# Computational identification of RNA and protein components from the signal recognition particle

Magnus Alm Rosenblad
*Dept. of Medical Biochemistry*

2005

The Sahlgrenska Academy
AT 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 summarises the accompanying papers. These papers have already been published or are in manuscripts at various stages (in press, submitted or accepted).


Department of Medical Biochemistry
The Sahlgrenska Academy at Göteborg University
Sweden

Printed by Intellecta Docusys, Göteborg, 2005

ISBN 91-628-6416-5

# Abstract

Magnus Alm Rosenblad **Computational identification of RNA and protein components from the Signal Recognition Particle** *Department of Medical Biochemistry, The Sahlgrenska Academy, Göteborg University, Medicinaregatan 9A, Box 440, SE-403 50 Göteborg, Sweden*

**Problem.** The signal recognition particle (SRP) is a ribonucleoprotein particle that targets proteins to the endoplasmic reticulum in eukaryotes, to the plasma membrane in Archaea and Bacteria and to the thylakoid membrane in chloroplasts of photosynthetic organisms. It has one RNA component and 1–6 proteins. The eukaryotic particle is composed of one S domain responsible for signal recognition and one Alu domain responsible for translation elongation arrest. In many phylogenetic groups the SRP is not characterized. Therefore, we aim to identify SRP component genes by computational screening of a large number of organisms where genomic information is available.

**Methods.** For the protein gene identification, we relied on methods based on primary sequence alignments (BLAST, FASTA), profile searches (PSI-BLAST, HMMER, Profilescan), and secondary structure prediction (PSI-Pred). The main focus in this work is the identification of SRP RNA. It is highly diverse in its structure and has a low primary sequence conservation between different phylogenetic groups. As a consequence, standard sequence analysis tools, such as BLAST, are not useful. We have developed a tool for the identification of SRP RNA (SRPscan) using algorithms for pattern matching and covariance analysis of secondary structures.

**Results.** We have carried out an extensive inventory of SRP components by screening available genomic sequences. As a result we have identified a large number of novel genes. The protein and RNA sequences are presented in the SRP database (SRPDB). We have identified full or partial SRP RNA genes in virtually all organisms where genomic sequences of nearly full genome coverage are available, and the findings have led to a proposal of a new nomenclature for SRP RNA.

In an analysis of bacterial RNAs we found species with an unusual URRC tetraloop and we identified an RNA from deeply branching gram-negative bacterium *Thermotoga* that is of the gram-positive Bacillus type. It was previously believed that chloroplasts do not have an SRP RNA. However, we have shown that chloroplast genomes of red algae or red algal origin, as well as some green algae, encode a bacterial type SRP RNA.

Eukaryotic SRP RNAs are highly divergent in their structures, mainly in the Alu domain. Based on an analysis of fungal RNAs we were able to present a novel secondary structure model of these RNAs. Analysis of eukaryotic RNAs includes a number of unexpected findings. In the fungal groups Basidiomycota and Zygomycota the SRP RNA has an Alu domain that conforms to the standard mammalian SRP RNA structure. The external loop of helix 8 is a tetraloop as a rule, but in several protists this sequence is a pentaloop. Finally, we suggest that some eukaryal species like Microsporidia might lack an SRP Alu domain.

**Conclusion.** By computational screening of genomic sequences we have identified a large number of novel SRP RNA and proteins components. The results of these studies provide significant insights into the structure, function and phylogeny of the SRP.

**Key words:** signal recognition particle, SRP, RNA secondary structure, non-coding RNA
**ISBN** 91-628-6416-5 Göteborg 2005

# List of separate publications

**I.** **Prediction of signal recognition particle RNA genes**
Regalia,M., Rosenblad,M.A., Samuelsson,T.
*Nucleic Acids Research*, Vol. 30, Nr. 15, Aug 2002.

**II.** **SRPDB: Signal Recognition Particle Database**
Rosenblad,M.A., Gorodkin,J., Knudsen,B., Zwieb,C. and Samuelsson,T.
*Nucleic Acids Research*, Vol. 31, Nr. 1, 2003.

**III.** **Identification and comparative analysis of components from the signal recognition particle in protozoa and fungi**
Rosenblad,M.A., Zwieb,C. and Samuelsson,T.
*BMC Genomics*, 2004, 5:5.

**IV.** **Identification of chloroplast signal recognition particle RNA genes**
Rosenblad,M.A. and Samuelsson,T.
*Plant and Cell Physiology* 45(11) November 2004

**V.** **A nomenclature for all signal recognition particle RNAs**
Zwieb,C., van Nues,R.W., Rosenblad,M.A., Brown,J.W. and Samuelsson,T.
*RNA* **11 (1)** January 2005.

## Table of contents

# Introduction

## Signal Recognition Particle

The signal hypothesis, first formulated in 1971 by Blobel and Sabatini and developed by Blobel and Dobberstein in 1975, showed how proteins that are destined for export from the cell entered the secretory apparatus. Secretory and plasma membrane proteins are synthesized with an amino-terminal ‘signal’ peptide that directs the nascent chain to the membrane. On reaching the membrane, the signal sequence facilitates the co-translational translocation of the polypeptide across the membrane, where the signal peptide is usually cleaved off. The signal recognition particle (SRP) plays an important part in this machinery.

SRP is a ribonucleoprotein particle found in all three kingdoms of life. It binds to the signal peptide emerging from the exit site of the ribosome and then directs the ribosome-nascent-chain-SRP (RNC-SRP) complex to the membrane. When the RNC-SRP complex has been formed there is an arrest in protein synthesis elongation activity. The complex eventually binds to the membrane-anchored SRP receptor (SR) in a GTP-dependent manner. SRP and SR is then released from the complex. Protein synthesis is resumed and the peptide is translocated into or across the membrane through a channel that is part of the translocon. [1]

SRP targets proteins to the endoplasmic reticulum (ER) i eukaryotes, to the plasma membrane in Archaea and Bacteria, and to the thylakoid membrane in chloroplasts of photosynthetic organisms. No SRP or SRP receptor homologues have been detected in mitochondria. Proteins targeted by SRP may be integral membrane proteins or secretory proteins in eukaryotes, but in *E.coli* the SRP pathway is used mainly for integral inner-membrane proteins.

Where as SRP usually has part in a co-translational mechanism, it has recently been shown that SRP also mediates post-translational targeting of tail-anchored proteins to the ER in eukaryotes [2]. Furthermore,  in chloroplasts there is a post-translational SRP

pathway for nuclear encoded proteins imported to the chloroplast destined for the thylakoid membrane [3].

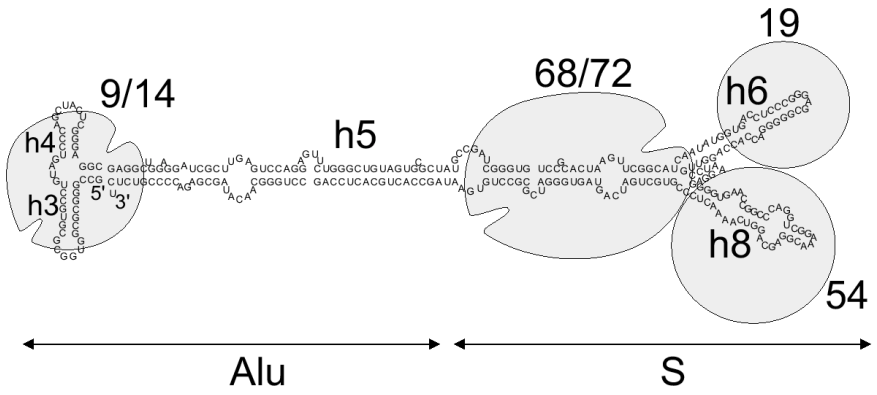


**Figure 1.** *The mammalian SRP. Proteins SRP9/14 etc. binds to different parts of the backbone RNA. Helices of RNA are labeled h3, h4 etc. Alu and S domains indicated below SRP.*

## The components of SRP

SRP is a ribonucloprotein particle, with 1–6 proteins and one RNA molecule. Mammalian SRP can be experimentally separated into two functional domains by micrococcal nuclease: the Alu and S domains [4]. The Alu domain (also called “small domain”) comprises the Alu part (helices 1–4 and part of helix 5) of SRP RNA, and proteins SRP9/14. The domain was named 'Alu' domain because it is evolutionary related to the Alu repetitive element in higher animals, in turn named after its *Alu*I cleavage site. The S-domain (“large domain”) contains part of helix 5 and helices 6–8 (or just helix 8) and proteins SRP19, 54 and SRP 68/72 (or just SRP54). (Fig. 1, X)

The elongation arrest function resides in the Alu-domain. The SRP–ribosome (elongation arrested)–signal sequence structure has recently been solved by cryo-electron microscopy and shows that the signal sequence binding site of SRP binds to the exit site and that the Alu domain binds to the elongation factor binding site [5].

Reconstitution experiments using eukaryotic SRP indicate that SRP19 is required for SRP54 to associate with SRP RNA, SRP9/SRP14 are required for elongation arrest, and

chemical modification of SRP68/SRP72 prevents the close interaction of SRP with SR and inhibits translocation-promoting activity.

The proteins (SRP9/14, SRP19/54, SRP68/72) are named after their molecular masses. In bacteria SRP54 is referred to as Ffh, an abbreviation of "fifty-four homologue". Properties of the different SRP proteins are summarized in Table 1.

| COMPONENT | DOM. | PROPERTIES | KINGDOMS |
|---|---|---|---|
| **SRP RNA** | | The backbone of SRP to which proteins bind. | All kingdoms |
| **SRP9*** | Alu | Binds to SRP14 to form a heterodimer that binds to the Alu domain of SRP RNA. | Eukaryotes (except some protozoa) |
| **SRP21*** | Alu | The fungal SRP9 homologue, exact binding/function not clear. Binds to SRP14. | Fungi |
| **SRP14*** | Alu | Binds to SRP9 and Alu domain SRP RNA. | Fungi, Eukaryotes (except some protozoa) |
| **SRP68** | S | Binds to the SRP RNA helix 5, then SRP72 binds to complex. | Eukaryotes |
| **SRP72** | S | Binds to SRP68-SRP RNA. SRP68/72 facilitates binding to the ribosome. | Eukaryotes (not identified in all protozoa so far) |
| **SRP19** (Sec65 in yeast.) | S | Required for the binding of SRP54 to SRP RNA. Binds to the tips of helix 6 and helix 8 causing a conformational change in the asymmetric bulge/loop in helix 8 which favours binding to SRP54. | Eukaryotes, Archaea |
| **SRP54**<br>**Ffh**<br>**cpSRP54** | S | The most important component of the SRP as it is conserved in all domains of life, including chloroplasts. It comprises an aminoterminal domain (N), a central GTPase domain (G) and a methionine rich terminal domain (M) that binds to the signal sequence (nascent chain) and anchors SRP54 to SRP RNA. The NG domains are also in the SRP receptor (eukaryal SRalfa, bacterial FtsY).<br>cpSRP54 is a Ffh homologue. | Eukaryotes, Archaea, Bacteria, Chloroplasts |

**Table 1.** *Properties of SRP components. * indicates that SRP9/14/21 are homologues, i.e. they share a common ancestor. cpSRP43 not included in table.*

The SRP54 protein and the SR alpha subunit of the SRP receptor (SR), FtsY in bacteria, are GTPases and the GTPase activity of these proteins play an important role in the targeting process. GTP binding to SRP54 is required for targeting to the SR and in an unusual mechanism the GTPase activities of SRP54 and SR alpha are reciprocally

stimulated so as to dissociate SRP from SR once the signal peptide has been delivered to the translocon.

## The different SRP types

One can distinguish two main types of the SRP: one in higher eukaryotes with an RNA of approx. 300 nt and 6 proteins, and one minimal in bacteria with an RNA of approx. 100 nt and a single protein. There are however several variants of these SRP types (Table 2). Different SRP RNAs are shown in Fig. 2 and 3.

**Eubacteria.** The small bacterial SRP (4.5S RNA + Ffh) in *E.coli* and most other eubacteria acts together with the SRP receptor FtsY, and plays a role in the insertion of certain plasma membrane proteins and may also be involved in protein secretion. However, *E. coli* has other parallel pathways (for instance the Sec pathway) for translocation of proteins. A translational arrest associated with SRP mediated targeting has not been identified in *E. coli*.

The Bacillus type SRP RNA (with an Alu domain, Fig. 2) is found in *Bacillales* (incl. *Bacillus, Listeria, Staphylococcus* and others) and *Clostridia* (incl. *Clostridium* and *Thermoanaerobacter*). The difference to the archaeal SRP is that the helix 6 is missing in Bacillus type SRP RNA and that it lacks SRP19.

It is not clear whether the Bacillus type SRP arrests translation. It has been suggested that the Bacillus protein HBsu folds into a structure similar to the SRP9/14 heterodimer and might serve as a functional analogue to the eukaryotic SRP9/14 heterodimer, which is involved in translational arrest acitivity. HBsu is essential for vegetative growth, whereas a truncated SRP RNA, lacking its Alu domain, does not affect vegetative growth. The Alu domain is, however, required for spore formation [6].

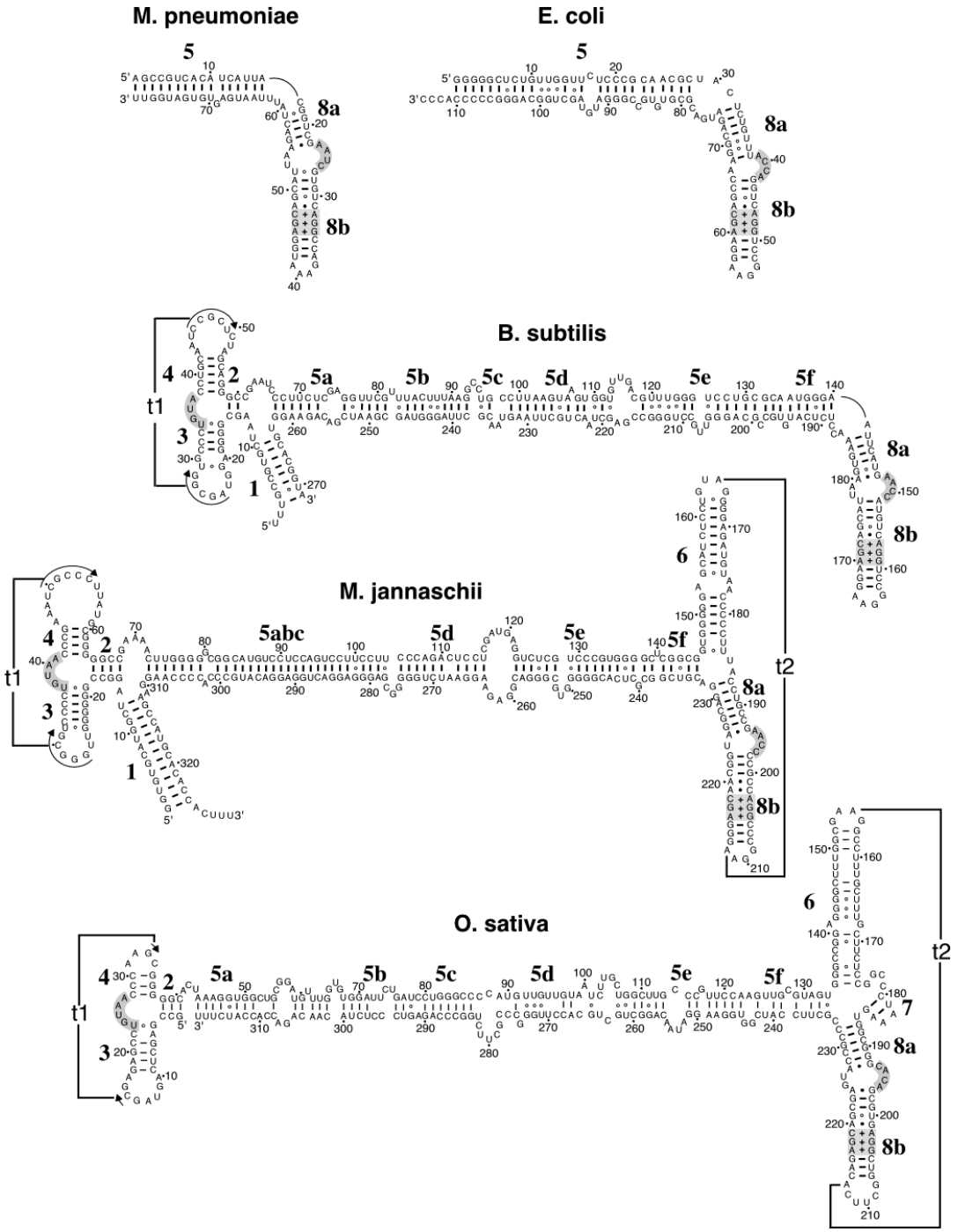


**Figure 2.** *SRP RNA of Bacteria, Archaea* (M.jannaschii) *and Plants* (O.sativa). *Helices are numbered. Tertiary interactions indicated by t1 and t2. A conserved motif is found in region 5e. The Alu domain UGUNR motif is shaded.*

**Archaea.** In mammals SRP19 is needed for binding of SRP54 to the RNA. However, in Archaea, SRP54 has been shown to bind to SRP RNA in the absence of SRP19. No protein associated with the Alu domain of the RNA has been reported. The helix 1 in SRP RNA is only found in Archaea and the Bacillus type SRP RNA. (*M.jannaschii* and *B.subtilis* in Fig. 2)

**Eukaryotes.** The "standard" SRP in eukaryotes is the mammalian SRP with a 7S RNA and proteins SRP9/14, SRP68/72 and SRP19/54. Examples of RNA are shown in Fig.2 and 3. There are, however, many different variations in the eukaryal SRPs, where the *Saccharomyces* SRP is the most divergent with an RNA of >500 nucleotides (main insertions being in helix 5 but also in helix 7) and SRP21 replaces SRP9. All Ascomycota fungi have a small Alu domain RNA with no helix 3 or 4. Another special case is the SRP in Trypanosomatids. There is evidence that there is an extra "tRNA-like" RNA in the SRP of the Trypanosomatids *Trypanosoma brucei* and *Leptosomas collosoma*, referred to as sRNA-76 or sRNA-85 respectively [7, 8]. Because the Alu domain of Trypanosoma SRP RNA is unusually small it has been proposed that the extra tRNA-like RNAs substitute for parts of the Alu domain that are missing.

**Chloroplast SRP (cpSRP)**. The cpSRP pathways are used for the insertion of integral membrane proteins into the thylakoid membrane and have been investigated in higher plants. The *post*-translational cpSRP pathway is involved in targeting of members of the nuclear encoded light-harvesting chlorophyll *a*/*b*-binding proteins (LHCP), and it consists of cpSRP54/43. It does not need an RNA component and does not seem to have one. The *co*-translational cpSRP, used for chloroplast encoded proteins, consists (at least) of cpSRP54. So far no RNA has been experimentally verified in the co-translational cpSRP. The cpSRP43 protein is unique to plants and has an unknown origin.

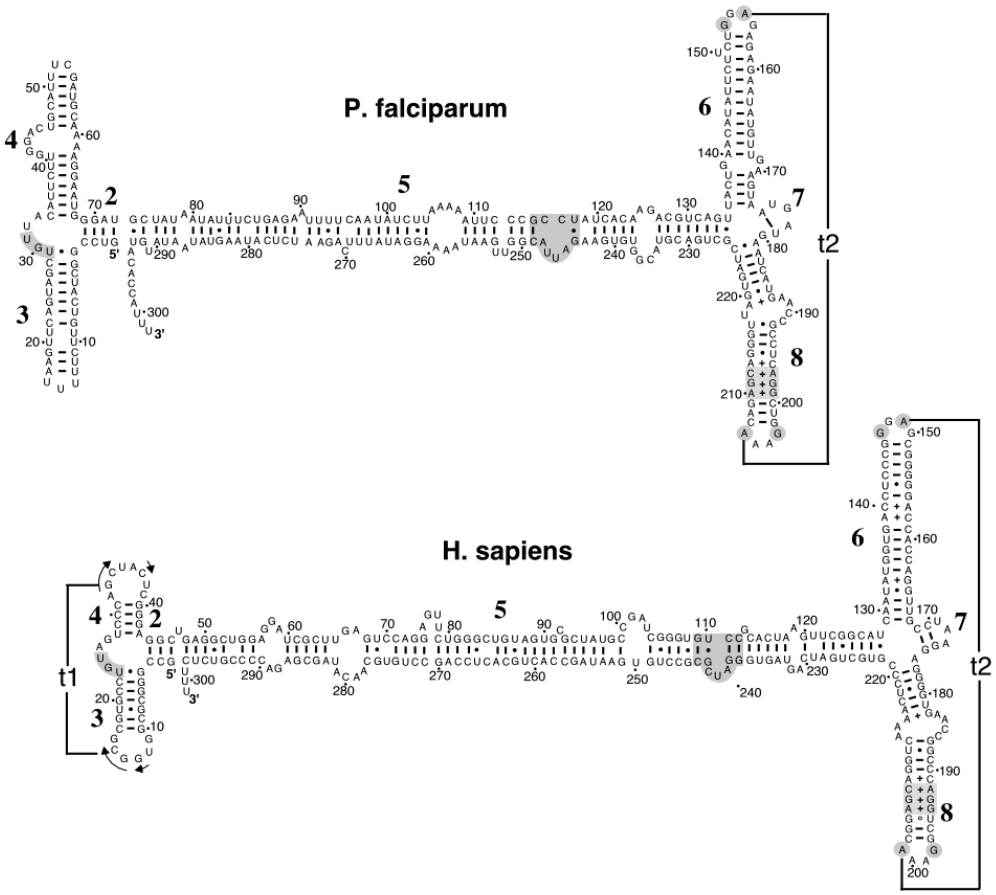


**Figure 3.** *SRP RNA of human and Plasmodium, the latter with long helices 3 and 4 (Paper III). UGU of the UGUNR Alu motif is shaded (between helices 3 and 4). The conserved helix 5e motif is shaded. Tertiary interactions indicated by t1 and t2.*

## Organisation of SRP genes: pseudogenes, multiple gene copies

Pseudogenes are similar in sequence to normal genes, but they usually contain mutations that disrupt expression or function. The genome-wide scans in human, mouse, fly, worm and yeast for pseudogenes are all focused on protein coding genes (http://bioinfo.mbb.yale.edu/genome/pseudogene).

Several RNA genes have many gene copies. *E.coli* has seven copies of its rRNA genes and human cells contain about 200 rRNA gene copies per haploid genome. The human pol III–transcribed 5S rRNA of the large ribosomal subunit is, however, present in 2000 copies arranged in a single cluster far from all other rRNA genes.[9]

The SRP RNA gene occurs both as single and in multiple copies. The largest number found so far is eight in *Arabidopsis thaliana* (Paper I). Promoter analysis may be used to distinguish these pseudogenes from true genes, as in the case with *Arabidopsis thaliana.*

SRP protein genes are found as single copies and no pseudogenes have been identified so far.

## Promoters

**Promoters in Prokaryotes.** In prokaryotes the genes are organized in operons with a common upstream promoter, and they are all transcribed by the same RNA polymerase.

Promoter specificity is accomplished through the polymerase sigma subunit. While the basic principles of promoter recognition and geometry of RNA polymerase are similar for all eubacteria, the actual sequences of promoter elements might not be. In some cases a promoter analysis may be useful in the prediction of an SRP RNA gene but this is not generally applicable.

**Promoters and polymerase in chloroplasts.** The plastid genome is known to be transcribed by a plastid-encoded prokaryotic-type RNA polymerase (PEP) and by at least one nucleus-encoded phage-type RNA polymerase (NEP). No chloroplast-specific studies of promoters have been published so far. However, promoters in cyanobacteria have been examined [10].

**Promoters in Eukaryotes.** There are three transcription systems, Pol I, Pol II and Pol III. in eukaryotes, and SRP RNA is transcribed from DNA to RNA by RNA polymerase III. The 519-bp-long SRP RNA *SCR1* gene in *S. cerevisiae* is the longest known pol III transcriptional unit. RNApol III also transcribes 5S rRNA (all other rRNA are transcribed by pol I), tRNA, 7SK RNA and U6 RNA.

Promoter elements located upstream as well as downstream of the transcription start site were found in human SRP RNA (Ullu & Weiner 1985; Bredow 1990). Contrary to the human genes, plant SRP RNA gene transcription only requires an upstream promoter composed of a TATA box and an upstream stimulatory element called USE (identical to that of all plant U-snRNA gene promoters). Yet another promoter organization is found in the SRP RNA genes of protozoans of the family Trypanosomatidae, whose transcription depends on the A- and B-blocks of a divergently oriented, companion tRNA gene

positioned 100 bp upstream of the transcription start site. The SRP RNA genes of the yeasts *Schizosaccharomyces pombe* and *S. cerevisiae* both contain intragenic sequences resembling the A- and B-blocks, but an upstream TATA box has been shown to play an essential transcriptional role in the *S. pombe* SRP RNA gene [11].

An important property of the upstream promoter sequences is that they are not conserved between organisms or even between pol III–transcribed genes of the same organism, and that the degree by which they determine gene activity varies greatly. For this reason they are of limited value in the prediction of pol III genes.

### RNA transcription and processing

RNA processing of SRP RNA have so far only been identified in bacteria. The small SRP 4.5S RNA of many eubacteria is cleaved by RNase P near the 5’ end after transcription [12]. The *B. subtilis* SRP RNA is cleaved at both the 5' and the 3' end by RNase III [13]. The SRP RNA is transcribed as a 354-nucleotide transcript which is cleaved to a 275-nucleotide intermediate with a two-base 3' overhang.

Ribosomal RNA and other structural RNAs — such as tRNAs and snRNAs — are known to be extensively modified. Internal subsequences, introns, that are removed (“spliced”) are of several kinds, for instance selfsplicing group I and II introns, and those that are removed by the spliceosome (“spliceosomal” intron). More than 1200 introns have been documented at over 150 unique sites in the small and large subunit ribosomal RNA genes (as of February 2002). No introns have so far been identified in pre-SRP RNA.

### Non-coding RNAs

The “central dogma” was sketched by James Watson as early as 1952, stating that there must be a coding RNA (“messenger RNA”) that is passed from the DNA to the protein synthetic machinery found in the cytoplasm. The heart of this machinery is the ribosome, a ribonucleoprotein complex composed of stable ribosomal RNAs (rRNA) and several proteins. Another RNAs was predicted by Francis Crick's “adaptor” hypothesis: transfer RNA (tRNA).

| Process | Example | Type | Function |
|---|---|---|---|
| Transcription | 184-nt *E. coli* 6*S* | sRNA | Modulates promoter use |
| | 331-nt human 7SK | | Inhibits transcription elongation factor P-TEFb |
| | 875-nt human SRA | | Steroid receptor coactivator |
| Gene silencing | 16,500-nt human *Xist* | | Required for X-chromosome inactivation |
| | ~100,000-nt human *Air* | | Required for autosomal gene imprinting |
| Replication | 451-nt human telomerase RNA | | Core of telomerase and telomere template |
| RNA processing | 377-nt *E. coli* RNase PRNA | | Catalytic core of RNase P |
| | 186-nt human U2 snRNA | snRNA | Core of spliceosome |
| RNA modification | 102-nt *S. cerevisiae* U18 C/D snoRNA | snoRNA | Directs 2'-*O*-ribose methylation of rRNA |
| | 189-nt *S. cerevisiae* snR8 H/ACA snoRNA | snoRNA | Directs pseudouridylation of target rRNA |
| RNA stability | 80-nt *E. coli* RyhB sRNA | sRNA | Targets mRNAs for degradation? |
| mRNA translation | 87-nt *E. coli* DsrA sRNA | sRNA | Activates translation by preventing formation of an inhibitory mRNA structure |
| | 22-nt *C. elegans lin-4* miRNA | miRNA (stRNA) | Represses translation by pairing with 3' end of target mRNA |
| Protein stability | 363-nt *E. coli* tmRNA | sRNA | Directs addition of tag to peptides on stalled ribosomes |
| Protein translocation | 114-nt *E. coli* (SRP) 4.5*S* RNA, 300 nt *H.sapiens* (SRP) 7S RNA | sRNA, scRNA | Component of SRP central to protein translocation across membranes |
| Posttranscriptional gene silencing (PTGS), "immune system" | 25 nt antisense, in plants (animals) and fungi transformed with foreign or endogenous DNA | siRNA, miRNA? | Nucleotide sequence-specific defense mechanism that can target both cellular and viral mRNAs |

**Table 2.** *Examples of cellular processes that involve non-coding RNAs. sRNA = small RNA (bacteria); snRNA = small nuclear RNA; snoRNA = small nucleolar RNA; scRNA = small cytoplasmic RNA; miRNA = micro RNA; siRNA = small interfering RNA.*

As more non-mRNA molecules were discovered, the term non-coding RNA (ncRNA) was coined (Olivas 1997). Where as a large number of ncRNAs now are known, the only ncRNAs that are found in all domains of life are rRNA, tRNA, SRP RNA and RNaseP RNA. Examples of ncRNAs are listed in Table 2.

## Gene identification

Five criteria of gene identification are in common use, but their application is not straightforward. Most of them are for protein genes.

1. **Open reading frames (ORFs).** An ORF is a string of codons bounded by start and stop signals, where codons are nucleotide triplets encoding amino acids. For ncRNAs there is no equivalent to ORFs.

2. **Sequence features.** Once an ORF is identified, codon bias is often used to determine whether the ORF is a gene. The value of this measure stems from the fact that genes, particularly highly expressed genes, exhibit biased nonrandom use of codons. However, for many genes, the bias is weak. In only a few examples has it been possible to detect ncRNA genes on the basis of nucleotide frequencies [14].

3. **Sequence conservation.** In contrast to focusing on an individual DNA sequence, genes can be identified by comparing multiple sequences among organisms. It requires sequences of related organisms that are separated by appropriate evolutionary distances. In the case of ncRNAs, the primary sequence is not conserved and one must therefore search for conserved secondary structures.

4. **Evidence for transcription.** Another approach for identifying genes is to search for RNA or protein expression, for instance by mining databases for EST support. EST databases contain sequences that represent partial mRNA sequences and contain very little information on ncRNAs.

5. **Gene inactivation.** One method for studying a gene's function is to mutate or inactivate its product. This can be done both for protein and ncRNA genes.

In the list above, the last three (3, 4 and 5) are applicable to RNA genes, and 3) should be modified to include secondary structure. For some RNA genes, the primary sequence

may be similar enough to identify homologues, if the evolutionary distance is very small. This is the case for mammalian SRP RNA.

Beyond the five criteria, there are additional issues in gene identification such as overlap, alternative splicing, and pseudogenes. All these all applicable to RNA genes, but alternative splicing is very uncommon (an example is the Sphinx ncRNA)[15].

In gene prediction we may also make use of promoters and terminator signals. As more genomes are fully annotated, it is also possible to use synteny to identify homologues.

### Tools for the identification of proteins and protein genes

Common tools for protein gene prediction are Genscan [16] and Genewise [17]. Searches at the protein level may be carried out using BLAST and PSI-BLAST [18], HMMER (Pfam) [19] and Profilesearch [20].

The identification of protein genes is even more powerful in cases where a protein matches a Pfam model because then a Pfam model search may be combined with gene prediction algorithms. Such a strategy is implemented in Genewise. Genewise can be thought of as considering every possible gene prediction in a genomic sequence and comparing each one to the protein profile-HMM.

The existing gene prediction methods are almost exclusively tailored for protein gene prediction. The reason for this is not only that the RNA genes are fewer and therefore not as “important” as the protein genes, but also that the properties of those genes are different.

The most important property of protein coding genes is that there is no relationship between a subsequence and another downstream subsequence. In most RNA genes, however, subsequences have to be matched (base paired) with other subsequences further downstream, resulting in a structure of a higher order. (Fig. 4)

So far RNA gene identification has had to be tailored for the individual RNA species, while the protein gene identification programs are general in the sense that they look for *all* proteins.

## Computational RNA identification

All standard programs used in annotation projects are tailored for protein gene identification, for instance finding coding exons or RNApol-II promoters. However, ncRNAs cannot be predicted by such criteria. Furthermore, promoters and terminators may not be the same for different ncRNAs, and not even the same for a specific ncRNA in different organisms. Simple sequence similarity (by BLAST) works for some RNA genes only if the species compared are closely related.

The first successful non-coding RNA gene-finding programs focused on ncRNAs with conserved primary sequence and strong secondary structure, such as tRNAs. Examples are tRNAscan[21] and Pol3scan[22], the latter recognizes the eukaryotic internal control regions that are typical of tRNA. Another approach was to let the user input patterns for secondary structure, including conserved primary sequence, two examples are RNAMOT[23] and RNAbob (Eddy 1996, unpublished).

Later, comparative sequence analysis was used to build probabalistic models of the structure of interest: RNACAD[24] and COVE[25] were published simultaneously and implemented, or was equivalent to, secondary structure profiles (SCFGs) described by Searls 1988[26-28] in the context of DNA sequence analysis. Besides the problem of gathering enough sequences to make an alignment that contains statistically significant signals, such as a conserved helix with compensatory base changes, the algorithms suffer from a time complexity of $> O(n^3)$, meaning that a ten-fold increase in sequence length increases the search time by at least a factor 1000.

Several programs use hybrid approaches where a fast program is first used to filter out interesting sequences from a longer sequence, often a complete genome, and then a second more computationally demanding program is used to further analyze the extracted subsequences. The first program to use this was tRNAscan-SE[29].

An important difference between different RNA prediction programs is whether they consider *pseudoknots* or not. A pseudoknot is the result of base-pairing between regions already part of a secondary structure, for instance base-pairing between two loops that each are part of a hairpin. Programs that use SCFGs (as COVE) cannot identify these.

Although theoretically important, pseudoknots may often be excluded in the analysis, as is the case for SRP RNA, or dealt with when evaluating found candidates.

The only ncRNA prediction programs that are commonly used are tRNAscan-SE, Pol3scan and Infernal (the new version of COVE), which is used by Rfam in an semi-automated procedure to mine the sequence databases for ncRNA homologues [30]. Rfam is the RNA equivalent to the Pfam database.

Rfam continuously searches all available nucleotide sequences in Genbank and EnsEMBL and incorporates hits above a certain threshold. This procedure is the same as in the protein family database Pfam [19]. To avoid the computational complexity of SCFGs, Rfam only analyzes sequences with a low-scoring BLAST match to a known ncRNA.

It is important to note, in the case of Rfam, that the result is highly dependent on the seed alignments used, and that there are no special expert groups that are responsible for the specific families. Thus, the sequences in Rfam may be different from the sequences in the specialized databases, such as the Signal Recognition Particle Database (SRPDB). For instance, in Rfam the Bacillus type SRP RNAs are grouped together with the smaller eubacterial SRP RNAs, with the result that the Bacillus type RNAs are not correctly predicted.

A list of available programs and algorithms related to RNA structure prediction is presented in Table 3.

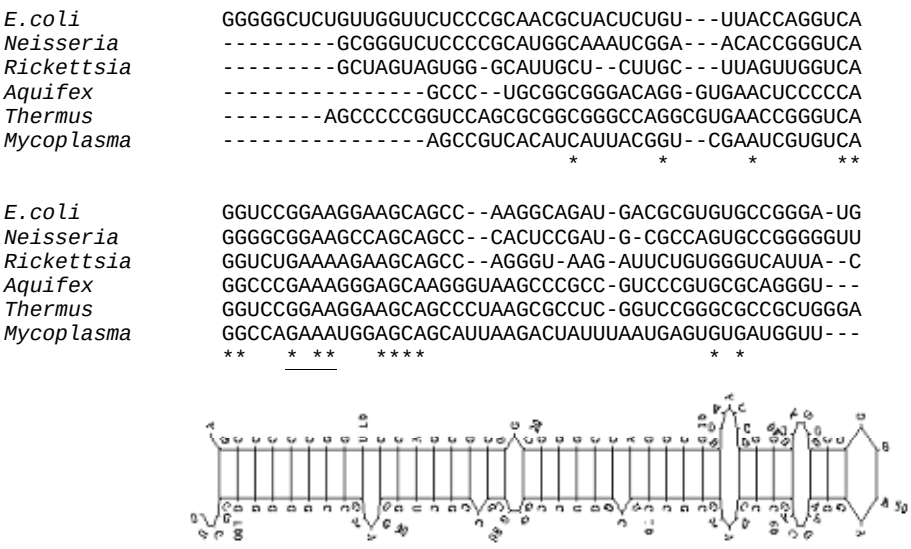


**Figure 4.** *Six bacterial SRP RNAs aligned. Only a few nucleotides (*) are conserved. Tetraloop nucleotides underlined. Secondary structure that match all sequences shown below.*

**Table 3.** *Programs related to RNA identification and structure prediction.*

| | | |
|---|---|---|
| **Pattern matching (CFG)** | | |
| | RNAMOT | Gautheret, 1993 |
| | PatScan | Dsouza, 1997 |
| | RNAbob | Eddy, 1997 |
| | Palingol | Billoud, 1996 |
| | Overseer | Sibbald, 1992 |
| | CITRON | Lisacek el al., 1994 |
| | HyPa | Gräf et al., 2001 |
| | rnaforester | Höchsmann et al., 2003 |
| **Extended pattern matching** | | |
| | PatSearch | Pesole, 2000 |
| | RNAmotif | Macke, 2001 |
| **Secondary structure profiles (SCFG)** | | |
| | RNAcad | Sakakibara, 1994 |
| | COVE/Infernal | Eddy & Durbin, 1994/2003 |
| | QRNA | Rivas & Eddy, 2001 |
| *incl. substitution matrices* | Rsearch | Klein & Eddy, 2003 |
| **Pattern or SCFG-style including pseudoknots** | | |
| | Imatch, bmatch | Tabaska et al., 1998 |
| | [Algorithm] | Rivas & Eddy, 1999 |
| | [Algorithm] | Akutsu, 2000 |
| | ERPIN | Gautheret, 2001 |
| | [Algorithm] | Dirks, 2003 |
| | ILM | Ruan et al., 2004 |
| **Energy based folding** | Mfold | Zuker, 1989- |
| | (no name) | Le & Zuker, 1991 |
| | RNAfold | Hofacker |
| | RNAstructure | Mathews, 2003 |
| | Sfold | Ding & Lawrence, 2003 |
| **Consensus secondary structure prediction** | | |
| | RAGA, PRAGA | Notredame et al., 1997 |
| | Foldalign | Gorodkin et al., 1997, 2005 |
| | Construct | Lück et al., 1999 |
| | X2s | Juan & Wilson, 1999 |
| | RNAGA | Chen et al., 2000 |
| | Dynalign | Mathews, 2002 |
| | CARNAC | Perriquet et al., 2002 |
| | Pfold | Knudsen & Hein, 2002 |
| | GPRM | Hu, 2002 |
| | RNAalifold | Hofacker et al., 2003 |
| | Stemloc | Holmes, 2003 |
| | Pmcomp | Hofacker et al., 2004 |
| | RNAProfile | Pavesi et al., 2004 |
| | BayesFold | Knight et al. 2004 |
| *incl. pseudoknots* | MIfold | Freyhult et al., 2004 |
| | hxmatch | Witwer et al., 2004 |
| **Hybrid approaches** | | |
| | tRNAscan-SE | Lowe & Eddy, 1999 |
| | Slash | Gorodkin, 2001 |
| | RNAmotif | Macke, 2001 |
| | SRPscan | Regalia, 2002 |
| | ARAGORN | Laslett & Canback 2004 |
| | (HMM filter + CM) | Weinberg & Ruzzo, 2004 |
| **New ncRNA, multiple align.** | | |
| | DDBRNA | di Bernardo et al., 2003 |
| | MSARI | Coventry et al., 2004 |
| | RNAz | Washietl et al., 2005 |
| **Others** | | |
| | MARNA | Siebert & Backofen, 2003 |

# Materials and methods

## Materials

Nucleotide, genome and protein sequences were primarily obtained from NCBI (and EMBL), including the recently launched TraceDB with shotgun reads from a rapidly growing number of genome sequencing projects. In addition, we have received permission from a number of sources (for example TIGR) to download their preliminary genomic contigs and assemblies, including protein gene predictions. Some genome projects that were first only accessible through web-based interfaces, for instance a web-based BLAST interface or a download interface for single sequences, have later been published and updated (for instance the *Saccharomyces* genomes from MIT and Washington University).

New genome assemblies are routinely searched for protein genes by using gene identification programs like Genscan, Genewise and fgenesh, thereby creating searchable lists of putative protein homologues. For completed chromosomes, the identified ORFs (Open reading Frames), hypothetical proteins and proteins are deposited in the protein databases such as “nr” at NCBI, where precomputed BLAST searches are linked to from the protein entries, making identification of homologues simple. However, for some genome projects these searches have not-as-yet being made as the contigs are considered too short, and in these cases we have searched the contigs ourselves by using BLAST (tblastn for a protein query and genomic sequence databases).

During our work many preliminary assemblies have been updated or finalized, making a comprehensive list of our complete material hard to maintain due to the large number of sequences and the different locations where they have been obtained. We have therefore not included such a list in this work. Found sequences will have to be updated as genomes are finalized, and put in SRPDB with the accession number of the updated sequences.

3D-structures of a few SRP RNAs are found in the Nucleic Acids Database (NDB) [31], where a secondary structure plot with 3D interactions marked may be produced

automatically (by RNAview [32]). These structures have been used to examine the interactions between nucleotides in non-consensus tetraloops of helix 8.

## Methods

### SRP protein component identification

Most SRP proteins may be identified by using primary sequence alignments to known proteins, eg. by using BLAST. Many of the SRP protein may also be identified with the use of Pfam models. There are such models for SRP9, SRP14, SRP19 and SRP54. SRP54 is the most conserved of all of the proteins and is easily identifiable. The SRP68 and SRP72 proteins have been identified using standard methods for pairwise alignment, such as BLAST, although the proteins are often hard to identify. In several cases TBLASTN was used with known proteins as query against unfinished or unannotated genome sequences.

### Multiple sequence alignments

CLUSTALW was used for both protein and nucleotide sequence multiple alignments. For protein sequences also HMMALIGN was used for alignment to a Pfam model or to a model constructed by us HMMBUILD. COVE/Infernal may be used in a similar way to align RNA sequences to a secondary structure profile.

### RNA secondary structure searches

Examples of programs that search sequences for matches to a secondary structure *pattern* are PatScan and RNAbob. PatScan is slower than RNAbob and has a different syntax for specifying the secondary structure pattern to be searched for. Both programs give a yes/no output, returning a sequence if a match is found. We have used RNAbob which searches a bacterial genome in seconds. A eukaryal genome may take hours to search.

For secondary structure profile (covariance model) creation we have used COVE or a further development of COVE, Infernal. COVE may be used with a set of unaligned or aligned sequences as input, to create a profile. Although it is possible to specifify base-pairings in the alignments used as input, we have not done so as it has not been necessary and as it is hard to reliably predict base-pairings when few sequences are known.

However, when more sequences with known structure are identified, this will be the preferred way of creating profiles as it simultaneously will predict the secondary structure.

The only program available that search a database for matches to a single query based on both primary sequence and secondary structure is Rsearch [33]. It could be called the “BLAST equivalent” for RNA and likewise outputs a pairwise alignment, but with an secondary structure added to the alignment. We have only occasionally used this program in our searches, but it may be well suited for further analysis of relationships between SRP RNA sequences.

### Secondary structure prediction of sequences

Folding of a single sequence is done by several programs: RNAfold that is part of the Vienna package[34], MFOLD [35] and RNAstructure [36].

MFOLD uses a dynamic programming approach to calculate the optimal (and sometimes suboptimal) folding based on the minimal free energy of the folded RNA. RNAfold calculates the partition function introduced by McCaskill [37], to determine the optimal folding. No suboptimal foldings are reported. Results of these programs are similar, if not identical. It should be noted that there are several examples (rRNA, SRP RNA) where the known structure is different from the optimal MFE folding. In both MFOLD and RNAfold, the user may input constraints that force or prevent certain base-pairings, making the folding conform to known features.

A MFE folding may also be simultaneously calculated for multiple sequences, thus producing a consensus secondary structure with an optimal MFE. RNAalifold [34] is such a program.

### Conserved SRP RNA motifs in used in searches

Only a few nucleotides are strictly preserved in all found SRP RNAs:

(i) The non-canonical basepair A/C and the 5' G directly downstream of that A in the symmetrical loop of helix 8 (Fig. 5). The base-pairs at these positions form a flattened minor groove which is the most important binding site for SRP54/Ffh [38].

(ii) The first, last nucleotides of the external loop of helix 8 are almost always G, A.

(iii) For SRP RNAs with an Alu domain (all except the small bacterial 4.5S SRP RNA) the UGUNR motif (Fig. 7), or UAUNR in some Archaea, is found in all SRP RNAs (a few exceptions exist).

(iv) In SRP RNAs with a helix 6 (all except the bacterial SRP RNAs) the A in the third position of the tetraloop of helix 6 (this nucleotide is involved in tertiary interaction with the last nucleotide in the helix 8 external loop) is conserved.

**Helix 8 stem-loop.** The GNRA external tetraloop (or GNNNNA in fungi and plants) has a 3 base-pair stem followed by a highly conserved symmetrical loop that is essential for the binding of SRP54/Ffh, typically the first pairs are GA, GG, AC, CA with a YR pair closing the loop. Together with another stem of 3–4 base pairs, this is the most highly conserved part of SRP RNA and we have used it in all pattern-based searches.

**Transcription termination signals.** Most bacterial SRP RNA gene candidates have a U-rich sequence at their 3' ends. The U-rich sequence is a useful source of information and in Paper I it was included to significantly improve the covariance model. In *Saccharomyces* genes there is a similar U-rich sequence, more extended than in most bacteria. For many other organisms, we have used the same motif to predict the 3' end of the SRP RNA.

**Conserved architecture.** All candidates are evaluated in the context of their phylogenetic group. For instance, in Archaea we require the presence of helices 6 and 8.

The pseudoknot in the Alu domain of mammals, where base pairs are formed between the loops of helices 3 and 4, is not used as it is not necessary for reliable detection of SRP RNAs, and as it does not seem to be conserved in all species with an Alu domain.

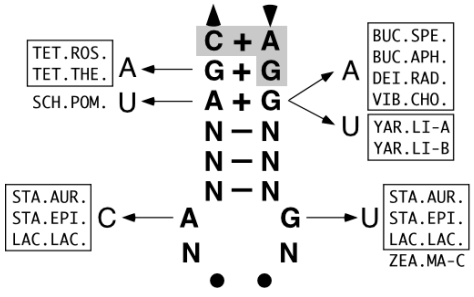


**Figure 5.** *Conserved part of helix 8. Invariant nucleotides shaded. Non-consensus variants shown with arrows and abbreviations of organism names. N=[AUCG] Circles indicate the possibility of additional nucleotides.*

## SRP RNA gene prediction procedure

For the prediction of SRP RNA genes we initially tried a pattern matching approach. However, results were not encouraging due to the great variability of the primary sequences. Also, the nature of pattern matching is non-probabilistic, i.e only a yes/no answer is presented. COVE, a more computationally demanding approach, was then considered since it can use multiple alignments of the so far known sequences. To use this on complete genomes was, however, not effective and barely feasible.

**Sequence prediction.** As descibed in Paper I, a filter (helix 8 secondary structure patterns used with RNAbob) was therefore applied that extracts possible candidate subsequences, which we then searched with the appropriate COVE model(s). The searches rarely produced more than one candidate. The COVE output score, which is dependent on the sequences used in the model creation, was not used as the only criterion for accepting a candidate (which is the case in Rfam). For organisms where the full profile did not match any sequence, we used Infernal in "local" search mode, where parts of the profile may be aligned to the sequence.

**Secondary structure prediction.** When a sequence or set of sequences have been found, a secondary structure may be predicted by optimizing the minimal free energy (MFE) of the sequence(s). We used MFOLD [35] and RNAalifold [39] which will predict a secondary structure on the basis of a multiple alignment of sequences. RNAalifold was used to provide support for the MFOLD predictions of the *Saccharomyces* SRP RNA sequences found in Paper III and to refine the proposed consensus structure obtained by manual editing of the structures predicted by MFOLD.

We named our approach for prection of SRP RNA genes SRPscan and introduced a web-based version (http://bio.lundberg.gu.se/srpscan/), in which the prediction is made automatically by combining the different programs used. Constraints for the MFOLD prediction were obtained by extracting the positions of helices with canonical base-pairings (A/T, G/C, G/T) contained in a second COVE search using the program "coves" with the output option "-m" to produce a secondary structure plot with base-paired nucleotides.

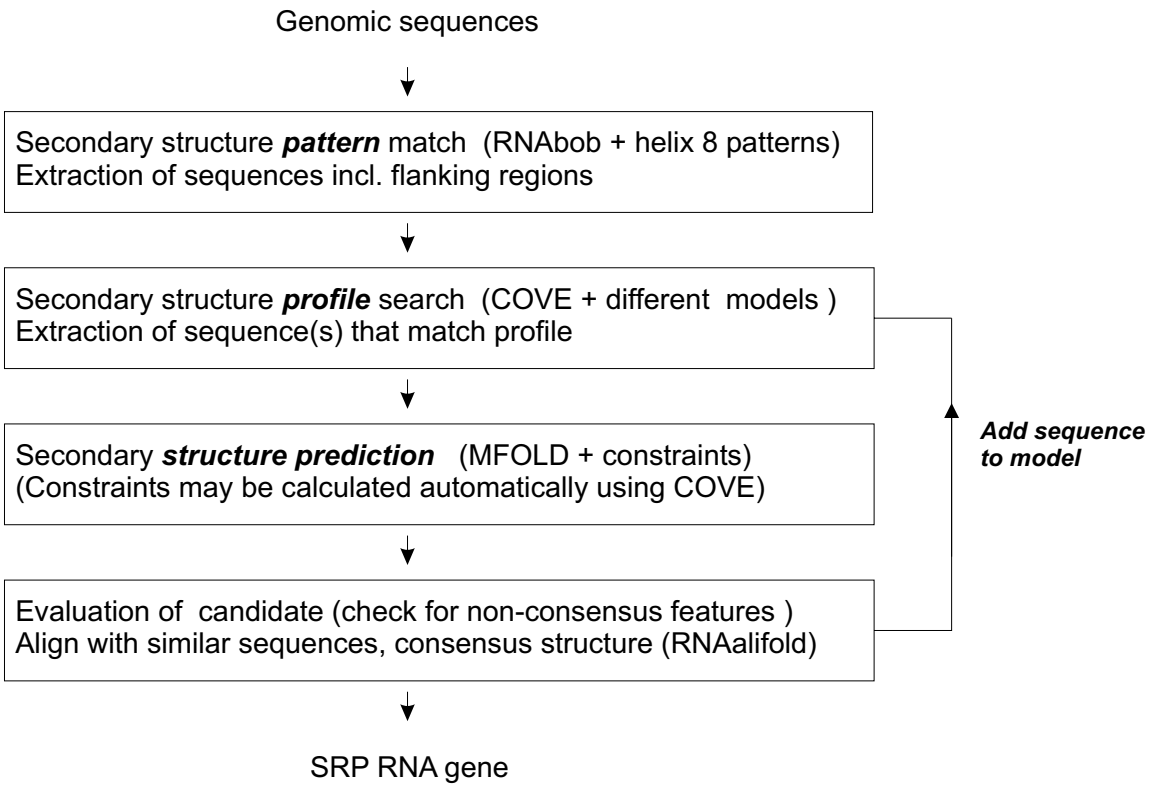


**Figure 6.** *SRP RNA gene prediction procedure, SRPscan. For organisms where the exact helix 8 pattern did not extract sequences where an SRP RNA candidate could be found, we first used patterns with mismatches allowed. For organisms where we failed to match the full profile to any sequence, we used Infernal in local search mode to match parts of the profile to candidate sequences, which were manually evaluated by using MFOLD and structure analysis based on known SRP RNAs.*

**Synteny and promoters.** For organisms with fully annotated genomes, one may distinguish pseudogenes, or provide more evidence for a prediction, by using synteny, eg. comparing the genomic location of candidate genes. In mammals, this may used to distinguish the expressed SRP RNA from the pseudogenes. Analysis of synteny was also used in the identification of chloroplast SRP RNA (Paper IV). Promoter analysis was used in Paper I (*Arabidopsis*) and Paper IV (chloroplasts).

# Results and discussion

We screened available genome sequences for SRP RNA and protein components (Fig.12). Numerous novel genes were identified as described in Papers I–IV.

### Design of SRP RNA

To summarize our findings I will focus on SRP RNA and its two domains: the Alu domain, comprising helices 1,2,3,4, part of helix 5 and in some species the new helices 9 and 12, and the S domain with the rest of helix 5, helices 6, 7, 8 and in some species the new helices 10 and 11.

The Alu domain has been shown to be very diverse, starting with some fungi SRP RNAs that have only helix 2, some other fungi that have helix 2 plus the new helices 9 and 12, and many lower eukaryotes which have helix 2 and 3 but a smaller helix 4, which in some species is more akin to a loop. An exception is found in *Plasmodium* species which have longer helices 3 and 4. (Fig. 7)

The rest of the eukaryotes have a rather well conserved Alu domain with helices 2, 3, 4, all similar to the mammal Alu domain. A variation of this “consensus” Alu domain is found in Archaea and some Bacteria, which also have a helix 1.

There is a correlation with SRP proteins in that eukaryotes with a non-consensus Alu domain often seems to lack proteins SRP9/14, or as in fungi, have a SRP21 protein instead of SRP9. (Table 4)

The S domain is much more conserved than the Alu domain. However, some fungi have an extra helix inserted between helix 6 and 8, and some fungi have an extra helix between helix 5 and helix 6. Helix 8 is highly conserved. In most species it has an external tetraloop. However, in a few instances this loop has 6 bases (plants and some fungi) or 5 bases (unpublished).

The nomenclature of SRP RNA has not been consistent. For instance, authors have used different schemes for naming domains and helices. Our detailed studies of yeast and protist SRP RNAs (Paper III) provided additional information on the design of SRP

RNAs. For these reasons we proposed a new nomenclature for SRP RNA, which will facilitate the identification and analysis of all other SRP RNA. (Paper V)

An SRP RNA has been identified in all eukaryal groups from which full genomic sequences are available, even though some are partial. The only organism that seems to lack an SRP RNA is the parasitic archaeon *Nanoarcheum equitans,* which also lacks SRP54 and thus a nuclear encoded SRP altogether. A special case is the remains of the primary host genome in organisms with a secondary endosymbiont plastid. The only nucleomorph genome that has been sequenced, from *Guillardia theta,* does not encode any SRP components.

The individual groups will be discussed in the following.

### Archaea and Eubacteria

**Archaea.** Archaeal SRP RNA lacks helix 7 but has the helix 1. The identified SRP RNAs (26 in SRPDB as of Dec 2004) conform to the canonical archaeal SRP RNA structure, with a few exceptions in the Crenarcheota group: *Aeropyrum pernix,* and *Pyrobaculum aerophilum* (not yet deposited). For these two RNAs the structure of the Alu domain is difficult to predict (they may have a helix instead of the UGUNR motif). The UGUNR motif is UAUNR in several Archaea, and in some even CNNNR.

**Eubacteria.** We have found an SRP RNA in all complete eubacterial genomes that we have screened. In this case our method to predict SRP RNAs was highly specific and sensitive. One unexpected finding was a novel UGAC helix 8 external loop found in *Lactococcus lactis* and *Staphylococcus aureus* and *epidermis*. Furthermore, *Thermotoga maritima* has a Bacillus type SRP RNA, the first Gram-negative bacteria with that type of SRP RNA. *Thermotoga* belongs to one of the deepest and most slowly evolving lineages of bacteria. (Paper I)

The UGAC loop is found in organisms from two different branches of Firmicutes, showing that the change from GGAA to UGAC has occurred twice during evolution.

This change in loop sequence is consistent with the structure of *E. coli* SRP RNA [38] (NDB id: PR0021) that shows the interaction between the first (G) and last (A)

nucleotides in the tetraloop (*trans* Hoogsteen/Sugar edge). The same tertiary structure may be achieved by U and C in those positions, according to isostericity matrices published by Leontis *et al.* [40]. In fact, according to these isostericity matrices, the two found consensus pairs, G/A and T/C, may be N/A (N=[ATCG]) and H/C (H = not G).

All organisms belonging to *Bacillales* (incl. *Bacillus*, *Listeria*, *Staphylococcus* and others) and *Clostridia* (incl. *Clostridium* and *Thermoanaerobacter*) have been found to have the Bacillus type SRP RNA.

### Chloroplasts

We have identified six SRP RNA from chloroplasts of red algae or of red algal origin (such as *Guillardia theta*), two from green algae and one from *Mesotigmatales*. These results strongly suggest that the SRP RNA form a complex with cpSRP54, similar to the eubacterial SRP. The prediction of these genes is supported by analysis of synteny and upstream promoter sequences. (Paper IV)

Interestingly, a few of these RNAs show non-consensus features as a non-canonical A/A base pair in the stem next to the tetraloop of helix 8, which has so far only been observed in *Mycoplasma* (G/A), and a UAAA helix 8 external loop. This loop is compatible with the three dimensional structure as the U/A have the same type of binding as a G/A, as noted for the URRC loop in bacteria. It also shows how the URRC type may have arisen by mutations from GRRA to URRA, and then URRC. The cpSRP RNA of *C.merolae* is the shortest SRP RNA found so far (61 nts).

We also showed that the the absence of SRP RNA in higher plants and *Chlamydomonas* are correlated with amino acid substitutions in cpSRP54 that are likely to affect RNA binding.

### Metazoa

SRP in metazoans is very conserved with all organisms having SRP9/14, SRP68/72 and SRP19/54 and the standard SRP RNA. More than 30 SRP RNAs have been identified, noteable new organisms include *Brugia* (nematode), *Schistosoma* (trematode), *Schmidtea* (turbellaria), *Nematostella*, *S.purpuratus* and *Ciona* (urochordata), and *Branchiostoma* (cephalochordata) (unpublished).

| Kingdom or Group | Organisms in group | RNA type | RNA length | Alu domain proteins | S domain Helix 5 proteins | S domain Helix 6/8 proteins | Includes results from |
|---|---|---|---|---|---|---|---|
| Bacteria | Most | 4.5S | 75–104 | – | – | Ffh | Paper I |
| | G+ *Bacillales, Clostridia* and G- *Thermotoga* | 6S, No helix 6,7 Alu with H1 | 270 | Hbsu (Bacill.)* | – | Ffh | Paper I |
| Archaea | All | 7S, Small helix 7, Alu with H1 | 300 | – | – | SRP19<br>SRP54 | Paper I |
| Eukarya | Protozoa *Trypanosomatids* | 7S + sRNA? | 280 | – | SRP68<br>SRP72 | SRP19<br>SRP54 | |
| | Protozoa *Plasmodium* | 7S, long H3, H4 | 300 | SRP9<br>SRP14 | SRP68<br>SRP72? | SRP19<br>SRP54 | Paper III |
| | Protozoa *Ciliophora* | 7S, No H4 | 275 | ? | SRP68<br>SRP72? | SRP19<br>SRP54 | Paper III |
| | Microsporidia | Alu unclear | 266 | ? | ? | SRP19<br>SRP54 | Paper III<br>Thesis |
| | Ascomycota | No H3, H4 | 300 | SRP21<br>SRP14 | SRP68<br>SRP72 | SRP19<br>SRP54 | Paper III |
| | *Saccharomyces* | No H3, H4 | 400–600 | SRP21<br>SRP14 | SRP68<br>SRP72 | SRP19/<br>Sec65<br>SRP54 | Paper III<br>Paper V |
| | Basidiomycota | 7S | 300 | SRP9<br>SRP14 | SRP68<br>SRP72 | SRP19/<br>Sec65<br>SRP54 | Thesis |
| | Mammals | 7S | 300 | SRP9<br>SRP14 | SRP68<br>SRP72 | SRP19<br>SRP54 | |
| Photo-synthetic organisms | Chloroplasts in red algae and some green algae | 4.5S | 67–102 | – | – | cpSRP54 +<br>cpSRP43/54 | Paper IV |
| | Chloroplasts in higher plants | Not probable | | – | – | cpSRP54 +<br>cpSRP43/54 | Paper IV |

**Table 4.** *Overview of SRP design. Comments: '*' Have not yet been found in all organisms in the group. '?' indicate that homology searches have not yet identified homologues. '–' not applicable or does not exist; cpSRP54 = Ffh homologue; The microsporidian* E.cuniculi *has the smallest eukaryal SRP. "sRNA" stands for the tRNA-like sRNA-76 and sRNA-85. The cpSRP43/54 is the post-translational SRP.*

### Plants and green algae

We identified a number of novel plant SRP RNAs. As for many fungi, the SRP RNA of higher plants have a 6 nts helix 8 external loop. In contrast, the green algal SRP RNA from *Chlamydomonas reinhardtii* and *Volvox carteri* have a tetraloop. Interestingly, the tetraloop in *Volvox* is UAAC, the only eukaryote with a non-GNRA tetraloop (unpublished).

Higher plants have multiple copies of SRP RNA genes. For instance, *Arabidopsis* has eight copies, including two pseudogenes (Paper I). Only two of these were previously known, and of the eight copies two were predicted to be pseudogenes as they lack the upstream promoters (USE and TATA, [41]) identified in the other SRP RNAs.

### Rhodophyta

The only branch in which there are genomes published, and where we have not found a nuclear encoded SRP RNA so far, is in red algae (*Cyanidioschyzon merolae* is the only published genome so far). As the S domain proteins, SRP19/54 and SRP68/72, are found, it is unlikely that this organism lacks an SRP RNA. Probably some part of the genome in the assembly is missing.

*C.merolae* SRP54 is similar to *Arabidopsis* SRP54, but also to *C.elegans* SRP54. SRP19 has similarities to *Arabidopsis* SRP19, to yeast Sec65 and also to SRP19 of higher eukaryotes. This is clearly consistent with the deep branching of red algae in the plant group.

### Heterokonta

Heterokonta (stramenopiles) is a very diverse group. It includes many photosynthetic organisms such as brown algae and diatoms (bacillariophyta), and non-photosynthetic organisms such as water molds (oomycetes). The photosynthetic organisms have a red algal secondary plastid.

The *Phythophtora* species (*P.ramorum* and *P.sojae*) were the first heterokonts to be sequenced. The SRP RNA of these species are of the standard eukaryotic type, but with a novel pentaloop instead of the usual helix 8 external tetraloop: GUUAA, GAUAA. (A partial *P.infestans* SRP RNA has GUAAA.) We have now identified this pentaloop in four other species. Otherwise, it resembles the human SRP RNA (unpublished).

*Phythophtora* has homologues to SRP54, SRP19 and SRP68/72, but no homologues to SRP9/14 have been identified so far.

In the diatoms *Thalassiosira* and *Phaeodactylum* we have identified an SRP RNA, but it is difficult to predict the folding of the Alu domain (unpublished). As for *Phythophtora*, SRP54/19 are identified in *Thalassiosira*, but it is not possible to conclude that the SRP68/72 exist. No putative homologues for SRP9/14 were found.

### Protozoa

A number of novel RNA genes were found by screening protozoan genomes. They conform to the structure of the mammalian typ RNA but also show a large degree of variation in their Alu domain (Paper III). More recently we have screened additional genomes (unpublished).

**Alveolata (Apicomplexa, Ciliophora).** In almost all of the cases we have been able to identify an SRP RNA that conforms to other found SRP RNA in related organisms, for instance the apicomplexans *Cryptosporidium* and *Toxoplasma*, and the similar *Perkinsus*, resemble *E.tenella* SRP RNA. The ciliophores *Oxytricha* and *Tetrahymena* both have the disrupted helix 4 region without base pairing. (Fig. 7) An alternative fold with a helix 4 is also possible for *Oxytricha*, but it looks less convincing when compared to Alu domains of other alveolates. With only a handful of sequences available, no thorough comparative analysis may be done. The disrupted helix 4 region is also found in the apicomplexa *Theileria* (Paper III).

*Plasmodium* is unusual in that it has long helices 3 and 4, a finding which is supported by SRP RNAs from eight *Plasmodium* species. (Fig. 3, 7)

**Euglenozoa.** The SRP RNA of *Entosiphon sulcatum* has a short helix 4. Interestingly, the SRP RNA gene was found in a cluster of 5S rRNA, U1, U2 and U5 snRNA, a gene organization found in *T.brucei* and *Leishmania* (Paper III). Very few genome and protein sequences are available from *E.sulcatum*, and as a result the SRP proteins of this organism are unknown.

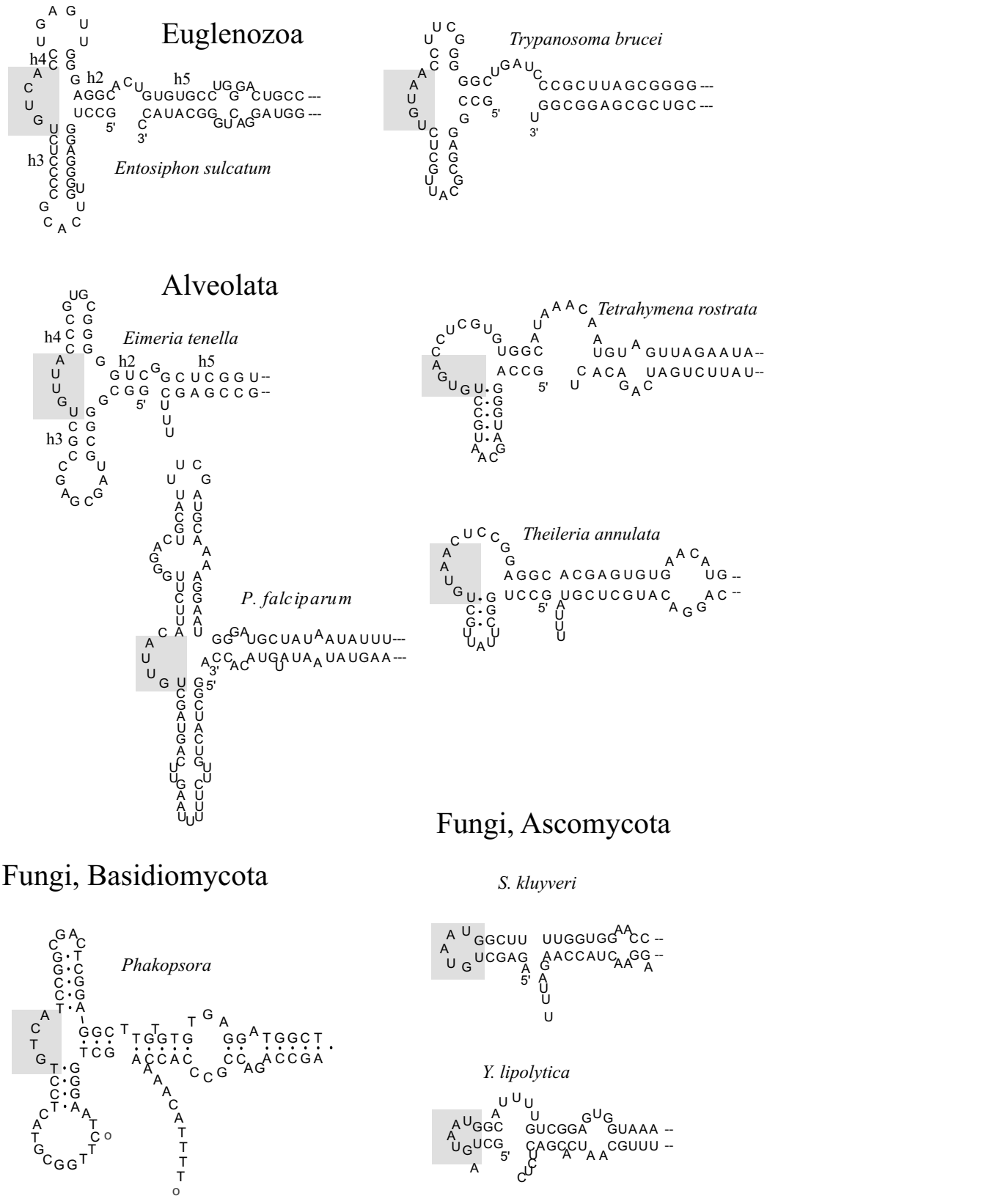


**Figure 7.** *Examples of Alu domains in Protozoa and Fungi. UGUNR motif shaded.*

**Entamoeba.** Genomic reads for five different *Entamoeba* species are available, and the previously mentioned pentaloop of helix 8 is found in three of these species. However, the two remaining species have the usual tetraloop (unpublished). In *E.histolytica* all proteins except SRP9/14 have been identified. (Paper III)

**Mycetozoa.** The genome (34 Mbases) of *Dictyostelium discoidum* is almost completed. *Dictyostelium* SRP RNA was identified in collaboration with Fredrik Söderbom *et al*. [42] who showed that this RNA is indeed expressed.The SRP RNA is of the standard metazoan type, but with a somewhat smaller helix 4, a small helix 7, and a CUAA helix 6 tetraloop. SRP9 and SRP19/54 are most similar to *Arabidopsis* SRP9/19/54, while SRP14 is more similar to human SRP14. These observations are consistent with the idea that *Dictyostelium* is grouped in between plants and animals. Homologues to SRP68/72 could be identified, using *Arabidopsis* SRP68/72 as query.

**Diplomonadida and Parabasala.** The most deeply branching eukaryotic groups are the diplomads and parabasala. In *Trichomonas vaginalis* (parabasala), we have identified an SRP RNA with a pentaloop (GUAAA) in the external loop of helix 8 (unpublished).

Apart from SRP54 and probably SRP19, it is unclear if *T.vaginalis* has any more SRP proteins. Since the genome is not yet assembled, protein searches are hard to perform.

Interestingly, the identified SRP54 is most similar to metazoan SRP54, which is also the case for *Giardia lamblia* SRP54. In *Giardia* (a diplomonad) we also identified SRP19 and putative homologues of SRP68/72. Interestingly, SRP19 is most similar to the archaeal SRP19, consistent with the fact that *Giardia* is a deeply branching eukaryote.

However, we have not been able to identify the Alu domain of SRP RNA in *Giardia*. As for some other protozoans, *Giardia* seems to lack SRP9/14. A putative termination signal (TTCCTTT) is located directy 3' of the S domain in the SRP RNA and directly upstream of the S domain is a sequence that resembles a transcription start signal identified in other Giardia genes (AATYAAAAA) [43]. This could indicate that the SRP RNA lacks the Alu domain altogether, as discussed for Microsporidia below.

## Fungi

It was previously known that the fungal SRP has diverged strongly compared to other phylogenetic groups, for instance that the helices 3 and 4 are missing in the SRP RNA. Our results not only confirm this, but also show that the variation within the group is more extensive than expected.

Our analysis of fungal SRPs gave a number of interesting results. Phylogenetic analysis gave rise to a secondary structure model for the Saccharomyces type RNA where a significant insertion took place, giving rise to the helices 9, 11, 12 and a long helix 7.

Furthermore, a novel helix is inserted between helices 5 and 6 in some Ascomycota. Finally, the organisms belonging to Basidiomycota and Zygomycota have an SRP RNA that resemble the standard eukaryotic SRP RNA and SRP9/14/68/72 proteins that are more similar to the metazoan counterparts than to the Ascomycota proteins.

Thus there are threee different types of SRP in fungi:

1. a Saccharomyces type with large SRP RNA and a SRP21 highly diverged from SRP9,
2. a smaller version with a smaller SRP RNA and a SRP21 not as diverged from SRP9,
3. a standard metazoan SRP, but with SRP19/54 more similar to the fungal Sec65/SRP54.

**Ascomycota.** In almost all Ascomycota, we have been able to identify SRP RNA, the exception being *Magnaporthe grisea*.

Starting with *S.cerevisiae* SRP RNA sequence, we obtained a secondary structure by first identifying the SRP RNA sequence of close relatives. By comparative analysis, we found a consensus structure of *S.cerevisiae, S.paradoxus, S.kudriavzevii, S.mikatae, S.bayanus, S.castellii* and *S.kluyveri*. Later, SRP RNA was identified in other species closely related to Saccharomyces: *Kluyveromyces waltii, Eremothecium gossypii* (Paper V) and *K.lactis, C.glabrata* and *K.yarrowia* (unpublished), all supporting the previously predicted structure. So far, our predictions of *Kluyveromyces waltii* and *Eremothecium gossypii* SRP RNA sequences have been experimentally verified (collaboration with Rob van Nues, Newcastle, unpublished). (Fig. 8)

In the group with a smaller SRP RNA, *Yarrowia lipolytica* and *S.pombe* were previously known and we identified similar SRP RNAs in *Neurospora crassa* (Pezizomycotina) and *Candida albicans* (Saccharomycetina).

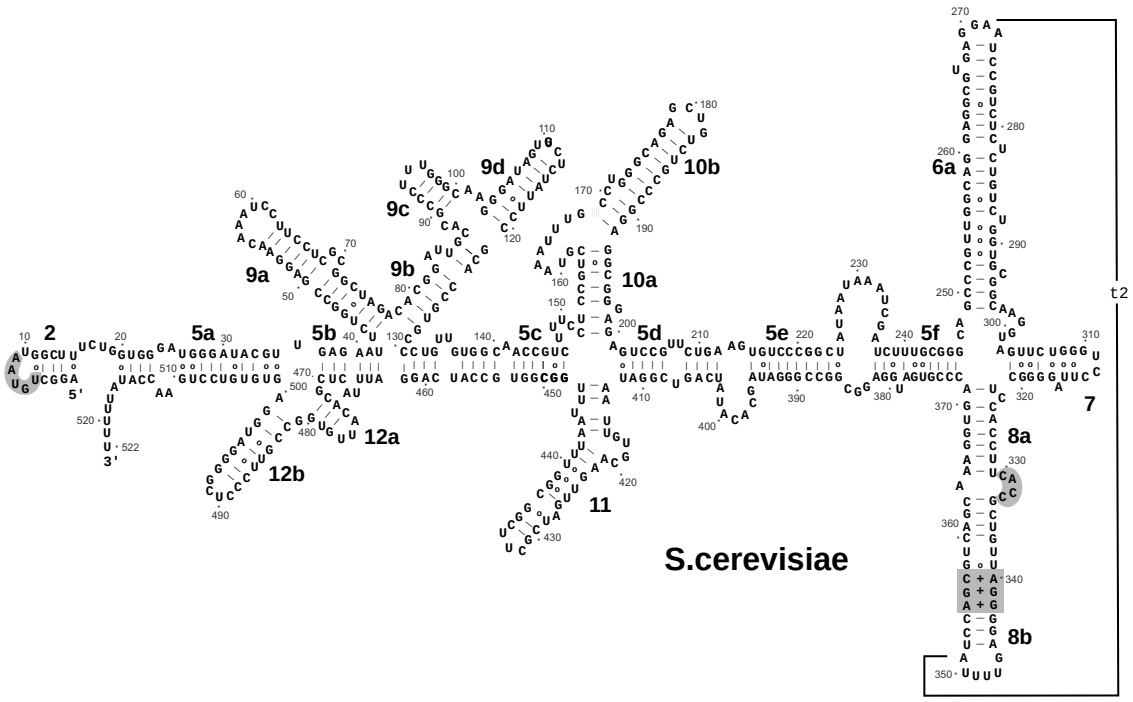


**Figure 8.** *Secondary structure of* Saccharomyces cerevisiae *SRP RNA as a result of comparative analysis with other Saccharomyces SRP RNAs. Helices numbered according to nomenclature proposal.*

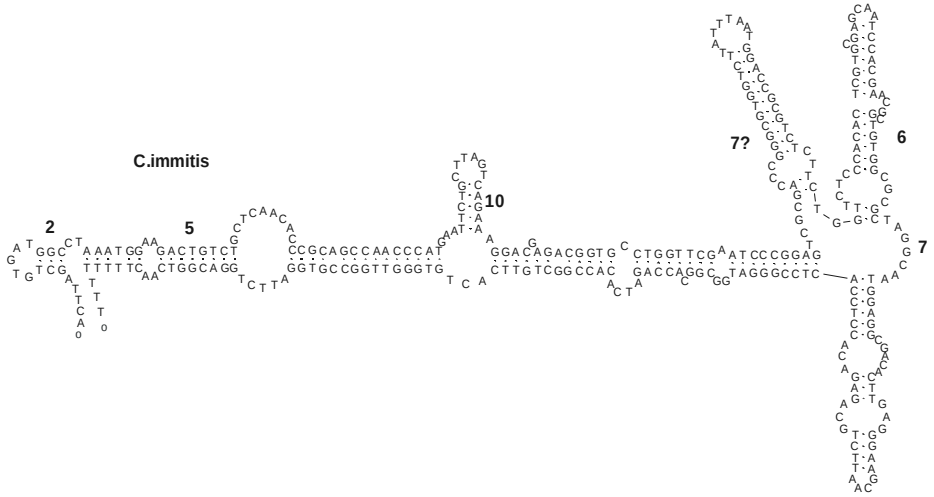


**Figure 9.** Coccidioides immitis *with a novel helix inserted between helices 5 and 6. A similar insertion is found in* Histoplasma. *Both are relatives of* Aspergillus.

Noteably, the helix 8 external loop is frequently a hexaloop in Ascomycota, but both loops exist in Saccharomycetina and Pezizomycotina, showing that the change from four to six nucleotides has occurred several times, as for the helix 8 pentaloops found in protozoa.

Other Saccharomycetina species where we have recently identified SRP RNA are *Lodderomyces elongisporus*, which has a helix 10, and *Candida dubliniensis*, *Debaryomyces hanseni, Clavispora lusitania* and *Pichia guilliermondii* (unpublished).

Furthermore, in species closely related to *Neurospora crassa* (Pezizomycotina, Sordariomycetes), we have identified SRP RNA in *Podospora anserina*, *Chaetomium globosum*, *Hypocrea jecorina*, and *Trichoderma reesei,* which all have a large bulge instead of a helix 10 that is found in *N.crassa*, and in *Gibberella zeae. Sclerotinia sclerotiorum* (Pezizomycotina; Leotiomycetes) SRP RNA is similar (unpublished).

Surprisingly, in the Eurotiomycetes group of Pezizomycotina (to which *Aspergillus* belongs), a new helix (not included in Paper V) is inserted between helix 5 and helix 6 in *Coccidioides* and *Histoplasma* (Fig. 9). The insertion is easily identified in sequence alignments with *Aspergillus.* One may speculate that if the long helix 7 in Saccharomyces SRP RNA stabilizes the S domain, this helix in *Coccidioides* and *Histoplasma* could be a functional equivalent.

**SRP21.** As the SRP21 protein is unique in yeast SRP we decided to examine its relationship to other SRP proteins. Using profile-based searches we showed that SRP21 is related to SRP9 from metazoans. Therefore it is probable that SRP9 and SRP21 share a common ancestor, as is the case for SRP9 and SRP14. The secondary structure of the SRP21 homologues were predicted using PSI-PRED, and a $\alpha\beta\beta\beta\alpha$ structure similar to that of SRP9/14 was also found in the SRP21 proteins. Although available experimental evidence suggests otherwise [44], these findings show that SRP21 might be part of the Alu domain in Ascomycota. (Paper III)

**Basidiomycota, Zygomycota.** We have failed to identify a SRP RNA in the almost complete genome sequences from the Basidiomycota *Ustilago, Coprinus*, *Cryptococcus*, *Phanerochaete* (and in addition *Laccaria* and *Puccinia* in TraceDB). However, in

*Phakopsora pachyrhizi* and *meibomiae* we have identified a standard eukaryotic SRP RNA with a short helix (helix10 in the nomenclature proposal) protruding from helix 5. The 12 nucleotides that differed between the two species were all found in bulges, loops or were compensatory base changes. Furthermore, a similar SRP RNA was found in the Zygomycota *Rhizopus oryzae*. (Fig. 10) As indicated above these RNAs have an Alu domain characteristic of metazoan RNAs.

Putative SRP9/14/68/72 homologues have also been identified in Basidiomycota, even though all of the proteins are not yet identified in all of the organisms. Interestingly, the SRP19/54 proteins resemble the Ascomycota SRP19(Sec65)/54, but SRP9/14/68/72 are more similar to the metazoan proteins, in accordance with the fact that we have identified a standard SRP RNA in *Phakopsora*.

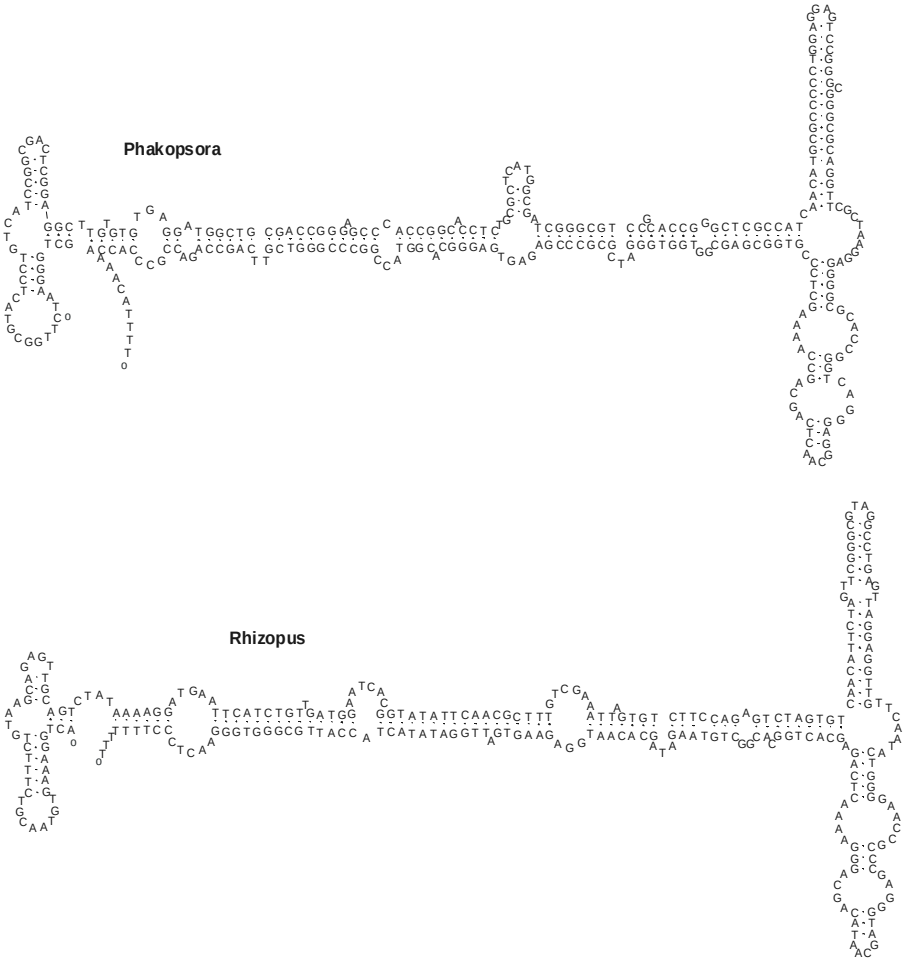


**Figure 10.** *SRP RNAs from* Phakopsora *(Basidiomycota) and* Rhizopus *(Zygomycota).*

## Microsporidia

Microsporidia are intracellular parasites with remarkably small genomes of 2.3–20 megabases. Initial phylogenetic analyses placed Microsporidia basal to most other eukaryotes, but it is now shown that they have evolved within fungi and probably has a zygomycete-like ancestor [45]. Microsporidia have a very compact genome. The genome organisation evolves slowly, but genes seem to evolve rapidly [46].

Sequences for three different microsporidia are available: *Encephalitozoon cuniculi*, assembled genome, 11 chromosomes, 2.9 Mbases; *Nosema locustae* (now *Antonospora locustae*), contigs, 2.145 Mbases; *Spraguea lophii*, genome survey, single pass reads, 120 Kbases sample of the genome (approximately 1.9%), genome estimated to 6.2 Mbases [http://jbpc.mbl.edu/Spraguea-HTML/].

In the *E.cuniculi* genome SRP54 and SRP19 are the only SRP proteins that have been identified, and SRP54 in *N.locustae* is easily identified due to the high sequence similarity to *E.cuniculi* SRP54.

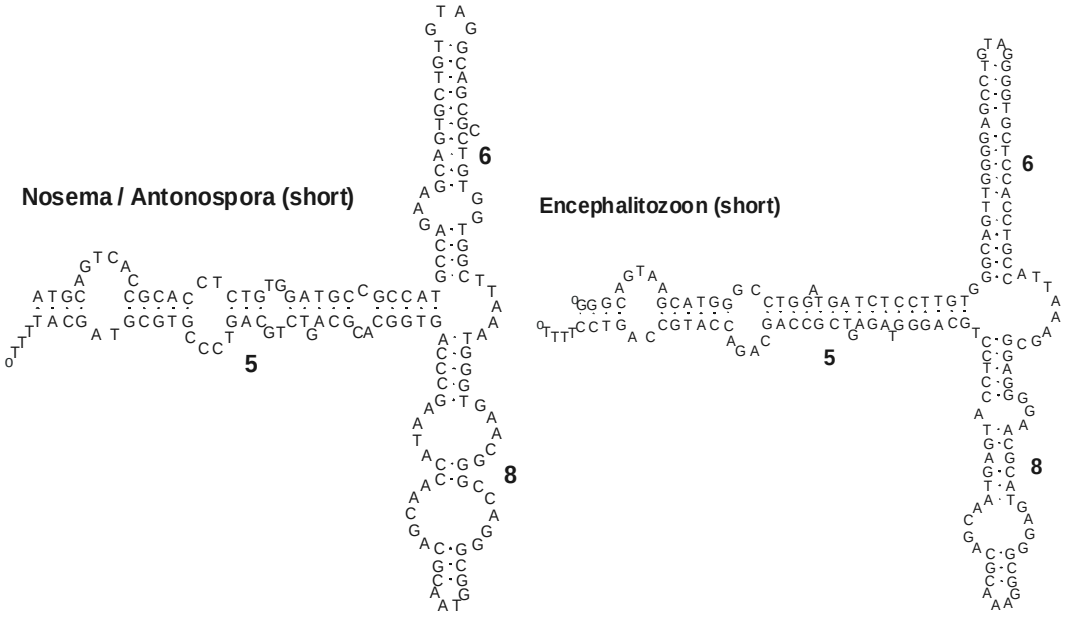


**Figure 11.** *Putative structure of* N.locustae *and* Encephalitozoon *SRP RNA. The termination signal (TTTT) at the 3' end is conserved in* Spraguea.

We found an SRP RNA candidate in *E.cuniculi* (Paper III) that has an S domain that conforms to known S domains, except for the eukaryotic helix 5 motif that seems to be an asymmetrical internal loop of 2 vs 4 nts or a bulge of 2 nts instead of the 3 nts bulge found in almost all eukaryotes. Directly upstream of the S-domain a helix 10 could be formed, as seen in many fungi SRP RNA.

The tentative model of SRP RNA has a small Alu domain that has some differences compared to other small Alu domains in ascomycota fungi. The loop includes the first U of the UGUNR, and the helix 2 is just 2 basepairs. The length between helix 8 and the 3' end is 56 nts, which is somewhat short if compared to *S.pombe*, *Neurospora* and *Yarrowia* (67–75 nts). This adds to the uncertainty of the proposed model.

Genomic sequences from the microsporidia *Nosema locustae* and *Spraguea lophii* have recently been made available and, as in *E.cuniculi*, we could identify a well conserved S domain with a rather high sequence similarity in helix 8, helix 7, and an identical helix 6 external loop. However, no UGUNR motif or standard Alu domain, as found in *Rhizopus*, could be identified in *Nosema* and *Spraguea*. Furthermore, there is a potential transcription termination signal at the 3' end of the S domain in all these species.

It is therefore tempting to speculate that these organisms have an SRP RNA consisting only of a S-domain, with or without an additional helix 10. (Fig. 11)

### The SRP proteins SRP9/14, SRP68/72

As noted above, some primitive eukaryotes seems to lack SRP9/14: *Encephalitozoon*, *Trypanosoma*, *Leishmania*, *Theileria*, *Entamoeba*, *Phythophtora*, *C.merolae*, *Giardia* and *Trichomonas*. In the case of microsporidia SRP, even SRP68/72 seem to be lacking, making it a “minimal” eukaryal SRP. This is in accordance with the fact that the archaeal SRP only contains SRP19/54 in addition to SRP RNA. However, it should be noted that SRP68/72 are more divergent and harder to identify than the other SRP proteins. More work clearly needs to be done before ruling out the presence of SRP68/72 homologues.

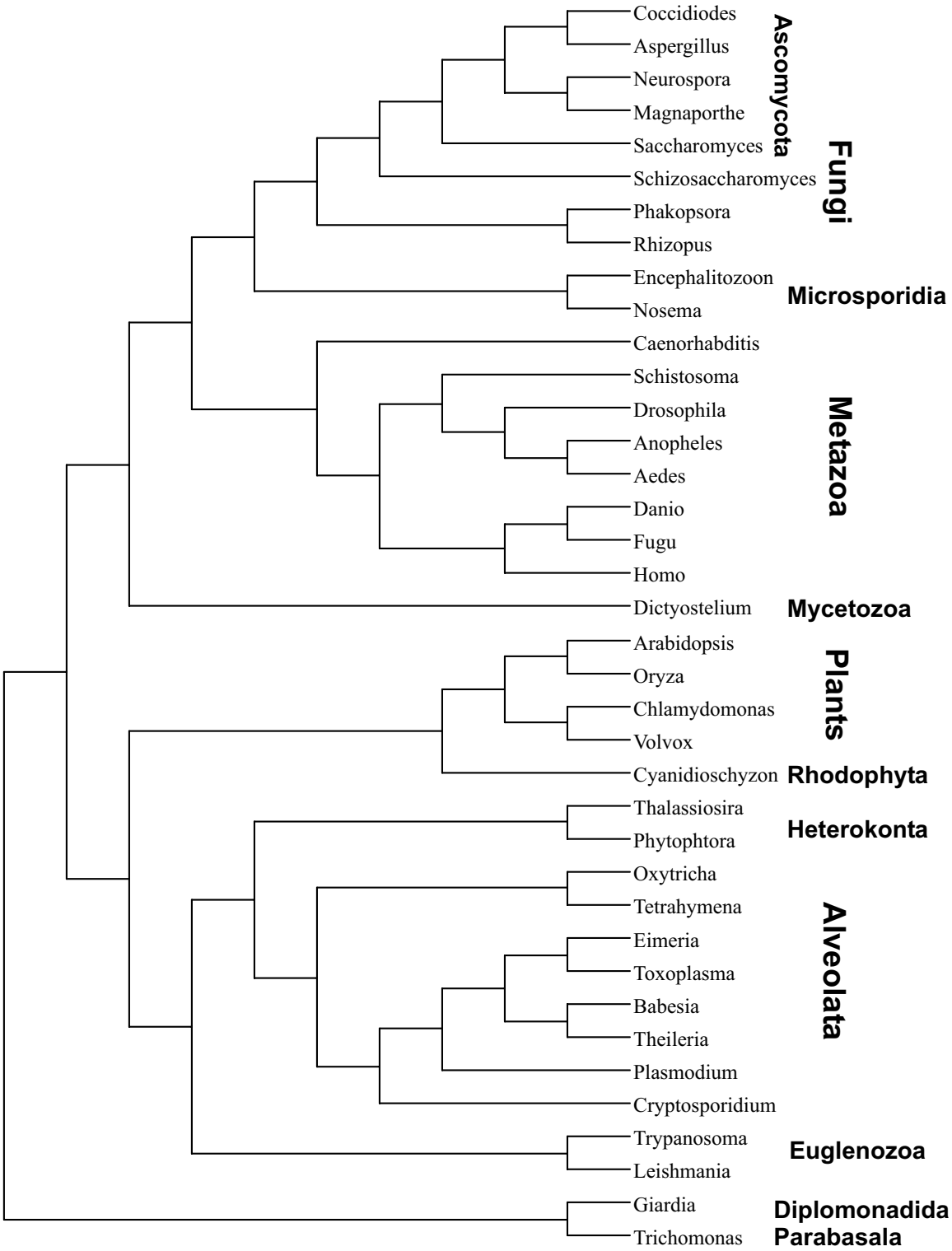


**Figure 12.** *Phylogenetic tree of most organisms searched for SRP components. Based on Baldauf et al., Science, Vol. 290, 3 Nov 2000, who used multiple proteins in the analysis. Saccharomyces branch includes all Saccharomycetina, as* C.albicans, E.gossypii, Kluyveromyces, *etc.*

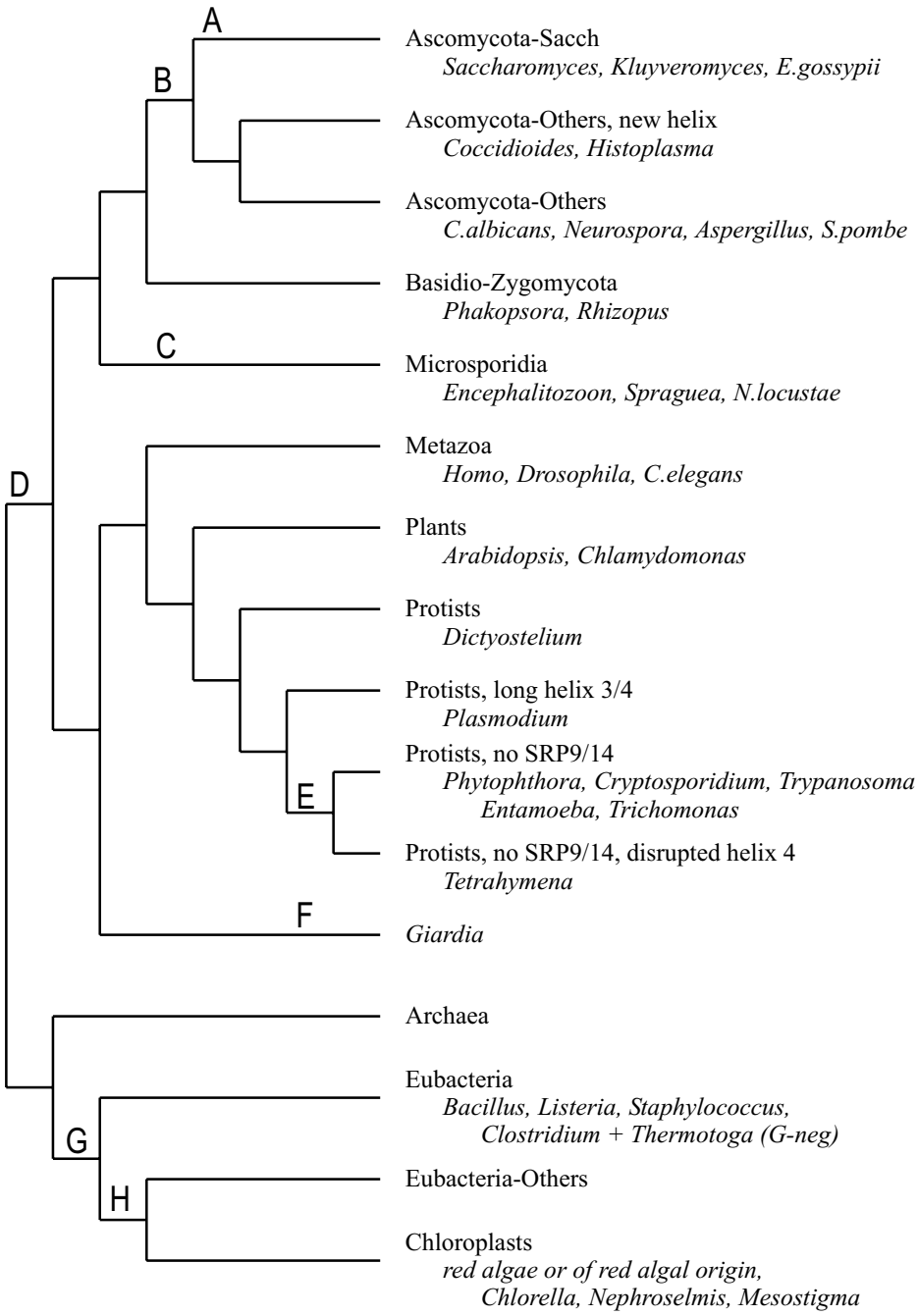


**Figure 13.** *Tree of the different types of SRP with changes marked with A, B, C ...* ***A*** *= insertions in helix 5;* ***B*** *= loss of helices 3, 4 (small Alu), SRP9>>SRP21;* ***C*** *= loss of SRP9/14, SRP68/72;* ***D*** *= loss of helix 1, addition of SRP9/14, SRP68/72;* ***E*** *= loss of SRP9/14;* ***F*** *= loss of SRP9/14,* Giardia *placed separately as Alu domain is unclear;* ***G*** *= loss of helix 1, loss of SRP19, SRP54>>Ffh;* ***H*** *= loss of Alu domain. Metazoa, Plants and Dictyostelium could be grouped together as SRP is similar. Pentaloop: Trichomonas, Entamoeba, Phythopthora; Hexaloop: Ascomycota, Plants.*

## Evolution of SRP

Based on our results it is tempting to speculate on the evolution of SRP. Before the divergence of Bacteria and Archaea an ancestral SRP might have been composed of an archaeal-type RNA together with SRP19/54. In Bacteria SRP19 and helix 6 of SRP RNA were first lost to yield the Bacillus type SRP. In most eubacteria the SRP RNA then lost helices 1, 2, 3 and 4, and a large portion of helix 5.

The first eukaryotic SRP may have had a SRP similar to the archaeal SRP, with the subsequent addition of SRP9/14 to a smaller Alu RNA domain without helix 1 and smaller helices 3 and 4, and the addition of SRP68/72. The SRP9/14 were lost in some protozoans, and the helix 4 of the Alu domain has changed in some species. Whether the ancestor of the most deeply branching eukaryotes, such as *Trichomonas* and *Giardia*, ever acquired SRP9/14 is unclear.

In the fungi branch, including Microsporidia, SRP19 diverged (resulting in Sec65), but microsporidian SRP19 still resemble both protozoan and archaeal SRP19. The eukaryotic SRP RNA and SRP9/14 (and SRP68/72) proteins were kept in Basidiomycota and Zygomycota.

In Ascomycota, the Alu domain of SRP RNA lost helices 3 and 4, and SRP9 mutated into SRP21. In Saccharomycetina branches the SRP RNA gained additional helices 9, 11 and 12, plus the longer extra helix 7. A similar long extra helix may be found in some Pezizomycotina, but inserted between helices 5 and 6. Addition of helix 10 to SRP RNA ocurred in several fungi branches. In Microsporidia, SRP9/14 and SRP68/72 were lost, and SRP RNA might have lost all of its Alu domain.

In several groups, mutations in the external loop of helix 8 gave rise to penta- (Heterokonta, Entamoeba, Parabasala) or hexaloops (many Ascomycota, Plants). Occasional mutations of this tetraloop from GNRA to URRC are found in bacteria and green algae. (A tree depicting the evolution is in Fig. 13.)

The details of SRP evolution outlined above may naturally be subject to discussion. Nevertheless, it is clear that SRP has been subject to a large number of interesting changes during evolution, and our studies provide insights into the evolution of ribonucleoprotein complexes in general.

# Conclusion

We have developed methods to identify SRP components by analyzing genome sequences data. We have used these to identify a large number of novel genes. All kingdoms of life and many different phylogenetic groups are represented. Together, this information gives important clues to the phylogeny, structure and function of SRP. Although a lot of effort has been put into SRP research over the last decade, the present work has provided surprising evidence on the degree of variation of SRP. As SRP RNA is one of four non-coding RNAs found in all domains of life (the others being rRNA, tRNA and RNaseP RNA), these results may also provide important clues to the evolution and function of non-coding RNAs in general.

### Future projects

Experimental verification of some protist SRP RNAs needs to done to decide on the sequence and structure of SRP RNA (especially in Microsporidia, Rhodophyta and *Giardia*), since these seem to differ too much from the known SRP RNAs to be reliably predicted.

As the chloroplasts of higher plants, including some green algae, seem to lack SRP RNA, experimental studies on how SRP in chloroplasts function should give valuable information on how ribonucleoprotein particles evolve and on RNA–protein interactions.

A project in collaboration with Jeremy Brown and Rob van Nues of CMB at The Medical School, University of Newcastle has been started in which we will combine mutational studies and bioinformatics to elucidate the secondary structure of fungal SRP RNAs.

Although our method of finding SRP RNA works very well for species where the SRP RNA of close relatives already is known, more divergent SRP RNAs are hard to detect using an automated procedure with pattern matching and covariance models, since they diverge too much from the current models. Also, the secondary structure (and tertiary interactions) is sometimes hard to predict, as not enough sequences with a probable similar struture are available for a thorough comparative analysis. For these reasons, we

will develop a procedure to automatically identify a full or partial SRP RNA based on new sequence and structure alignments for all identified SRP RNAs.

Finally, we want to investigate the evolution of SRP and to elucidate the functionally important features using bioinformatics tools. For instance, it is of interest to study the protein and RNA motifs involved in Alu domain assembly and the translational arrest function. This includes the tRNA-like RNA being an SRP component in certain protozoa [8] and the putative bacterial Alu protein (HBsu) of the *Bacillus* type SRP [6]. As the number of completed genomes grows, it will also be possible to characterize SRP components through the analysis of promoters, terminators and synteny.

# Acknowledgements

This work has been done in collaboration with my supervisor Tore Samuelsson, and to some extent my former colleague Marco Regalia, who assembled the first alignments and produced the first covariance models, and Christian Zwieb who is the main keeper of the SRPDB. Lately, Rob W. van Nues and Jeremy D. Brown in Newcastle were also part of the team. However, without the almost infinite patience of Tore Samuelsson this work had not been finished, thank you all (including my family and the Rosenblad foundation)!

# References


1. Koch, H.G., M. Moser, and M. Muller, *Signal recognition particle-dependent protein targeting, universal to all kingdoms of life.* Rev Physiol Biochem Pharmacol, 2003. **146**: p. 55-94.
2. Abell, B.M., et al., *Signal recognition particle mediates post-translational targeting in eukaryotes.* Embo J, 2004. **23**(14): p. 2755-64.
3. Li, X., et al., *A chloroplast homologue of the signal recognition particle subunit SRP54 is involved in the posttranslational integration of a protein into thylakoid membranes.* Proc Natl Acad Sci U S A, 1995. **92**(9): p. 3789-93.
4. Gundelfinger, E.D., et al., *Structure and evolution of the 7SL RNA component of the signal recognition particle.* Embo J, 1984. **3**(10): p. 2325-32.
5. Halic, M., et al., *Structure of the signal recognition particle interacting with the elongation-arrested ribosome.* Nature, 2004. **427**(6977): p. 808-14.
6. Nakamura, K., et al., *Bacillus subtilis histone-like protein, HBsu, is an integral component of a SRP-like particle that can bind the Alu domain of small cytoplasmic RNA.* J Biol Chem, 1999. **274**(19): p. 13569-76.
7. Beja, O., E. Ullu, and S. Michaeli, *Identification of a tRNA-like molecule that copurifies with the 7SL RNA of Trypanosoma brucei.* Mol Biochem Parasitol, 1993. **57**(2): p. 223-9.
8. Liu, L., et al., *The trypanosomatid signal recognition particle consists of two RNA molecules, a 7SL RNA homologue and a novel tRNA-like molecule.* J Biol Chem, 2003. **278**(20): p. 18271-80.
9. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J.D., *Molecular Biology of the Cell, 3rd ed.* 1994.
10. Vogel, J., et al., *Experimental and computational analysis of transcriptional start sites in the cyanobacterium Prochlorococcus MED4.* Nucleic Acids Res, 2003. **31**(11): p. 2890-9.
11. Dieci, G., et al., *Intragenic promoter adaptation and facilitated RNA polymerase III recycling in the transcription of SCR1, the 7SL RNA gene of Saccharomyces cerevisiae.* J Biol Chem, 2002. **277**(9): p. 6903-14.
12. 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.
13. Oguro, A., et al., *Bacillus subtilis RNase III cleaves both 5'- and 3'-sites of the small cytoplasmic RNA precursor.* J Biol Chem, 1998. **273**(31): p. 19542-7.
14. Schattner, P., *Searching for RNA genes using base-composition statistics.* Nucleic Acids Res, 2002. **30**(9): p. 2076-82.
15. Wang, W., et al., *Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4448-53.
16. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.

17. Birney, E. and R. Durbin, *Using GeneWise in the Drosophila annotation experiment.* Genome Res, 2000. **10**(4): p. 547-8.
18. 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.
19. Sonnhammer, E.L., et al., *Pfam: multiple sequence alignments and HMM-profiles of protein domains.* Nucleic Acids Res, 1998. **26**(1): p. 320-2.
20. Womble, D.D., *GCG: The Wisconsin Package of sequence analysis programs.* Methods Mol Biol, 2000. **132**: p. 3-22.
21. Fichant, G.A. and C. Burks, *Identifying potential tRNA genes in genomic DNA sequences.* J Mol Biol, 1991. **220**(3): p. 659-71.
22. Pavesi, A., et al., *Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions.* Nucleic Acids Res, 1994. **22**(7): p. 1247-56.
23. Laferriere, A., D. Gautheret, and R. Cedergren, *An RNA pattern matching program with enhanced performance and portability.* Comput Appl Biosci, 1994. **10**(2): p. 211-2.
24. Sakakibara, Y., et al., *Stochastic context-free grammars for tRNA modeling.* Nucleic Acids Res, 1994. **22**(23): p. 5112-20.
25. Eddy, S.R. and R. Durbin, *RNA sequence analysis using covariance models.* Nucleic Acids Res, 1994. **22**(11): p. 2079-88.
26. Searls, D.B., Proc Natl Conf Artif. Intell., 1988: p. 386-391.
27. Searls, D., *The Linguistics of DNA.* Am. Sci., 1992. **80**: p. 579-591.
28. Searls, D.B., *The language of genes.* Nature, 2002. **420**(6912): p. 211-7.
29. 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.
30. Griffiths-Jones, S., et al., *Rfam: annotating non-coding RNAs in complete genomes.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D121-4.
31. Berman, H.M., et al., *The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids.* Biophys J, 1992. **63**(3): p. 751-9.
32. Yang, H., et al., *Tools for the automatic identification and classification of RNA base pairs.* Nucleic Acids Res, 2003. **31**(13): p. 3450-60.
33. Klein, R.J. and S.R. Eddy, *RSEARCH: finding homologs of single structured RNA sequences.* BMC Bioinformatics, 2003. **4**(1): p. 44.
34. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Res, 2003. **31**(13): p. 3429-31.
35. Zuker, M., *On finding all suboptimal foldings of an RNA molecule.* Science, 1989. **244**(4900): p. 48-52.
36. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7287-92.
37. McCaskill, J., *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* Biopolymers, 1990. **29**: p. 1105-1119.
38. Batey, R.T., et al., *Crystal structure of the ribonucleoprotein core of the signal recognition particle.* Science, 2000. **287**(5456): p. 1232-9.

39. 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.
40. Leontis, N.B., J. Stombaugh, and E. Westhof, *The non-Watson-Crick base pairs and their associated isostericity matrices.* Nucleic Acids Res, 2002. **30**(16): p. 3497-531.
41. Heard, D.J., et al., *An upstream U-snRNA gene-like promoter is required for transcription of the Arabidopsis thaliana 7SL RNA gene.* Nucleic Acids Res, 1995. **23**(11): p. 1970-6.
42. Aspegren, A., et al., *Novel non-coding RNAs in Dictyostelium discoideum and their expression during development.* Nucleic Acids Res, 2004. **32**(15): p. 4646-56.
43. Adam, R.D., *Biology of Giardia lamblia.* Clin Microbiol Rev, 2001. **14**(3): p. 447-75.
44. Mason, N., L.F. Ciufo, and J.D. Brown, *Elongation arrest is a physiologically important function of signal recognition particle.* Embo J, 2000. **19**(15): p. 4164-74.
45. Keeling, P.J., *Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia.* Fungal Genet Biol, 2003. **38**(3): p. 298-309.
46. Slamovits, C.H., B.A. Williams, and P.J. Keeling, *Transfer of Nosema locustae (Microsporidia) to Antonospora locustae n. comb. based on molecular and ultrastructural data.* J Eukaryot Microbiol, 2004. **51**(2): p. 207-13.

# Appendix