: Rime of the ancient mutators Proc. Natl. Acad. Sci. USA **93**: 7427-7431.

Reith, M. and Munholland, J. (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Rep **13**: 333-335.

Rossmanith, W. and Karwan, R. M. (1998) Characterisation of human mitochondrial RNase P: novel aspects in tRNA processing. Biochem. Biophys. Res. Commun. **247**: 234-241.

Saitou, N. and Nei, M. (1987) The Neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406-425.

Sbisà, E., Pesole, G., Tullo, A. and Saccone, C. (1996) The evolution of the RNase P- and RNase MRP-associated RNAs: Phylogenetic analysis and nucleotide substitution rate. J Mol Evol **43**:46-57.

Schmitt, M. E., Bennett, J. L., Dairaghi, D. J. and Clayton, D. A. (1993) Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison. The FASEB Journal 7:208-213.

Schuster, P., Fontana, W., Stadler, P. and Hofacker, I. (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc. R. Soc. Lond. B **255**: 279-284.

Scott, W. G. and Klug, A. (1996) Ribozymes: structure and mechanism in RNA catalysis. TIBS **21**: 220-223.

Shapiro, B. and Zhang, K. (1990) Comparing multiple secondary structures using tree comparison. CABIOS **6**:309-318.

Shinozaki, K. and 22 other authors (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal **5**: 2043-2049.

Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. J Mol Biol **147**:195-197.

Smith, D. and Pace, N. R. (1993) Multiple magnesium ions in the Ribonuclease P mechanism. Biochemistry 32: 5273-5281.

Stams, T., Niranjanakumari, S. Fierke, C. A. and Christianson D. W. (1998) Ribonuclease P protein structure: Evolutionary origins in the translational apparatus. Science **280**: 752-755.

Sulo, P., Groom, K. R., Wise, C., Steffen, M. and Martin, N. (1995) Successful transformation of yeast mitochondria with RPM1: an approach for in vivo studies of mitochondrial RNase P RNA structure, function and biosynthesis. Nucleic Acids Research **23**: 856-860.

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

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **25**:4876-4882.

Tonges, U., Perrey, S. W., Stoye, J. and Dress, A. W. (1996) A general Method for Fast Multiple Sequence Alignment. Gene **172**: GC33-GC41.

Tranguch, A. J. and Engelke, D. R. (1993) Comparative structural analysis of nuclear RNase P RNAs from yeast. The Journal of Biological Chemistry **268**:14045-14053.

True, H. L. and Celander, D. W. (1996) Ribonuclease P of *Tetrahymena thermophila*. The Journal of Biological Chemistry **271**: 16559-16566.

Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T. and Sugiura, M. (1994) Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci USA **91**: 9794-9798.

Warnecke, J. M., Fürste, J. P., Hardt, W-D., Erdmann, V. A. and Hartmann, R. K. (1996) Ribonuclease P (RNase P) RNA is converted to a Cd^{2+} -ribozyme by a single RNase P-phosphorothioate modification in the precursor tRNA at the RNase P cleavage point. Proc. Natl. Acad. Sci. USA **93**: 8924-8928.

Weeks, K. M. and Cech, T. R. (1996) Assembly of a Ribonucleoprotein catalyst by tertiary structure capture. Science **271**: 345-348.

Winker, S. and Woese, C. R. (1991) A definition of the domains *Archaea, Bacteria* and *Eucarya* in terms of small subunit ribosomal RNA characteristics. System. Appl. Microbiol. **14**: 305-310.

Wischmann, C. and Schuster, W. (1995) Transfer of *rps10* from the mitochondrion to the nucleus in *Arabidopsis thaliana*: evidence for RNA-mediated transfer and exon shuffling at the integration site. FEBS Letters **374**: 152-156.

Wise, C. A. and Martin, N. C. (1991) Dramatic size variation of Yeast mitochondrial RNA's suggests that RNase P RNA's can be quite small. The Journal of Biological Chemistry **266**: 19154-19157.

Woese, C. R. (1987) Bacterial Evolution. Microbiology Reviews **51**:221-271.

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

Wool, I. G. (1996) Extraribosomal functions of ribosomal proteins. TIBS **21**:164-165.

Yuan, Y. and Reddy, R. (1991) 5' flanking sequences of human MRP / 7-2 RNA gene are required and sufficient for the transcription by RNA polymerase III. Biochimica et Biophysica Acta **1089**: 33-39.

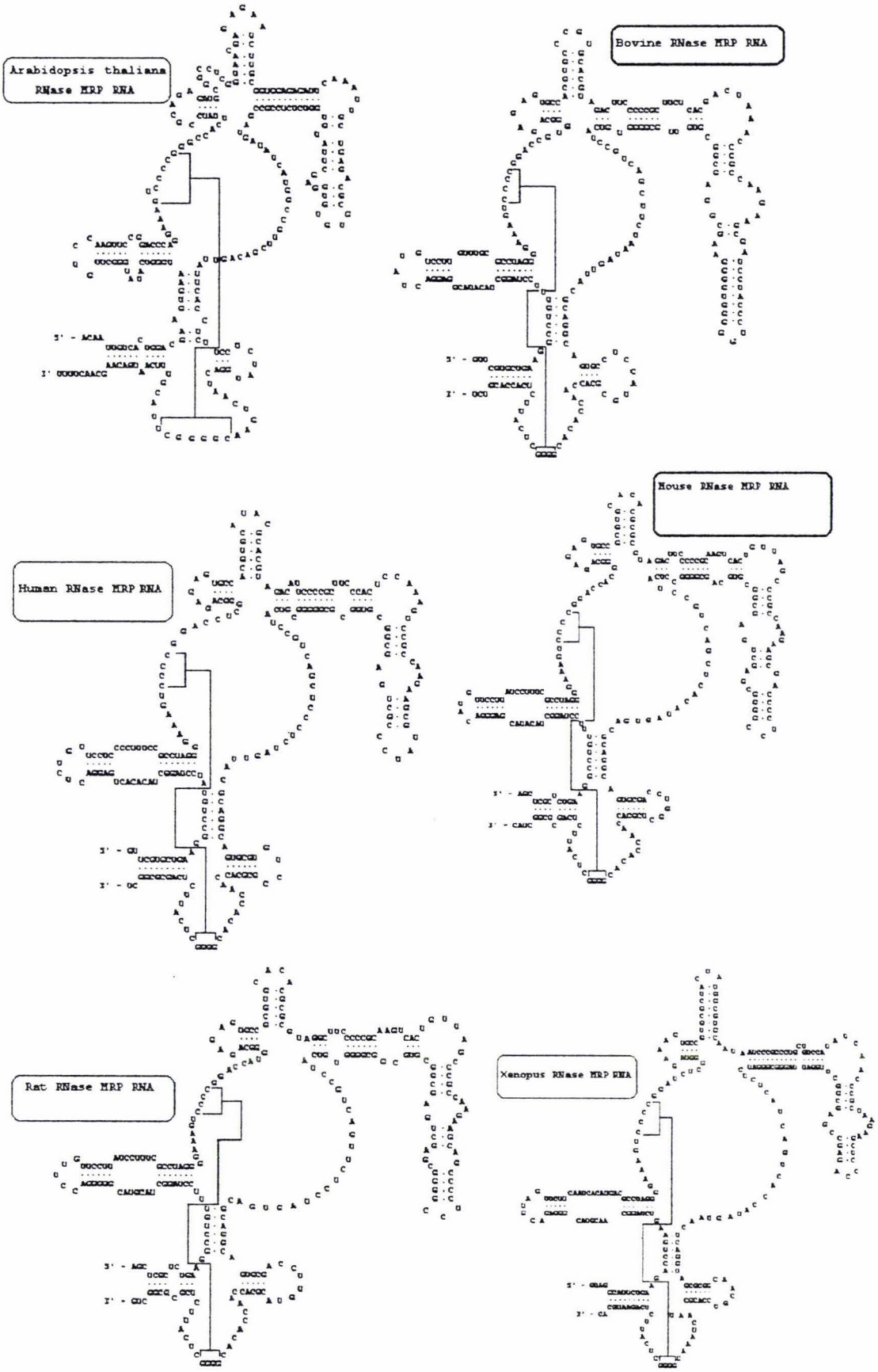
Zardoya, R. and Meyer, A. (1996) Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol. Biol. Evol **13**: 933-942.

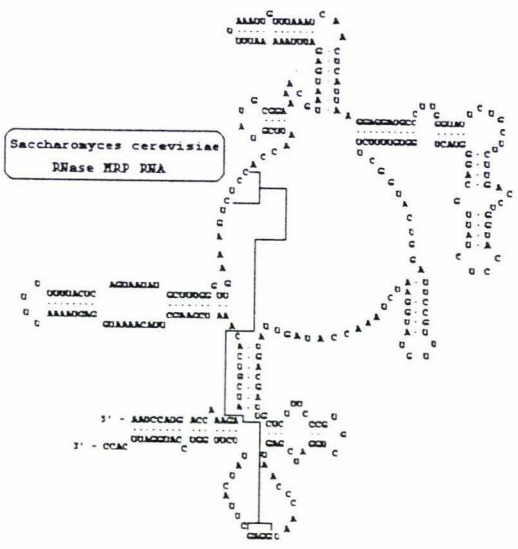
Ziehler, W. A., Lee, Y., Pagan-Ramos, E. and Engelke, D. R. (1997) An active domain of the nuclear RNase P RNA. Nucleic Acids Symp. Ser. **36**: 45-48.

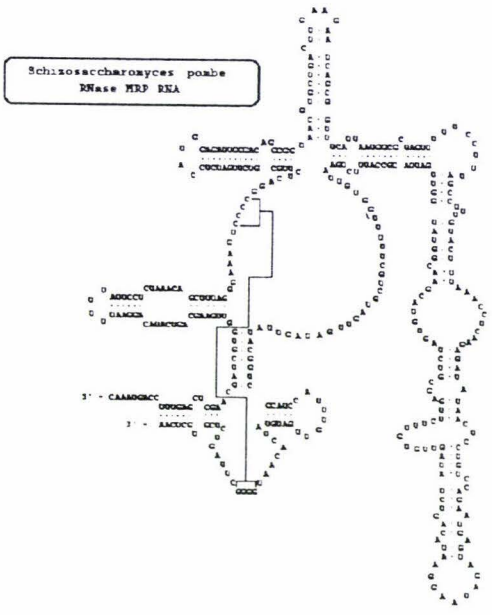
Zimmerly, S., Gamulin, V., Burkard, U. and Söll, D. (1990) RNA component of RNase P in *Schizosaccharomyces* species. FEBS **271**: 189-193.

Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. Science **244**: 48-52.

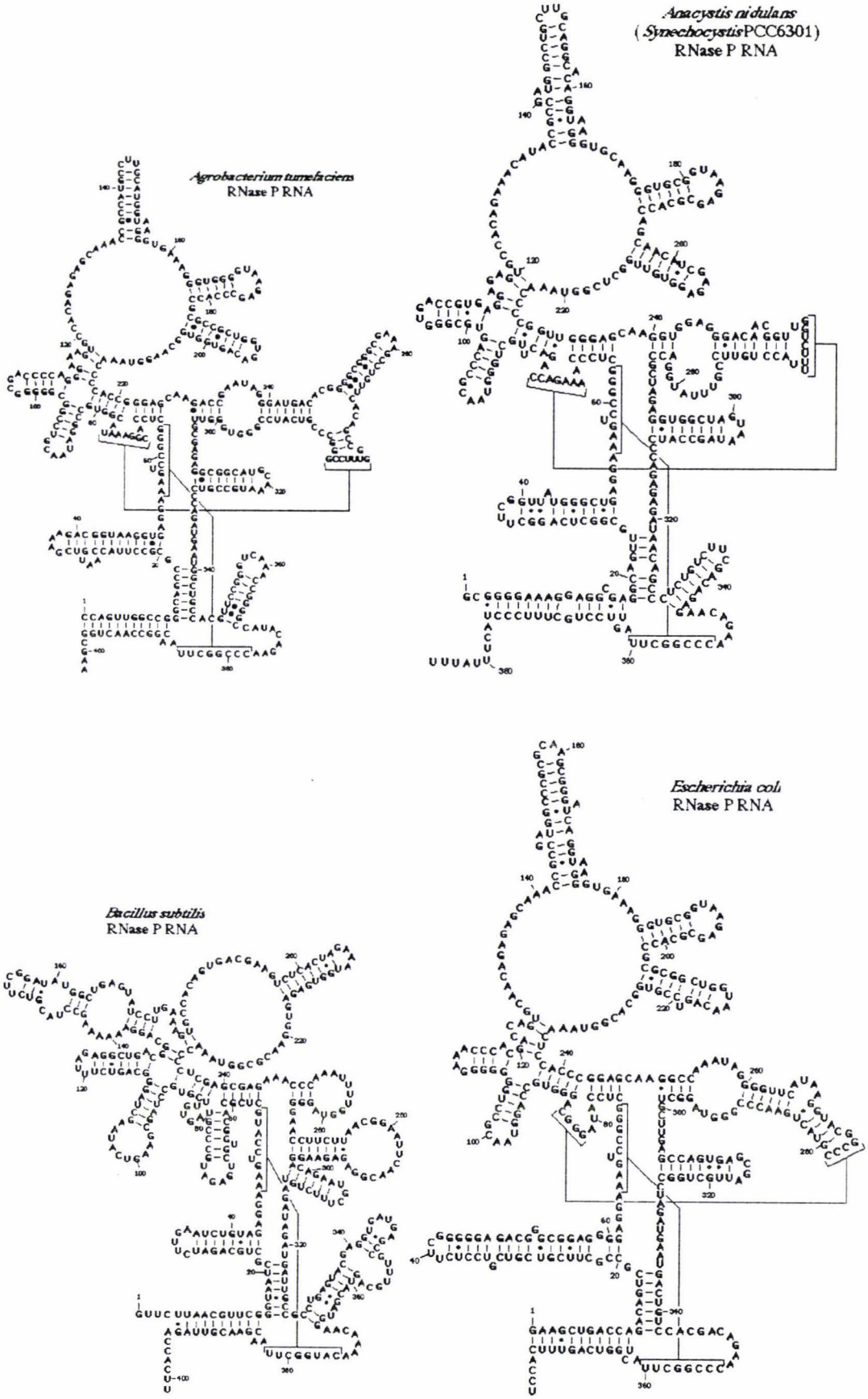
Appendix 1A: Biological secondary structures of mrpRNA

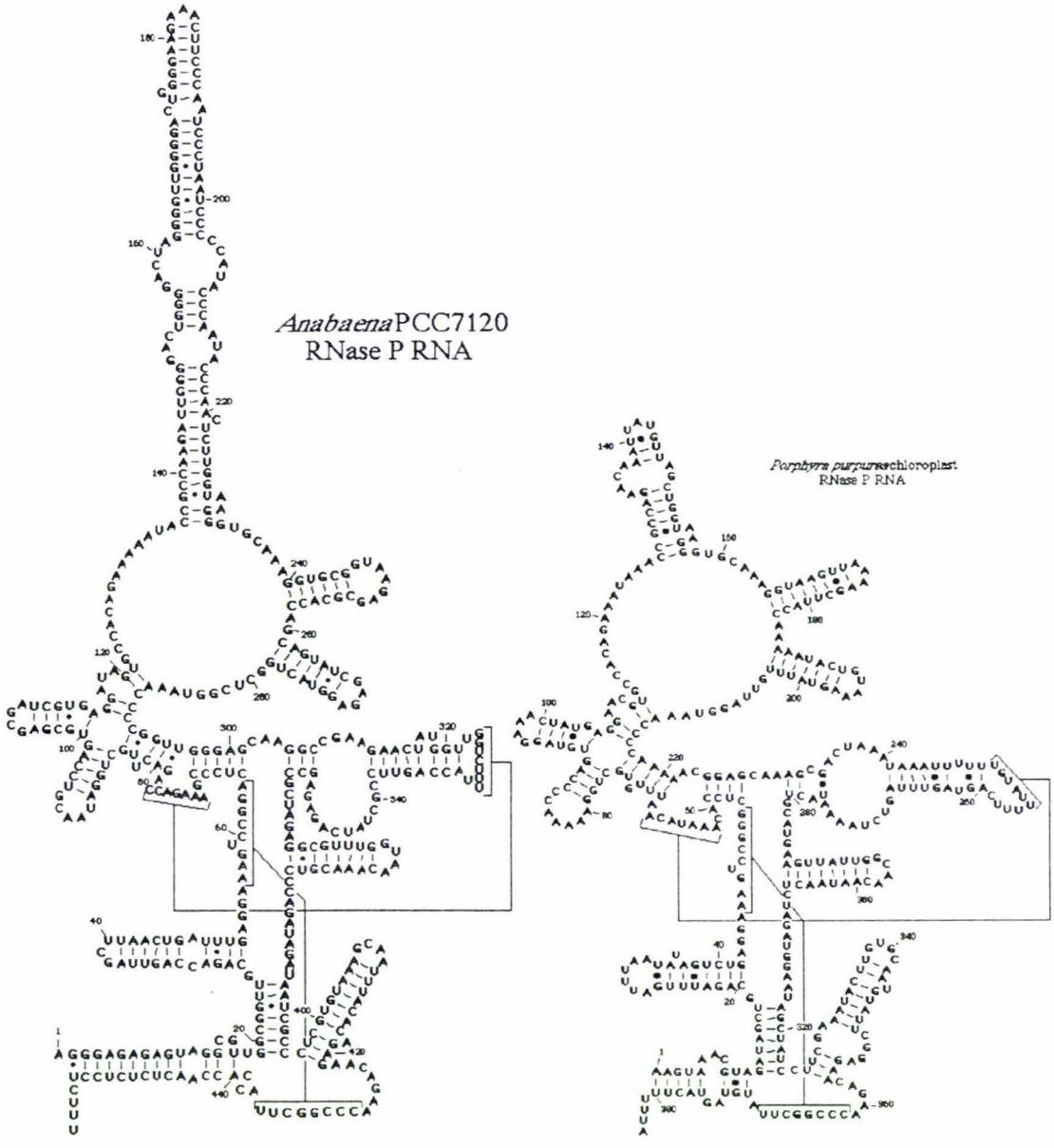


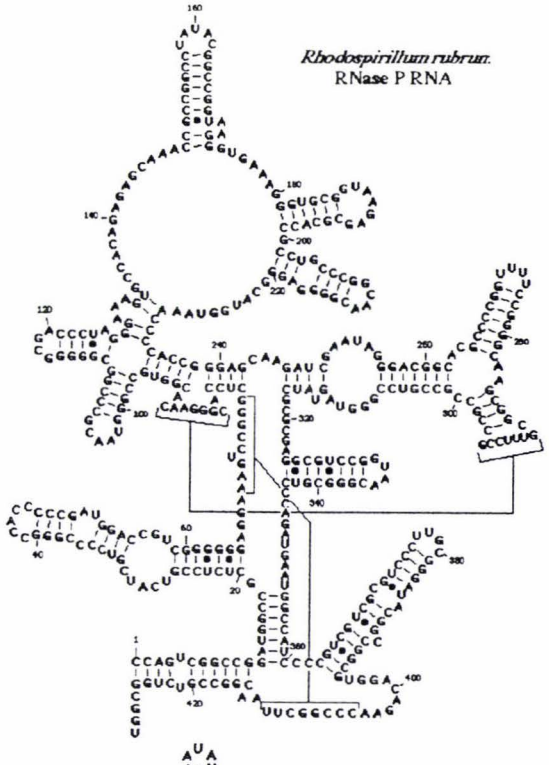

Saccharomyces cerevisiae
RNase MRP RNA

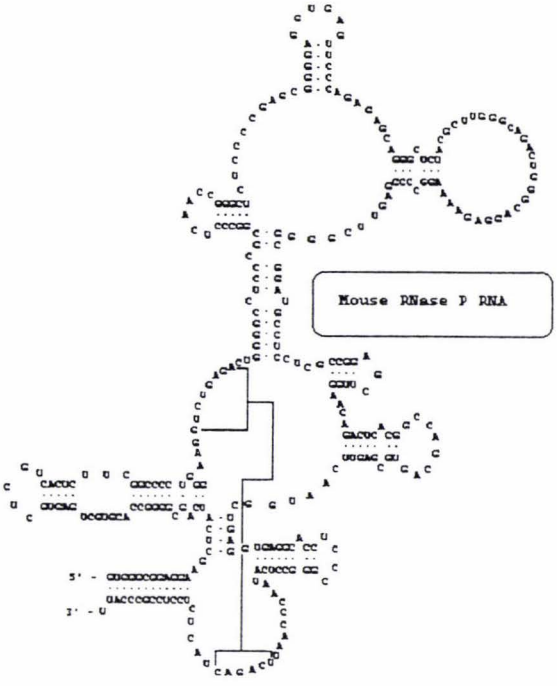

Schizosaccharomyces pombe
RNase MRP RNA

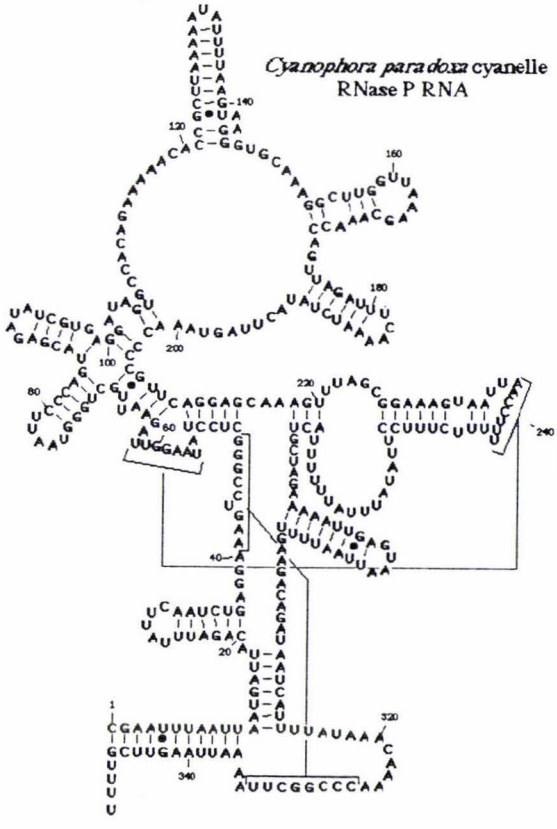
Appendix 1B: Biological secondary structures of pRNA

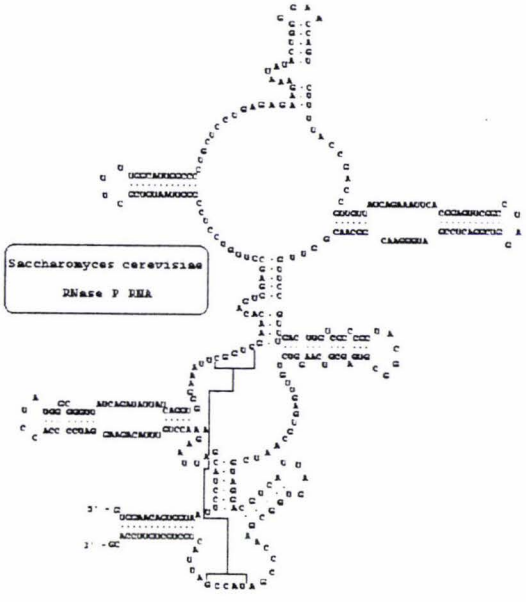


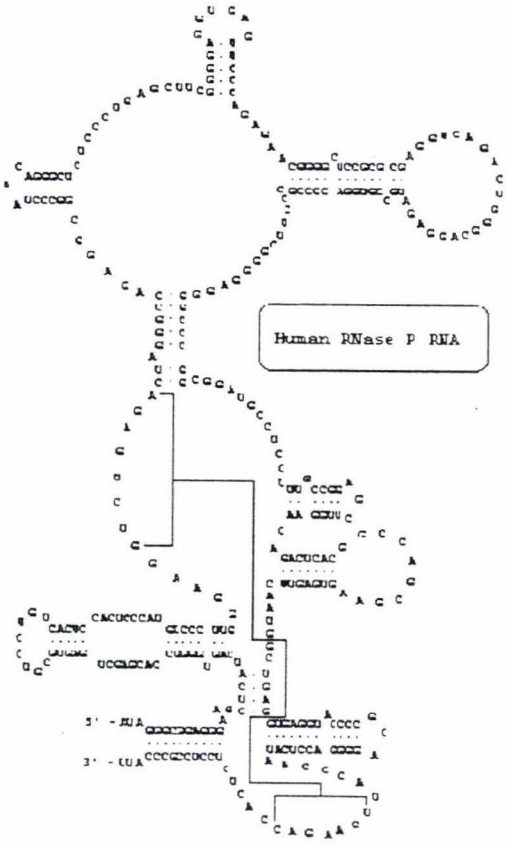

Anabaena PCC7120
RNase P RNA
Porphyra purpurea chloroplast
RNase P RNA

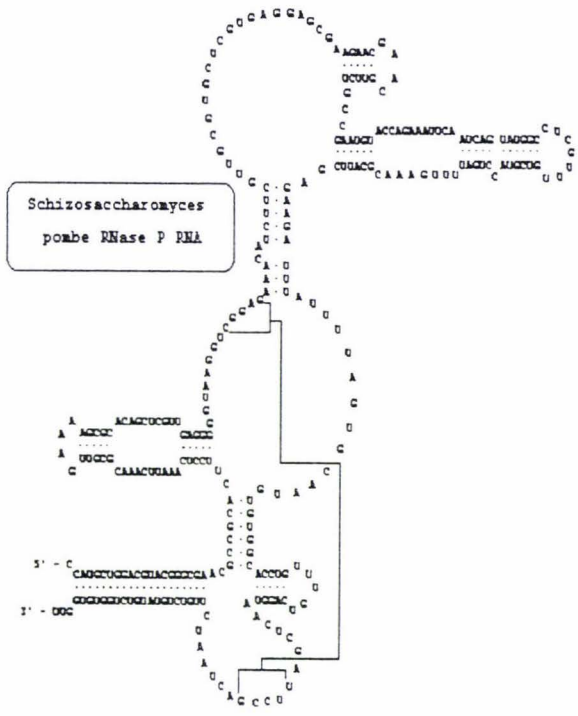

Rhodospirillum rubrum
RNase P RNA


Mouse RNase P RNA

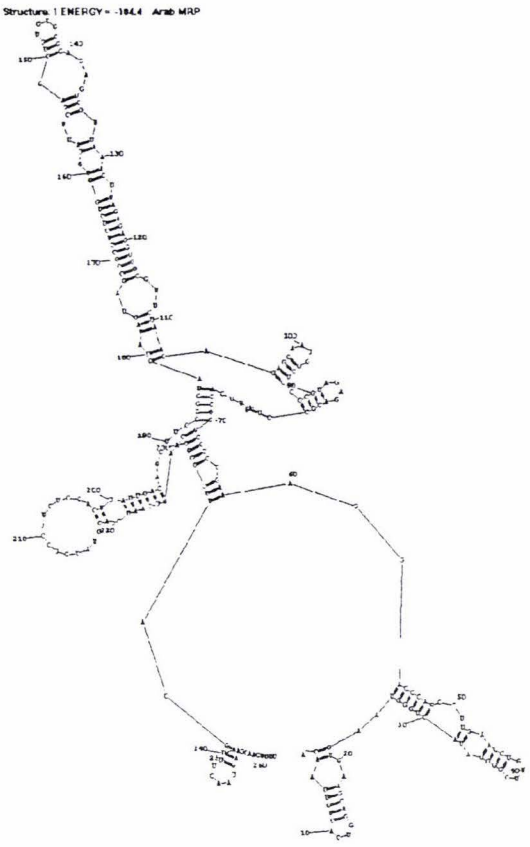

Cyanophora paradoxa cyanelle
RNase P RNA


Saccharomyces cerevisiae
RNase P RNA

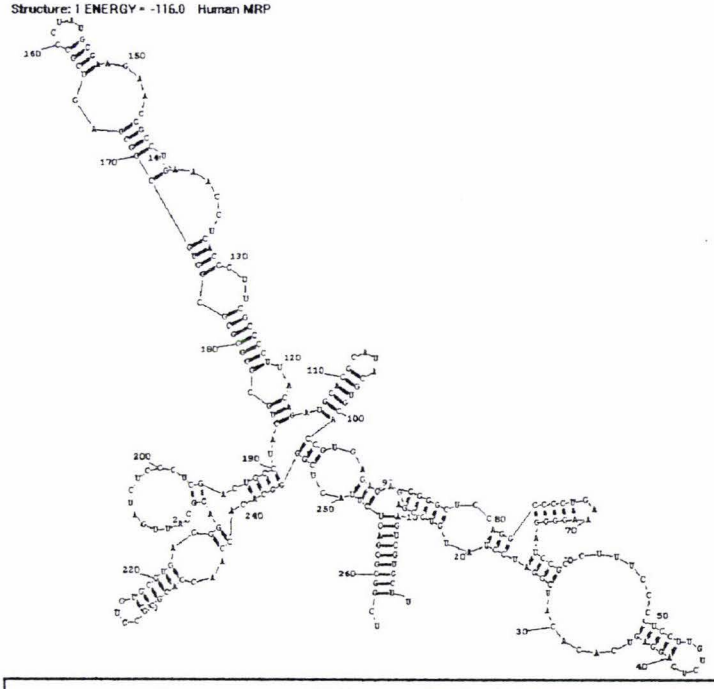

Human RNase P RNA


Schizosaccharomyces
pombe RNase P RNA

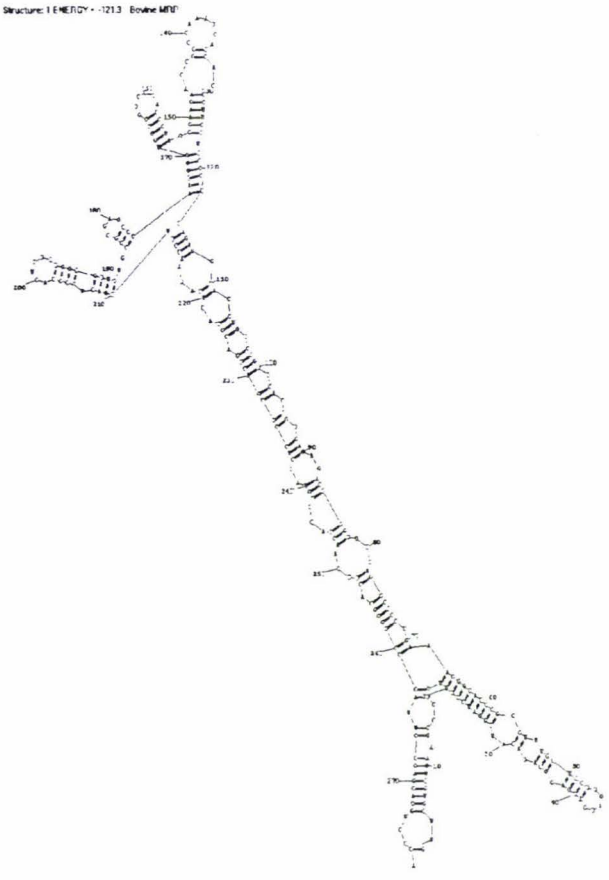
Appendix 1C: RNAstructure (Mfold) secondary structures of mrpRNA



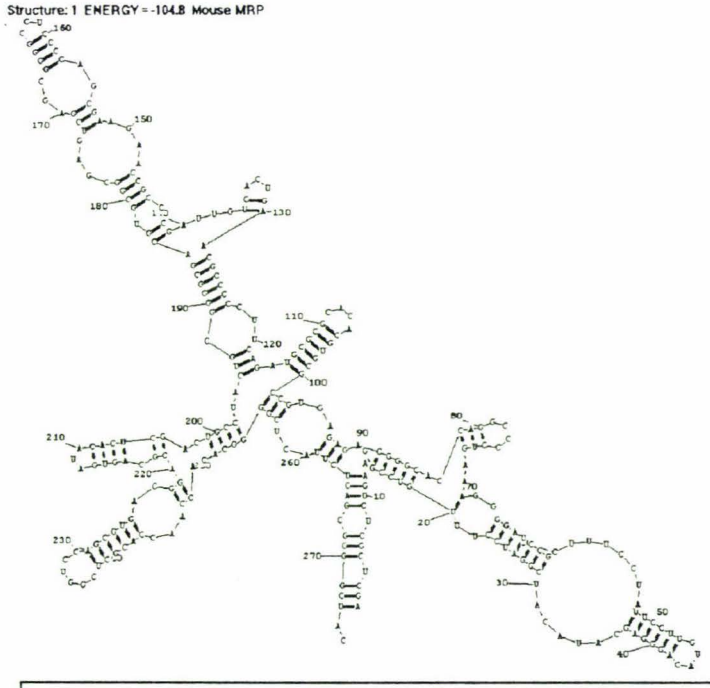
Arabidopsis mrpRNA - mfold (RNAstructure)



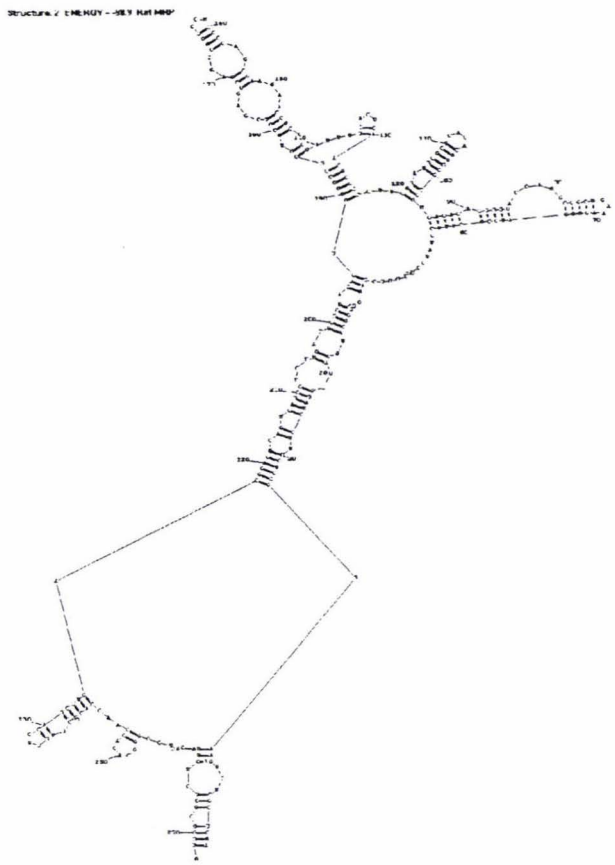
Human mrpRNA - mfold (RNAstructure)



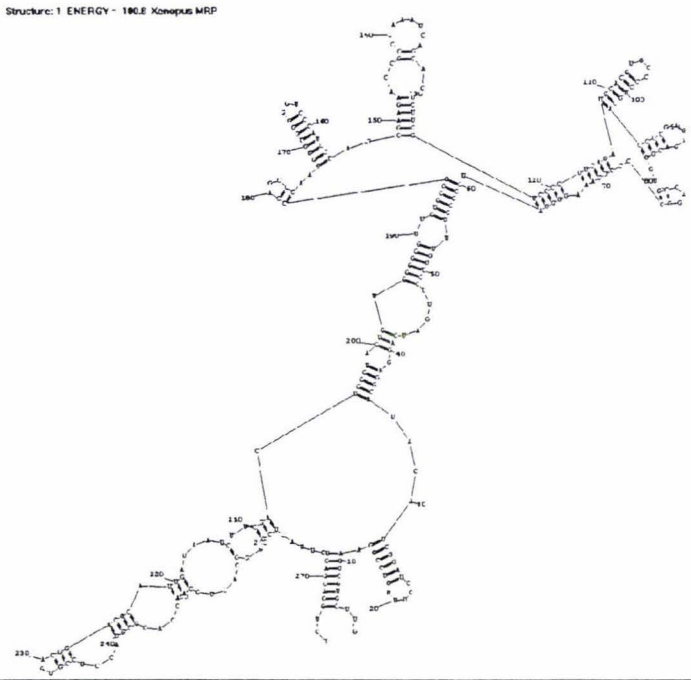
Bovine mrpRNA - mfold (RNAstructure)



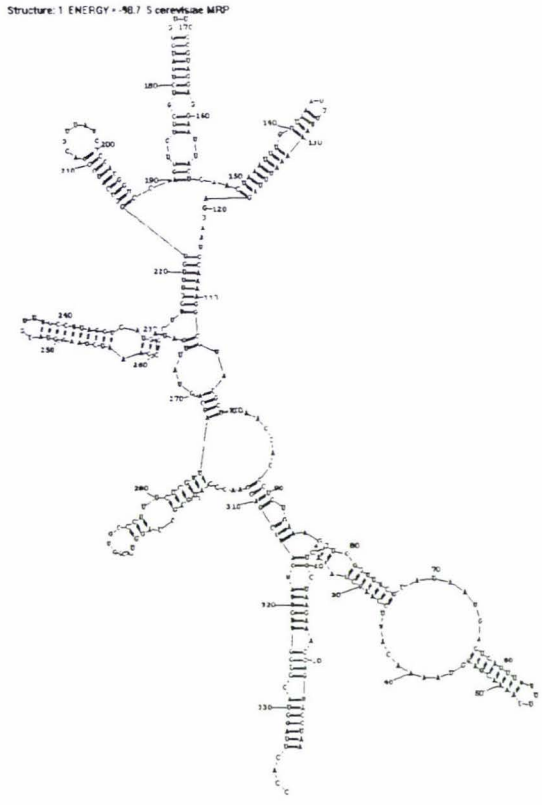
Mouse mrpRNA - mfold (RNAstructure)



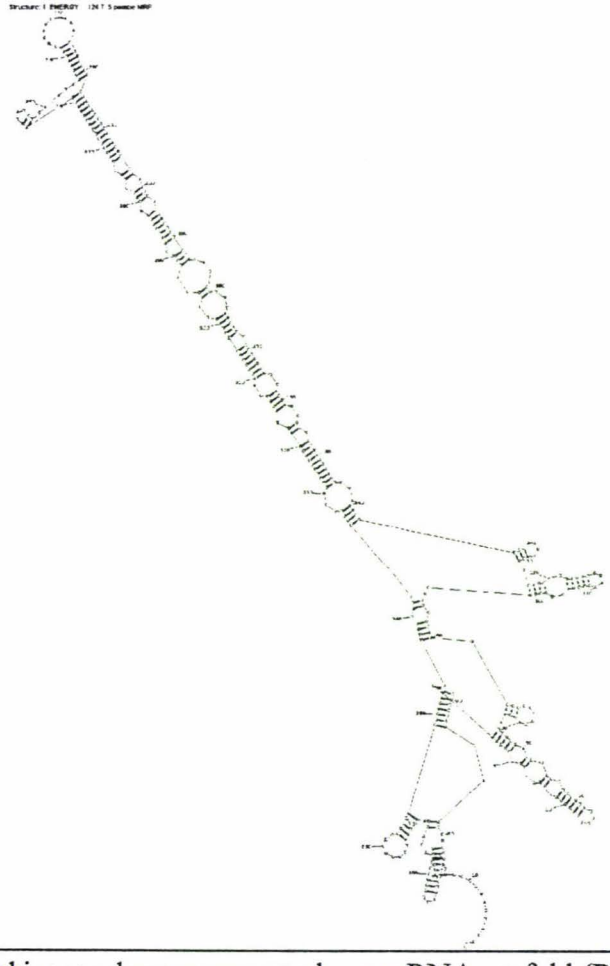
Rat mrpRNA - mfold (RNAstructure)



Xenopus mrpRNA - mfold (RNAstructure)

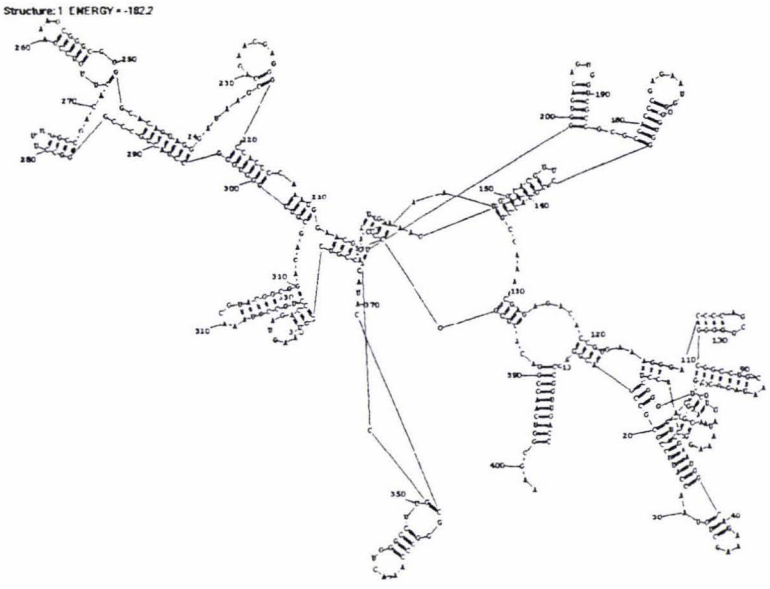


Saccharomyces cerevisiae mrpRNA - mfold (RNAstructure)

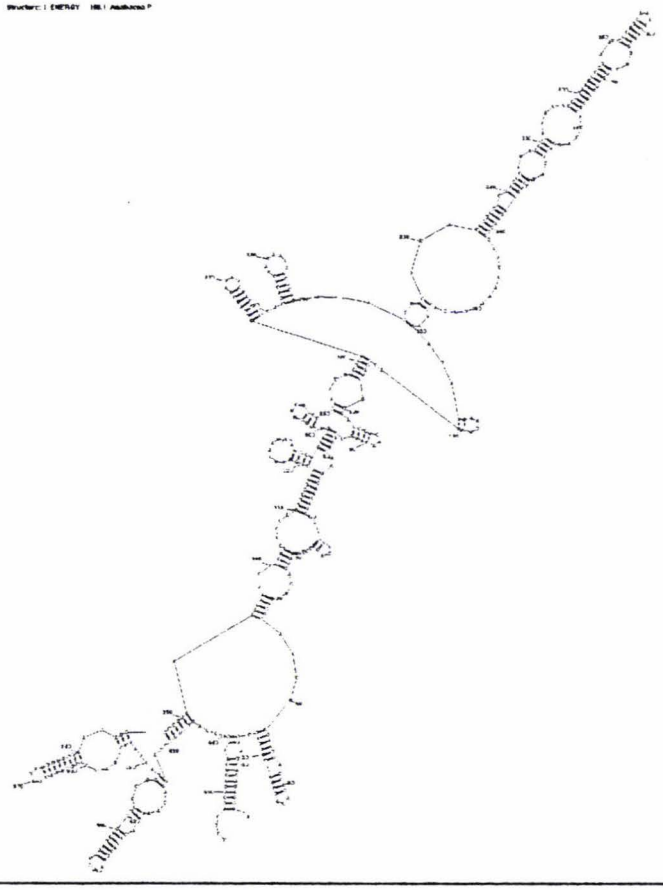


Schizosaccharomyces pombe mrpRNA - mfold (RNAstructure)

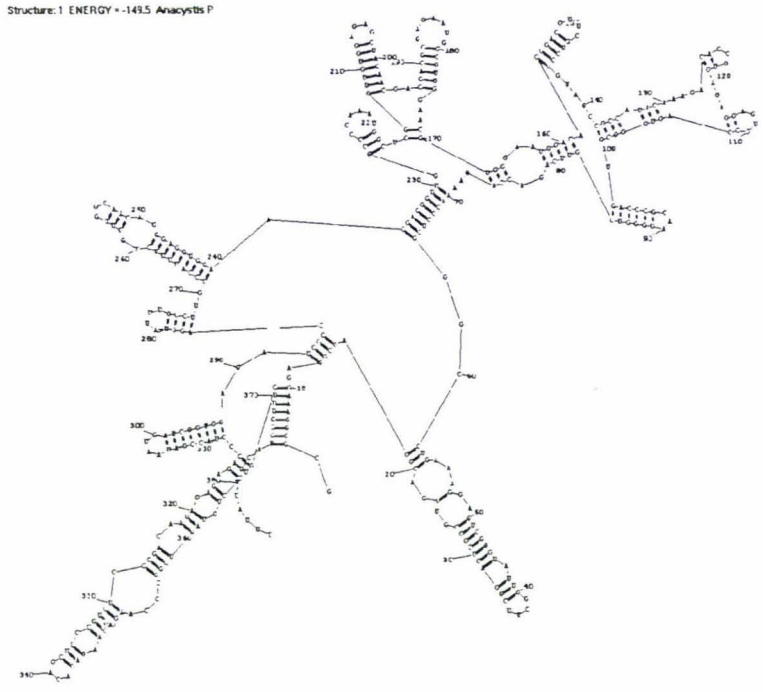
Appendix 1D: RNAstructure (Mfold) secondary structures of pRNA



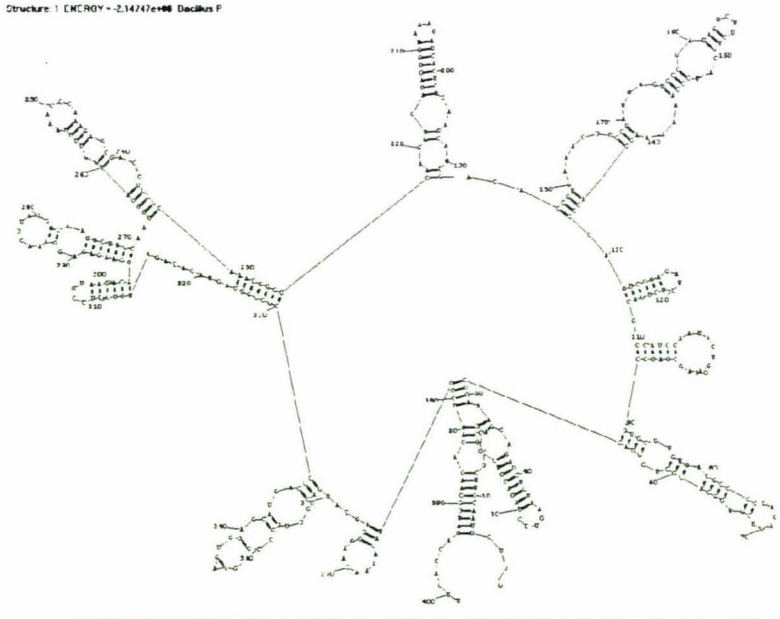
Agrobacterium tumefaciens pRNA - mfold (RNAstructure)



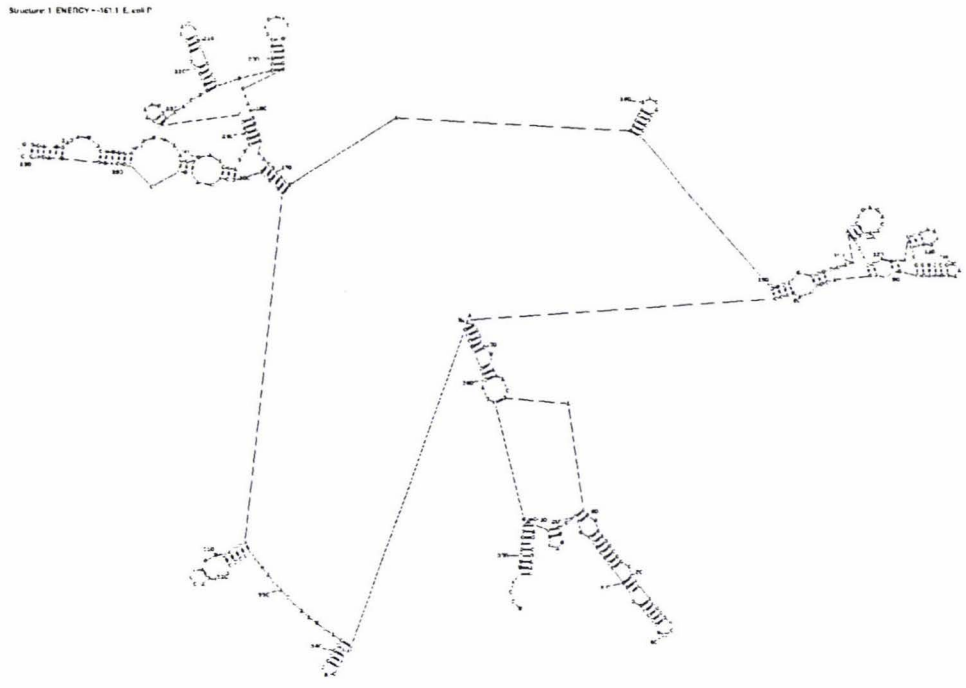
Anabaena pRNA - mfold (RNAstructure)



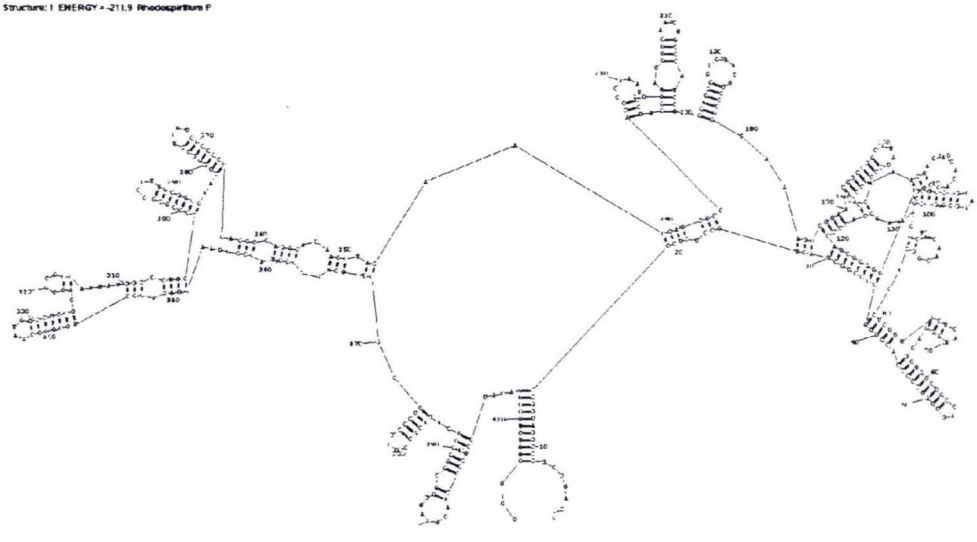
Anacystis nidulans pRNA - mfold (RNAstructure)



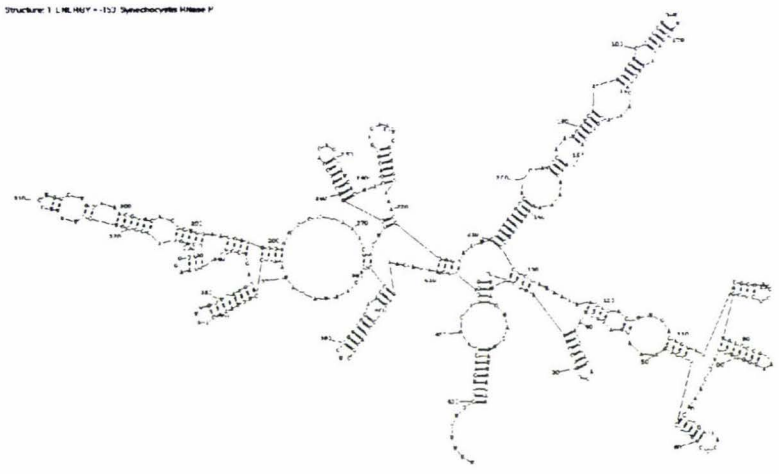
Bacillus subtilis pRNA - mfold (RNAstructure)



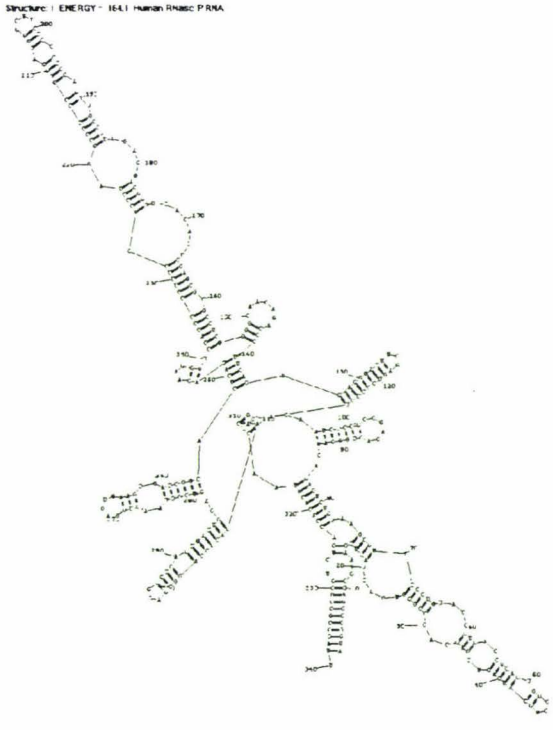
E.coli pRNA - mfold (RNAstructure)



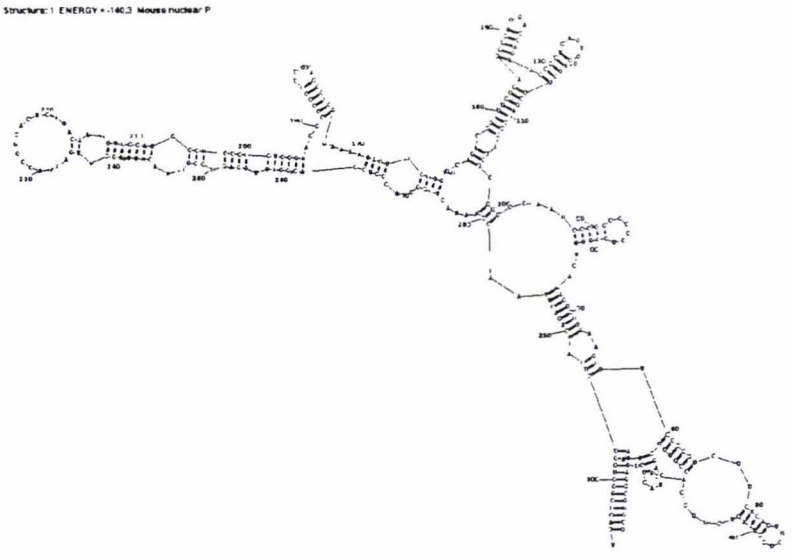
Rhodospirillum rubrum pRNA -mfold (RNAstructure)



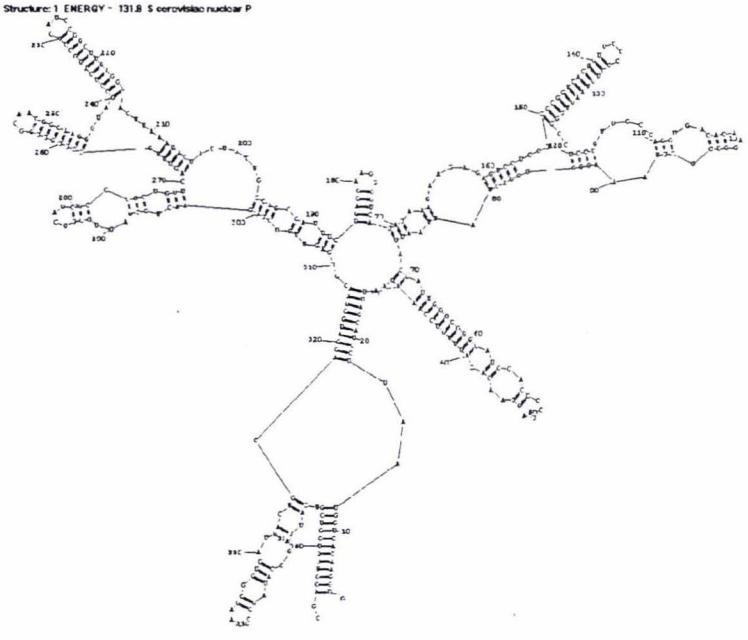
Synechocystis pRNA - mfold (RNAstructure)



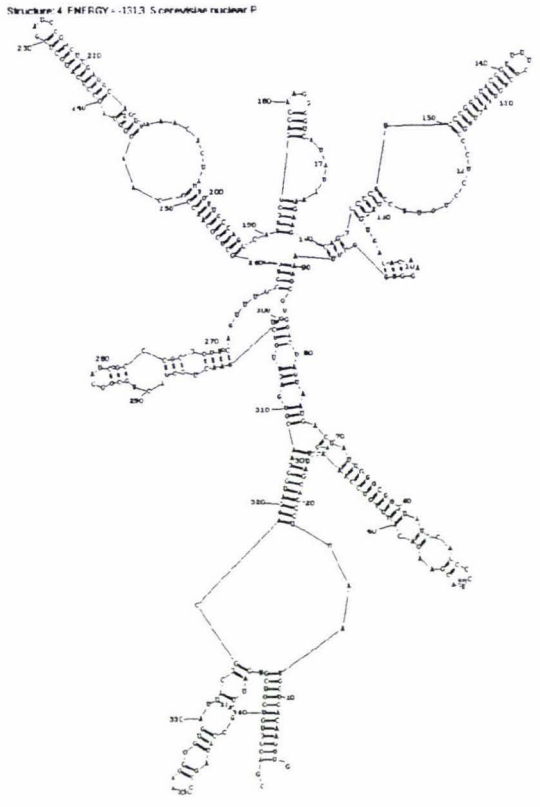
Human nuclear pRNA - mfold (RNAstructure)



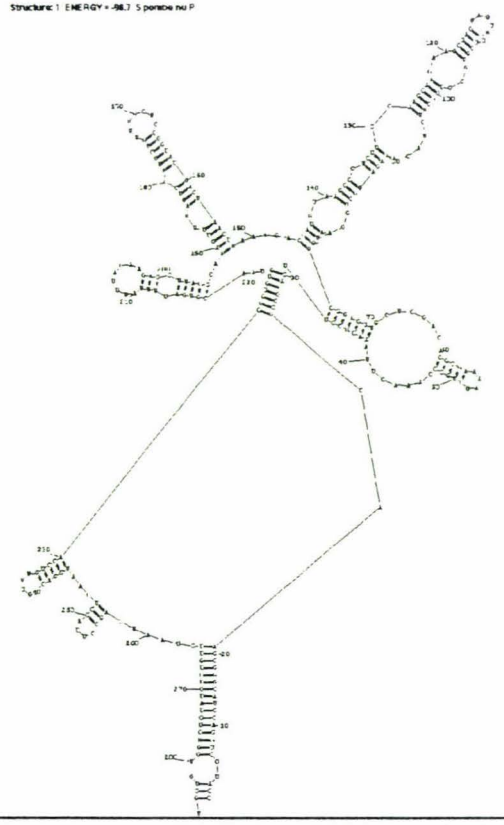
Mouse nuclear pRNA - mfold (RNAstructure)



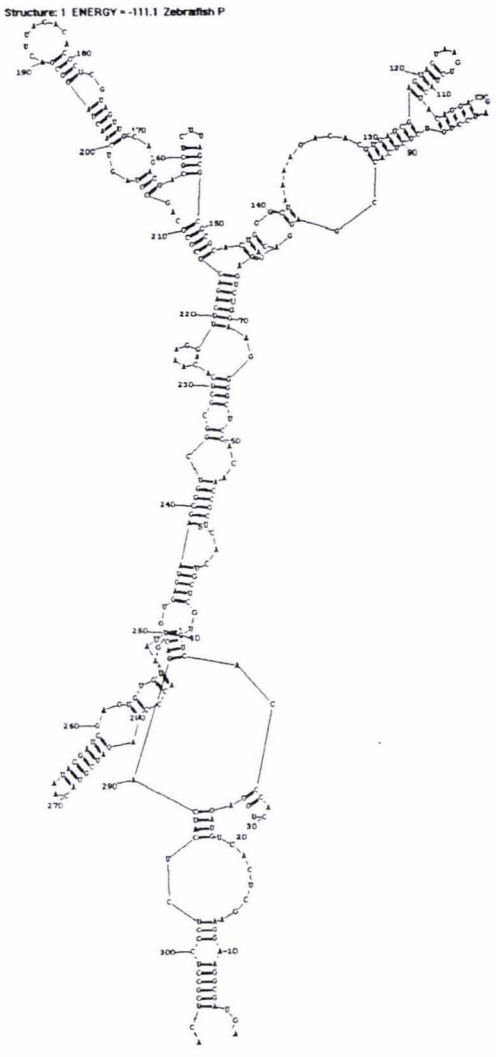
Saccharomyces cerevisiae nuclear pRNA - mfold (RNAstructure)



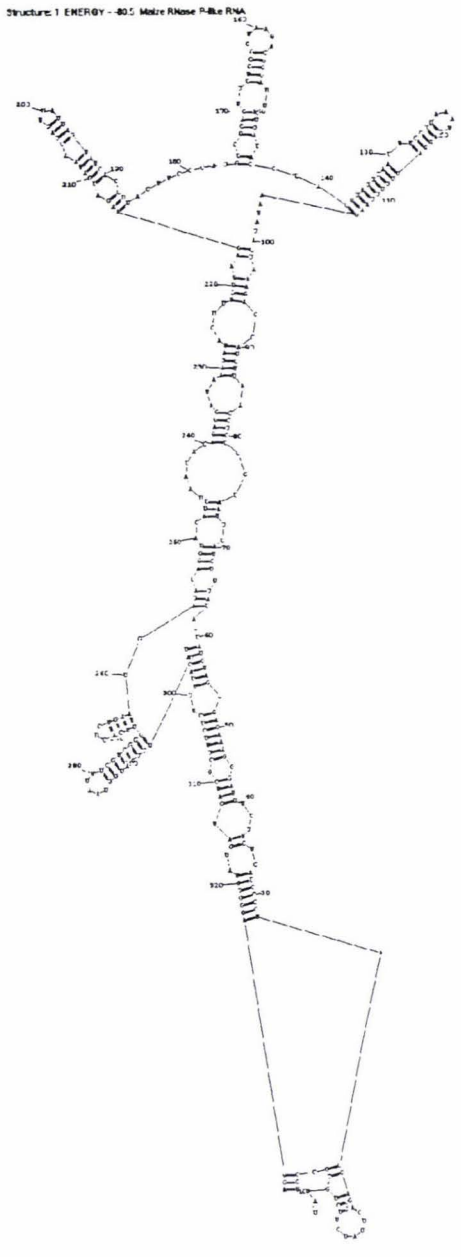
Saccharomyces cerevisiae pRNA - mfold (RNAstructure) (structure 4)



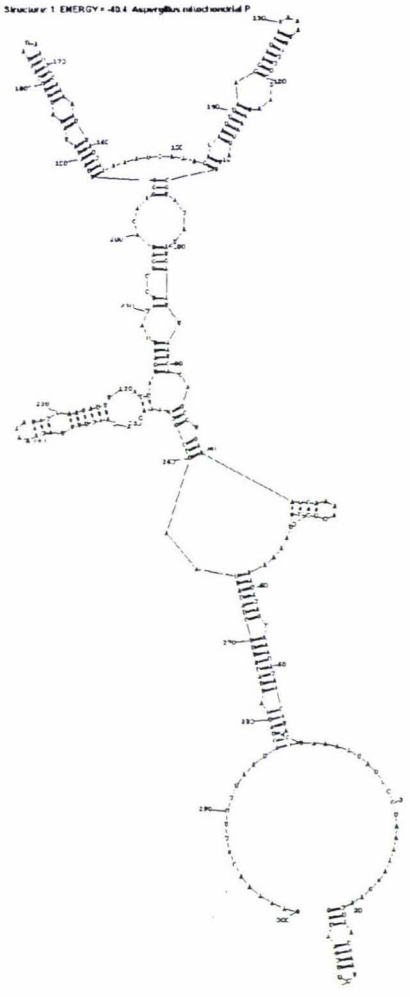
Schizosaccharomyces pombe pRNA - mfold (RNAstructure)



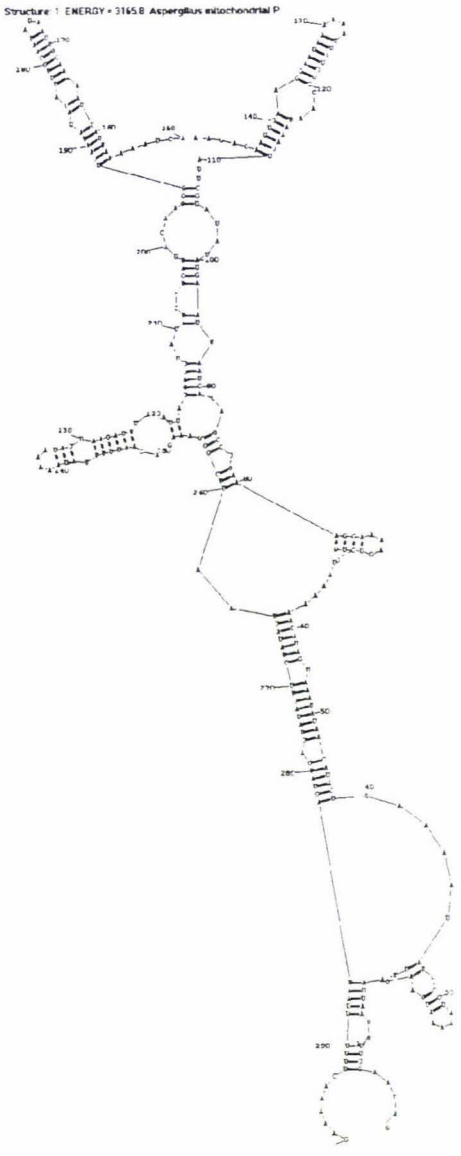
Zebrafish nuclear pRNA - mfold (RNAstructure)



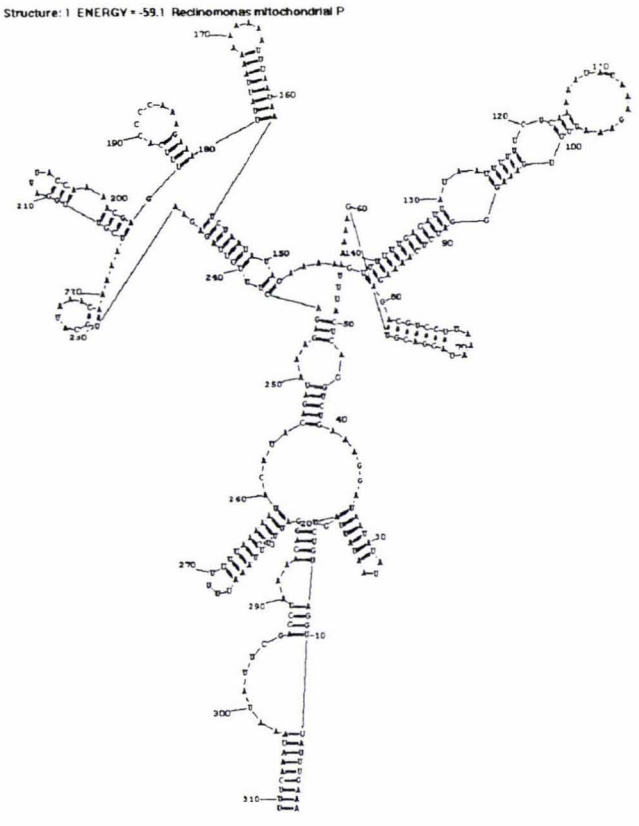
Maize chloroplast P-like - mfold (RNAstructure)



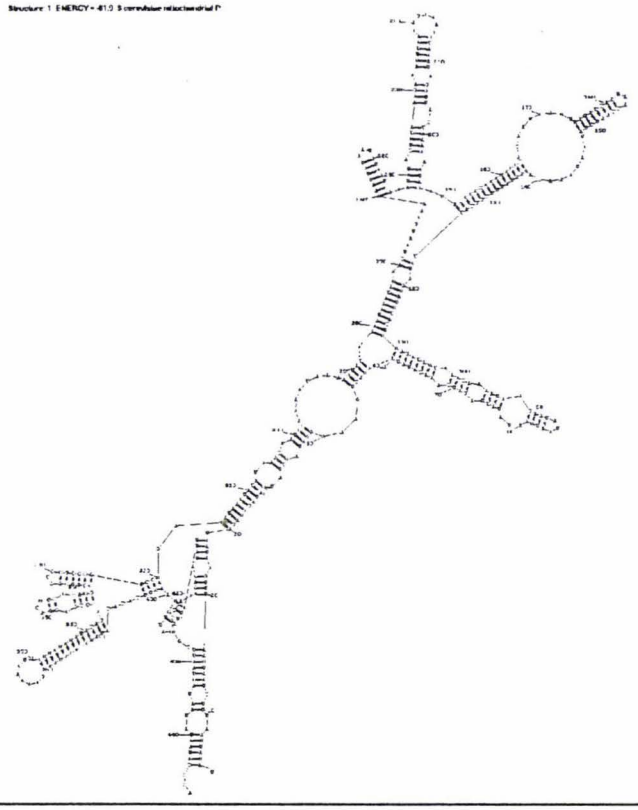
Aspergillus nidulans mitochondrial pRNA - mfold (RNAstructure)



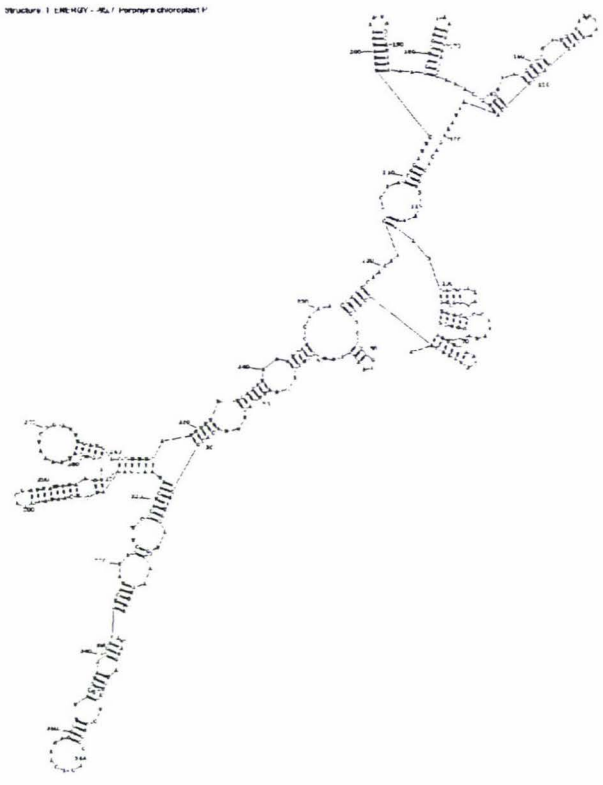
Aspergillus nidulans mitochondrial pRNA -mfold (RNAstructure) 5' - 3' corrected



Reclinomonas americana mitochondrial pRNA - mfold (RNAstructure)

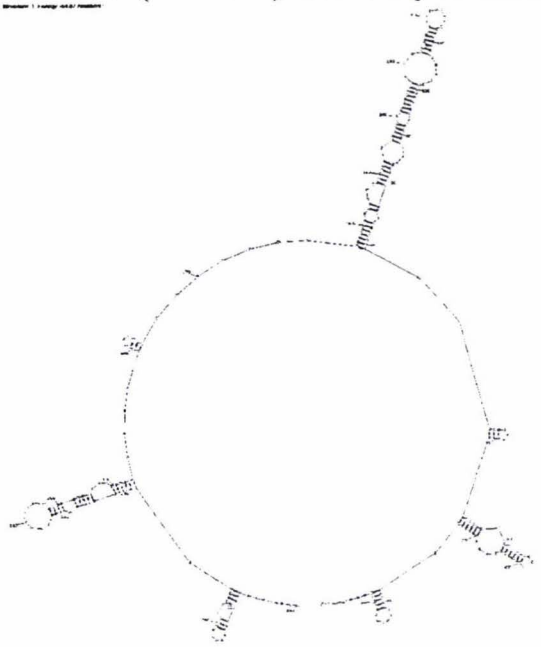


Saccharomyces cerevisiae mitochondrial pRNA - mfold (RNAstructure)

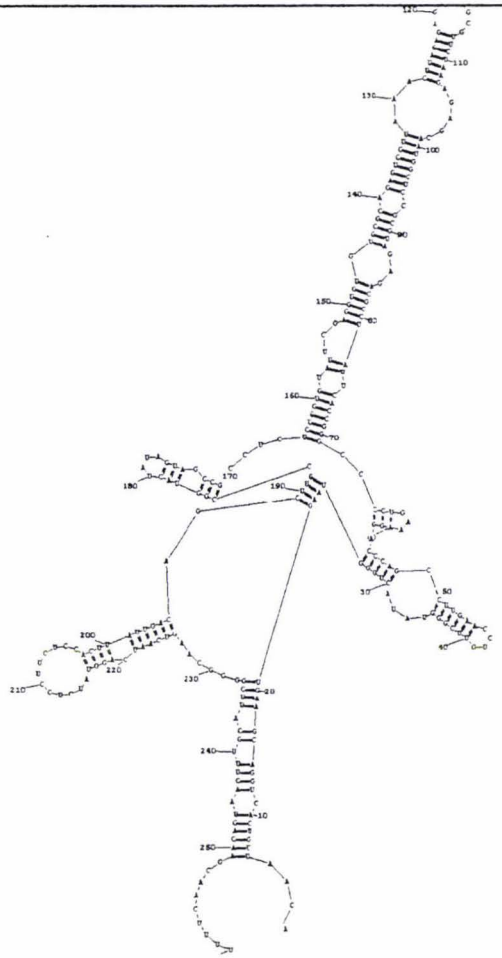


Porphyra purpurea chloroplast pRNA - mfold (RNAstructure)

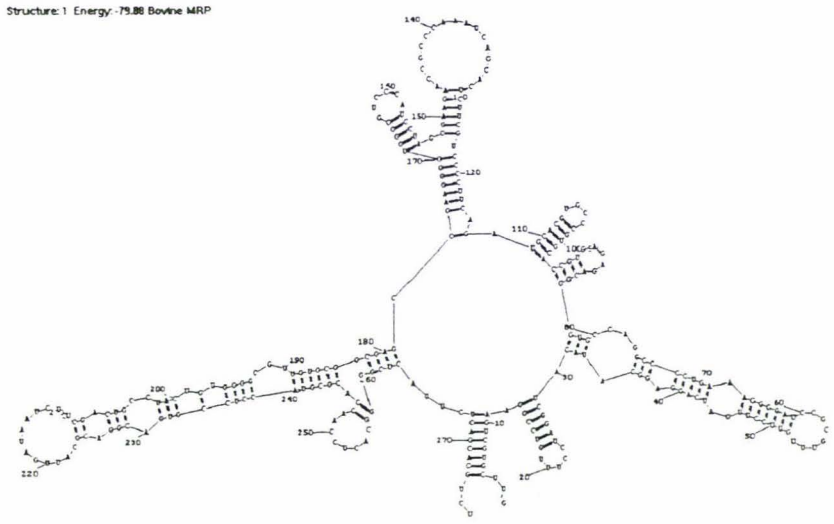
Appendix 1E: RNAdraw (RNAfold) secondary structures of mrpRNA



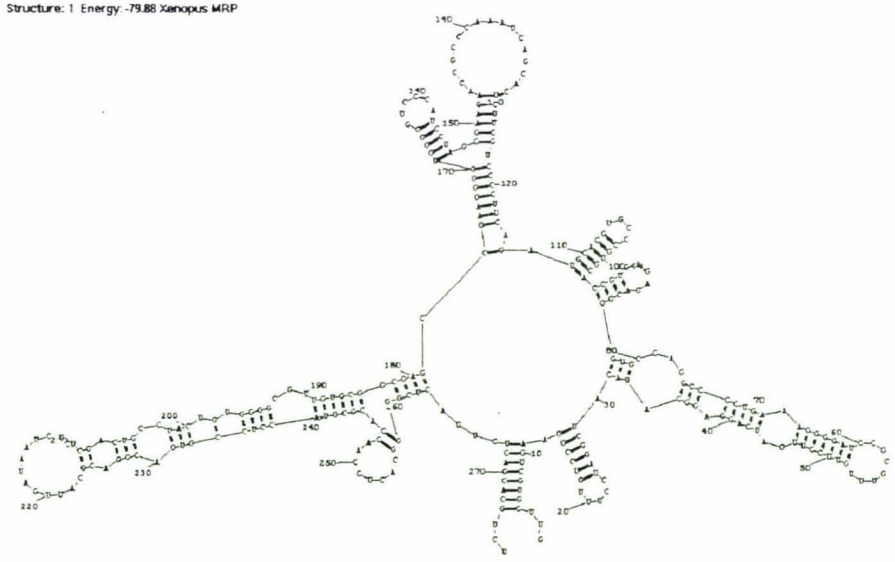
Arabidopsis mrpRNA - RNAfold



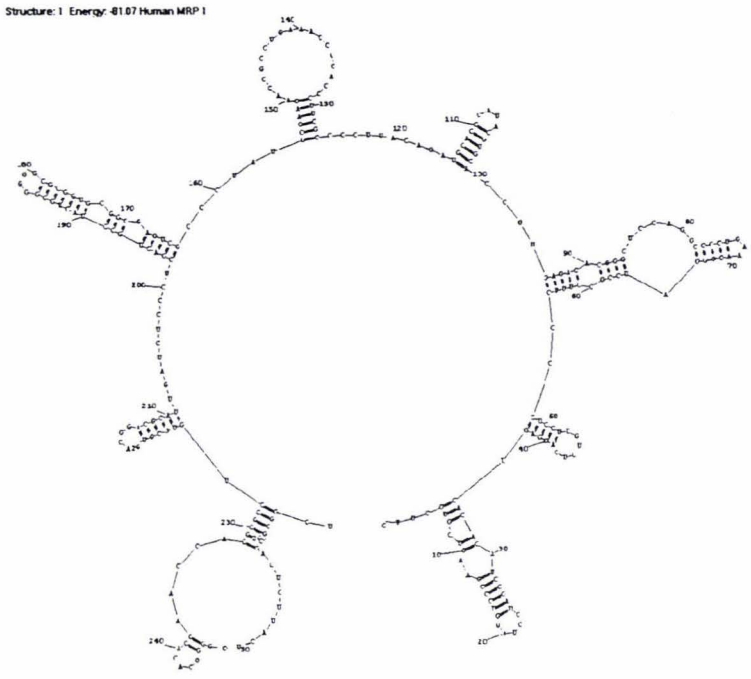
Arabidopsis mrpRNA - RNAfold
5' - 3' corrected



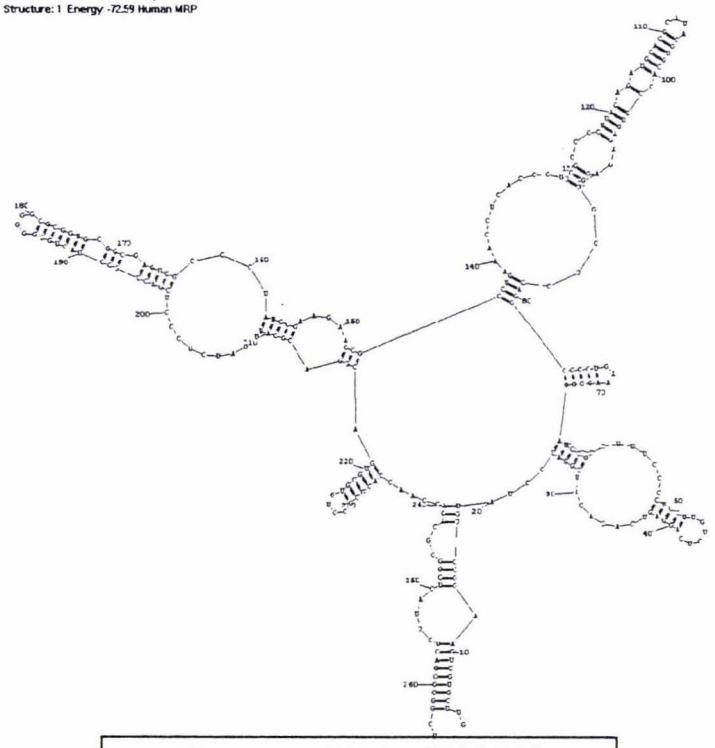
Bovine mrpRNA - RNAfold



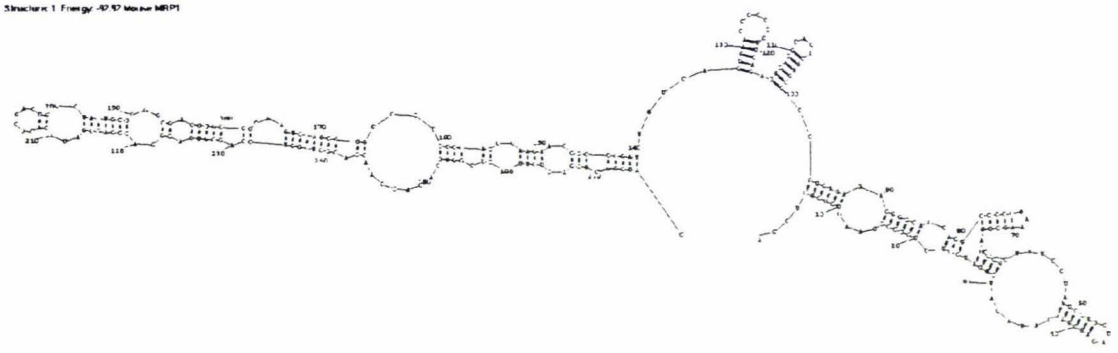
Xenopus mrpRNA - RNAfold



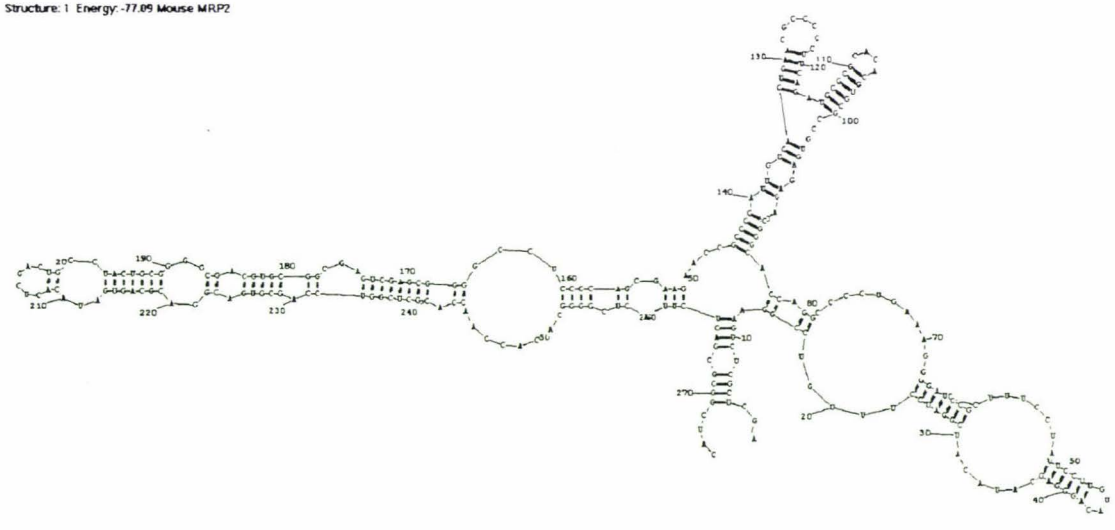
Human mrpRNA - RNAfold



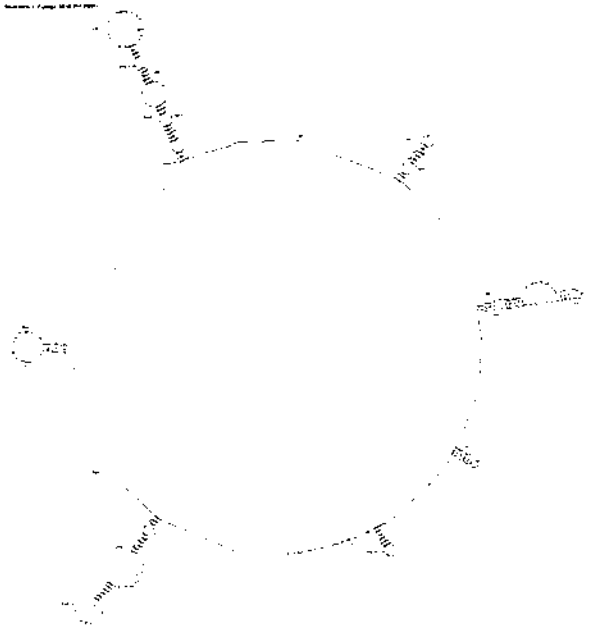
Human mrpRNA - RNAfold
5' - 3' corrected



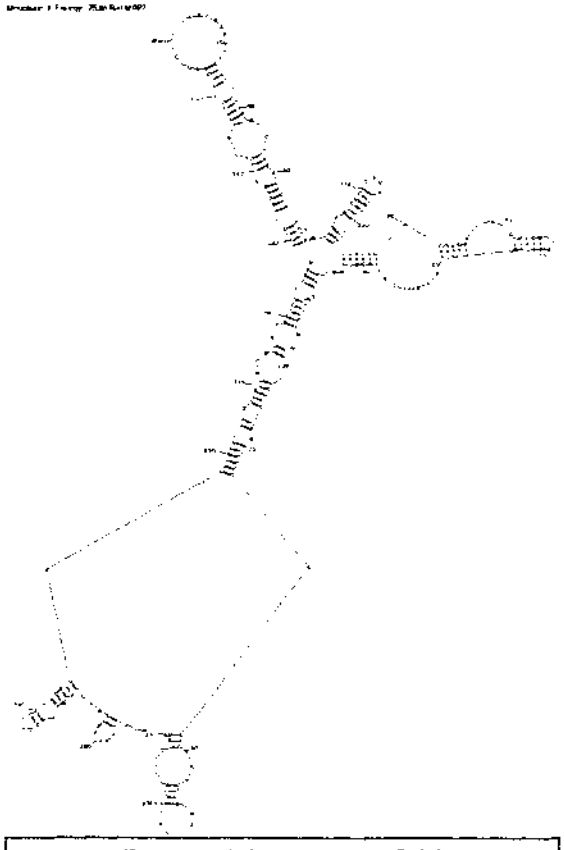
Mouse mrpRNA - RNAfold



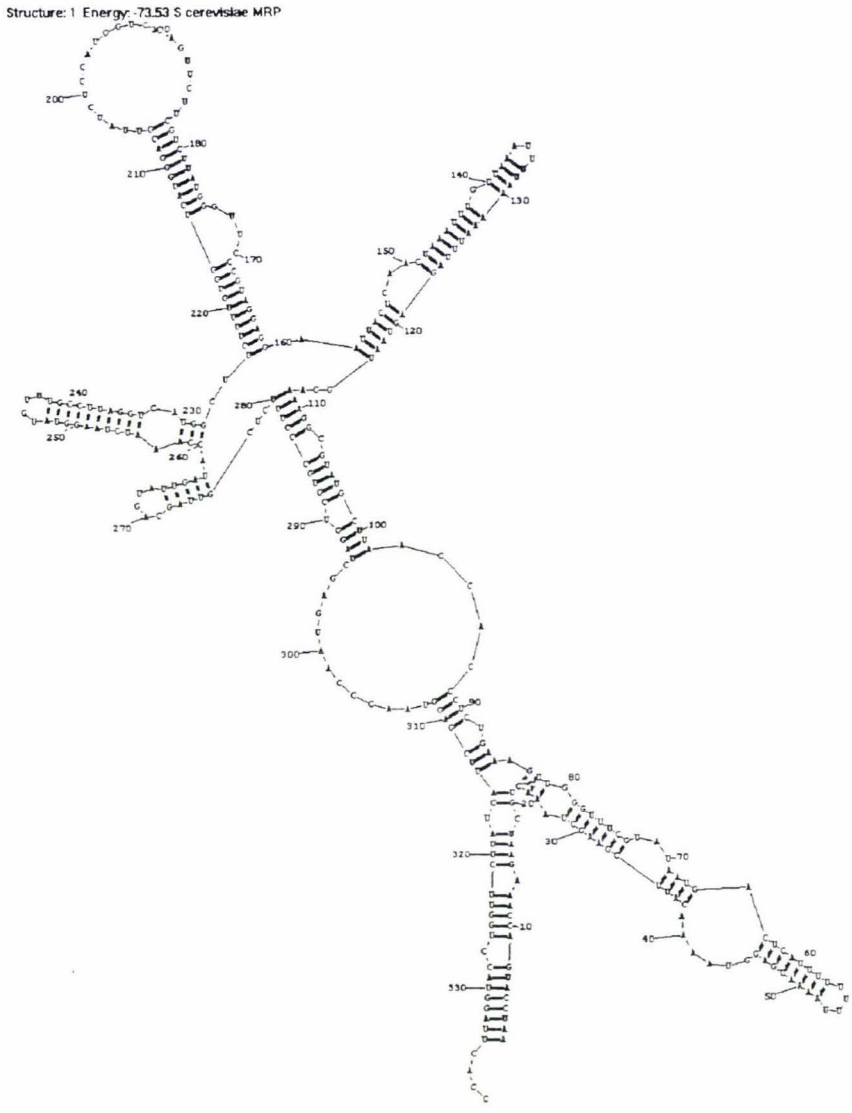
Mouse mrpRNA - RNAfold
5' - 3' corrected



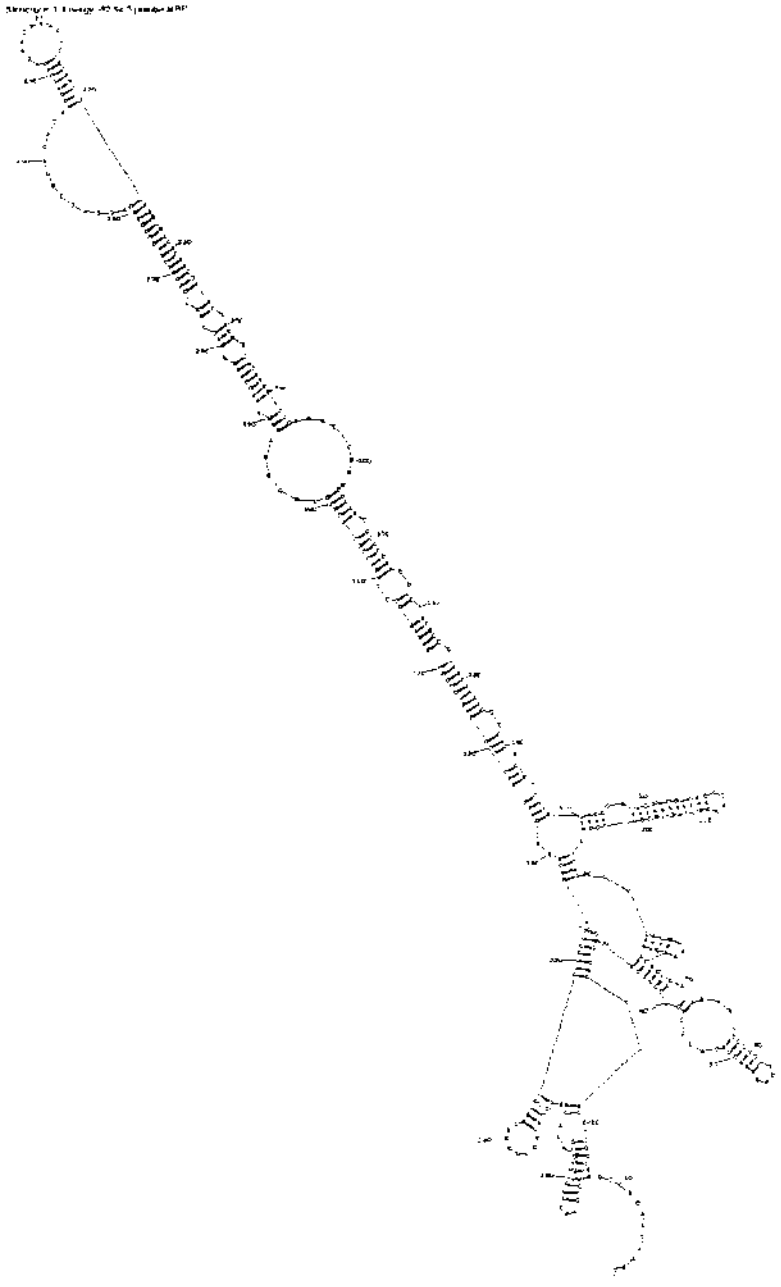
Rat mrpRNA - RNAfold



Rat mrpRNA - RNAfold
5' - 3' corrected

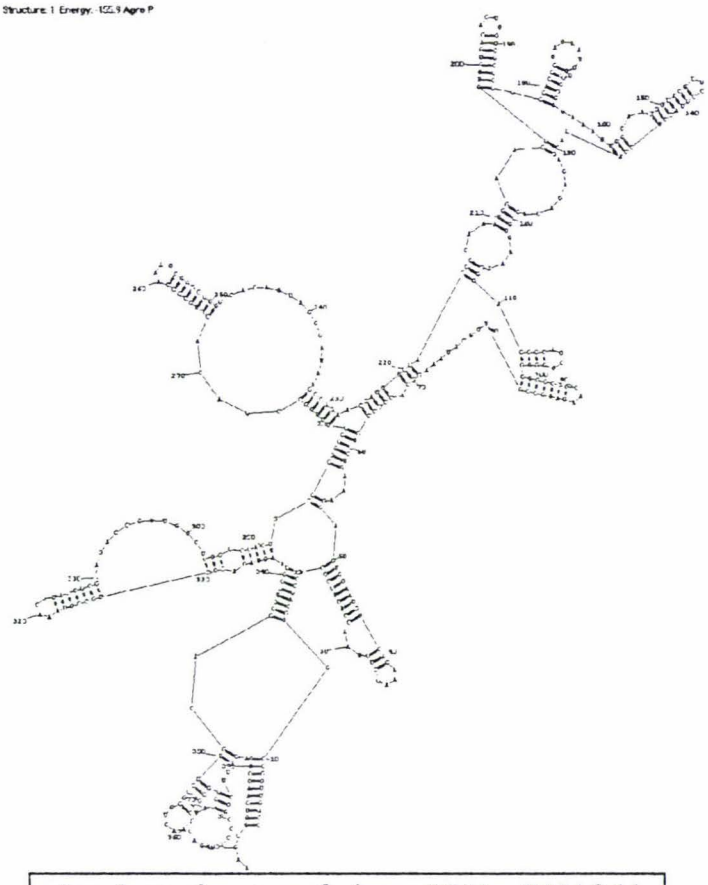


Saccharomyces cerevisiae mrpRNA - RNAfold

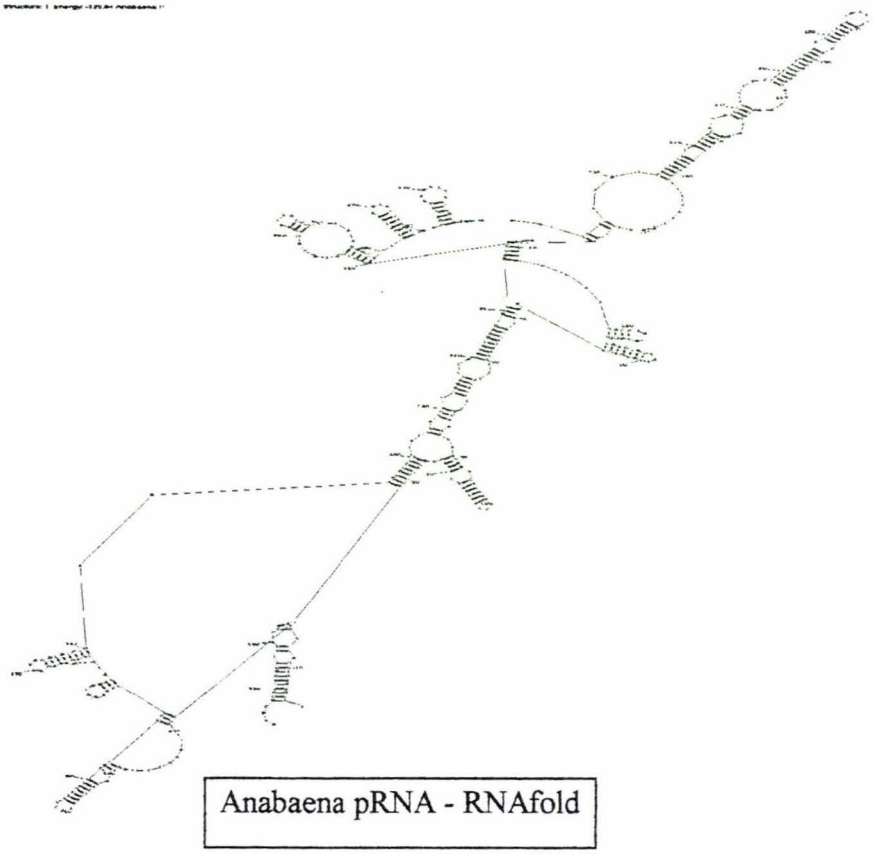


Schizosaccharomyces pombe mrpRNA - RNAfold

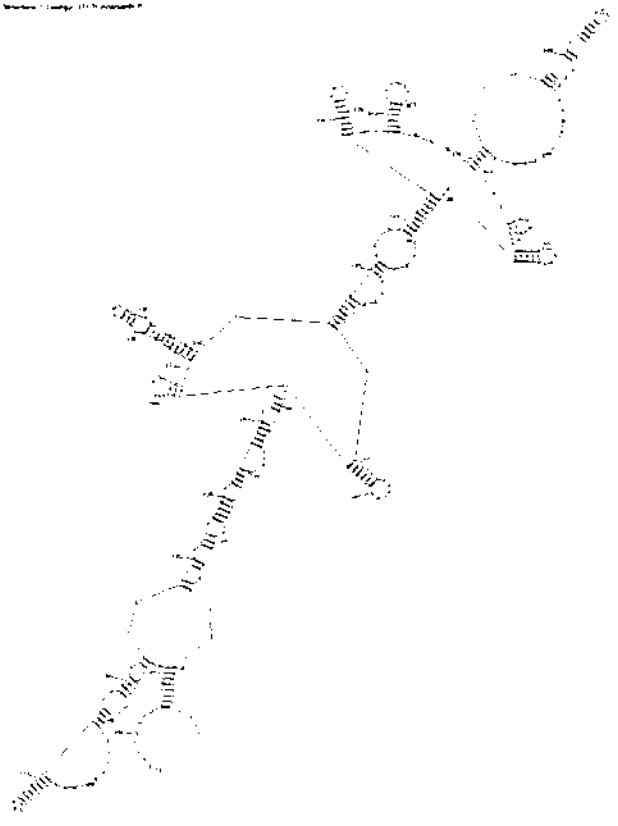
Appendix 1F: RNAdraw (RNAfold) secondary structures of pRNA



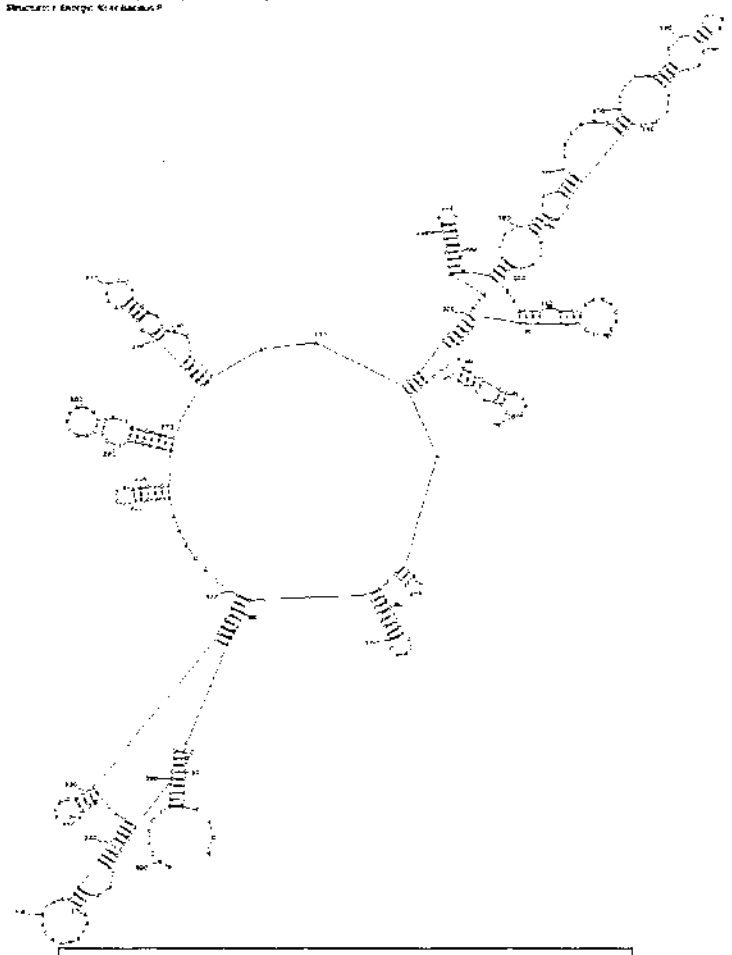
Agrobacterium tumefaciens pRNA - RNAfold



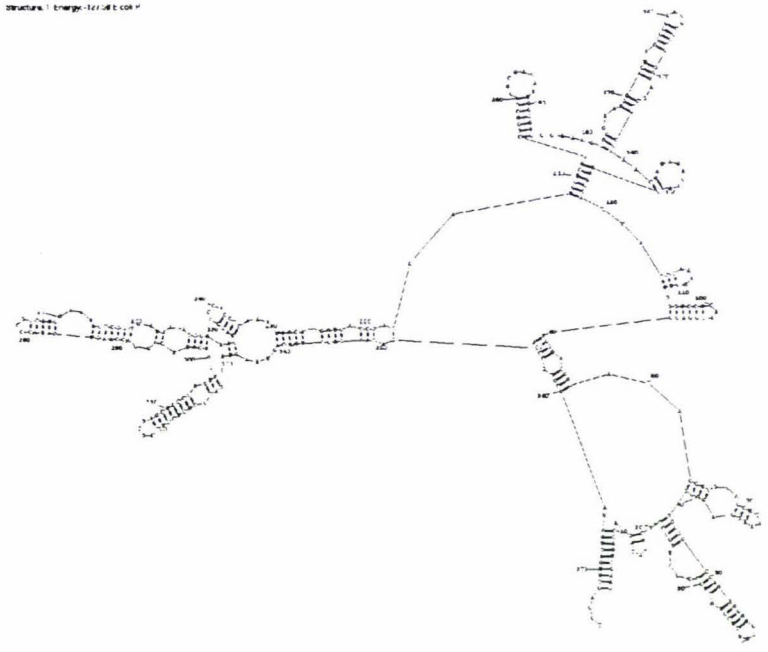
Anabaena pRNA - RNAfold



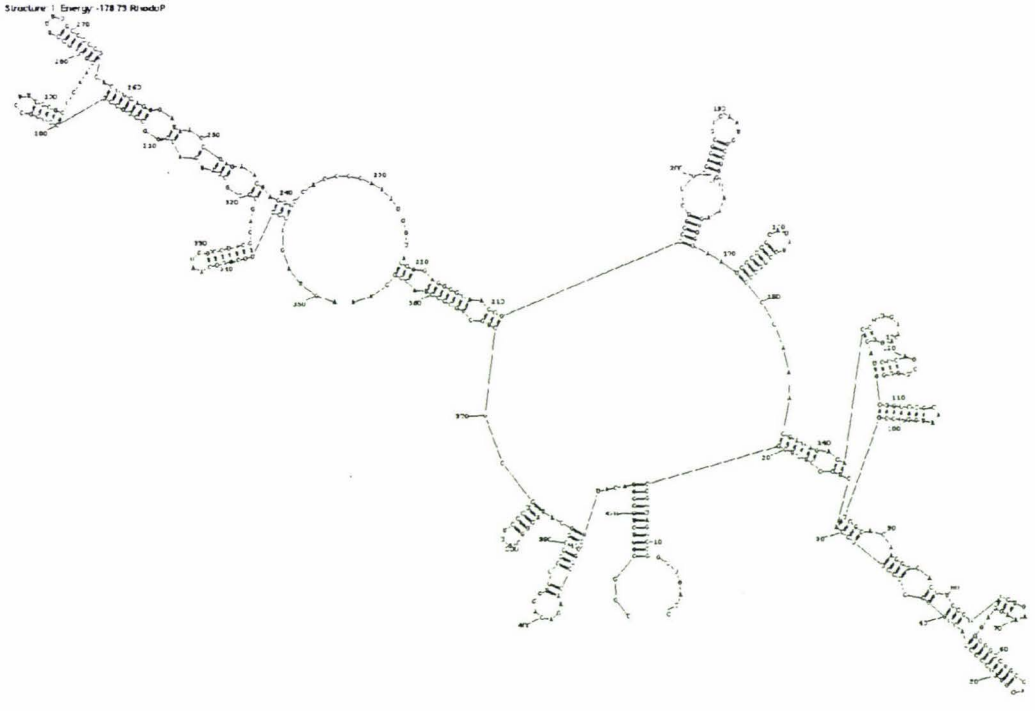
Anacystis pRNA - RNAfold



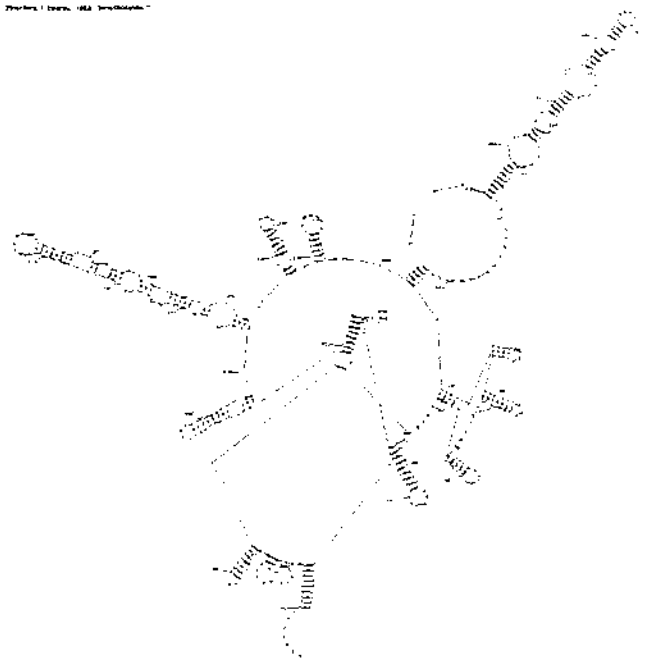
Bacillus subtilis pRNA - RNAfold



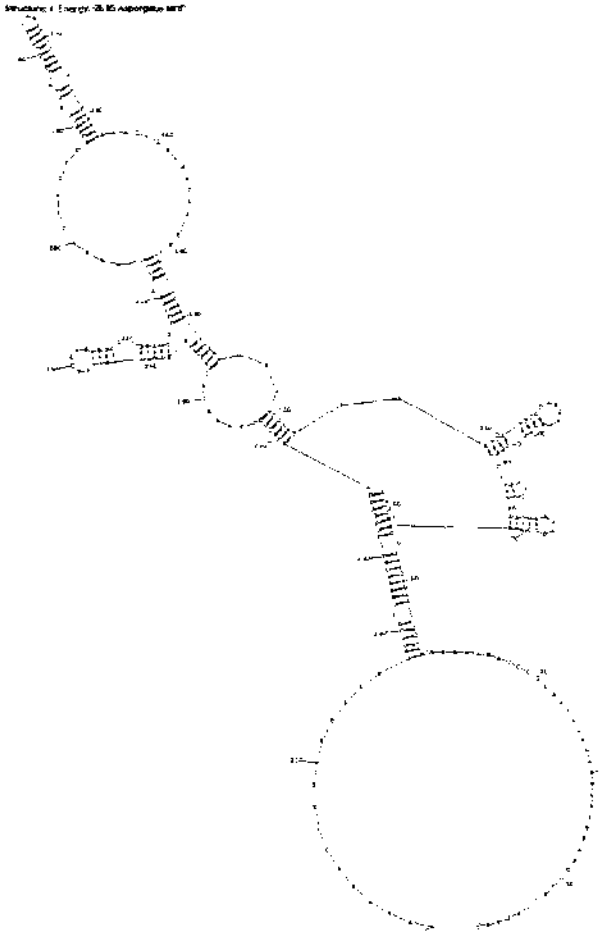
E. coli pRNA - RNAfold



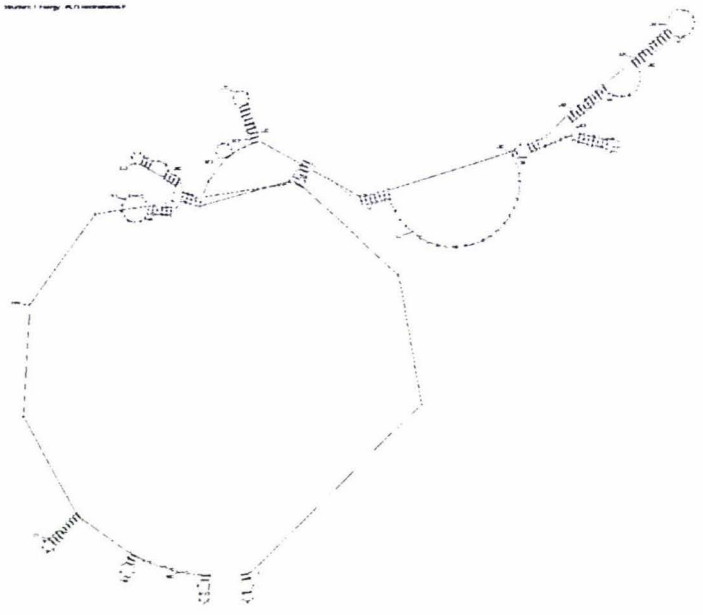
Rhodospirillum rubrum pRNA - RNAfold



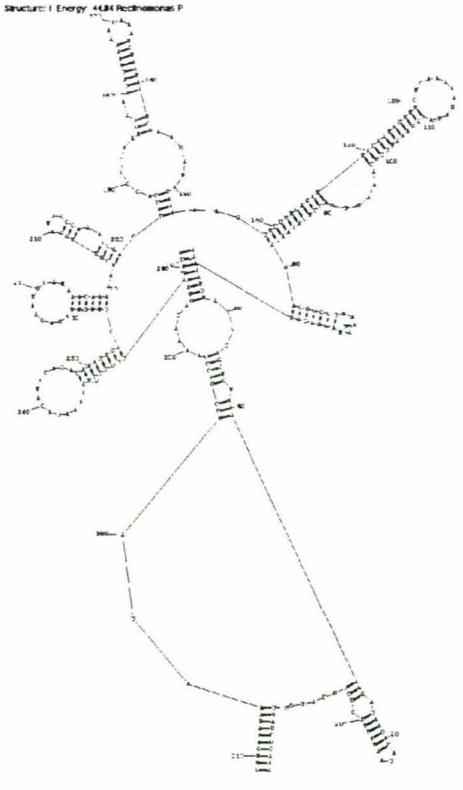
Synechocystis pRNA - RNAfold



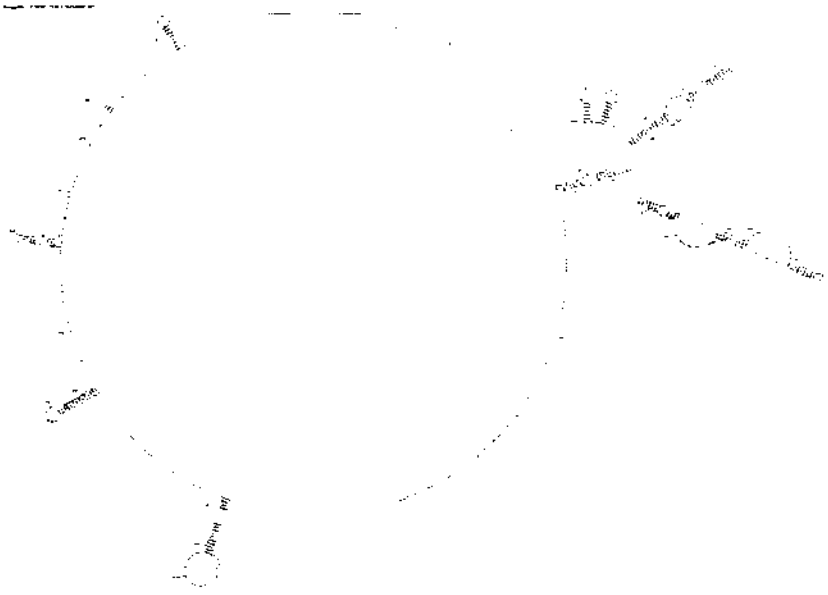
Aspergillus mitochondrial pRNA - RNAfold



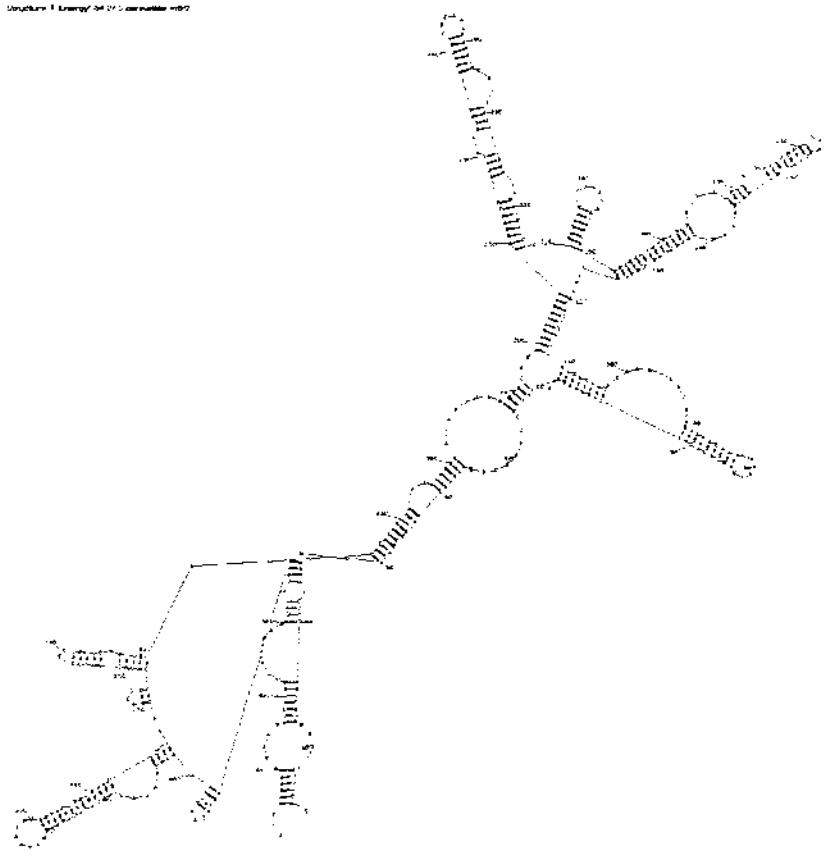
Reclinomonas americana mitochondrial pRNA - RNAfold



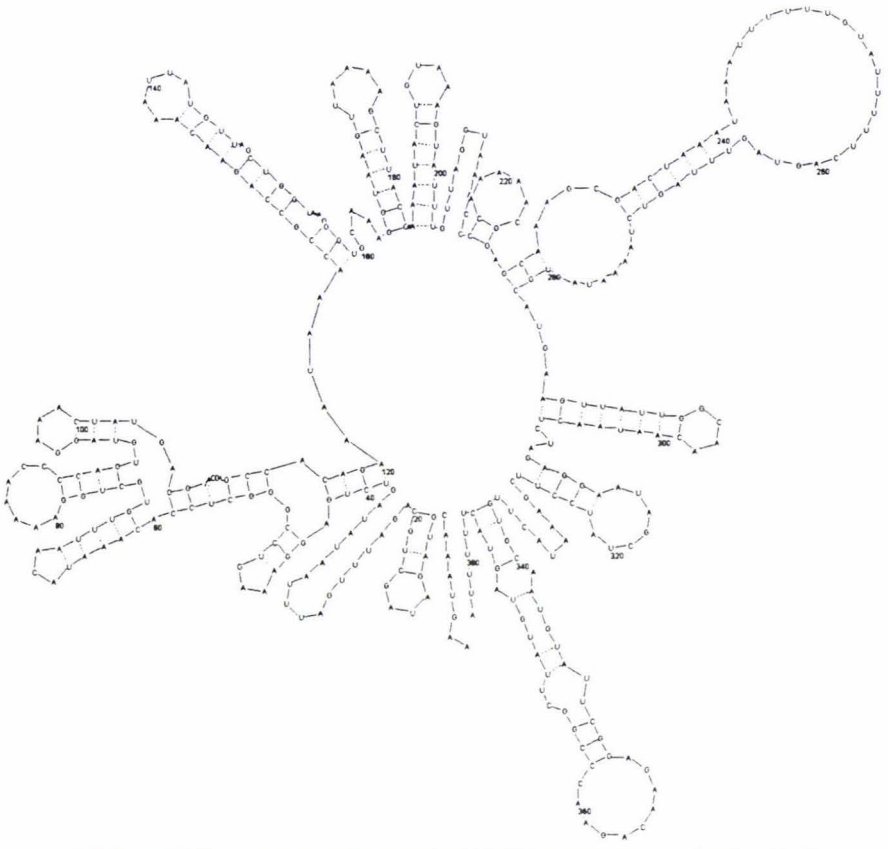
Reclinomonas americana mitochondrial pRNA - RNAfold
5' - 3' corrected



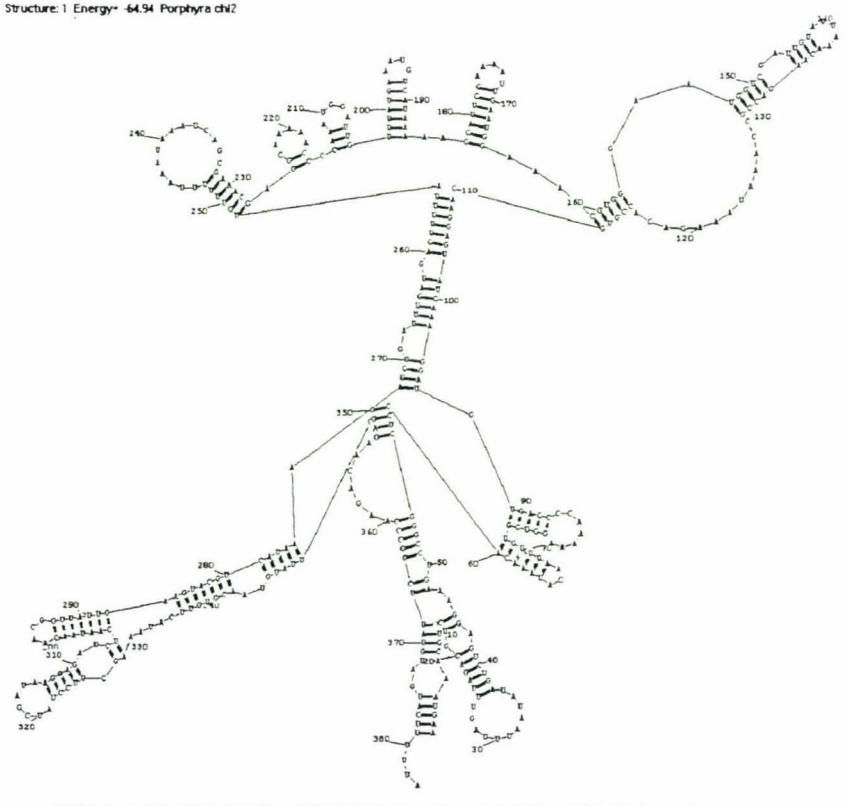
Saccharomyces cerevisiae mitochondrial pRNA - RNAfold



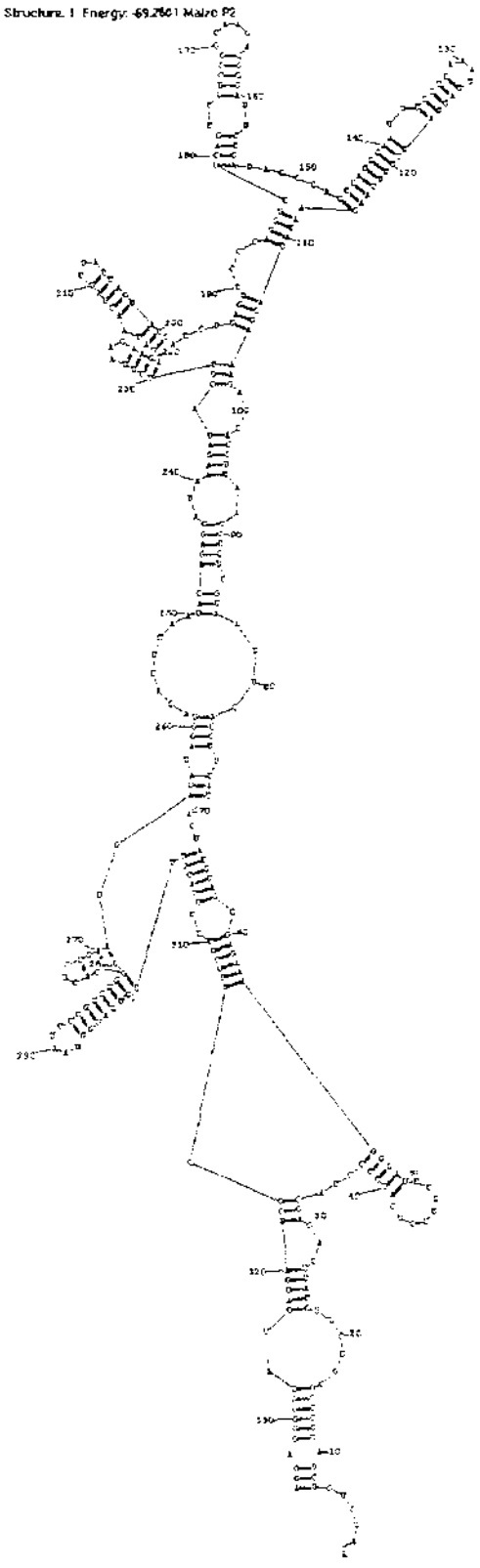
Saccharomyces cerevisiae mitochondrial pRNA - RNAfold
5' - 3' corrected



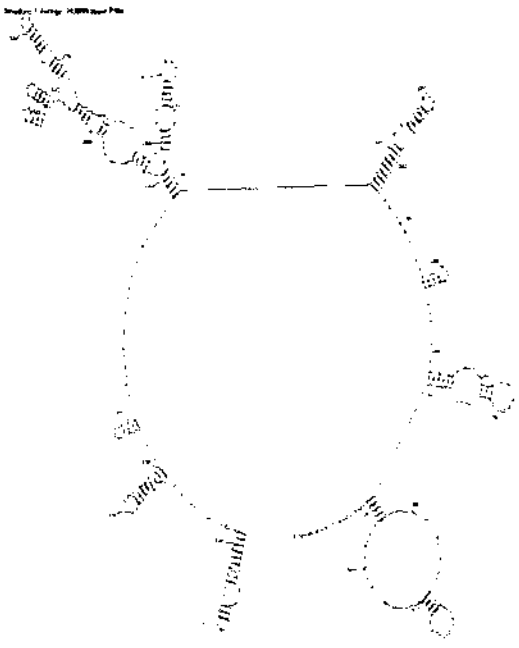
Porphyra purpurea chloroplast pRNA - RNAfold



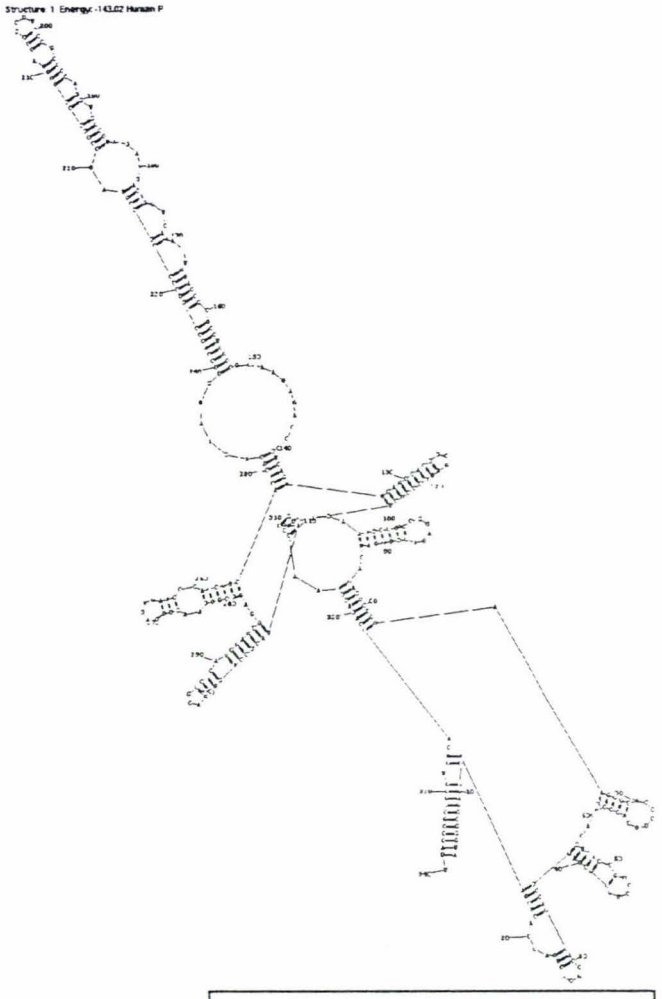
Porphyra purpurea chloroplast pRNA - RNAfold
5' - 3' corrected



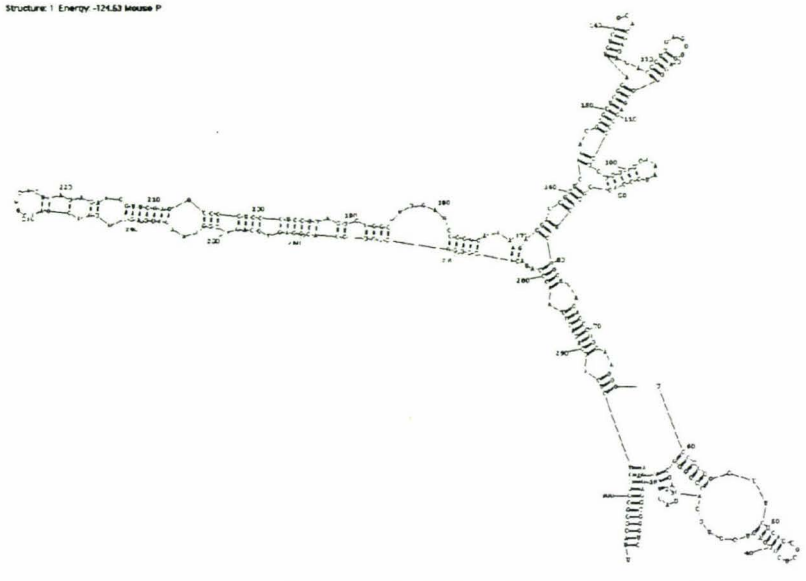
Maize chloroplast P-like - RNAfold
5' - 3' corrected



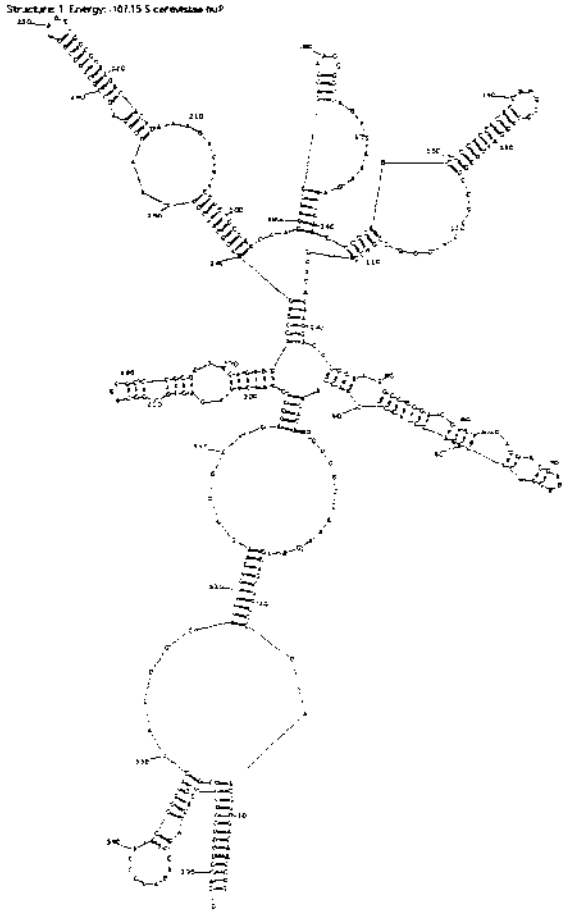
Maize chloroplast P-like -RNAfold



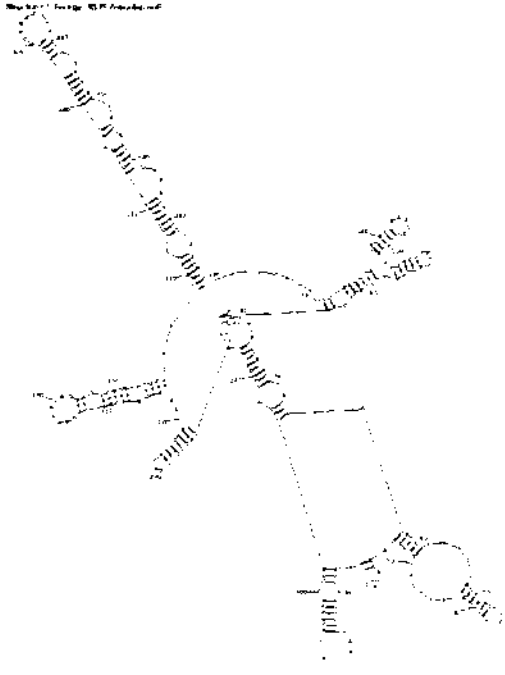
Human nuclear pRNA - RNAfold



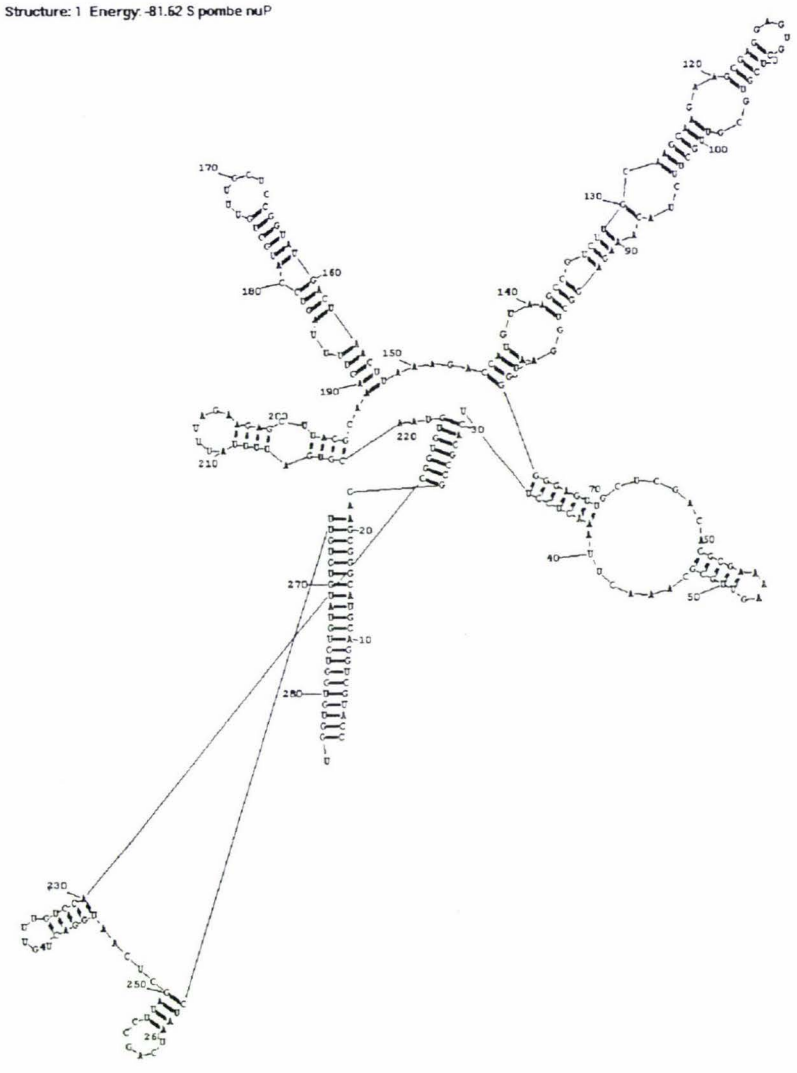
Mouse nuclear pRNA - RNAfold



Saccharomyces cerevisiae nuclear pRNA - RNAfold



Zebrafish nuclear pRNA - RNAfold



Schizosaccharomyces pombe nuclear pRNA - RNAfold

Appendix 2: Bracket Notation

Human mrpRNA

GUUCGUGCUGAAGGCCUGUAUCCUAGGCUACACACUGAGGACUCUGUUCCUCCCCUUUCCGCCUAGGGGA
AAGUCCCCGGACCUCGGGCAGAGAGUGCCACGUGCAUACGCACGUAGACAUUCCCCGCUUCCCACUCCAA
AGUCCGCCAAGAAGCGUAUCCCGCUGAGCGGCGUGGCGCGGGGGCGUCAUCCGUCAGCUCCCUCUAGUUA
CGCAGGCAGUGCGUGUCCGCGCACCAACCACACGGGGCUCAUUCUCAGCGCGGCU

Biological secondary structure
...((((.((((..((((((..(((((((.......((((((....))))))........)))))))........((((........((((.....))))((((((....)))))...(((.
..((((((....(((........((((.....(((..((((...))))...)))..)))).)))..)))))).)))...................)))))).((((((......)))))
).........)))).......)))).))))....

RNAstructure (Mfold) secondary structure
.(((((((((..(((((..(((.((((........(((((......)))))........))).((((((...))))))))...))))).(((...((((((((((....)))))).((
(..(((((((...((((.......(.((((.....((((.....))))..)))))))))).)))))))).)))..((((..((............)).((..((((((....))))))...
))..))))))))...)))))))))))))..

RNAdraw (RNAfold) secondary structure
.((((...))))((((((.....))))))........(((((......)))))..(((((.((((.(((((...)))))........)))).)))))....((((((....)))))).(
((..(((((((...((((..........(((....((((.....))))....))))))).))))))).)))..................((((((.....)))))).((((((.....(
(.....))...........))))))..

Human nuclear pRNA

AUGGGCGGAGGGAAGCUCAUCAGUGGGGCCACGAGCUGAGUGCGUCCUGUCACUCCACUCCCAUGUCCCU
UGGGAAGGUCUGAGACUAGGGCCAGAGGCGGCCCUAACAGGGCUCUCCCUGAGCUUCAGGGAGGUGAGUU
CCCAGAGAACGGGGCUCCGCGCGAGGUCAGACUGGGCAGGAGAUGCCGUGGACCCCGCCCUUCGGGGAGG
GGCCCGGCGGAUGCCUCCUUUGCCGGAGCUUGGAACAGACUCACGGCCAGCGAAGUGAGUUCAAUGGCUG
AGGUGAGGUACCCCGCAGGGGACCUCAUAACCCAAUUCAGACCACUCUCCUCCGCCCAUU

Biological secondary structure
((((((((((((..((((((.(.((((((.......(((((.....)))))...))))))).))...(((((((...(((((.(((.(((((((......)))))..........((
(((......)))))........(((.(((...........................))).)))..........))))).)))))....((((...))))....(((((.((........)).)))
))........)))).((((((.((....)))))))).........)))))).....)))))))))))).

RNAstructure (Mfold) secondary structure
((((((((((((.((((((.....(((((...(((..((((((.....).)))))..)))....))))).)))).((((((((..(((((((......)))))))...(((.((
((((((((...)))))))))((((((.(((........((((((.((((((.......(((((.....(((((.((.((((.(((....))).)))))))))))..))))).)))
)))))))))))))....))))))((((((..(((.....)))..))))))..((((((((.((((...))))))))))))..))).)))))))).)).))))))))))))).

RNAdraw (RNAfold) secondary structure
((((((((((((.(((((....((...))...)))))((((((........))))))...((((((.......)))))).((((((((.....((((((...))((((((...))))))((
(((((((...))))))))((((((((...........((((((((..((((((...((...((((.....((((..((.((((.(((....))).)))).)))))))..))))))))))))
).)))))))......))))))))(((((..(((.....)))..)))))(..((((((((.((((...))))))))))))..))).))))))))....))))))))))))).

E. coli pRNA

GAAGCUGACCAGACAGUCGCCGCUUCGUCGUCGUCCUCUUCGGGGGAGACGGGCGGAGGGGAGGAAAGUC
CGGGCUCCAUAGGGCAGGGUGCCAGGUAACGCCUGGGGGGGAAACCCACGACCAGUGCAACAGAGAGCAA
ACCGCCGAUGGCCCGCGCAAGCGGGAUCAGGUAAGGGUGAAAGGGUGCGGUAAGAGCGCACCGCGCGGCU
GGUAACAGUCCGUGGCACGGUAAACUCCACCCGGAGCAAGGCCAAAUAGGGGUUCAUAAGGUACGGCCCG
UACUGAACCCGGGUAGGCUGCUUGAGCCAGUGAGCGAUUGCUGGCCUAGAUGAAUGACUGUCCACGACAG
AACCCGGCUUAUCGGUCAGUUUCACCU

Biological secondary structure

((((((((((((((((.(((((((((...(((....)))))))))))....(((.(((((((((.....(((((((((.....)))))(((((....)))).)((
...((.............(((((((((...)))))))..)).....(((((.....)))))).(((((((....)))))))).....)))))))))))...((((..
....(((((((...(((((((....)))))))...(((.....)))..)))))))....)))).....(((((((....)))))))).....)))))))..(((((
((.....)))))).....)))))))..)))))))))))....

RNAstructure (Mfold) secondary structure
((((((((((.((...)).((.((((((((((.(((((....)))))))).)))))))..)).((.(((.(((((..((((..(((.((.(((...((...((((((.(((....
)))....)))).)).))).....)))..)))))).))))(((((....))))).(((((((..(((((...(((((((.....)))(((((((((((....))).)))))...)))
)..)))))))))..(((...(((.....(((((((....((((((...))))))))))).)))..)))))))(((((((....)))))))).((.....)).(((((...
)))))..)))))))).))))))))))....

RNAdraw (RNAfold) secondary structure

(((((((((.....)))).))))).(((((.(((((...))))).)))))..............(((((((.....(((...))))).....)))))))(((....))).
.....(((.....))).(((...(((..(((((....)))))))).))).((((((((..(((((((.....)))))).(((.((((((..(((((((((.....)))
)))..)))(((..(((.....(((((((....((((((...))))))))))).....)))))))).))))))..)))...(((((((..((((....((((...))))).....
....)))))))))))))))))))

Appendix 3: Input matrices for Neighbor

The Neighbor program has been developed for use with sequence alignment data, especially that developed with other programs in the Phylip package, such as DNAdist. It was found that the matrix output from the RNAdistance program gave strange results when directly input into the Neighbor program. Often taxa would be missing and some taxa had extremely short branch lengths when the data suggested otherwise. A modification of the RNAdistance matrices was required for consistent Neighbor results.

An example is given below for the mrpRNA sequences and biological secondary structures.

Matrix output from DNAdist program

```
 8
Humanmrp
Bovinemrp    0.2264
Mousemrp     0.2161  0.2096
Ratmrp       0.2018  0.2105  0.0650
Xenopmrp     0.2264  0.0000  0.2096  0.2105
Arabmrp      1.2195  1.1386  1.1723  1.1570  1.1386
Sceremrp     1.1629  1.0571  1.2255  1.1575  1.0571  1.3947
Spombemrp    1.2866  1.1028  1.2444  1.1748  1.1028  1.6714  1.1826
```

```
Original matrix output from RNAdistance

> f   8
42
40 34
45 31 9
35 61 51 56
94 100 92 97 101
110 140 132 139 117 148
159 151 167 160 154 193 205
```

```
Modified matrix output from RNAdistance
   8
HumanMRP
BovineMRP  0.41
MouseMRP   0.39 0.34
RatMRP     0.44 0.31 0.09
XenopusMRP 0.34 0.61 0.51 0.56
ArabMRP    0.93 1.00 0.92 0.97 1.01
ScereMRP   1.11 1.40 1.32 1.39 1.17 1.48
SpombeMRP  1.58 1.51 1.67 1.60 1.54 1.93 2.05
```

Appendix 4: Computer Program Parameters

Divide and Conquer

Divide-and-Conquer Multiple Sequence Alignment (DCA) is a program for producing fast, high quality simultaneous multiple sequence alignments. The general idea of DCA is that each sequence is cut in two, behind a suitable cut position somewhere close to its midpoint. This way, the problem of aligning one family of (long) sequences is divided into the two problems of aligning two families of (shorter) sequences, the prefix and the suffix sequences. This procedure is re-iterated until the sequences are sufficiently short - say, shorter than a pregiven stop size L, so that they can be aligned optimally. Finally, the resulting short alignments are concatenated, yielding a multiple alignment of the original sequences.

Parameters:

Substitution matrix: The matrix used in this study was the following called 'RNA' which was optimised for working with the pRNA and mrpRNA sequences.

```
6
- -  0
A A  0
G G  0
U U  0
C C  0
- A  2
- G  2
- U  2
- C  2
A G  1
A U  2
A C  2
G U  2
G C  2
U C  1
```

Gap parameters: The cost of a gap is computed by the formula where is the "gap initiation cost" and is the "gap extension cost". The default settings were used in his study.

Free shift: DCA by default does not penalise gaps at either end of the sequences to make the compensation of differences in the length of the sequences free of charge. This free shift option can be deactivated as was in this study.

Approximate cut positions/FDCA: The most time consuming phase of the search for cut positions can be deactivated so that the sequences are cut at approximate slicing positions, which generally yields slightly less accurate but often much faster alignments. The optimal splicing sites were used for this study.

Recursion stop size: The recursion stop size can be set to any number. Too large an L can result in very long running times and very big memory usage. Too small can result in empty subsequences at the end of the iteration which may lead to bad alignments. The stopsize was set to 20 for this study.

The program is available from: http://bibiserv.TechFak.Uni-Bielefeld.DE/dca/

References for this program:

Dress, A.W.M. Füllen G., Perrey S. W. (1995) A Divide and Conquer Approach to Multiple Alignment. Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB 95), AAAI Press, Menlo Park, CA, USA, 107-113.

Tönges, U. Perrey, S. W. Stoye, J. Dress A.W.M. (1996) A General Method for Fast Multiple Sequence Alignment Gene 172:GC33-GC41.

Perrey, S. W. and Stoye J. (May 1996) Fast Approximation to the NP-hard Problem of Multiple Sequence Alignment. Information and Mathematical Sciences Reports, Series B: 96/06 (ISSN 1171-7637).

Stoye, J. Perrey, S. W. Dress A.W.M. (1997) Improving the Divide-and-Conquer Approach to Sum-of-Pairs Multiple Sequence Alignment. Appl. Math. Lett. 10 67-73.

Stoye, J. (1997) Divide-and-Conquer Multiple Sequence Alignment Dissertation Thesis. Universität Bielefeld, Forschungsbericht der Technischen Fakultät, Abteilung Informationstechnik, Report 97-02. (ISSN 0946-7831).

Dialign

DIALIGN is a novel alignment method developed by Burkhard Morgenstern et al. (1996). While standard alignment methods rely on comparing single residues and imposing gap penalties, DIALIGN constructs alignments by comparing whole segments of the sequences. No gap penalty is employed. This point of view is especially adequate if sequences are not globally related but share only local similarities as is the case in genomic DNA sequences and in many protein families and what was shown with the mrpRNA sequences. Burkhard Morgenstern using the default parameters for this program kindly did the alignments.

This program is available at:

http://bibiserv.TechFak.Uni-Bielefeld.DE/dialign

Reference for this program:

Morgenstern, B. Dress, A. and Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. USA 93: 12098 - 12103

ClustalX (v1.64b)

ClustalX is a new windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments, and analysing the results. A distance is calculated between every pair of sequences and these are used to construct the phylogenetic tree that guides the final multiple alignment. The scores are calculated from separate pairwise alignments. These can be calculated using two methods: dynamic programming (slow but accurate) or by the method of Wilbur and Lipman (extremely fast but approximate). The dynamic programming option was used in this study. All parameters were used at the default settings:

This program can be accessed from:

http://www.hgmp.mrc.ac.uk/Menu/Help/clustalx.html#T

Reference for this program:

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876–4882.

Phylip package (DNAdist and Neighbor)

This is a free package of programs for inferring phylogenies and carrying out certain related tasks. At present it contains 30 programs, which carry out different algorithms on different kinds of data. (c) Copyright 1986-1995 by Joseph Felsenstein and the University of Washington.

DNADIST. Computes four different distances between species from nucleic acid sequences. The distances can then be used in the distance matrix programs. The distances used in this study are the Jukes-Cantor formula, and one based on Kimura's 2-parameter method.

NEIGHBOR. An implementation by Mary Kuhner and John Yamato of Saitou and Nei's "Neighbor Joining Method," and of the UPGMA (Average Linkage clustering) method. Neighbor Joining is a distance matrix method producing an unrooted tree without the assumption of a clock. The neighbor-joining method only was used with the default settings.

The Vienna RNA package

The Vienna RNA Package consists of a code library and several stand-alone programs for the prediction and comparison of RNA secondary structures.
This package can be downloaded from:
Http://www.tbi.univie.ac.at/

RNAfold:
RNAfold reads RNA sequences and calculates their minimum free energy (mfe) structure, partition function (pf) and base pairing probability matrix. It returns the mfe structure in bracket notation, its energy, the free energy of the thermodynamic ensemble and the frequency of the mfe structure in the ensemble to stdout. The calculation of mfe structures is based on dynamic programming algorithm originally developed by M. Zuker and P. Stiegler. The partition function algorithm is based on work by J.S. McCaskill.
Parameters:
-Ttemp Rescale energy parameters to a temperature of temp C. The default is 37C which was used in this study.
References for this program:

I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster (1994), Fast Folding and Comparison of RNA Secondary Structures, Monatshefte f. Chemie 125:167-188.

M. Zuker, P. Stiegler (1981) Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information, Nucl. Acid Res 9: 133-148.

RNAdistance:
RNAdistance reads RNA secondary structures and calculates one or more measures for their dissimilarity, based on tree editing or string alignment. This study only used the tree editing option. It accepts structures in bracket format, where

matching brackets symbolise base pairs and unpaired bases are represented by a dot '.', or coarse grained representations where hairpins, interior loops, bulges, multiloops, stacks and external bases are represented by (H), (I), (B), (M), (S), and (E), respectively. These can be optionally weighted. Full structures can be represented in the same fashion using the identifiers (U) and (P) for unpaired and paired bases, respectively. We call this the HIT representation (Homeomorphically irreducible tree). For example the following structure consists of two hairpins joined by a multiloop:

.((..(((...)))..((..)))).	full structure (usual format);
(U)((U2)((U3)P3)(U2)((U2)P2)P2)	HIT structure;
((H)(H)M) or ((((H)S)((H)S)M)S)	coarse grained structure;
(((((H3)S3)((H2)S2)M4)S2)E2)	weighted coarse grained.

Parameters:

The aligned structures are written to file, if specified.

-D[fhwcFHWC] use the full, HIT, weighted coarse, or coarse representation to calculate the distance. Capital letters indicate string alignment otherwise tree editing is used. Any combination of distances can be specified. The default if 'f'.

-Xm calculate the distance matrix between all structures. The output is formatted as a lower triangle matrix.

-B [file] Print an "alignment" with gaps of the structures, to show matching substructures.

References for this program:

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

Shapiro B A, Zhang K (1990) Comparing multiple RNA secondary structures using tree comparison, CABIOS 6: 309-318.

Fontana W, Konings D A M, Stadler P F, Schuster P, (1993) Statistics of RNA secondary structures, Biopolymers 33:1389-1404.

I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster (1994) Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie 125: 167-188.

TreeView(Win32) (v1.40)

TreeView is a simple program for displaying phylogenies on Apple Macintosh and Windows PCs. TreeView allows you to create publication quality trees from PAUP files, either directly, or by generating graphics files for editing by other programs. The program currently reads trees with up to 500 taxa.

TreeView comes in four versions, one for standard Macs, one for Power Macs and two for Windows (16 and 32 bit). The Windows versions require either Windows 3.1 or later (Win16) or Windows 95 or Windows NT (Win32). TreeView is free, but please register your copy.
The program is available at:

http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Reference for this program:

Page, R. D. M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

RNAstructure and Mfold

RNAstructure was written by David H. Mathews and Mark E. Burkard for Isis Pharmaceuticals and is made available for the RNA community by Isis.
It is available on the Turner Lab Homepage at: http://rna.chem.rochester.edu.

The other modern implementation of the algorithm is mfold, available on the World Wide Web at Michael Zuker's homepage at: **http://www.ibc.wustl.edu/~zuker**. The parameters described below are exactly the same when using the Mfold program.

RNAstructure

RNAstructure uses the Zuker algorithm (Zuker 1989) for predicting RNA secondary structure. This is based on free energy minimisation using the nearest neighbor parameters of Doug Turner and coworkers derived largely from optical melting experiments.

Predicting a structure is a two-step process. The first step is a recursive algorithm that generates an optimal structure and a series of structures that are called sub-optimal structures (because their free energy is not as favourable as the free energy of the optimal structure). The number of sub-optimal structures generated is controlled by two parameters entered by the user, maximum structures and percentage sort. A third parameter entered is Window Size

The second step is a re-ordering of the favourability of the structures. The energy of each structure is re-calculated using an energy function that includes coaxial

stacking of helixes and a Jacobson-Stockmayer function for determining the free energy of large loops. The output is then sorted according to the recalculated free energy.

Parameters:

Max % Energy Difference: Sets the percent deviation from the lowest free energy allowed for the structures output. For example if the lowest-free energy (DG037) structure is -100 kcal/mol, and the Max % Energy Difference is 10, any structures with an energy of -90 kcal/mol or greater is rejected (greater means less negative).

The max energy difference for this study was set at 10.

Max number of structures: Sets an absolute upper limit on the number of structures that can be generated. A maximum of 1000 structures can be generated, if you choose a larger value, 1000 will be used. The max No of structures for this study was set at 5.

Window size: This parameter controls how different the sub-optimal structures must be from each other. A small window size allows very similar structures to be generated while a larger window size requires them to be more different. The window size for this project was set at 7.

The temperature was set at the default of 37°C for this study.

Output from these programs is in the form of 'ct' files, which can be drawn into diagrammatic representations of the structures.

References for this program:

Jaeger, J. A. Turner D. H. and Zuker M. (1989) Improved Predictions of Secondary Structures for RNA. Proc. Natl. Acad. Sci. USA, Biochemistry, 86:7706-7710.

Mathews, D. H. Andre, T. C. Kim, J. Turner, D. H. and Zuker M. (1997b). An Updated Recursive Algorithm for RNA Secondary Structure Prediction with Improved Free Energy Parameters In Press.

Zuker M. (1989) On Finding All Suboptimal Foldings of an RNA Molecule. Science, 244: 48-52.

RNAdraw V1.1b

The optimal structure/basepair-probability matrix/heat curve calculation algorithms were all ported directly from the Vienna RNA package V1.1 (1).

Temperature is set at default of 37°C for this study. The pf-scale variable is used to scale the floating-point values generated during matrix calculation. If you get a floating-point error during calculation, you can try to raise/lower this value. Otherwise,

the value should be left at the default of 1.04. Setting this value too high/low will cause a calculation error. This and the other parameters are set at default settings.

The latest information about RNAdraw is available at the RNAdraw homepage, located at: http://mango.mef.ki.se/~ole/rnadraw/rnadraw.html

References for this program:

I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster (1994) Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie 125: 167-188.

Snfold

Developed by R. F. Pointon to combine a sequence shuffle routine with the Mfold folding program.

The input is as follows

Number of shuffles

Sequence

Program is as follows:

```
echo shuffleNfold program


echo -n "Enter number of times to shuffle:"
set number=$<

echo "Enter the sequence (one line of input, no spaces):"
shuffle -n $number >shuffle.dat
echo "file shuffle.dat generated"

echo "removing old results.b file - new results will be here"
echo > results.b

@ line=1
while ( $line <= $number )
  l2mfold $line shuffle.dat >infile
  mfold SEQ=infile T=25 MAX=1
  ct2bracket infile.ct >>results.b
  rm -f infile*
□[7m--More--(88%)□[m
   rm -f log-infile
  echo "Done " $line "/" $number
  @ line++
end
```

The output file contains the number of random structures required with one structure per line. This file can be directly used in the RNAdistance program.

Rsnfold

Developed by R. F. Pointon to combine a sequence shuffle routine with the RNAfold folding program (from the Vienna RNA package).

The input is as follows

Number of shuffles

Sequence

Program is as follows:

```
echo RNAfold shuffleNfold program

echo -n "Enter number of times to shuffle:"
set number=$<

echo "Enter the sequence (one line of input, no spaces):"
shuffle -n $number >shuffle.dat
echo "file shuffle.dat generated"

echo "removing old results.b file - new results will be here"
echo > results.b

RNAfold -T 25 <shuffle.dat >output.txt

rm -f rna.ps

@ line=1
@ pos=2
while ( $line <= $number )
  tail +$pos output.txt | head -1 >temp.txt
  awk '{ print $1 }' temp.txt >>results.b
  rm -f temp.txt
  @ line++
  @ pos+=2
end

rm -f output.txt

echo "Finished"
```

The output file contains the number of random structures required with one structure per line. This file can be directly used in the RNAdistance program.

Pairs

Developed by R. F. Pointon to calculate the percentage of pairing within a secondary structure. The input of the structure is in the bracket notation with one structure per line. The output shows the percentage of pairing per line of structure.

The program is written in C code and is as follows:

```
- reads lines from "stdin" (thats either the console or files redirected by
"<", etc)
-  calls "getseq" which checks if a line is a valid sequence of .)(
characters.
```

- walks along the sequence determines length and number of brackets.
- calculates the percentage and prints the result.

```
---CUT "pairs.c"--
#include <stdio.h>
#include "read.h"

#define maxstr 1000

int len,cnt,line;
char buff[maxstr];
char *seq;

main()
{ line=0;
  while(fgets(buff,maxstr,stdin))
  { line++;
    seq=getseq(buff);
    if(seq[0]==0)
    { fprintf(stderr,"ERROR - bad input on line %d\n%s\n",line,buff);
      exit(-1);
    }
    for(len=0,cnt=0;;)
    { switch(seq[len])
      { case '(':
        case ')':
          cnt++;
        case '.':
          len++;
          continue;
        case '\0':
        case '\n':
          break;
      }
      break;
    }
    printf("%d\n",(100*cnt)/len);
  }
}
---CUT---
```

here is the C code from "read.h" & "read.c"
- it just checks that there are no illegal characters and that the brackets balance.

```
---CUT "read.h"---
#ifndef _read_header_
#define _read_header_

char *getseq(char *s);

#endif
---CUT "read.c"---
char *getseq(char *s)
{ int i,j,start,finish;
  i=0;
  start=-1;
  finish=-1;

  while(start==-1)
```

```
 switch(s[i])
 { case '.':
  case '(':
  case ')':
  case '\0':
    start=i;
    break;
  default:
    i++;
 }

while(finish==-1)
 switch(s[i])
 { case '.':
  case '(':
  case ')':
    i++;
    break;
  default:
    finish=i;
 }

j=0;
for(i=start;i<finish;i++)
 switch(s[i])
 { case '.':
    break;
  case '(':
    j++;
    break;
  case ')':
    j--;
    if(j<0)
      j=20000;
 }
if(j!=0)
 return "";

 s[finish]='\0';
 return &s[start];
}
---CUT---
```

(to actually compile the code into a program you could run, you would type
gcc -o pairs pairs.c read.c
which produces the program "pairs")

Ssearch from the FASTA package

This program scans a protein or DNA sequence library for similar sequences using the rigorous Smith-Waterman algorithm (Smith and Waterman, J. Mol. Biol. (1983) 147:195-197. The sequence to be searched for is designated the query sequence. This program was used to search organellar genomes for potential pRNA sequences. Ssearch was run in the interactive mode, which prompts for the file names of the query sequence and the library. The FASTA programs, including Ssearch use a standard text

format sequence file. Lines beginning with '>' or ';' are considered comments and ignored; sequences can be upper or lower case, blanks, tabs and unrecognisable characters are ignored.

Penalty for the first residue in a gap (-12 by default).

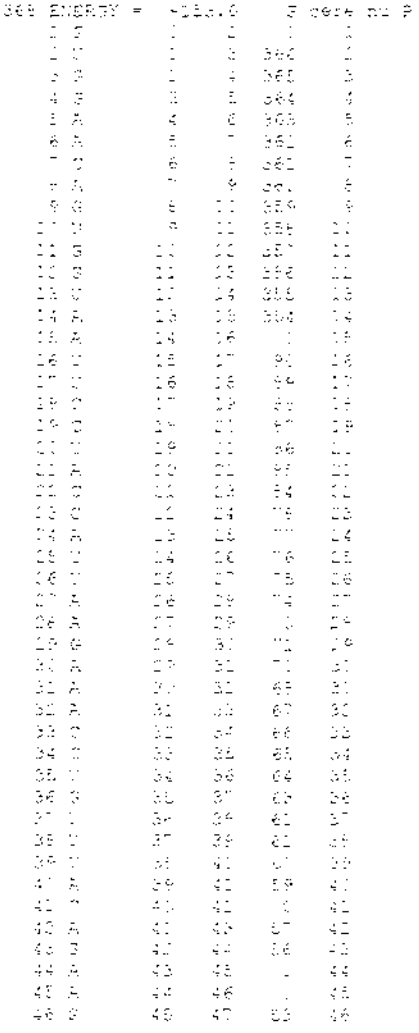
Penalty for additional residues in a gap (-2 by default).

This package was written by Bill Pearson, wrp@virginia.EDU

Ct2bracket

This program was written by R. F. Pointon at Massey University to convert the "ct" output files from the RNA folding programs into bracket notation for input into RNAdistance.

The ct files are as follows using the first 46 nucleotides from the *Saccharomyces cerevisiae* nuclear pRNA as an example:



```
Converted to bracket notation.

.((((((((((((.((((((((((.((((((((((.((
..((....))..)).)))))))))).)..)))))).....)))
))))..(((((.........(((.(((.((((((..(((((((
((((......)))))))))))).(((.((((..((((.......
((((....))))))))))...(((....)))).))))...((((
(((((((((....))))))))))).))))...))))..((((((
((..((...(((((..(((.(((....))))...))))...))))
)...))...)))))))))))))).........)))....))).)
))))).....))))))))))))..
```