Computational identification of non-coding RNAs

Paul Piccinelli

Göteborg 2006

Institute of Biomedicine at Sahlgrenska Academy
Göteborg University

A doctoral thesis at a university in Sweden is produced either as a monograph or as a collection of papers. In the latter case, the introductory part constitutes the formal thesis, which summarizes the accompanying papers. These have either already been published or are in the form of manuscripts at various stages (in press, submitted or accepted).

Department of Medical Biochemistry and Cell Biology,
Institute of Biomedicine, Sahlgrenska Academy
at Göteborg University, Box 440, SE-405 30 Göteborg.

Printed by Intellecta Docusys, Göteborg 2006
ISBN 978-91-628-7056-0 (91-628-7056-0)

Abstract

Paul Piccinelli **Computational identification of non-coding RNAs**
Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy at Göteborg University, Box 440, SE-405 30 Göteborg.

Problem. A large amount of genomic information is now becoming available. Suitable bioinformatic tools to organize and analyze this vast amount of information are therefore important. In the case of protein genes, the majority of these may be correctly identified using standard search methods that are based on sequence alignment. However, a different problem is presented when analysing non-coding RNA genes, since for their identification it is essential to take into consideration secondary structure features. Secondary structure is not only important for non-coding RNA genes, but it is also important in the regulation of gene expression. This work is concerned with the development of methods for ncRNA prediction and the application of these methods to identify specific ncRNA families.

Methods. Bioinformatic methods are used to identify ncRNA genes and ncRNA regulatory motifs. These methods include *de novo* methods, statistical profiles for primary sequence and secondary structure, sequence homology methods and minimum free energy methods. For protein gene identification we have used primary sequence alignments and profile searches and for protein classification we have used phylogenetic methods.

Results. RNase P and RNase MRP are two related ribonucleoprotein particles involved in RNA processing. We have used an approach based on conserved sequence elements to computationally analyze various eukaryotic genomic sequences for P and MRP RNA genes. We have found over 100 novel sequences, all able to fold into the consensus secondary structure of P and MRP RNAs. These genes reveal further evidence of the evolutionary relationship between these RNAs.

We also performed a computational analysis of the P/MRP protein subunits in eukaryotic organisms. A number of novel homologues were identified and we found novel orthologous relationships between fungal and metazoan proteins. Our results further emphasize a structural and functional similarity between the yeast and human P/MRP complexes.

The iron responsive element (IRE) is an RNA hairpin structure located in certain genes that are post-transcriptionally regulated in response to iron. We have found more than 90 novel sequences with the characteristics of known IREs. We have found evidence that the ferritin IRE represents the ancestral version of this type of translational control.

Finally, ncRNA genes in yeast have been predicted using two the *de novo* methods, RNAz and QRNA. A number of predicted candidates have been selected for experimental testing and more candidates will be tested.

Conclusions. We have used different bioinformatic methods to identify ncRNAs in a variety of organisms and report on several ncRNA sequences not previously reported. These novel RNA sequences make it possible to better predict the structure of these RNAs as well as to better understand their evolution and function. To further understand the structure and evolution of the RNases P and MRP we also analyzed the protein composition of these enzymes. Together, these new predictions aid to better understand the structure, function and evolution of RNase P and MRP.

Keywords: RNase P, RNase MRP, IRE, non-coding RNA, secondary structure
ISBN 978-91-628-7056-0 (91-628-7056-0) Göteborg 2006

Papers

I. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes
Paul Piccinelli, Magnus Alm Rosenblad, Tore Samuelsson.
Nucleic Acids Res. 2005 Aug 8;33(14):4485-95.

II. Inventory and analysis of the protein subunits of the ribonucleases P and MRP provides further evidence of homology between the yeast and human enzymes
Magnus Alm Rosenblad, Marcela Dávila López , Paul Piccinelli and Tore Samuelsson.
Nucleic Acids Res. 2006 Sep 34(18):5145-5156.

III. Evolution of the iron responsive element
Paul Piccinelli and Tore Samuelsson.
Submitted for publication.

IV. Hunting for non-coding RNA genes in yeast
Paul Piccinelli, Jonathan Esguerra, Anders Blomberg, Tore Samuelsson.
In manuscript.

Table of contents

Introduction

According to the central dogma of molecular biology DNA acts as a template for the production of an RNA transcript and that transcript in turn specifies a protein. At the same time, numerous transcripts have been found that exert their function without ever producing proteins. These RNAs are referred to as non-coding RNAs (ncRNAs), in contrast to mRNAs that specify proteins. For many years it was believed that ncRNAs are rare, and that they, like tRNAs and rRNAs, are molecules that just aid in the production of proteins. The development of the ncRNA field was slow for many years because RNAs were difficult to study experimentally. Furthermore, genome sequences were lacking as well as adequate bioinformatic approaches to identify and analyze RNAs computationally. Consequently, identification of novel ncRNA species and their functional role occurred by chance rather than by systematic screens. However, it eventually became apparent that there are numerous ncRNAs, and that their cellular functions are varied and important [1-5]. In the past few years, new experimental strategies and computational methods have been developed demonstrating that the number of ncRNAs in genomes of model organisms is much higher than was previously anticipated. In addition to tRNA and rRNA there are a number of other large ncRNA families that have emerged during recent years, as described further below.

Non-coding RNA genes

Most genes in higher eukaryotes contain introns. Intron elimination and ligation of the exons take place with the help of a ribonucleoprotein (RNP) complex called the spliceosome [6]. There is growing evidence that the main catalytic function in the spliceosome is in fact performed by RNA components, i.e. that the spliceosome is a ribozyme [7-9]. The spliceosomal RNA U1 has an additional function in the regulation of transcriptional initiation [10].

The signal recognition particle (SRP) is a ribonucleoprotein particle (RNP) that targets nascent proteins to the ER membrane. In the process, protein synthesis is arrested when the SRP binds to the N-terminal signal of the nascent protein chain [11]. The SRP have been identified in all three domains of life [12]. The eukaryotic SRP consists of a 300-nucleotide 7S RNA and six proteins: SRP9, SRP14, SRP19, SRP54, SRP68 and SRP72. The archaeal

SRP consists of a 7S RNA and homologues of the eukaryotic SRP19 and SRP54 proteins. In most bacteria, the SRP consists of 4.5S RNA and the Ffh protein, a homologue of the eukaryotic SRP54 protein.

The snoRNAs are involved in alteration and cleavage of nascent rRNA transcripts in both eukarya and archaea [13, 14]. There are two classes of snoRNA identified: The C/D box snoRNAs direct 2´-O-methylation of the ribose, while the H/ACA box snoRNAs guide the conversion of uridine nucleotides to pseudouridine [15-18]. In addition to their roles in rRNA maturation, snoRNAs also target spliceosomal RNA. These snoRNAs perform their function in the Cajal bodies; for this reason they are sometime referred to as scaRNAs (small Cajal-body associated RNAs) [19].

MicroRNAs (miRNAs) form a class of non-coding RNA genes whose products are small single-stranded RNAs with a length of about 22nt. These are involved in the regulation of translation and degradation of mRNAs [20]. miRNAs are transcribed as ~70nt precursors and subsequently processed by the Dicer enzyme to give a ~22nt product. The products have regulatory roles through their complementarity to mRNA. MicroRNAs have been identified in both multi-cellular animals and plants.

Telomerases are specialized RNPs that cap chromosome ends that are essential for genome stability and cellular proliferation [21]. Sequence loss during replication is prevented with a particular mechanism in organisms that have linear chromosomes [22]. In most of these organisms, the telomerase expand chromosome ends by iterative reverse transcription of the telomerase RNA [23]. Telomerase RNAs are very different in sequence and structure between vertebrates, ciliates and yeasts, but they share a 5´ pseudoknot structure close to the template sequence.

In the present work we have studied in greater detail the family of RNAs formed by RNase P and MRP RNAs. This area is described further below.

RNase P and RNase MRP

Ribonuclease P (RNase P) and ribonuclease MRP (RNase MRP) are two related ribonucleo-protein particles (consist of both RNA and protein subunits) involved in RNA processing

[24]. The universal function of RNase P is to carry out an important step in pre-tRNA processing. This enzyme is responsible for the removal of a 5′ leader of precursor tRNAs (pre-tRNAs), by catalyzing the hydrolysis of a specific phosphodiester bond that leaves a phosphate at the 5′ end of the mature tRNA and a hydroxyl group at the 3′ end of the leader. The RNase P is found in all living cells throughout the three kingdoms and in mitochondria and chloroplasts as well [25, 26]. Other activities have been reported for RNase P in bacteria which involves recognition and cleavage of non-tRNA substrates including some viral tRNA-like structures and antisense phage RNAs [27, 28], the pre-SRP RNA of *E.coli* (4.5S RNA) [29], the pre-tmRNA of *E. coli* (Sa10 RNA) [30], a few polycistronic mRNAs [31] and the B_{12} riboswitch of *E. coli* and *B. subtilus* [32].

RNase MRPs (Mitochondrial RNA Processing) are nucleoproteins found only in eukaryotes. The enzyme was initially described as an endoribonuclease with ability to cleave a mitochondrial transcript in vitro, which was consistent with a role in the formation of a primer for the initiation of mitochondrial DNA replication [33]. However, the majority of RNase MRP RNAs have been localized to the nucleolus [34, 35], the enzyme seem to play the most essential role in pre-ribosomal RNA (pre-rRNA) processing where it specifically cleaves 27SA pre-rRNA at a site A3 within the first internal transcribed spacer (ITS 1), which is necessary for generating the mature 5.8S rRNA [36-38]. RNase MRP has also been implicated in cell cycle regulation in yeast. Thus, the enzyme cleaves the CLB2 mRNA and as a result allows for rapid degradation and completion of mitosis [39, 40]. MRP RNA is also associated with the genetic disease cartilage hair hypoplasia (CHH) where mutations in certain parts of the RNA are found in patients affected. CHH is inherited in an autosomal recessive manner characterized by unequal short-limbed dwarfism, sparse hair, impaired immunity and anemia [41].

The RNA component. The RNA subunit of bacterial RNase P has an important role in the enzymatic reaction of. All bacterial RNase P RNAs studied so far are ribozymes since they can recognize and cleave substrates of pre-tRNA without the help of the protein subunit under high ionic strength in vitro [42-44]. Also certain archaeal RNase P RNAs are catalytically active in very high salt concentrations [45]. In contrast, eukaryotic RNase P RNAs cannot act as ribozymes in vitro although they are structurally similar to bacterial P RNA [26].

Comparative analyzes of the RNA subunits of RNase P from all organisms reveals similarities at both primary and secondary structure level, indicating that all known RNase P RNAs contain a similar core structure [46], also shared with RNase MRP RNA [47].
The bacterial P RNA is organized into two independently folded domains, one catalytic domain (C) and one specificity (S) domain [48]. The S domain identifies the TψC loop of pre-tRNA, whereas the C domain recognizes the acceptor stem and the 3´ CCA trailer sequence [49]. Based on primary and secondary structure comparison, the eukaryotic RNAs seem to have a similar organization where the C domain is referred to as domain 1 and the S domain is referred to as domain 2 (Fig.1). Two structures of a the bacterial P RNA S domain have recently been reported [50, 51] as well as the structure of a full-length RNA with both C and S domains [52].

Comparison of the primary and secondary structures of P RNA from bacteria to those of eukaryotes has revealed both similarities and discrepancies. A universal consensus structure [46, 53-55] comprise five critical regions, termed CR-I through CR-V, with conserved nucleotides and several stems (P1, P2, P3, P4, P7, P10/11, and P12) (Fig.1). The P4 helix, which is formed by the base pairing of CR-I and CR-V, is suggested to be the catalytic center of the bacterial enzymes [56-59]. For eukaryal RNase P RNA, mutagenesis of the conserved nucleotides within and near P4 has suggested an important role in pre-tRNA binding, catalysis and maturation [60]. CR-II and CR-III, together with P10/P11 and P12, form a domain in eukaryotes in which the internal loop containing the CR-II and CR-III regions is essential for yeast viability and RNase P activity [60]. The CR-II region contains a consensus sequence AGARA, which is conserved in many species and similar to the bacterial CR-II consensus [54, 61]. The results of mutational studies of the AGARA sequence in CR-II of the yeast P RNA suggest that it has a function in magnesium utilization [60, 62]. Furthermore, steady-state kinetic studies of the mutant complexes point to an important role for CR-II in catalytic efficiency [61]. Mutagenesis of the CR-IV region of the yeast RNase P RNA results in a large decrease in turnover rate but no significant changes in substrate binding [60]. This would suggest that the CR-IV region has a role in the catalytic function. However, in bacteria this region seems to have a role in substrate recognition [43, 48, 58].

P RNA from both eukaryotes and bacteria contain a P3 stem. However, whereas the bacterial P3 is rather small the eukaryotic P3 is more extended and has an internal loop [46, 54]. P3 has been proposed as a protein binding site in bacterial RNase P holoenzyme [63]. In yeast RNase

P RNA, certain residues in P3 intra-loop seems important in pre-tRNA processing and mutations of these nucleotides interrupt the interaction between the RNA subunit and the Pop1 protein subunit [62]. In human RNase P RNA an analogous P3 stem has been found to bind specifically to a 40-kDa protein [64] and is also suggested to be involved in correct subcellular localization of the RNA subunit [34].

Additional elements in the eukaryotic consensus structure may well have counterparts in the bacterial consensus structure. For instance, helices eP8 and eP9 could be the equivalent to the corresponding helices in bacterial RNase P RNA, but the variability in this region of eukaryal RNase P RNA sequences leaves the homology uncertain. Most fungal eP8 stem-loop structures have a NUGA loop sequence, whereas most of the eP9 hairpin loops contain GNRA tetraloops [54]. Studies of ribosomal RNA and other sources have suggested that a GNRA tetraloop could enhance the stability of an RNA duplex by acting as site for intra-molecular or inter-molecular RNA-RNA interactions [65-67].

A number of secondary structure elements are shared between the RNA subunits of RNase P and RNase MRP (Fig.1). Very similar P4 helices may be formed in RNase P and MRP RNAs [68] and the sequences of the regions CR-I and CR-V are conserved during evolution, suggesting that P4 in RNase MRP may also contribute to the catalytic center of RNase MRP [69]. Another conserved region in domain 1 is CR-IV with the consensus sequence AGNNNNA for P RNA and AGNNA for MRP RNA. In P3, several residues in the internal loop and flanking base pairs are conserved [62]. Experiments in yeast have shown that the P3 helix in the two enzymes could be exchanged without loss of function or specificity [70], suggesting that the evolution of this part of the RNA is constrained by binding to a protein component [62]. In contrast to P RNA, MRP RNA does not appear to have the conserved sequence motifs CR-II and CR-III in domain 2 that are characteristic of RNase P RNA (Fig.1).

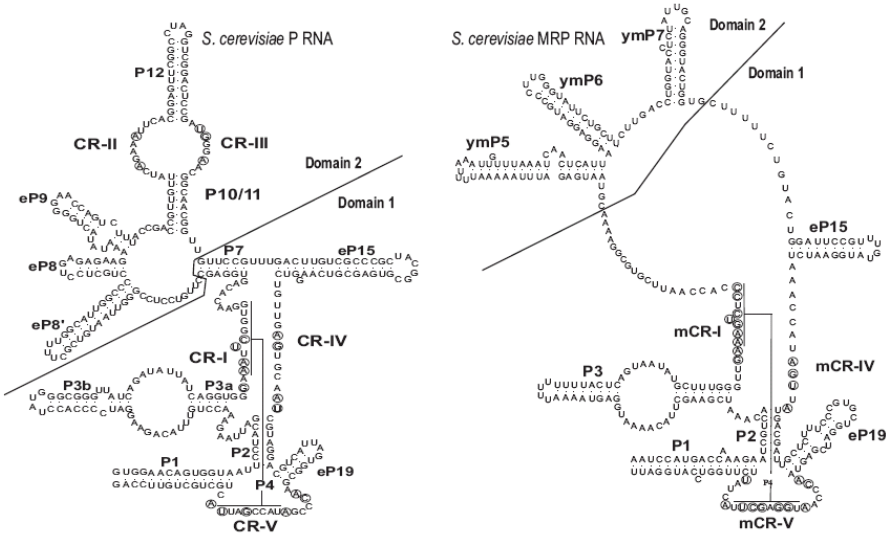


Figure 1. Models of eukaryotic RNase P and MRP RNAs. Nucleotides that are conserved in all known eukaryotic P and MRP RNAs, respectively, are encircled. The eukaryotic RNAs seem to have a similar organisation as bacterial P RNA, domain 1 and 2, where bacterial catalytic domain resembles domain 1. Eucaryotic P RNA sequence consensus consist of regions CR-I-CR-V and also several stems; P1, P2, P3, P4, P7, P10/11, and P12. In domain 1 secondary structure elements P1-P4 are shared between P and MRP RNAs. In addition regions CR-I, CR-IV and CR-V are conserved in both RNAs.

Protein components. The RNases P and MRP both have an RNA molecule and one or several protein subunits [71]. The bacterial RNase P has only a single small protein whereas the archaeal and eukaryal nuclear RNase P enzymes have a larger number of protein subunits [26]. Furthermore, a number of protein components are shared between RNases P and MRP, although the eukaryal RNase P have acquired additional protein subunits during evolution [27]. Surprisingly, there is no obvious sequence homology between the bacterial protein subunit and archaeal/eukaryal proteins and as a consequence the evolutionary link between the proteins is not clear at the moment. In eukaryotes the RNases P and MRP have almost identical protein content, although they have different substrate specificities [71, 72].

Although the bacterial RNase P RNA can function as a ribozyme *in vitro* the cleavage rate of pre-tRNA is enhanced 20-fold by the protein moiety [73]. Also some archaeal RNase P RNAs show enzymatic activity under high salt conditions [45], but seem structurally defective in the

absence of protein. In contrast, a catalytic activity has not been demonstrated for the nuclear RNase P RNA, suggesting a more important role for the protein moiety [53].

In *Saccharomyces cerevisiae* at least ten protein components of RNase MRP have been identified so far (Table 1). The protein composition of RNase MRP closely resembles that of RNase P as eight proteins are shared between RNase MRP and RNase P; Pop1, Pop3, Pop4, Pop5, Pop6, Pop7, Pop8, and Rpp1 [74]. The proteins Snm1 [75] and Rmp1 [76] are the only protein components that specifically associate with RNase MRP RNA (Table 2). Conversely, Rpr2p has been identified as a protein unique to the RNase P complex [72].

Ten proteins are believed to be stably associated with human RNase MRP and RNase P, Rpp14, Rpp20, Rpp21, Rpp25, Rpp29, Rpp30, Rpp38, Rpp40, hPop1 and hPop5 [77, 78] (Table 1). Recent work has demonstrated however a preferential association of hPop4, Rpp21, and Rpp14 with RNase P and only a transient association of Rpp25 and Rpp20 with RNase MRP [79]. At least six of the P/MRP subunits appear to be homologous to the subunits identified in *S. cerevisiae*, Pop1 (hPop1), Pop4 (Rpp29/hPop4), Pop5 (hPop5), Pop7 (hPop7/Rpp20), Rpp1 (Rpp30), Rpr2 (Rpp21) [26] (Table 1). More recently the human Rpp38 was also suggested to be a functional homologue of yeast Pop3 [80]. Comparative studies show that archaeal RNase P has at least four protein subunits homologous to eukaryotic RNase P/MRP proteins, Pop4, Pop5, Rpp1, and Rpr2 (Rpp21) [81].

Structural models based on protein-RNA and protein-protein interactions have been proposed for human and yeast [74, 77, 78, 82]. Many of the interactions in these models have also been found in archaeal holoenzymes [83-85]. In the human RNase P the RNA molecule has been shown to interact with Rpp29, Rpp30, Rpp21 and Rpp38 [77]. Furthermore, hPop1, Rpp20, Rpp21, Rpp25, Rpp30 and Rpp38 interact directly with the RNA subunit of human RNase MRP. [78, 86]. For the yeast MRP there is evidence that the RNA interacts with the protein subunits Pop1 and Pop4 [74].

Table 1. Previously reported MRP/P protein relationships of yeast and human.

RNase P/MRP subunits		Further relationships & particle specific subunits
H. sapiens	*S. cerevisiae*	
hPop1	Pop1	
Rpp38	Pop3	
Rpp29 (hPop4)	Pop4	
hPop5	Pop5	
	Pop6	
Rpp20	Pop7	ALBA domain
	Pop8	
Rpp14		Pop5 paralogue
Rpp21(Rpp2)	Rpr2	only in RNase P
Rpp25		ALBA domain
Rpp30	Rpp1	
Rpp40		
	Snm1	only in RNase MRP
	Rmp1	only in RNase MRP

Regulatory RNA motifs

In addition to ncRNA genes there is also another level of RNA function presented by functional motifs within the mRNA of protein-coding genes. A few examples are discussed below.

A RNA structural element referred to as IRES (Internal Ribosome Entry Site) assists in cap-independent initiation of translation starting at an internal initiation codon. IRESs occur in several types of viruses, but also a limited number of eukaryotic mRNAs can be translated by internal ribosome entry. Most mRNAs with IRES encode regulatory proteins such as growth factors and transcription factors. Studies have reported that under stress conditions, where cap-dependent translation is blocked, translation of specific mRNAs is enabled through IRES elements [87]. IRESs are also involved in alternative initiation of translation. For example, the human fibroblast growth factor 2 contains 5 different translation initiation codons. A cap-dependent process initiates translation initiation of the codon closest to the 5´ end, whereas initiation of the remaining codons depends on the IRES [88].

Selenocysteine insertion sequences (SECIS) are located in the coding region of some eubacterial mRNAs and in 3´ untranslated regions of some mRNAs in archaea and eukaryotes where it incorporate selenocysteins at UGA codons (usually encodes stop) in these proteins [89]. In eubacteria, such a sequence occurs as a hairpin structure of conserved length soon

after the UGA codon. In archaea, the primary sequence rather than the secondary structure is conserved. The hairpin varies in stem length, occurrence of internal loops and size of the hairpin loop, but it has a very conserved sequence motif in the helix adjacent to the apical loop. In eukaryotes the SECIS element is characterized by a specific secondary structure, while only shorter sequence motifs are conserved. The secondary structure is composed of a long hairpin structure constructed from two or three consecutive helices [90].

Riboswitches are independent structural elements primarily found within the 5-UTRs of bacterial mRNAs, which, upon direct binding of small organic molecules, can trigger conformational changes. Riboswitches regulate several key metabolic pathways in bacteria including those for coenzyme B12, thiamine, pyrophosphate, riboflavin monophosphate, S-adenosylmethionine, as well as different amino acids [91, 92].

Most of the ncRNA genes and RNA motifs are collected in specific databases. The most comprehensive databases are Rfam [93, 94], the NONCODE database [95] and the RNAdb [96].

In the present work we have focused on a RNA structural element to be described below that is important in the regulation of iron metabolism in eukaryotes.

Iron metabolism and the IRE element

Iron deficiency is a worldwide health problem. Over the past four or five decades, much research has focused on the metabolic consequences of iron deficiency. Organisms have developed many responses to iron deficiency and iron repletion to keep various essential functions [97].

Iron is very appropriate as cofactor in enzymatic reactions due to its two stable oxidation states. Ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron have a suitable redox potential capable to drive a huge number of catalytic reactions. Besides this important use, free iron is a possible producer of hydroxyl and superoxide radicals in the presence of oxygen, which are highly toxic for almost every cell type [98].

A well established model has emerged how iron uptake, utilization and storage are

coordinately regulated. Vital components to help arrange a combined response to variations in iron availability are the iron regulatory proteins (IRP) [99, 100] and the *cis*-acting regulatory motifs, termed iron-responsive elements (IREs) [101]. This model exemplifies one of the earliest post-transcriptional control examples of gene expression through RNA–protein interactions.

IREs are found in the UTR-regions of associated transcripts encoding central components of iron metabolism and the citric acid cycle. The IREs are 26–30 nucleotide long hairpin-forming sequences with a CAGUGX terminal loop sequence, which is conserved in all IREs (Fig.2). X at position 6 can be either an A, C or U but never a G [102]. Basically there are two types of IRE sequences; the first type has a conserved C residue five bases upstream of the CAGUGX sequence creating a bulge in the hairpin while the second type has a UGC/C loop-bulge (Fig.2).

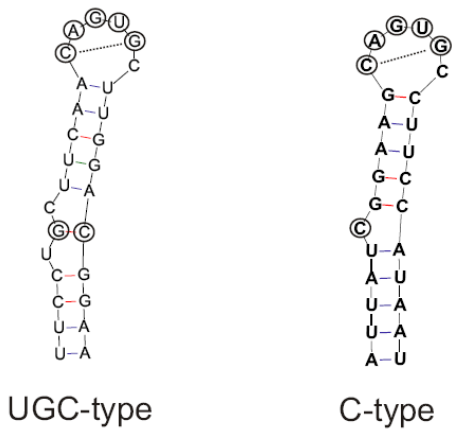


Figure 2. The two types of IRE. Conserved residues are encircled. Both structures share a highly conserved sequence of nucleotides in the stemloop. The suggested basepair interaction between C and G in the loop is shown with a dashed line.

NMR spectroscopy, nuclease and chemical probing have been used to characterize the IRE structure. The AGU of the CAGUGX hexaloop is exposed to solvent due to the C-G base-pairing [103] [104]. The G-C base pair within the internal loop of certain IRE creates a pocket of the large groove that selectively improves binding of IRPs [105]. The large groove of the

IRE stem is enlarged by distortions at the C-bulge or internal loop [106] creating specific base and ribose contact sites for protein [103].
Trans-acting factors in the translational regulation of mRNA involved in the iron homeostasis was first suggested by data obtained in the late 1970s [97]. Two proteins have been identified since then, the IRP1 and IRP2 proteins that specifically control the translation or turnover of IRE-containing mRNAs.

IRP1 exists in two forms. It is mainly present as cytosolic aconitase (cAcn), a protein characterized by a [Fe-S] cluster that catalyses the conversion of citrate to isocitrate under normal iron conditions [107]. However, low intracellular iron levels or oxidative stress produced by reactive oxygen [108] or nitrogen species [98] triggers loss of the [Fe-S] cluster and conversion to an IRE binding form. We still lack information about the cellular factors that trigger the switch between the two functionally different proteins and the detailed mechanism of the switch, but there is evidence that this conversion is associated with local structural changes affecting the entire protein [109].

A second IRE-binding protein, IRP2, was initially identified in rat hepatocytes, and subsequently cloned from a variety of mammalian tissues and cells [110]. IRP2 shares 62% amino acid sequence identity with IRP1 but lacks the [Fe-S] cluster. Each of the two IRPs has a 30% sequence identity to the mitochondrial form of aconitase (m-aco) [111]. Characteristic of IRP2 is a 73-amino acid insertion in its N-terminal region [112]. This region contains a cysteine-rich sequence that is known to be responsible for targeting the protein for degradation when cellular Fe levels are high [113]. IRP2 does not have aconitase activity, perhaps due to the lack of the [Fe-S] cluster, but is more sensitive to iron status than IRP1 and therefore seems to dominate post-transcriptional regulation of iron metabolism in mammals [107]. IRP2 seems to be expressed mainly in the forebrain and cerebellum whereas IRP1 expression mainly is in the kidney, liver and brown fat [107].

Both IRPs are cytosolic RNA-binding proteins that bind to and regulate the translation or stability of mRNA that contains IREs. In low-iron conditions, the IRP bind the IRE regions with high affinity. Certain mRNAs contains a single IRE near the 5′ end, usually within the 5′ UTR. When IRP bind to those IREs, IRP appears to block the ability of eIF4F to recruit the 40S subunit with its associated factors (the 43S pre-initiation complex) to the mRNA and as a consequence the translation is inhibited [114]. The best studied example is the ferritin chains

where the IRP binding leads to decreased iron storage [114]. The binding of IRP and ferritin IRE is the most efficient of all IRP-IRE interactions [105], probably because of a conserved internal loop involving UGC/C rather than the bulge C of all other IREs found so far (Fig.2). The mRNAs encoding ferritin, in vertebrates both H and L ferritin subunits, have an IRE in the 5′-UTR [115]. Elevated iron levels, in contrast, prevent IRP binding to ferritin IREs and consequently the chains become expressed and assemble into a typical 24-mer macromolecule with a large cavity that can store up to 4000 iron atoms.

In contrast to the IRP-IRE regulation in 5′ UTR, IRPs bound to IREs in 3′ UTR protects the mRNA from degradation, which leads to improved mRNA stability and enhanced protein synthesis. The best studied example is the transferrin receptor (Tfr), which together with the plasma protein transferrin (Tf) are involved in the main pathway by which all cells internalize iron [116]. Under high iron conditions, the IRP does not bind to the IRE regions of the Tfr and the Tfr mRNA is degraded. Since ferritin mRNA is translated under these conditions, the net result is the inhibition of further iron uptake and promotion of iron storage by the cell [98]

The number of mRNAs subject to IRP-mediated regulation has been growing lately (Table 2), indicating that a wide range of versatile regulation exists by the binding of the two different IRPs to unique target sequences which may or may not differ from the consensus IRE. These include mRNAs encoding proteins involved in iron storage (H- and L ferritin) and cellular iron internalization (transferrin receptor) previously mentioned, heme formation in erythroid cells (erythroid 5-aminolevulinate synthase), cellular iron uptake (divalent metal transporter-1 and transferrin receptor), iron export (ferroportin) as well as two tricarboxylic acid cycle enzymes, mitochondrial aconitase and the insect succinate dehydrogenase. Recently, IREs have been observed in four additional mRNAs. The mRNAs of glycolate oxidase [117], amyloid precursor protein [118], myotonic dystrophy kinase-related Cdc42-binding kinase α [119] and cell division cycle14A [120] and NADH dehydrogenase [121] (Table 2). These findings would suggest that the IRE/IRP system of regulation is more exploited than previously anticipated. However, more experimental work is required to clarify the role of these recently discovered elements.

Table 2. Proteins proposed to have an IRE in their transcripts.

Protein	Function	IRE type	IRE localization
Ferritin L-chain	cellular iron storage	UGC	5′-UTR
Ferritin H-chain	cellular iron storage	UGC	5′-UTR
Ferritin M-chain	cellular iron storage	UGC	5′-UTR
Transferrin receptor1	cellular iron internalization	C	3′-UTR
Ferroportin1	Iron transport in enterocytes	C	5′-UTR
DMT1 (divalent metal transporter)	Iron transport in enterocytes	C	3′-UTR
e-ALAS (erythroid aminolevulinate synthase)	heme synthesis pathway	C	5′-UTR
m-ACO (mitochondrial aconitase)	enzyme in citric acid cycle	C	3′-UTR
Succinate dehydrogenase	enzyme in citric acid cycle	C	5′-UTR
CDC14A	cell cycle progression	C	3′-UTR
MRCK α (myotonic dystrophy kinase-related Cdc42-binding kinase α)	cytoskeletal reorganization	C	3′-UTR
GOX (glucolate oxidase)	production of oxalate	C	3′-UTR
NADH dehydrogenase	involved in respiratory chain	other	5′-UTR
APP (amyloid precursor protein)	neurotoxic conditions	other	5′-UTR

A number of inherited disorders are associated with mutations in the genes subjected to IRP-mediated regulation. For example mutations found in the gene of ferroportin lead to an iron loading disorder while mutations found in the gene of DMT1 leads to severe hypochromic, microcytic anaemia [122]. In L-ferritin a heterogenous pattern of mutations in the IRE are associated with hereditary hyperferritinemia-cataract syndrome (HHCS) which is an autosomal dominant disorder characterized by bilateral cataracts and increased serum L-ferritin, in low iron conditions [123]. In the related H-ferritin a single point mutation found in the IRE of a Japanese family is associated with dominantly inherited iron overload responsible for tissue iron deposition [124].

Bioinformatic ncRNA methods

Prediction of ncRNA genes versus prediction of protein genes

The analysis of genomic sequence has focused on development of methods to identify and describe protein genes within the genomic sequence. This is usually based on the identification of conserved coding exons by comparative genome analysis or on computational gene prediction, which relies on gene-finding algorithms [125, 126]. Such gene-finding algorithms are designed to identify open reading frames (ORFs), polyadenylation signals, conserved promoter regions and splice sites typically associated with protein-coding genes. A popular *ab initio* genfinding tool is GENSCAN [127]. It uses a

general probabilistic model, which incorporates basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions.

NcRNAs that are conserved in sequence, for instance ribosomal RNAs, can easily be identified using sequence similarity search programs like BLAST [128] and FASTA [129]. Such programs can primarily be used to find orthologous ncRNAs in closely related species (see e.g. [130]). However, since RNA sequence often evolves much faster than structure in most cases, an approach based on primary sequence alignment is often inadequate for ncRNAs.

NcRNAs with little primary sequence conservation have been more difficult to predict. In particular it has been difficult to design *ab initio* methods that find ncRNA genes, including genes that encode previously unrecognized classes of ncRNAs. One possible method of gene finding is one where promoter and transcription termination signals are analyzed. Eukaryotic ncRNAs are transcribed by different polymerases: rRNAs by Pol I, small structural RNAs like tRNAs and 5S RNA by Pol III, and most other ncRNAs by Pol II. Some ncRNAs are not independently transcribed at all, such as vertebrate small nucleolar RNAs that are processed out of introns. Two successful screens for ncRNAs in *E. coli* used promoter and terminator identification combined with comparative genome analysis to identify conserved noncoding regions [131, 132].

In another type of method, statistical signals in splice sites were used to predict transcription units and is suggested to work on all types of genes, even ncRNA genes [133]. Yet another approach is to examine statistical signals such as base composition [134]. However, these methods are not sufficiently selective since they make use of properties also found in protein genes and they will therefore give rise to a high rate of false-positives.

An RNA molecule will fold into a tertiary structure guided by the primary sequence. The molecule is able to form intramolecular helices, giving rise to a 'secondary structure' as shown for tRNA in Fig.3 (Fig.3B). The tertiary structure (Fig.3C) has additional hydrogen bonds giving rise to a more compact structure. A detailed analysis of functional classes of RNAs shows that their secondary structures are very well conserved between species, indicating that the secondary structure is important for the function of the molecule. At the same time there is

little similarity at the sequence level. Therefore many successful ncRNA identification and screening methods have made use of secondary structure information in addition to primary sequence information.

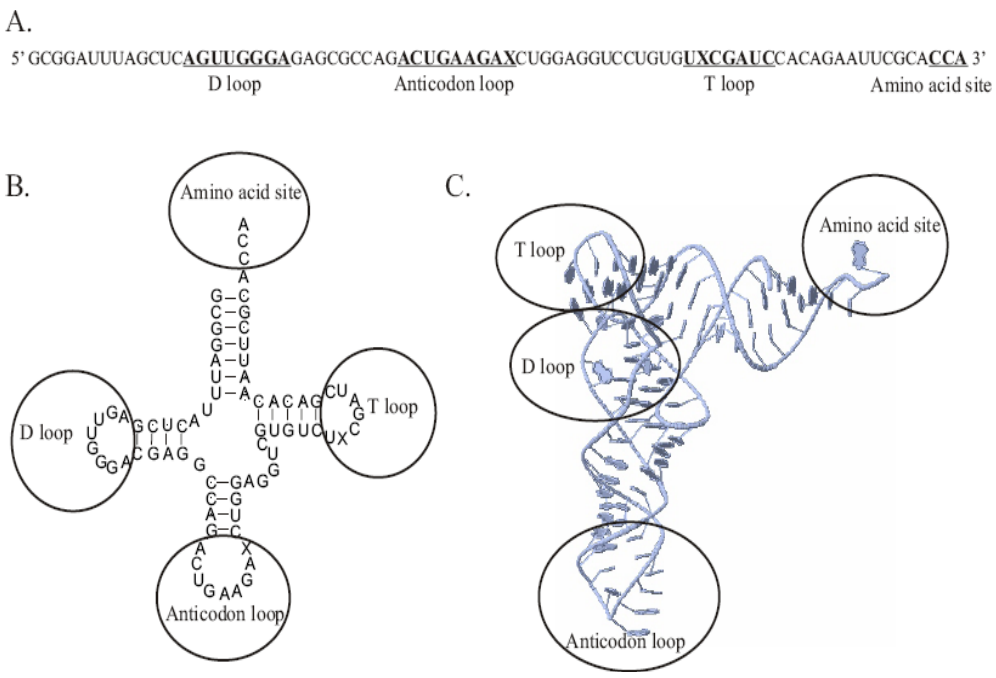


Figure 3. The different levels of tRNA structure. A) The primary sequence of a tRNA. B) The secondary structure of the sequence. C) The tertiary structure (PDB ID: 1EVV [135]).

Available ncRNA methods

There are a number of approaches and methods developed through the years that work reasonably well but they are not as reliable as protein gene finding methods. The simplest class of search tools uses regular or context free grammars (CFG) to describe RNA motifs that are known and well defined. PatSearch [136] and RNAMotif [137] are tools which allow the user to specify a given motif with a particular description language and offer search approaches to identify instances of the motif in a set of sequences. A problem with these tools is that the model is very sensitive to variations, and it is also very difficult to describe production rules for complicated motifs with a large number of exceptions.

With probabilistic models, such as stochastic context free grammars (SCFG), the user is able to assign probability distributions to production rules; noise in the dataset is handled easily because the model can adapt itself to variations. The main drawback of stochastic context free grammars is that most of the available implementations are highly demanding in terms of computational resources and as a result they are are not suitable for the analysis of whole genomes. A SCFG algorithm typically require $O(N^3)$ in memory and $O(N^4)$ in time complexity, where N is the length of the sequence. Still, there are many approaches that utilize SCFGs. An example is RSEARCH [138] that aligns an RNA query to target sequences, using SCFG algorithms to score both secondary structure and primary sequence alignment simultaneously. It is very time-consuming compared to sequence alignment methods like BLAST. For example, we noted that it took 2.9 CPU days to search a 113 nt RNA against a 2.1×10^7 nucleotide database [138].

An important category of RNA bioinformatics methods, are those that attempt to predict a secondary structure from a sequence. Most successful methods are based on the principle of finding a structure with minimal free energy (MFE). The stability of a secondary structure is measured as the amount of free energy released or used by forming base pairs. The more negative the free energy of a structure, the more likely is formation of that structure. To compute the minimum free energy of a sequence, empirical energy parameters are used. These parameters summarize free energy change (positive or negative) associated with all possible pairing configurations including base pair stacks and internal base pairs, internal bulges, hairpin loops, and various motifs which are know to occur with great frequency. Many well established algorithms are based on a thermodynamic model for the prediction of RNA secondary structures, returning a structure of minimal free energy called MFE-structure for short. The most well known MFE program is the MFOLD program by Zuker [139]. From a single sequence it calculates its MFE structure visualized in a planar graph representation.

Another important principle in RNA bioinformatics is that of covariation. Covariation may be defined as changes that maintain a basepairing pattern in double-stranded regions in the RNA molecule (Fig.4). Mutual information is a measure of covariation [140]. Programs such as RNA Structure Logo [141] and MatrixPlot [142] allow the user to display mutual information content for an alignment of RNA sequences. One important application of mutual information is that it may be used as evidence for a particular base pairing or secondary structure.

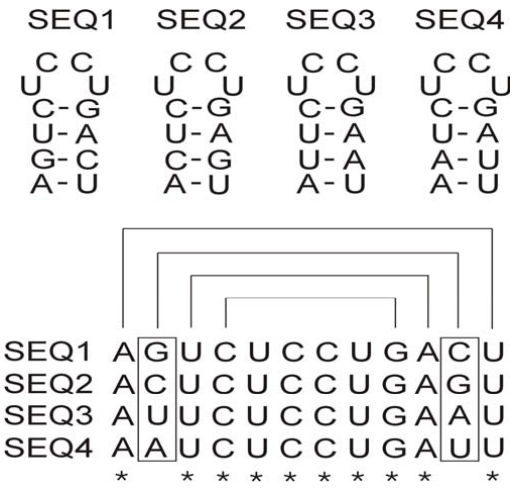


Figure 4. Four related sequences displaying covariation. The four sequences adopt the same secondary structure although the sequences are variant in two basepairing positions as indicated by the multiple alignment. The changes in these two positions are maintaining a basepairing pattern, thus not affecting the secondary structure.

In the method implemented in RNAALIFOLD [143], a folding prediction is achieved with a sequence profile, or multiple sequence alignment, instead of a single sequence. The principle of energy minimisation is combined with the principle of covariation.

Structural alignments are very useful in methods to identify new ncRNAs. Covariance models transform the information in a structural alignment into a probabilistic model capturing both primary consensus and secondary structure information through the use of SCFGs [144]. Covariance models are constructed from multiple sequence alignments and have high sensitivity, high specificity and general applicability to any RNA sequence family of interest. However, covariance model dynamic programming algorithms are very CPU intensive and are alone not suitable for genomic searches [144]. Covariance models of various ncRNA families can be found at the Rfam database [94]. The INFERNAL program package contains methods both for generating and searching covariance models [1].

Another type of comparative approach is to predict secondary structure from a set of unaligned sequences and in this case the algorithm searches over all possible foldings. Sankoff described such an algorithm, which is $O(N^{2m})$ in memory and $O(N^{3m})$ in time for m sequences of length N [145]. Since this algorithm is quite unappealing when it comes to longer sequences, approaches based on this algorithm have included heuristics in some parts.

FOLDALIGN [146], the first practical algorithm of this type, was developed specifically for the identification of local motifs in RNA sequences, where the motif is composed of both sequence and structure constraints. It uses a dynamic programming algorithm that is guaranteed to find the highest scoring local alignment between two sequences, or between a sequence and an alignment of other sequences. Another Sankoff algorithm approach is Dynalign [147]. The Dynalign algorithm takes advantage of both free energy minimization and comparative sequence analysis to predict RNA secondary structure for 2 sequences. It can improve the accuracy of secondary structure prediction compared to standard free energy minimization methods that consider the structure of only one sequence.

Current approaches for ncRNA gene identification at a genomic scale can be divided into two classes regardless of method: approaches to detect new members of already known and well-characterized ncRNA families, and attempts to predict any ncRNA genes so that novel families of ncRNAs can also be found (*de novo* methods).

Specialized programs have been developed that are suitable for screens at a genomic scale to detect members of particular ncRNA families. Examples of this such programs are miRseeker for microRNAs [148], tRNAscan-SE for tRNAs [149], snoScan for box C/D snoRNAs [150], fisher for box H/ACA snoRNAs [151], as well as a srpscan for SRP RNAs [152]. MiRseeker uses a comparative approach where pairwise alignments of closely related species are analyzed together with MFE folding and a set of rules derived from previously known microRNAs. Both tRNAscan-SE and srpscan uses a similar approach where the initial part is a fast, first-pass prefilter to identify candidate ncRNAs based on simple constraints. The second part uses a highly selective RNA covariance model, which is used to search the sequences derived from the first step. Snoscan first uses a greedy search algorithm based on common motifs in snoRNAs that rapidly scan for 2′-O-methylation guide snoRNA candidates in the genome sequence. An integrated model consisting of both SCFG and HMM then scores the sequences based on the sequence features specific to this RNA gene family. Fisher applies a fast primary sequence search using important H/ACA class snoRNA features together with methods that are based on properties of predicted minimum free-energy (MFE) secondary structures.

Each of the methods described above are restricted to one type of ncRNAs, but attempts have been made to develop identification methods for all possible ncRNAs. Two of these, QRNA and RNAz have been used in our work and will be described here in some detail.

QRNA makes a prediction of ncRNA based on pairwise alignments [153]. It compares the score of three distinct models of sequence evolution to decide which one describes best the given alignment: a pair SCFG is used to model the evolution of secondary structure, a pair hidden Markov model (HMM) describes the evolution of protein coding sequence, and a different pair HMM implements the independent model of a sequence with an evolutionary random pattern not consistent with either a secondary structure or protein coding sequence.

QRNA is currently limited to pairwise alignments, and rather slow for ncRNA gene prediction at a genomic scale. A program similar to QRNA, which tests for complementary mutations in three-sequence multiple alignments, is ddbRNA [154]. It searches for common stems in the multiple alignments in a greedy fashion. The assessment of the significance of the conserved structure is based on shuffled alignments.

The program RNAz makes a prediction of ncRNA based on multiple sequence alignments [155]. It uses two independent criteria for classification: a z-score measuring thermodynamic stability of individual sequences, and a structure conservation index obtained by comparing folding energies of the individual sequences with the predicted consensus folding. The two criteria are then combined to detect conserved and stable RNA secondary structures with high sensitivity and specificity. Yet another application suitable for multiple alignments is MSARI [156]. The approach uses information from a larger set of sequence-aligned orthologs to detect significant ncRNA secondary structures. Primary sequence alignments are often inaccurate. In MSARI, one part of the method tries to correct errors in multiple alignments through energy minimisation calculations.

Methods

Sequence homology methods

In some cases when the evolutionary distance between two species is limited, a simple sequence similarity searches by BLAST [128] or FASTA [129] is sufficient for RNA gene identification. Usually these searches are suitable as a first step when comparing closely related RNA genes.

Pattern matching and covariance models

For the identification of P/MRP RNA as well as IRE we used a combination of pattern searches and secondary structure profile searches with cmsearch of the Infernal package [1, 93]. Nuclear P RNA and MRP RNA sequences are poorly conserved in sequence. However, three conserved regions are shared; CR-I, CR-IV and CR-V. For nuclear P RNA there are also conserved elements in the domain 2 to take into account; CR-II and CR-III. Therefore, for the identification of P and MRP RNA we used a pattern based on consensus features including the CR-I, CR-IV and CR-V motifs as well as base-pairing rules consistent with the helix P2. When a P or MRP RNA gene was not found using these patterns new searches were carried out where mismatches were allowed. After the pattern matching procedure, sequences fitting the secondary structure template were further analyzed with Rfam covariance models. High-scoring candidates were further analyzed for characteristics typical for P/MRP RNA secondary structure; base pairing between the CR-I and CR-V motifs, presence of CR-IV as well as the helices P1, P2 and P3. In the case of P RNA the CR-II and CR-III motifs should also be present (Fig.1).

Also IREs were identified using a combination of pattern matching and covariance models. To identify as many potential IREs as possible we primarily searched available mRNA sequences. In case there was no available mRNA, genomic sequences was searched for regions homologous to available proteins/mRNAs. Whenever an IRE candidate was found in a genomic sequence it was checked for reasonable proximity to the protein/mRNA match. Candidate sequences were checked for conserved primary sequence motifs and the ability to fold into a secondary structure typical for the iron responsive element (Fig.2).

Profile HMMs of highly conserved regions in P and MRP RNA

For prediction of P and MRP RNAs we also used profile HMMs created from CR-I and CR-V multiple alignments. We further analyzed all genomic sequences that contained the CR-I and CR-V motifs and where the distance between the two motifs is less than 3000 bases. Advantages of this method are that large genomes may be searched quickly (100 Mbases in a few minutes) and in a highly specific manner identifies the P and MRP RNA genes. Candidates identified in the search based on HMM profiles were further analyzed to check that other conserved features of the RNA were present.

Identification of protein homologues

An efficient method for protein identification is PSI-BLAST (Position Specific Iterative BLAST) [157, 158]. PSI-BLAST can repeatedly search the target databases, using a multiple alignment of high scoring sequences found in each search round to generate a new more sensitive scoring matrix able to find distantly related sequences that are sometimes missed in a BLAST search. Multiple PSI-BLAST searches with different query sequences were carried out in order to identify as many homologues as possible belonging to a certain protein family. The NCBI Genbank protein set was used as the primary source [159], but additional proteins were identified from individual genome projects or identified from TBLASTN searches of genome sequences. Whenever relevant, these novel sequences were included in the set of sequences used as database in the PSI-BLAST search.
We also used profile HMMs at the Pfam database [84] for Pop1, Pop3 (Rpp38), Pop5, Rpp14, Rpp20, Rpp25, Rpp40, Rpr2 (Rpp21) to identify homologues. In cases where available Pfam models were not sufficient or present, new models were created from multiple alignments and used with the HMMER package to find additional homologues.

To identify homologues to previously known proteins whose mRNAs are known to contain IREs we mainly used BLAST to search the NCBI Genbank set of proteins. Some gene sequences that were not in Genbank were identified by Genewise [160] Genewise uses a combination of comparative analysis (aligns proteins to genomic sequences) together with statistical signals to predict genes.

For classification of proteins we also made use of phylogenetic analysis, including methods of parsimony, maximum likelihood and neighbour-joining.

ncRNA prediction using *de novo* methods

As opposed to the methods that detect new members of already known ncRNA families described previously (IRE and MRP/P RNA identification), we have also used two *de novo* methods, QRNA [153] and RNAz [155], to computationally screen the *S.cerevisae* genome for ncRNAs. These methods are described above in the 'Introduction' section.

Results and discussion

Several ncRNA genes and motifs have been identified using computational methods. On the one hand we have successfully identified ncRNA that belongs to already known families; on the other hand we have used *de novo* methods to find novel ncRNA.

The following known ncRNA families have been studied using new methods:

- P and MRP RNA (paper I), two related ncRNAs genes.
- IRE (paper III), a cis-regulatory mRNA element.

In another project we looked further at the protein subunits of P/MRP particles (paper III). Here we carried out a computational analysis of the protein subunits in a broad range of eukaryotic organisms using profile-based searches and phylogenetic methods.

We have also screened the yeast genome for novel ncRNAs using *de novo* methods (paper IV) that employ general characteristics that apply to almost all known ncRNAs.A selection of candidates from both methods were experimentally tested.

Ribonucleases MRP and P

Using a computational approach we identified more than 100 novel P and MRP RNAs (Paper I). All the RNA sequences have a conserved structural design in domain 1 containing the P1, P2, P3 and P4 helices as well as the CR-I, CR-IV and CR-V regions. In addition RNase P RNA has conserved elements CR-II and CR-III, not present in MRP RNA. Furthermore we observed that the domain 2 of nuclear P RNA contains helices eP8, eP9 and an extra helix 5´ of the eP8 helix, here referred to as eP8´. The three helices eP8´, eP8, and eP9 are highly variable in sequence. For the MRP sequences identified, three helices in domain 2 are found. The stem-loop structure in domain 2 that follows CR-I is conserved with the consensus sequence 'GARAR', in some cases it is also a tetraloop with the consensus GARA.

We believe that the P/MRP RNAs identified are true genes since they all have the consensus properties of such RNAs and some of our predictions of protozoan RNAs were verified experimentally (Fredrik Söderbom, personal communication and Marquez *et al*. [161]). In

some organisms we failed to identify a P or MRP RNA gene. This might be due to incomplete genome sequencing or, alternatively, the gene could be very different from previously known P/MRP RNA genes.

To further examine the relationship of RNase P and MRP RNA to the protein subunits a range of eukaryotic organisms where further analyzed with respect to RNase P and MRP protein subunits. Several homologues were found that were not previously reported (Paper II). An overview of the phylogenetic distribution of the RNA and protein components will be presented below.

Phylogenetic distribution of RNase P and MRP RNA and protein components

Fungi. A number of Saccharomycotina P RNAs were identified, all similar to the well-studied *S.cerevisiae* RNA. The Pezizomycotina species all have a relatively long eP8´ helix. In *Aspergillus* the P3 helix has been extended as compared to other fungi. We identified an RNA also in *Trichoderma reseii* that has an extra helix in the eP8/eP8´ region.
In the Basidiomycota group we found a P RNA in three species. Interestingly, the *Phanerochaete* RNA is 1143 nucleotides long and seems to have a large insertion in the eP15 region, similar to *Candida glabrata*, where a very long P RNA sequence has been identified [162].

MRP RNA was identified in several fungal groups. To the previously known Saccharomycotina we also identified MRP RNA in the Pezizomycotina group. In these organisms the domain 2 is very large, forming a long helical region. Furthermore, MRP RNA was identified in the Basidiomycota organisms. Here the domain 2 seems to be built from three different helices but it is not clear whether they are related to the three helices of other MRP RNAs, or to the ymP5, ymP6 and ymP7 helices of *S. cerevisiae*.

The protein subunits Pop1, Pop3, Pop4, Pop5, Pop6, Pop7, Pop8, Rpp30, Rpp21, Snm1 and Rmp1 are in all Saccharomycotina organisms except *Yarrowia*. Pop6 and Pop8 are the only proteins that seem to be exclusive to Ascomycota. Basidiomycota organisms seem to have a smaller set of proteins than Ascomycota as we have failed to identify Pop6, Pop7, Pop8, Snm1 and Rmp1 homologues.

Microsporidia. The microsporidia P RNAs identified are relatively small with an atypical small helix 3 which is lacking the internal loop characteristic of that helix in most other P RNAs. As with P RNA microsporidia MRP RNAs are exceptionally small in size, in particular the P3 helix and domain 2. It seems as if microsporidia MRP RNAs represent 'minimal' forms of the RNA.

Proteins identified in microsporidia include Pop1, Pop4, Pop5, Rpp1, Rpr2, and Rmp1. Both the small size of the RNAs and the small set of proteins is consistent with the fact that the microsporidia have unusually small genomes (2.5 million bases) and may be considered minimal eukaryotes. Thus, they have retained only the genes that are critical for function and a large number of RNA and protein genes have been reduced in size as compared to other organisms.

Plants, green and red algae. So far, an RNase P RNA has not been identified in plants. In our searches for eukaryotic P and MRP RNAs we failed to identify a P RNA in the plants, green algae and in red algae, suggesting that RNase P is missing in these organisms. Conversely MRP RNA, which previously has been found in Arabidopsis thaliana, was identified also in rice and in green and red algae. These findings indicate that MRP RNA is ubiquitous in the plant group.

One major difference between protein distribution in the plants as compared to other organisms is that Rpp21 (Rpr2) is missing in the plants. Rpp21 is considered specific to RNase P and is not found in MRP, which is in agreement with the observation that a RNase P RNA is not identified in the plants.

There is evidence that Rpp29 (Pop4) and Rpp14 are specific to RNase P and are not present in the human RNase MRP [79]. If this applies to the plant group our failure to identify Rpp14 in plants is consistent with its absence in MRP. On the other hand, if an RNase P is missing in the plant group, the fact that a Rpp29 homologue is present seems to be in conflict with the suggestion that Rpp29 is specific to RNase P.

Insects and nematodes. P RNAs were identified in a number of Drosophila species. The secondary structure that is consistent with all these Drosophila is also consistent with P RNA sequences from other insects. However, only the Drosophila RNAs have an eP8´ helix.

Previously a P RNA sequence was identified in C. elegans. Here we identified homologues also in 2 other nematodes.

As with P RNA we found MRP RNA genes from different *Drosophila* species and other insects. In nematodes we were also able to detect an MRP RNA. The RNA structures of nematodes and insects are very similar to vertebrate MRP RNAs, with the exception of an extended helix eP9 in *Drosophila* structures.

A comparison of the MRP RNAs from plants, red algae, heterokonta, vertebrates, insects and nematodes show that they all may be folded into very similar structures. As discussed below the fungi are different, particularly in domain 2.

The protein repertoire of nematodes and insects are similar to that of humans, the only exception is that an Rpp38 homologue is not found in the genomes.

Other eukaryotic groups. Among the protists we identified a P RNA gene from a number of species. Many of the smallest PRNAs detected are found in this group like *Giardia lamblia*, *Babesia bovis* and *Theileria annulata*. RNase P genes were also identified in a range of *Plasmodium* species. They are larger than the other protist P RNAs and even if they show a large variation in size they are all closely related. A characteristic of *Plasmodium* species is that the helices P3, eP8′, P12 and eP19 have grown considerably with long AU-rich stretches.

Vertebrate MRP RNA sequences that were not previously described were identified in fishes. As with P RNA, we found a MRP candidate in many different *Plasmodium* species. The large degree of variation in sequence of P RNA in the different *Plasmodium* species has no equivalent in the evolution of MRP RNA sequences. The protist group Euglenozoa, with Leishmania and Trypanosoma, is the only phylogenetic group where no RNase P or MRP RNA could be identified.

In vertebrates we found protein homologues for Pop1, Rpp29 (Pop4), Pop5, Rpp20 (Pop7), Rpp30 (Rpp1), Rpp21 (Rpr2), Rpp14, Rpp25, Rpp38, and Rpp40. The set of proteins in the protist group alveolata is similar to that of Metazoa, with the exception that Rpp14, Rpp38 and possibly Rpp25 are missing. In *Giardia lamblia* we failed to identify Rpp20 and Rpp25 (Table 3), proteins proposed to interact with helix P3 [79, 86]. Since the helix P3 is very small

as well as the entire domain 2, the difference in protein repertoire might be related to these differences. The small protein composition in *G. lambia* is similar to the situation in microsporidia organisms and together these ribonucleases represent complexes with a very small subset of proteins (Table 3).

As with P and MRP RNA we were not able to demonstrate a single RNase P or MRP protein subunit in Euglenozoa. Therefore, it seems highly likely that a RNase P/MRP is missing, at least of the type found in other eukaryotes.

Protein relationships in RNases P and MRP

As referred to above, the proteins Pop1, Pop4, Pop5 and Rpp1 are all widely distributed, and were found in all phylogenetic groups except for the Euglenozoa group, where no P and MRP RNA could be identified (Table 3).

Pop3 homologues have previously been reported in *S. cerevisiae* and *C. albicans* [27, 163]. Additional fungal homologues were found indicating that Pop3 is ubiquitous in this group. When Pop3 is used as query in profile-based searches, the L7Ae/L30e domain containing Rpp38 proteins are identified as previously reported [80]. Phylogenetic analyzes further support an orthology relationship between Pop3 and Rpp38 (Table 3).

Human Rpp20 was previously proposed to be the homologue of yeast Pop7 [164] and from searches that we carried out it seems highly likely that proteins Rpp20 and Pop7 are orthologues (Table 3).

Rpp25 and Pop7/Rpp20 were previously shown to be evolutionary related [165]. We also found novel homologues in green algae, heterokonts, *Caenorhabditis* and two different proteins related to Rpp25 in fishes (Fig. 5). A protein related to human Rpp25, referred to as C9orf23 (Chromosome 9 open reading frame 23 protein) was previously reported [166]. We found Rpp25 and C9orf23 paralogues to exist in all vertebrates, including fishes (Fig. 5) while most non-vertebrates had only one Rpp25 homologue. It seems that a gene duplication event took place at a point of evolution close to the development of Deuterostomia, giving rise to the Rpp25 and C9orf23 protein.

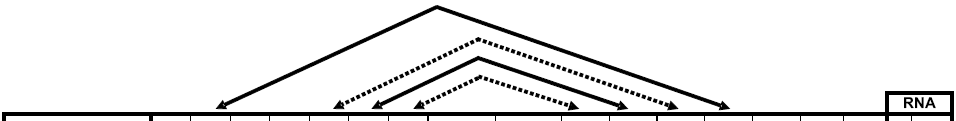
It has previously been noted that Pop5 is homologous to Rpp14 [167]. Pop5 is found in all groups while Rpp14 is found only in Metazoa. A number of obvious Pop5 proteins were erroneously annotated as Rpp14 in protein sequence databases. For this reason a phylogenetic analysis was necessary to classify Rpp14 and Pop5 homologues correctly (Fig. 4).

We also found evidence that Pop8 is related to the Pop5/Rpp14 proteins. This relationship was identified by carrying out PSI-BLAST searches where we included previously unrecognized homologues from Saccharomycotina and Pezizomycotina. Based on these results an attractive possibility is that Pop8 is the fungal orthologue to Rpp14 (Table 3).

Pop6 was previously identified in *S. cerevisiae* [72]. Pop6 homologues were found in a number of Saccharomycotina but not in other fungi. Homology between Rpp25 and Pop6 was suggested from profile-based searches (Table 3). The relationship between Pop6 and Rpp25 is consistent with available protein-RNA and protein-protein interaction data where human MRP Rpp25 seems to have a role similar to that of yeast RNase P Pop6 [74, 78].

The addition of two novel protein relationships between fungal and metazoan protein families (Pop8/Rpp14 and Pop6/Rpp25) means that all the shared fungal RNases P and MRP proteins now have metazoan homologues, suggesting that the metazoan and fungal RNase P and MRP complexes are very similar in terms of both RNA and protein subunits. For other known relationships we have broaden the phylogenetic range and identified further support for previous observations.

Table 3. Simplified phylogenetic distribution of MRP/P RNA and proteins. Gray boxes and white boxes symbolize presence and absence of protein subunits respectively. The arrows denote orthologue relationships between the different groups. The dotted arrows are relationships not described before.



																	RNA	
Phylogenetic groups	**Pop1**	**Pop3**	**Pop4**	**Pop5**	**Pop6**	**Pop7**	**Pop8**	**Rpp1/p30**	**Rpr2/p21**	**Rpp14**	**Rpp20**	**Rpp25**	**Rpp38**	**Rpp40**	**Snm1**	**Rmp1**	**P**	**MRP**
Fungi	■		■	■				■	■						■	■	■	■
Protozoa	■		■	■				■	■			■		■			■	■
Metazoa	■		■	■				■	■	■	■	■	■	■			■	■
Plants & algae	■		■	■				■				■	■					■
Minimal composition																		
Microsporidia	■		■	■				■	■							■	■	■
Giardia lamblia	■		■	■				■	■								■	
Euglenozoa												■						

Structural and evolutionary relationship between P and MRP RNA

P and MRP RNAs are structurally similar and many observations suggest that they are evolutionary related. P and MRP RNA sequences from one organism typically share sequence elements in the P3 region, most often in the lower strand of the internal loop of P3. Analysis of the novel RNAs identified in Paper I reveal notable cases of such sequence conservation (paper I).

Usually the P and MRP RNA genes appear in different locations in the genome. However, in certain protists the two genes appear in tandem with a spacer of approximately 50 nucleotides and are more similar as compared to P/MRP RNA pairs in other organisms. This indicates an early gene duplication in these ancient organisms and supports the close P and MRP RNA relationship.

A K-turn motif was found within the helix P12 in a large number of P and MRP RNA sequences (Fig.5), a motif previously shown to be present in other ribonucleoproteins [168, 169] and for P and MRP RNAs discussed only in the context of human MRP RNA [170]. A K-turn like motif, referred to as a K-loop, is found in most fungal MRP RNAs except for the Saccharomycotina group. In a K-loop the G-A base pairs are connected directly with a loop instead of a longer helical region.

K-turns present in ribosomal RNAs are known to interact with a number of different ribosomal proteins, including the archaeal L7Ae. The archaeal L7Ae protein binds to K-turn [169] as well as K-loop [168] structures and there is evidence that L7Ae is a subunit of the archaeal RNase P [171, 172]. Since Rpp38/Pop3 is related to L7Ae, it is possible these proteins bind to K-turns of P and MRP RNAs (Fig.5). To support this notion Rpp38 is known to interact with the P12 helix [86] [78].

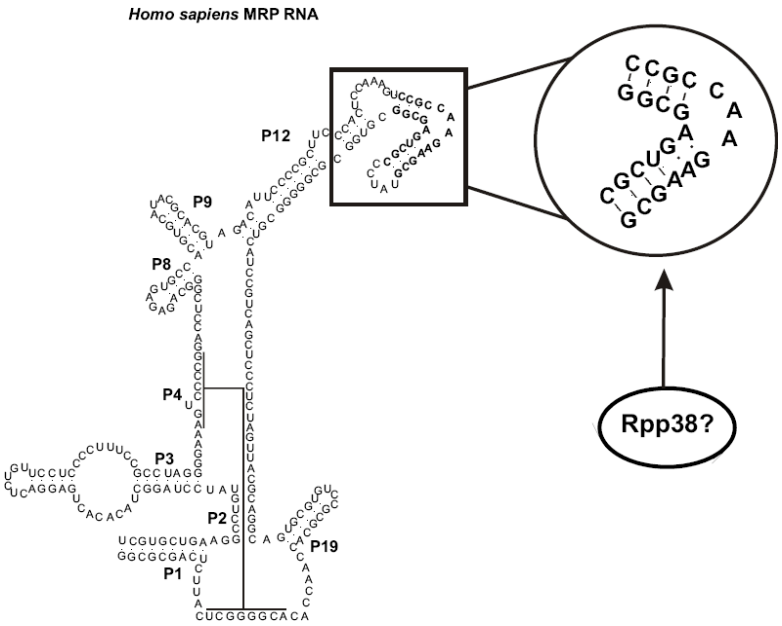


Figure 5. K-turn motif in MRP RNA. The K-turn motif is present in the P12 helix of several P and MRP RNAs identified. The P12 region is marked with a rectangle and the motif is highlighted in the circle. Rpp38 might interact with the the k-turn motif located in P12.

The iron responsive element

The iron responsive element (IRE) is an RNA structural motif critical for regulation of many proteins involved in iron metabolism. Typical features of the IRE include an apical loop sequence (CAGUGN) and two stems separated by either a bulge/loop (UGC/C) or a C-bulge. Based on these features we constructed a method to computationally identify IREs. A large number of novel IRE sequences were identified (paper III). The IRE occurs in a number of mRNA encoding iron-related proteins and these proteins and their phylogenetic distribution will be discussed below as well as the distribution of the IRE.

Ferritin

Ferritin is an almost ubiquitous multi-protein complex higly conserved from bacteria to man. In metazoa, ferritins consists of heavy (H) and light (L) chains. An additional chain form (M) has been found in Xenopus [173]. For all vertebrate organisms where we found more than one ferritin chain, an IRE was identified in the corresponding mRNA. In insects there are two

ferritin chains as well but only the heavy chain (HCH) has an IRE [174, 175], the only exception is seen in butterflies and moths where both chains have the IRE [176, 177]. The fact that IREs are identified in lower metazoa such as Sponges, Cnidaria strongly suggest that the IRE is a very early metazoan invention. We found an IRE in all metazoa, the only exceptions are the worms *C. elegans* [178] and *S. mansoni* [179].

As a rule the ferritin IRE is of the UGC/C type.Variants of the UGC sequence occur in some organisms but the G in second position is always conserved and is able to pair with C in the bulge on the opposing strand which is also supported by structural studies of selected IREs [102, 106]. Another conserved basepair in some UGC/C type IREs is a G-A pair, close to the UGC/C-bulge. This non-canonical pair is seen in invertebrates but not found in vertebrate ferritin IREs or in any non-ferritin IREs.

The C-bulge type is typically associated with non-ferritin transcripts described to contain IRE, exceptions are found in insect and in the Southeast Asian horseshoe crab ferritins where the non-typical C-bulge structure is present.

Transferrin receptor

The mRNA of mammalian transferrin receptor 1 (Tfr1) is unique since it is the only mRNA with multiple IREs (5 IREs in the 3´-UTR). A number of non-vertebrates proteins show a high sequence similarity to Tfr1 and other Tfr1-related proteins but do not clearly associate with Tfr1 in a phylogenetic analysis. At any rate, IREs are absent in the non-vertebrate mRNAs related to Tfr.

We report a number of vertebrate Tfr IREs that were not previously described, mainly in fishes. Fishes are the only group where the multiple IREs are present in two Tfr1 transcripts, Tfr1a and Tfr1b, where Tfr1a is more similar to mammalian Tfr1. The Tfr1a which is found in all vertebrates has a non-canonical apical loop sequence in IRE a (closest to the termination codon) while the IRE a in the fish-specific Tfr1b is more similar to the canonical sequence (CAGUGN). This implies that Tfr1b IREs represents an ancestral form. It is interesting that the non-canonical IRE has been preserved during vertebrate evolution. This suggests that it is associated with an important function.

All IRE sequences as well as their order is extremely conserved and in addition there are three conserved regions between IREs that are able to form short hairpin structures, as previously noted for the human Tfr1 mRNA [103, 180, 181, 182]. In one of these hairpins there are instances of compensatory mutations maintaining the structure and in the two other hairpins the sequences are all the same. These structures are presumably functionally important and may constitute protein binding sites.

IREs of other mRNAs

eALAS. eALAS exist in two forms, H (housekeeping form) and E (erythroid form), both types are present in vertebrates [183]. To previously identified IREs [184] we found additional elements in the E chain transcripts of lower vertebrates such as fishes, frog and sea squirt but not in any organisms outside vertebrates.

Ferroportin. Ferroportin is found in vertebrates, in many invertebrates, in plants, and in some fungal groups. However, a search of IREs in ferroportin mRNAs revealed such elements only in vertebrates.

DMT1. The protein is found in all chordata organisms but only mRNAs from mammals have an IRE. In frog an IRE-related sequence is found with a single nucleotide change in the hexa-loop (CAGUGN). Sequences found in fishes are also similar but the loop sequence is disrupted as well as the hairpin structure characteristic of IRE. This suggests that typical IRE in mammals evolved from a non-IRE sequence by a number of mutations in fish/frog sequences.

Succinate dehydrogenase. An IRE in *Drosophila* mRNA of succinate dehydrogenase, a universal Krebs cycle enzyme, has been reported [185]. The element of succinate dehydrogenase is found in all available *Drosophila* species but nowhere else, suggesting that the IRE in succinate dehydrogenase mRNA is restricted to *Drosophila*.

Some IREs seem to be restricted to higher animals. These are IREs that are not well studied and in case they are involved in transcriptional regulation they appeared at a late stage in evolution. Examples are Cdc14a [117] and the human MRCKα [118].

Evolution of IRE

With bioinformatic methods we have identified over 90 novel IRE sequences in a wide range of metazoan organisms. With a few exceptions, the IRE of ferritin mRNAs is ubiquitous in the metazoa group. In other mRNAs, IREs appeared for the first time at the level of chordata (mACO and eALAS) and therefore we suggest that the IRE/IRP system first operated in ferritin mRNA (Fig.6). In general ferritin IREs are of the UGC/C type while all other mRNAs conform to the C-bulge type. However some ferritin IREs are of the C-bulge type, and a transition to a C-type IRE seems to have occurred more than once.

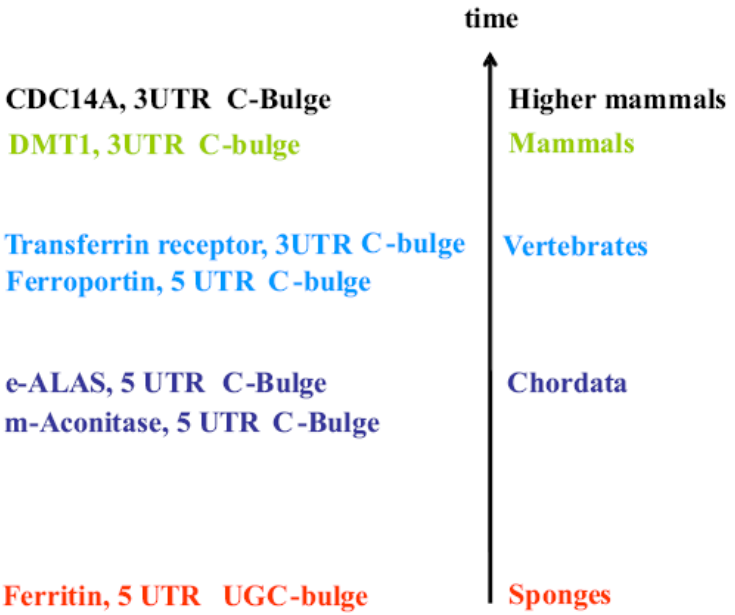


Figure 6. The evolution of the IRE element. IRE appears initially in ferritin mRNAs at the level of early metazoans. In other mRNAs, IREs are introduced first at a point close to chordata speciation or later. All mRNAs except the ferritin carries the C-bulged IRE.

Hunting for ncRNAs in yeast using *de novo* methods

We have conducted computational screens of yeast genomes for novel ncRNAs using two *de novo* methods, implemented into the QRNA and RNAz programs by Rivas *et al*. and Washietl *et al.* respectively [153, 155]. QRNA uses mutational patterns consistent with a conserved RNA secondary structure to discriminate functional RNAs from coding sequences and

random sequences while RNAz uses a combination of free energy and covariance to distinguish ncRNA genes from other genes. Both methods exploit for their prediction genome sequences of closely related organisms.

QRNA predictions

To predict novel ncRNAs, 6347 *S. cerevisiae* intergenic regions were blasted against the genomes of six members of the *Saccharomyces* family. Redundancy, overlaps and candidates with a log odds score below 5 were removed and resulted in 405 candidates.
During our work, a different result of a similar analysis was published by McCutcheon *et al.* [186] with many of our candidates not identified, the overlapping results was discarded. We also wanted to reduce the probability that some of the candidates are not ncRNAs but rather conserved secondary structures within UTR regions by removing sequences proximal to an ORF. A compilation of 245 candidates was finally obtained. From our QRNA candidates we selected 10 for experimental testing using three physiological windows: adaptation phase, exponential growth phase and stationary phase. However, for none of the 10 ncRNA candidates tested was there evidence of expression as judged by northern blots.

We also wanted to examine the possibility that some of the hypothetical ORFs (hORFs) in the current annotation of the yeast genome are mispredictions and actually correspond to ncRNA genes. We extracted from the *S. cerevisiae* genome 1234 hORFs and used them for pairwise comparison against *Saccharomyces* genomes. This resulted in a final list of 146 candidate sequences.

RNAz predictions

In the RNAz approach we made use of 5290 alignments of *S. cerevisiae, S. paradoxus, S. mikatae, and S. bayanus* produced by M. Kellis *et al.* [187], resulting in 1619 ncRNA candidates as predicted by RNAz. As with the QRNA intergenic screen, we discarded candidates that are not likely ncRNAs but rather conserved secondary structures within UTR regions. We also decided to remove sequnces that were poorly aligned. The fact that not all intergenic regions are present in the initial set of alignment and that only a limited number of high quality four-way alignments are available is certainly a limitation of our RNAz approach. From alignments that were approved, RNAz predicted 47 ncRNA candidates. From the final list of RNAz candidates, 20 sequences were experimentally tested. In addition to the

three physiological windows as used when testing QRNA candidates we also used a variety of inhibitors (NaCl, ethidium bromide, paraquat, caffeine and $CdCl_2$). A for the experimentally tested QRNA candidates we could not find any evidence of expression in any of the growth conditions.

Discussion

The negative results from the northern blot experiments could mean that QRNA and RNAz are not efficient in identifying such RNAs, at least given the conditions that we have used. However, it could also be that very few ncRNAs remain to be identified in *S. cerevisiae.* Another factor is the experimental setup which could miss out on true ncRNAs if they are expressed during other conditions than we used or if a ncRNA is expressed at a very low level.

An important issue in the computational screening is the selection of species. It is important that the species chosen for comparison is close enough to be accurately aligned but at the same time it is important that they are at a distance that allows for compensatory mutations. A limitation of the methods that we used is that alignments are based only on primary sequence, for this reason we are missing rapidly evolving RNAs.

The fact that none of the candidates from a yeast transcriptome analysis [188] are predicted by QRNA and RNAz is noteworthy. Such a result questions the methods used here but on the other hand the candidates from the tiling array might not be RNAs that depend on a conserved secondary structure for proper function.

Conclusions

We have used different bioinformatic methods to identify ncRNAs in a variety of organisms. These methods can be classified into two main approaches. First, there are methods that make use of consensus characteristics found in known ncRNA families to find novel ncRNAs (*de novo* methods). We have attempted to use such methods to identify novel ncRNA genes in yeast. Second, there are approaches that make use of characteristics of a specific ncRNA family to find new members of that family. We have used methods of that category to successfully identify a large number of previously unrecognized members of the RNA families RNase P and MRP as well as the iron responsive element. These novel RNA sequences make it possible to better predict the structure of these RNAs as well as to better understand their evolution and function. To further understand the structure and evolution of the RNases P and MRP we also carried out an extensive analysis of the protein subunits of these enzymes. A number of novel homologues were identified and these allowed conclusion as to the relationship between protein subunits. Importantly, the human and yeast enzymes are shown to be more related than previously thought.

Acknowledgements

Special thanks to Tore Samuelsson, my supervisor, for giving me this opportunity in the first place but also for believing in me when my first projects did not go as planned. A also want to thank my dear collegue Magnus Alm Rosenblad for his enthusiasm and good discussions relating to everything from RNAs to snus (snuff).
Moreover, I am very grateful to Anders Blomberg and the research school for funding my project and for letting me be part of his highly talented family.
Other collegues that I have to mention and that made my stay worth while are my hommies Jonathan Esguerra, Daniel Dalevi and Ernest Chi Fru. You guys are really great! Finally I want to thank my parents, my brother and all my friends (outside the university) for your endless support and understanding during my time here.

References

1. Eddy, S.R., *A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.* BMC Bioinformatics, 2002. **3**: p. 18.
2. Gottesman, S., *Stealth regulation: biological circuits with small RNA switches.* Genes Dev, 2002. **16**(22): p. 2829-42.
3. Huttenhofer, A., J. Brosius, and J.P. Bachellerie, *RNomics: identification and function of small, non-messenger RNAs.* Curr Opin Chem Biol, 2002. **6**(6): p. 835-43.
4. Mattick, J.S., *RNA regulation: a new genetics?* Nat Rev Genet, 2004. **5**(4): p. 316-23.
5. Storz, G., J.A. Opdyke, and A. Zhang, *Controlling mRNA stability and translation with small, noncoding RNAs.* Curr Opin Microbiol, 2004. **7**(2): p. 140-4.
6. Nilsen, T.W., *The spliceosome: the most complex macromolecular machine in the cell?* Bioessays, 2003. **25**(12): p. 1147-9.
7. Turner, I.A., et al., *Roles of the U5 snRNP in spliceosome dynamics and catalysis.* Biochem Soc Trans, 2004. **32**(Pt 6): p. 928-31.
8. Valadkhan, S. and J.L. Manley, *Splicing-related catalysis by protein-free snRNAs.* Nature, 2001. **413**(6857): p. 701-7.
9. Valadkhan, S. and J.L. Manley, *Characterization of the catalytic activity of U2 and U6 snRNAs.* Rna, 2003. **9**(7): p. 892-904.
10. Kwek, K.Y., et al., *U1 snRNA associates with TFIIH and regulates transcriptional initiation.* Nat Struct Biol, 2002. **9**(11): p. 800-5.
11. Keenan, R.J., et al., *The signal recognition particle.* Annu Rev Biochem, 2001. **70**: p. 755-75.
12. Rosenblad, M.A., et al., *SRPDB: Signal Recognition Particle Database.* Nucleic Acids Res, 2003. **31**(1): p. 363-4.
13. Dennis, P.P., A. Omer, and T. Lowe, *A guided tour: small RNA function in Archaea.* Mol Microbiol, 2001. **40**(3): p. 509-19.
14. Omer, A.D., et al., *Homologs of small nucleolar RNAs in Archaea.* Science, 2000. **288**(5465): p. 517-22.
15. Weinstein, L.B. and J.A. Steitz, *Guided tours: from precursor snoRNA to functional snoRNP.* Curr Opin Cell Biol, 1999. **11**(3): p. 378-84.
16. Bachellerie, J.P., J. Cavaille, and A. Huttenhofer, *The expanding snoRNA world.* Biochimie, 2002. **84**(8): p. 775-90.
17. Terns, M.P. and R.M. Terns, *Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin.* Gene Expr, 2002. **10**(1-2): p. 17-39.
18. Henras, A.K., C. Dez, and Y. Henry, *RNA structure and function in C/D and H/ACA s(no)RNPs.* Curr Opin Struct Biol, 2004. **14**(3): p. 335-43.
19. Kiss, T., *Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs.* Embo J, 2001. **20**(14): p. 3617-22.
20. Nelson, P., et al., *The microRNA world: small is mighty.* Trends Biochem Sci, 2003. **28**(10): p. 534-40.
21. Ferreira, M.G., K.M. Miller, and J.P. Cooper, *Indecent exposure: when telomeres become uncapped.* Mol Cell, 2004. **13**(1): p. 7-18.
22. Lingner, J., J.P. Cooper, and T.R. Cech, *Telomerase and DNA end replication: no longer a lagging strand problem?* Science, 1995. **269**(5230): p. 1533-4.
23. Kelleher, C., et al., *Telomerase: biochemical considerations for enzyme and substrate.* Trends Biochem Sci, 2002. **27**(11): p. 572-9.

24. Xiao, S., et al., *Characterization of conserved sequence elements in eukaryotic RNase P RNA reveals roles in holoenzyme assembly and tRNA processing.* Rna, 2005. **11**(6): p. 885-96.
25. Frank, D.N. and N.R. Pace, *Ribonuclease P: unity and diversity in a tRNA processing ribozyme.* Annu Rev Biochem, 1998. **67**: p. 153-80.
26. Xiao, S., et al., *Eukaryotic ribonuclease P: a plurality of ribonucleoprotein enzymes.* Annu Rev Biochem, 2002. **71**: p. 165-89.
27. Hartmann, E. and R.K. Hartmann, *The enigma of ribonuclease P evolution.* Trends Genet, 2003. **19**(10): p. 561-9.
28. Mans, R.M., et al., *Interaction of RNase P from Escherichia coli with pseudoknotted structures in viral RNAs.* Nucleic Acids Res, 1990. **18**(12): p. 3479-87.
29. Bothwell, A.L., R.L. Garber, and S. Altman, *Nucleotide sequence and in vitro processing of a precursor molecule to Escherichia coli 4.5 S RNA.* J Biol Chem, 1976. **251**(23): p. 7709-16.
30. Komine, Y., et al., *A tRNA-like structure is present in 10Sa RNA, a small stable RNA from Escherichia coli.* Proc Natl Acad Sci U S A, 1994. **91**(20): p. 9223-7.
31. Li, Y. and S. Altman, *A specific endoribonuclease, RNase P, affects gene expression of polycistronic operon mRNAs.* Proc Natl Acad Sci U S A, 2003. **100**(23): p. 13213-8.
32. Altman, S., et al., *RNase P cleaves transient structures in some riboswitches.* Proc Natl Acad Sci U S A, 2005. **102**(32): p. 11284-9.
33. Chang, D.D. and D.A. Clayton, *A novel endoribonuclease cleaves at a priming site of mouse mitochondrial DNA replication.* Embo J, 1987. **6**(2): p. 409-17.
34. Jacobson, M.R., et al., *Nuclear domains of the RNA subunit of RNase P.* J Cell Sci, 1997. **110 (Pt 7)**: p. 829-37.
35. Li, K., et al., *Subcellular partitioning of MRP RNA assessed by ultrastructural and biochemical analysis.* J Cell Biol, 1994. **124**(6): p. 871-82.
36. Chu, S., et al., *The RNA of RNase MRP is required for normal processing of ribosomal RNA.* Proc Natl Acad Sci U S A, 1994. **91**(2): p. 659-63.
37. Lygerou, Z., et al., *Accurate processing of a eukaryotic precursor ribosomal RNA by ribonuclease MRP in vitro.* Science, 1996. **272**(5259): p. 268-70.
38. Schmitt, M.E. and D.A. Clayton, *Nuclear RNase MRP is required for correct processing of pre-5.8S rRNA in Saccharomyces cerevisiae.* Mol Cell Biol, 1993. **13**(12): p. 7935-41.
39. Cai, T., et al., *The Saccharomyces cerevisiae RNase mitochondrial RNA processing is critical for cell cycle progression at the end of mitosis.* Genetics, 2002. **161**(3): p. 1029-42.
40. Gill, T., et al., *RNase MRP cleaves the CLB2 mRNA to promote cell cycle progression: novel method of mRNA degradation.* Mol Cell Biol, 2004. **24**(3): p. 945-53.
41. Ridanpaa, M., et al., *Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia.* Cell, 2001. **104**(2): p. 195-203.
42. Kazantsev, A.V., et al., *High-resolution structure of RNase P protein from Thermotoga maritima.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7497-502.
43. Loria, A. and T. Pan, *Modular construction for function of a ribonucleoprotein enzyme: the catalytic domain of Bacillus subtilis RNase P complexed with B. subtilis RNase P protein.* Nucleic Acids Res, 2001. **29**(9): p. 1892-7.
44. Niranjanakumari, S., et al., *Protein component of the ribozyme ribonuclease P alters substrate recognition by directly contacting precursor tRNA.* Proc Natl Acad Sci U S A, 1998. **95**(26): p. 15212-7.

45. Pannucci, J.A., et al., *RNase P RNAs from some Archaea are catalytically active.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 7803-8.
46. Chen, J.L. and N.R. Pace, *Identification of the universally conserved core of ribonuclease P RNA.* Rna, 1997. **3**(6): p. 557-60.
47. Forster, A.C. and S. Altman, *Similar cage-shaped structures for the RNA components of all ribonuclease P and ribonuclease MRP enzymes.* Cell, 1990. **62**(3): p. 407-9.
48. Loria, A. and T. Pan, *Domain structure of the ribozyme from eubacterial ribonuclease P.* Rna, 1996. **2**(6): p. 551-63.
49. Torres-Larios, A., et al., *Structure of ribonuclease P--a universal ribozyme.* Curr Opin Struct Biol, 2006. **16**(3): p. 327-35.
50. Krasilnikov, A.S., et al., *Basis for structural diversity in homologous RNAs.* Science, 2004. **306**(5693): p. 104-7.
51. Krasilnikov, A.S., et al., *Crystal structure of the specificity domain of ribonuclease P.* Nature, 2003. **421**(6924): p. 760-4.
52. Torres-Larios, A., et al., *Crystal structure of the RNA component of bacterial ribonuclease P.* Nature, 2005. **437**(7058): p. 584-7.
53. Altman, S., L. Kirsebom, and S. Talbot, *Recent studies of ribonuclease P.* Faseb J, 1993. **7**(1): p. 7-14.
54. Frank, D.N., et al., *Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA.* Rna, 2000. **6**(12): p. 1895-904.
55. Pitulle, C., et al., *Comparative structure analysis of vertebrate ribonuclease P RNA.* Nucleic Acids Res, 1998. **26**(14): p. 3333-9.
56. Christian, E.L., N.M. Kaye, and M.E. Harris, *Evidence for a polynuclear metal ion binding site in the catalytic domain of ribonuclease P RNA.* Embo J, 2002. **21**(9): p. 2253-62.
57. Haas, E.S., et al., *Further perspective on the catalytic core and secondary structure of ribonuclease P RNA.* Proc Natl Acad Sci U S A, 1994. **91**(7): p. 2527-31.
58. Harris, M.E. and N.R. Pace, *Identification of phosphates involved in catalysis by the ribozyme RNase P RNA.* Rna, 1995. **1**(2): p. 210-8.
59. Kazakov, S. and S. Altman, *Site-specific cleavage by metal ion cofactors and inhibitors of M1 RNA, the catalytic subunit of RNase P from Escherichia coli.* Proc Natl Acad Sci U S A, 1991. **88**(20): p. 9193-7.
60. Pagan-Ramos, E., Y. Lee, and D.R. Engelke, *Mutational analysis of Saccharomyces cerevisiae nuclear RNase P: randomization of universally conserved positions in the RNA subunit.* Rna, 1996. **2**(5): p. 441-51.
61. Pagan-Ramos, E., Y. Lee, and D.R. Engelke, *A conserved RNA motif involved in divalent cation utilization by nuclear RNase P.* Rna, 1996. **2**(11): p. 1100-9.
62. Ziehler, W.A., et al., *An essential protein-binding domain of nuclear RNase P RNA.* Rna, 2001. **7**(4): p. 565-75.
63. Biswas, R., et al., *Mapping RNA-protein interactions in ribonuclease P from Escherichia coli using disulfide-linked EDTA-Fe.* J Mol Biol, 2000. **296**(1): p. 19-31.
64. Yuan, Y., E. Tan, and R. Reddy, *The 40-kilodalton to autoantigen associates with nucleotides 21 to 64 of human mitochondrial RNA processing/7-2 RNA in vitro.* Mol Cell Biol, 1991. **11**(10): p. 5266-74.
65. Cate, J.H., et al., *Crystal structure of a group I ribozyme domain: principles of RNA packing.* Science, 1996. **273**(5282): p. 1678-85.
66. Gluck, A., Y. Endo, and I.G. Wool, *The ribosomal RNA identity elements for ricin and for alpha-sarcin: mutations in the putative CG pair that closes a GAGA tetraloop.* Nucleic Acids Res, 1994. **22**(3): p. 321-4.

67. Jaeger, L., F. Michel, and E. Westhof, *Involvement of a GNRA tetraloop in long-range RNA tertiary interactions.* J Mol Biol, 1994. **236**(5): p. 1271-6.
68. Li, X., et al., *Phylogenetic analysis of the structure of RNase MRP RNA in yeasts.* Rna, 2002. **8**(6): p. 740-51.
69. Li, X., et al., *Identification of a functional core in the RNA component of RNase MRP of budding yeasts.* Nucleic Acids Res, 2004. **32**(12): p. 3703-11.
70. Lindahl, L., et al., *Functional equivalence of hairpins in the RNA subunits of RNase MRP and RNase P in Saccharomyces cerevisiae.* Rna, 2000. **6**(5): p. 653-8.
71. Walker, S.C. and D.R. Engelke, *Ribonuclease p: the evolution of an ancient RNA enzyme.* Crit Rev Biochem Mol Biol, 2006. **41**(2): p. 77-102.
72. Chamberlain, J.R., et al., *Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP.* Genes Dev, 1998. **12**(11): p. 1678-90.
73. Guerrier-Takada, C., et al., *The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.* Cell, 1983. **35**(3 Pt 2): p. 849-57.
74. Houser-Scott, F., et al., *Interactions among the protein and RNA subunits of Saccharomyces cerevisiae nuclear RNase P.* Proc Natl Acad Sci U S A, 2002. **99**(5): p. 2684-9.
75. Schmitt, M.E. and D.A. Clayton, *Characterization of a unique protein component of yeast RNase MRP: an RNA-binding protein with a zinc-cluster domain.* Genes Dev, 1994. **8**(21): p. 2617-28.
76. Salinas, K., et al., *Characterization and purification of Saccharomyces cerevisiae RNase MRP reveals a new unique protein component.* J Biol Chem, 2005. **280**(12): p. 11352-60.
77. Jiang, T., C. Guerrier-Takada, and S. Altman, *Protein-RNA interactions in the subunits of human nuclear RNase P.* Rna, 2001. **7**(7): p. 937-41.
78. Welting, T.J., W.J. van Venrooij, and G.J. Pruijn, *Mutual interactions between subunits of the human RNase MRP ribonucleoprotein complex.* Nucleic Acids Res, 2004. **32**(7): p. 2138-46.
79. Welting, T.J., et al., *Differential association of protein subunits with the human RNase MRP and RNase P complexes.* Rna, 2006.
80. Dlakic, M., *3D models of yeast RNase P/MRP proteins Rpp1p and Pop3p.* Rna, 2005. **11**(2): p. 123-7.
81. Hall, T.A. and J.W. Brown, *Archaeal RNase P has multiple protein subunits homologous to eukaryotic nuclear RNase P proteins.* Rna, 2002. **8**(3): p. 296-306.
82. Jiang, T. and S. Altman, *Protein-protein interactions with subunits of human nuclear RNase P.* Proc Natl Acad Sci U S A, 2001. **98**(3): p. 920-5.
83. Kifusa, M., et al., *Protein-protein interactions in the subunits of ribonuclease P in the hyperthermophilic archaeon Pyrococcus horikoshii OT3.* Biosci Biotechnol Biochem, 2005. **69**(6): p. 1209-12.
84. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.
85. Hall, T.A. and J.W. Brown, *Interactions between RNase P protein subunits in archaea.* Archaea, 2004. **1**(4): p. 247-54.
86. Pluk, H., et al., *RNA-protein interactions in the human RNase MRP ribonucleoprotein complex.* Rna, 1999. **5**(4): p. 512-24.
87. Martineau, Y., et al., *Internal ribosome entry site structural motifs conserved among mammalian fibroblast growth factor 1 alternatively spliced mRNAs.* Mol Cell Biol, 2004. **24**(17): p. 7622-35.

88. Bonnal, S., et al., *A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons.* J Biol Chem, 2003. **278**(41): p. 39330-6.
89. Krol, A., *Evolutionarily different RNA motifs and RNA-protein complexes to achieve selenoprotein synthesis.* Biochimie, 2002. **84**(8): p. 765-74.
90. Fagegaltier, D., et al., *Structural analysis of new local features in SECIS RNA hairpins.* Nucleic Acids Res, 2000. **28**(14): p. 2679-89.
91. Brantl, S., *Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes.* Trends Microbiol, 2004. **12**(11): p. 473-5.
92. Nudler, E. and A.S. Mironov, *The riboswitch control of bacterial metabolism.* Trends Biochem Sci, 2004. **29**(1): p. 11-7.
93. Griffiths-Jones, S., et al., *Rfam: an RNA family database.* Nucleic Acids Res, 2003. **31**(1): p. 439-41.
94. Griffiths-Jones, S., et al., *Rfam: annotating non-coding RNAs in complete genomes.* Nucleic Acids Res, 2005. **33**(Database issue): p. D121-4.
95. Liu, C., et al., *NONCODE: an integrated knowledge database of non-coding RNAs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D112-5.
96. Pang, K.C., et al., *RNAdb--a comprehensive mammalian noncoding RNA database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D125-30.
97. Eisenstein, R.S., *Iron regulatory proteins and the molecular control of mammalian iron metabolism.* Annu Rev Nutr, 2000. **20**: p. 627-62.
98. Ponka, P., *Cell biology of heme.* Am J Med Sci, 1999. **318**(4): p. 241-56.
99. Butt, J., et al., *Differences in the RNA binding sites of iron regulatory proteins and potential target diversity.* Proc Natl Acad Sci U S A, 1996. **93**(9): p. 4345-9.
100. Henderson, B.R., E. Menotti, and L.C. Kuhn, *Iron regulatory proteins 1 and 2 bind distinct sets of RNA target sequences.* J Biol Chem, 1996. **271**(9): p. 4900-8.
101. Hentze, M.W., et al., *Identification of the iron-responsive element for the translational regulation of human ferritin mRNA.* Science, 1987. **238**(4833): p. 1570-3.
102. Addess, K.J., et al., *Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins.* J Mol Biol, 1997. **274**(1): p. 72-83.
103. Schlegl, J., et al., *Probing the structure of the regulatory region of human transferrin receptor messenger RNA and its interaction with iron regulatory protein-1.* Rna, 1997. **3**(10): p. 1159-72.
104. Ciftan, S.A., E.C. Theil, and H.H. Thorp, *Oxidation of guanines in the iron-responsive element RNA: similar structures from chemical modification and recent NMR studies.* Chem Biol, 1998. **5**(12): p. 679-87.
105. Ke, Y., et al., *Loops and bulge/loops in iron-responsive element isoforms influence iron regulatory protein binding. Fine-tuning of mRNA regulation?* J Biol Chem, 1998. **273**(37): p. 23637-40.
106. Gdaniec, Z., H. Sierzputowska-Gracz, and E.C. Theil, *Iron regulatory element and internal loop/bulge structure for ferritin mRNA studied by cobalt(III) hexammine binding, molecular modeling, and NMR spectroscopy.* Biochemistry, 1998. **37**(6): p. 1505-12.
107. Meyron-Holtz, E.G., et al., *Genetic ablations of iron regulatory proteins 1 and 2 reveal why iron regulatory protein 2 dominates iron homeostasis.* Embo J, 2004. **23**(2): p. 386-95.
108. Pantopoulos, K., et al., *Differences in the regulation of iron regulatory protein-1 (IRP-1) by extra- and intracellular oxidative stress.* J Biol Chem, 1997. **272**(15): p. 9802-8.

109. Brazzolotto, X., et al., *Human cytoplasmic aconitase (Iron regulatory protein 1) is converted into its [3Fe-4S] form by hydrogen peroxide in vitro but is not activated for iron-responsive element binding.* J Biol Chem, 1999. **274**(31): p. 21625-30.
110. Guo, B., Y. Yu, and E.A. Leibold, *Iron regulates cytoplasmic levels of a novel iron-responsive element-binding protein without aconitase activity.* J Biol Chem, 1994. **269**(39): p. 24252-60.
111. Kennedy, M.C., et al., *Purification and characterization of cytosolic aconitase from beef liver and its relationship to the iron-responsive element binding protein.* Proc Natl Acad Sci U S A, 1992. **89**(24): p. 11730-4.
112. Ke, Y., et al., *Internal loop/bulge and hairpin loop of the iron-responsive element of ferritin mRNA contribute to maximal iron regulatory protein 2 binding and translational regulation in the iso-iron-responsive element/iso-iron regulatory protein family.* Biochemistry, 2000. **39**(20): p. 6235-42.
113. Iwai, K., et al., *Iron-dependent oxidation, ubiquitination, and degradation of iron regulatory protein 2: implications for degradation of oxidized proteins.* Proc Natl Acad Sci U S A, 1998. **95**(9): p. 4924-8.
114. Muckenthaler, M., N.K. Gray, and M.W. Hentze, *IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F.* Mol Cell, 1998. **2**(3): p. 383-8.
115. Wallander, M.L., E.A. Leibold, and R.S. Eisenstein, *Molecular control of vertebrate iron homeostasis by iron regulatory proteins.* Biochim Biophys Acta, 2006. **1763**(7): p. 668-89.
116. Aisen, P., *Transferrin receptor 1.* Int J Biochem Cell Biol, 2004. **36**(11): p. 2137-43.
117. Sanchez, M., et al., *Iron regulation and the cell cycle: identification of an iron-responsive element in the 3'-untranslated region of human cell division cycle 14A mRNA by a refined microarray-based screening strategy.* J Biol Chem, 2006. **281**(32): p. 22865-74.
118. Cmejla, R., J. Petrak, and J. Cmejlova, *A novel iron responsive element in the 3'UTR of human MRCKalpha.* Biochem Biophys Res Commun, 2006. **341**(1): p. 158-66.
119. Kohler, S.A., E. Menotti, and L.C. Kuhn, *Molecular cloning of mouse glycolate oxidase. High evolutionary conservation and presence of an iron-responsive element-like sequence in the mRNA.* J Biol Chem, 1999. **274**(4): p. 2401-7.
120. Rogers, J.T., et al., *An iron-responsive element type II in the 5'-untranslated region of the Alzheimer's amyloid precursor protein transcript.* J Biol Chem, 2002. **277**(47): p. 45518-28.
121. Lin, E., J.H. Graziano, and G.A. Freyer, *Regulation of the 75-kDa subunit of mitochondrial complex I by iron.* J Biol Chem, 2001. **276**(29): p. 27685-92.
122. Anderson, G.J., et al., *Mechanisms of haem and non-haem iron absorption: lessons from inherited disorders of iron metabolism.* Biometals, 2005. **18**(4): p. 339-48.
123. Roetto, A., et al., *Pathogenesis of hyperferritinemia cataract syndrome.* Blood Cells Mol Dis, 2002. **29**(3): p. 532-5.
124. Kato, J., et al., *A mutation, in the iron-responsive element of H ferritin mRNA, causing autosomal dominant iron overload.* Am J Hum Genet, 2001. **69**(1): p. 191-7.
125. Roest Crollius, H., et al., *Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence.* Nat Genet, 2000. **25**(2): p. 235-8.
126. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
127. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.

128. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
129. Pearson, W.R., *Flexible sequence similarity searching with the FASTA3 program package.* Methods Mol Biol, 2000. **132**: p. 185-219.
130. Weber, M.J., *New human and mouse microRNA genes found by homology search.* Febs J, 2005. **272**(1): p. 59-73.
131. Argaman, L., et al., *Novel small RNA-encoding genes in the intergenic regions of Escherichia coli.* Curr Biol, 2001. **11**(12): p. 941-50.
132. Wassarman, K.M., et al., *Identification of novel small RNAs using comparative genomics and microarrays.* Genes Dev, 2001. **15**(13): p. 1637-51.
133. Lim, L.P. and C.B. Burge, *A computational analysis of sequence features involved in recognition of short introns.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11193-8.
134. Carter, R.J., I. Dubchak, and S.R. Holbrook, *A computational approach to identify genes for functional RNAs in genomic sequences.* Nucleic Acids Res, 2001. **29**(19): p. 3928-38.
135. Jovine, L., S. Djordjevic, and D. Rhodes, *The crystal structure of yeast phenylalanine tRNA at 2.0 A resolution: cleavage by Mg(2+) in 15-year old crystals.* J Mol Biol, 2000. **301**(2): p. 401-14.
136. Pesole, G., S. Liuni, and M. D'Souza, *PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance.* Bioinformatics, 2000. **16**(5): p. 439-50.
137. Macke, T.J., et al., *RNAMotif, an RNA secondary structure definition and search algorithm.* Nucleic Acids Res, 2001. **29**(22): p. 4724-35.
138. Klein, R.J. and S.R. Eddy, *RSEARCH: finding homologs of single structured RNA sequences.* BMC Bioinformatics, 2003. **4**: p. 44.
139. Zuker, M., *On finding all suboptimal foldings of an RNA molecule.* Science, 1989. **244**(4900): p. 48-52.
140. Wong A.K.C, C.D.K.Y., *An event-covering method for effective probabilistic inference.* Pattern Recognition, 1987. **20**(No. 2): p. 245-255.
141. Gorodkin, J., et al., *Displaying the information contents of structural RNA alignments: the structure logos.* Comput Appl Biosci, 1997. **13**(6): p. 583-6.
142. Gorodkin, J., et al., *MatrixPlot: visualizing sequence constraints.* Bioinformatics, 1999. **15**(9): p. 769-70.
143. Hofacker, I.L., M. Fekete, and P.F. Stadler, *Secondary structure prediction for aligned RNA sequences.* J Mol Biol, 2002. **319**(5): p. 1059-66.
144. Eddy, S.R. and R. Durbin, *RNA sequence analysis using covariance models.* Nucleic Acids Res, 1994. **22**(11): p. 2079-88.
145. Sankoff, D., *Simultaneous solution of the RNA folding, alignment, and protosequence problems.* SIAM J Appl Math., 1985(45): p. 810-825.
146. Gorodkin, J., L.J. Heyer, and G.D. Stormo, *Finding the most significant common sequence and structure motifs in a set of RNA sequences.* Nucleic Acids Res, 1997. **25**(18): p. 3724-32.
147. Mathews, D.H. and D.H. Turner, *Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.* J Mol Biol, 2002. **317**(2): p. 191-203.
148. Lai, E.C., et al., *Computational identification of Drosophila microRNA genes.* Genome Biol, 2003. **4**(7): p. R42.
149. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.* Nucleic Acids Res, 1997. **25**(5): p. 955-64.
150. Lowe, T.M. and S.R. Eddy, *A computational screen for methylation guide snoRNAs in yeast.* Science, 1999. **283**(5405): p. 1168-71.

151. Edvardsson, S., et al., *A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction.* Bioinformatics, 2003. **19**(7): p. 865-73.
152. Regalia, M., M.A. Rosenblad, and T. Samuelsson, *Prediction of signal recognition particle RNA genes.* Nucleic Acids Res, 2002. **30**(15): p. 3368-77.
153. Rivas, E. and S.R. Eddy, *Noncoding RNA gene detection using comparative sequence analysis.* BMC Bioinformatics, 2001. **2**: p. 8.
154. di Bernardo, D., T. Down, and T. Hubbard, *ddbRNA: detection of conserved secondary structures in multiple alignments.* Bioinformatics, 2003. **19**(13): p. 1606-11.
155. Washietl, S., I.L. Hofacker, and P.F. Stadler, *Fast and reliable prediction of noncoding RNAs.* Proc Natl Acad Sci U S A, 2005. **102**(7): p. 2454-9.
156. Coventry, A., D.J. Kleitman, and B. Berger, *MSARI: multiple sequence alignments for statistical detection of RNA secondary structure.* Proc Natl Acad Sci U S A, 2004. **101**(33): p. 12102-7.
157. Altschul, S.F. and E.V. Koonin, *Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases.* Trends Biochem Sci, 1998. **23**(11): p. 444-7.
158. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
159. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2006. **34**(Database issue): p. D16-20.
160. Birney, E., M. Clamp, and R. Durbin, *GeneWise and Genomewise.* Genome Res, 2004. **14**(5): p. 988-95.
161. Marquez, S.M., et al., *Structural implications of novel diversity in eucaryal RNase P RNA.* Rna, 2005.
162. Dujon, B., et al., *Genome evolution in yeasts.* Nature, 2004. **430**(6995): p. 35-44.
163. Dichtl, B. and D. Tollervey, *Pop3p is essential for the activity of the RNase MRP and RNase P ribonucleoproteins in vivo.* Embo J, 1997. **16**(2): p. 417-29.
164. Stolc, V., A. Katz, and S. Altman, *Rpp2, an essential protein subunit of nuclear RNase P, is required for processing of precursor tRNAs and 35S precursor rRNA in Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 1998. **95**(12): p. 6716-21.
165. Aravind, L., L.M. Iyer, and V. Anantharaman, *The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism.* Genome Biol, 2003. **4**(10): p. R64.
166. Guerrier-Takada, C., et al., *Purification and characterization of Rpp25, an RNA-binding protein subunit of human ribonuclease P.* Rna, 2002. **8**(3): p. 290-5.
167. Koonin, E.V., Y.I. Wolf, and L. Aravind, *Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.* Genome Res, 2001. **11**(2): p. 240-52.
168. Nolivos, S., A.J. Carpousis, and B. Clouet-d'Orval, *The K-loop, a general feature of the Pyrococcus C/D guide RNAs, is an RNA structural motif related to the K-turn.* Nucleic Acids Res, 2005. **33**(20): p. 6507-14.
169. Rozhdestvensky, T.S., et al., *Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea.* Nucleic Acids Res, 2003. **31**(3): p. 869-77.
170. Klein, D.J., et al., *The kink-turn: a new RNA secondary structure motif.* Embo J, 2001. **20**(15): p. 4214-21.
171. Fukuhara, H., et al., *A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3.* Biochem Biophys Res Commun, 2006. **343**(3): p. 956-64.

172. Terada, A., et al., *Characterization of the Archaeal Ribonuclease P Proteins from Pyrococcus horikoshii OT3.* J Biochem (Tokyo), 2006.
173. Dickey, L.F., et al., *Differences in the regulation of messenger RNA for housekeeping and specialized-cell ferritin. A comparison of three distinct ferritin complementary DNAs, the corresponding subunits, and identification of the first processed in amphibia.* J Biol Chem, 1987. **262**(16): p. 7901-7.
174. Charlesworth, A., et al., *Isolation and properties of Drosophila melanogaster ferritin--molecular cloning of a cDNA that encodes one subunit, and localization of the gene on the third chromosome.* Eur J Biochem, 1997. **247**(2): p. 470-5.
175. Georgieva, T., et al., *Iron availability dramatically alters the distribution of ferritin subunit messages in Drosophila melanogaster.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2716-21.
176. Pham, D.Q., et al., *Manduca sexta hemolymph ferritin: cDNA sequence and mRNA expression.* Gene, 1996. **172**(2): p. 255-9.
177. Zhang, D., et al., *Repression of Manduca sexta ferritin synthesis by IRP1/IRE interaction.* Insect Mol Biol, 2001. **10**(6): p. 531-9.
178. Gourley, B.L., et al., *Cytosolic aconitase and ferritin are regulated by iron in Caenorhabditis elegans.* J Biol Chem, 2003. **278**(5): p. 3227-34.
179. Schussler, P., et al., *Ferritin mRNAs in Schistosoma mansoni do not have iron-responsive elements for post-transcriptional regulation.* Eur J Biochem, 1996. **241**(1): p. 64-9.
180. Casey, J.L., et al., *Iron regulation of transferrin receptor mRNA levels requires iron-responsive elements and a rapid turnover determinant in the 3' untranslated region of the mRNA.* Embo J, 1989. **8**(12): p. 3693-9.
181. Koeller, D.M., et al., *A cytosolic protein binds to structural elements within the iron regulatory region of the transferrin receptor mRNA.* Proc Natl Acad Sci U S A, 1989. **86**(10): p. 3574-8.
182. Chan, L.N., et al., *Chicken transferrin receptor gene: conservation 3' noncoding sequences and expression in erythroid cells.* Nucleic Acids Res, 1989. **17**(10): p. 3763-71.
183. Sadlon, T.J., et al., *Regulation of erythroid 5-aminolevulinate synthase expression during erythropoiesis.* Int J Biochem Cell Biol, 1999. **31**(10): p. 1153-67.
184. Duncan, R., et al., *Phylogenetic analysis of the 5-aminolevulinate synthase gene.* Mol Biol Evol, 1999. **16**(3): p. 383-96.
185. Gray, N.K., et al., *Translational regulation of mammalian and Drosophila citric acid cycle enzymes via iron-responsive elements.* Proc Natl Acad Sci U S A, 1996. **93**(10): p. 4925-30.
186. McCutcheon, J.P. and S.R. Eddy, *Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics.* Nucleic Acids Res, 2003. **31**(14): p. 4119-28.
187. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.
188. David, L., et al., *A high-resolution map of transcription in the yeast genome.* Proc Natl Acad Sci U S A, 2006. **103**(14): p. 5320-5.