Lost in the RNA World:

Non-coding RNA and the Spliceosome in the Eukaryotic Ancestor

A thesis presented in partial fulfilment of the requirements for the degree of

PhD
in
Bioinformatics

at Massey University, Palmerston North,
New Zealand.

Lesley Joan Collins

2004

Abstract

The "RNA world" refers to a time before DNA and proteins, when RNA was both the genetic storage and catalytic agent of life; it also refers to today's world where non-coding RNA (ncRNA, RNA that does not code for proteins) is central to cellular metabolism. In eukaryotes, non-coding regions (introns) are spliced out of protein-coding mRNAs by the spliceosome, a massive complex comprised of five ncRNAs and about 200 proteins. This study examines the nature of the spliceosome and other non-coding RNAs, in the last common ancestor of eukaryotes, called here **the eukaryotic ancestor**. By looking at the differences between ncRNAs from diverse eukaryotic lineages, it may be possible to infer aspects of the eukaryotic ancestor's RNA systems.

Comparing ncRNA and ncRNA-associated proteins involves the evaluation of the available software to search newly available basal eukaryotic genomes (such as *Giardia lamblia* and *Plasmodium falciparum*). ncRNAs are not often found using sequence-similarity based software, thus specialist ncRNA-search software packages were evaluated for their use in finding ncRNAs. One such program is RNAmotif, which was further developed during this study (with the help of its principle programmer), and which proved successful in recovering ncRNAs from basal eukaryotic genomes. In a similar manner, sequence-based search techniques may also fail to recover proteins from distantly related genomes. A new protein-finding technique called "Ancestral Sequence Reconstruction" (ASR) was developed in this thesis to aid in finding proteins that have diverged greatly between distantly-related eukaryotic species.

A large amount of data was collected to investigate aspects of the eukaryotic ancestor, highlighting data management issues in this post-genomic era. Two databases were created P-MRPbase and SpliceSite to manage, sequence, annotation and results data from this project.

Examination of the distribution of spliceosomal components and splicing mechanisms indicate that not only was a spliceosome present in the eukaryotic ancestor, it contained many of the components found in today's eukaryotes. Splicing in the eukaryotic ancestor may have used several mechanisms and have already formed links with other cellular processes such as transcription and capping. Far from being a simple organism, the last common ancestor of living eukaryotes shows signs of the molecular complexity seen today.

Preface and Acknowledgements

"And so, it begins"- Babylon5

Bioinformatics has always held an interest for me, probably from when the first computer appeared in the corner of the laboratory back when I was starting as a technician in molecular biology. It always amazed me how much information was available, out there, if only you knew how and where to search for it. Many years, and a number of programming languages later, this project set out to explore the wealth of genomic information presently available, and to show that much can be learnt about biological function through the combination of biologically-based knowledge and computational analysis. Although this project was computer-based; I remain very much a molecular biologist. Instead of a "wet-lab", I now use the computer, unless, of course, I spill my coffee over the keyboard, then it becomes a wet-lab.

There are many people I would like to thank for their help and contribution during this project. First of all to my supervisors David Penny and Mike Hendy for keeping me on track and filling in my copious amounts of spare time with many interesting and profitable distractions.

On the computing side of things I would like to thank the nocturnal Tim White for all the programming and computing help. Many thanks also to the Helix parallel processing facility at Albany, especially James Chai, for great support and advice. I am also very grateful for the assistance of Tom Macke who, in answering an e-mail for help in less than two hours (and thus setting a new record for software support) was invaluable in adapting RNAmotif for genomic searches.

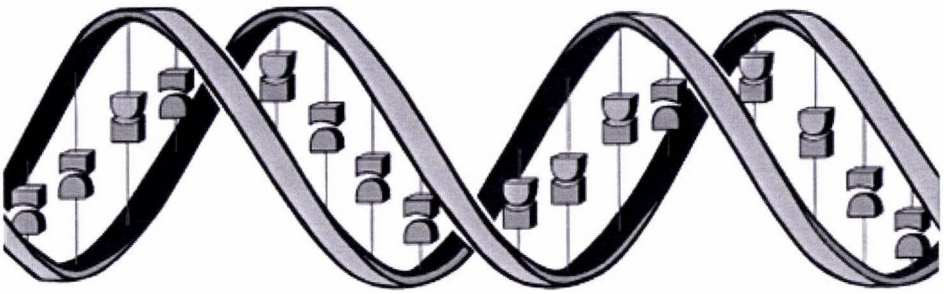
Thanks to the many people at the Allan Wilson Centre, especially Joy and Susan, for friendship, support and funding. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for the RT-PCR and sequencing of the *G. lamblia* ncRNA gene candidates over the course of some summers. A special thanks to Barbara Holland for critical reading of this thesis and the occasional stress-relieving cup of coffee.

Many thanks to Mitchell L. Sogin, and Andrew G. McArthur and their teams at the *Giardia lamblia* Genome Project, Marine Biological Laboratory at Woods Hole, for access to non-public data before the *G. lamblia* genome was publicly released.

Lastly, many, many thanks go to my family for all the craziness and grumpiness especially during the writing of this thesis. To my husband, Maurice who tried to help as much as possible, although not understanding a word of what I was researching, and to Shannen who taught me that a happy face drawn in the middle of one's work is not altogether a bad thing.

This project was funded partly by the Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution.

Table of Contents



Figures and Tables

Chapter 1: Lost in the RNA World – an Introduction

Chapter 2: Zen and the art of finding non-coding RNA genes

Manuscript Figures:

Chapter 3: Having a BLAST with Ancestral Sequences

Manuscript Figures:

Chapter 4: Splicing and the Spliceosome in the Eukaryotic Ancestor

Appendix E: Data management in the post-genomic era

Terminology

Alternative splicing: Process by which one pre-mRNA can be processed to form any one of a number of different mature mRNAs.

Bioinformatics: Information technology applies to the management and analysis of biological data.

Basal Eukaryote: A unicellular eukaryotic species not belonging to the crown group of eukaryotes.

BLAST: (Basic Local Alignment Search Tool) Method for rapid screening of nucleotide and protein databases.

Candidate sequence: Preliminary sequence recovered from a database with searching software.

Crown Eukaryote: An eukaryotic species belonging to either the animal, fungi or plant lineages.

Data-mining: Process by which useful data is extracted from a database.

Eukaryote: Organism with membrane-bound nuclei in its cell(s).

Eukaryotic Ancestor: The last common ancestor of living (extant) eukaryotes.

Excavate: Lineage of basal eukaryotes comprised of flagellate protozoa that contain a ventral feeding groove. This lineage included Diplomonads (*Giardia lamblia*) and Euglenozoa (*Trypanosoma brucei*).

Exon: Protein-coding region of a pre-mRNA.

Exon definition: Mechanism by which the boundaries between introns and exons are recognised by protein binding across the exon.

First Eukaryote: Theoretically, the first organism to envelop its nucleus in a membrane and distinguish itself from prokaryotes.

Intron: Non-coding region within a pre-mRNA. In eukaryotes introns are spliced out of the pre-mRNA by the spliceosome.

Intron definition: Mechanism by which the boundaries between introns and exons are recognised by protein binding across the intron.

LUCA: Last Universal common Ancestor: The last common ancestor of all living organisms.

Mitochondria: An organelle found in most eukaryotes that manufactures adenosine triphosphate (ATP) which is used as an energy source for the cell. Mitochondrial-like organelles present in some basal eukaryotes are hydrogenomes and mitosomes.

mRNA: (Messenger RNA) RNA transcribed from DNA as pre-mRNA which is then spliced to form the mature mRNA. Mature mRNA is then translated by the ribosome into protein.

ncRNA: (Non-coding RNA) RNA that does not code for proteins. Includes functional and sterile RNA.

Polyadenylation: The enzymatic addition of a sequence of 20 to 200 adenyl residues at the 3' end of an eukaryotic mRNA

PolyA tail: The string of 20 to 200 adenyl residues added to the 3' end of an eukaryotic mRNA by the process of polyadenylation. This region targets the mRNA to the ribosome prior to translation.

Polycistronic operon: One pre-mRNA transcript containing exons for more than one gene. In eukaryotes these genes are spliced using the SL-trans-splicing mechanism

pre-mRNA: (Preliminary mRNA) produced from DNA by transcription containing exons (protein-coding regions) and introns (non-coding regions).

Prokaryote: Unicellular organisms (bacteria and archaea) having cells lacking membrane-bound nuclei.

Py-tract: (Polypyrimidine Tract) Motif region near the 3' end of an intron with a high percentage of pyrimidines. This region binds to spliceosomal components during splicing.

Query sequence: Sequence used to search a target genome for candidate sequences.

Ribosome: Ribonucleoprotein complex responsible for translating mRNA into proteins.

RNA World: Hypothetical time in the evolution of early life, before DNA and proteins, where RNA was both the genetic storage and catalytic molecule.

RNP: (Ribonucleoprotein) A complex of ncRNA and proteins. RNPs mentioned in this study include snRNPs, RNaseP, the spliceosome and the ribosome.

rRNA: (Ribosomal RNA) ncRNA that together with proteins, comprise the ribosome.

Secondary structure: Structure formed with the folding of RNA. Helices (stems) are formed by the hydrogen bonding between certain pairs of nucleotides. Loops are single-stranded regions at the ends of stems.

SL-RNA: (Spliced Leader RNA) ncRNA used in trans-splicing to form the 5'end of the mature transcript.

snRNA: (Small nuclear RNA) group of ncRNAs that are components of the spliceosome.

Spliceosome: The ribonucleoprotein complex in eukaryotes that removes introns from a pre-mRNA, i.e. the site of eukaryotic splicing.

Splicing: The process by which introns are removed from a pre-mRNA.

Sterile RNA: Transcribed RNA that does not appear to have any function.

Target genome: Genomic database that is being searched by a particular method.

Trans-splicing: Splicing together of two independently transcribed mRNAs. One type of trans-splicing is SL-trans-splicing where an SL-RNA is joined to each exon in a polycistronic operon.

5'UTR: (read five prime untranslated region) region of mRNA before the start codon of the protein coding sequence, often contains the 5' cap.

3'UTR(read three prime untranslated region) region of mRNA after the stop codon of the protein coding sequence, contains the polyA-tail.

Measurement
av: Average
bp: Base-pair
kD: KiloDalton
nt: Nucleotide

Chapter 1 Lost in the RNA World – an Introduction

"It might look as if I am doing nothing, but at the cellular level I'm really very busy" – Anon

The "RNA world" is a term that originally referred to a time, before DNA and proteins when it was likely that RNA was both the genetic storage and catalytic agent of life (Gesteland et al. 1999). It is with a little tongue-in-cheek that I use this term to refer, not to a time in the past, but to the present. In eukaryotes RNA metabolism is involved with basic cellular process including transcription, translation and processing of other RNA molecules (Anantharaman et al. 2002) thus it can be easy indeed to get "lost" in the world of RNA. For convience, the world of RNA can be divided into two groups, messenger RNA (mRNA) which is the product of DNA transcription that is translated into protein (hence it is often called 'coding' RNA). Then there is non-coding RNA (or more accurately 'untranslated' RNA) which is transcribed from DNA but is not translated into protein. Non-coding RNA (ncRNA) includes catalytically-active functional RNAs which are an integral part of the flow of genetic information in eukaryotes, but also includes non-functional 'sterile' mRNA transcripts that are often produced by eukaryotic cells (Elmendorf et al. 2001; Lehner et al. 2002). Some ncRNAs such as snoRNAs and microRNAs can be contained in the intergenic regions (introns) between the coding regions (exons) and are released from the mRNA transcript during splicing (Tycowski and Steitz 2001). Thus one mRNA transcript may contain both coding and non-coding RNA (Figure 1.1).

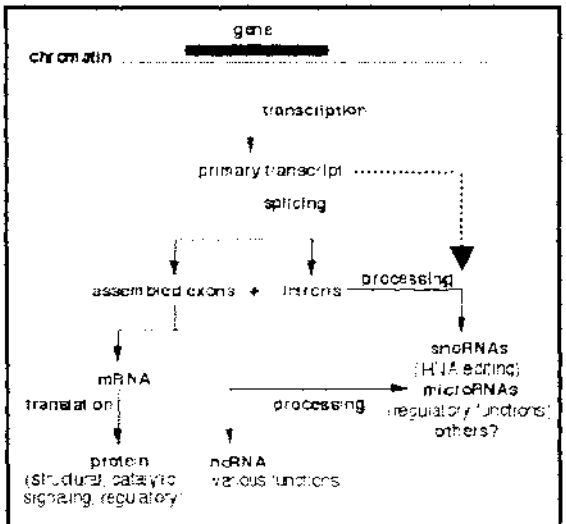


Figure 1.1: Based on a figure from Mattick (2003), showing how in eukaryotes, one transcript can result in both coding and non-coding RNA.

ncRNAs come in many flavours (Table 1.1) and range in length from 21-25 nucleotides (nt) for the MicroRNA family (development modulators), to > 10,000 nt for RNAs involved in gene silencing (Storz 2002). ncRNAs are also involved in gene regulation and the transport of mRNA from the nucleus to the cytoplasm. Often ncRNAs form part of RNA-protein complexes (Ribonucleoproteins, RNPs). One example is the ribosome itself which is comprised of ncRNA components (ribosomal RNA or rRNA) and 70-80 associated ribosomal proteins.

Knowledge of most ncRNAs has been limited largely to those from crown eukaryotes and even after the completion of many genome sequences, both the number and diversity of ncRNA genes remain largely unknown (Eddy 2001).

ncRNA		Function	Example
tRNA	Transfer RNA	Amino acid transfer	$tRNA_{ala}$
snRNA	Small Nuclear RNA	Spliceosome component	U5snRNA
snoRNA	Small Nucleolar RNA	rRNA modification	U18 C/D snoRNA
miRNA	Micro RNA	Gene regulation	*let-7* RNA, *lin-4* RNA
siRNA	Small interfering RNA	mRNA degradation	PSK132 RNA target: human protein serine kinase H1 - PSK
rRNA	Ribosomal RNA	Ribosome components	12S rRNA
RNaseP	Ribonuclease P	tRNA processing	H1RNA
RNaseMRP	Ribonuclease MRP	rRNA processing	Human RNaseMRP RNA
SRP-RNA	Signal Recognition Particle RNA	Protein secretion	7S RNA (SRP-RNA)
Telomerase RNA	Telomerase RNA	Telomeric DNA synthesis	Human telomerase RNA
rsRNA	Riboswitch RNA	Gene regulation	Winkler et al. 2004
grRNA	Gene regulatory RNAs	Regulate specific gene expression	Xist, mei, DGCR5

Table 1.1: Some of the functional ncRNAs found in eukaryotic cells.

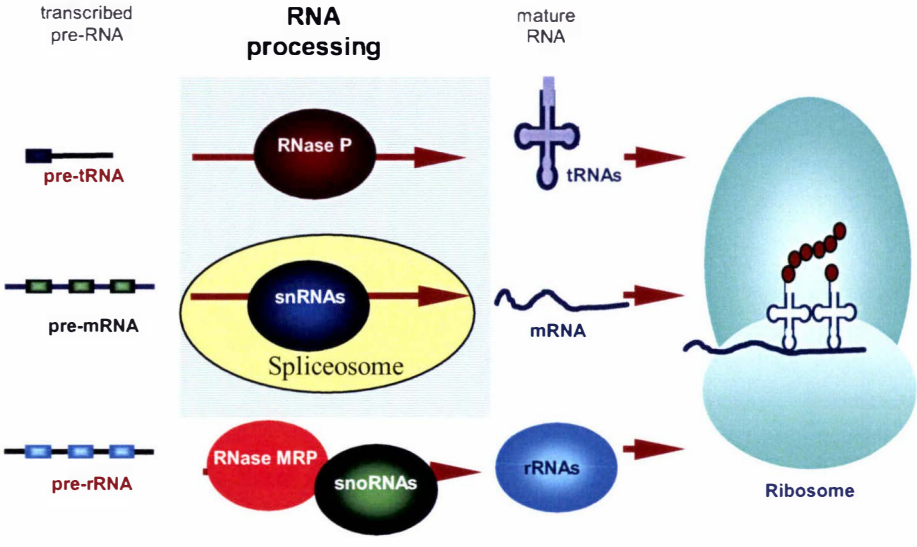


Figure 1.2: A representation of some of the RNA processing events in eukaryotes that lead to the translation of mRNA to protein within the ribosome (adapted with permission from Penny & Poole 1999). Not all features of the RNA processing events are shown, only key elements. The blue box represents the ncRNA-protein complexes used in this study. Precursor ribosomal RNA (rRNA) is processed by reactions involving the RNaseMRP ribonucleoprotein and various small nucleolar RNAs (snoRNAs). This rRNA then binds to ribosomal proteins to form the ribosome. Precursor transfer RNA (Pre-tRNA) is processed by the RNaseP ribonucleoprotein. tRNA can then deliver a designated amino acid to the ribosome. Messenger RNA (mRNA) is the transcript that codes for protein synthesis. In eukaryotes mRNA contains introns, non-protein coding sequences that must be removed by the large RNA-protein complex, the spliceosome (which contains snRNAs and many proteins), before proceeding to the ribosome where the transcript is used to produce the required protein.

Studies of RNA processing can cast light on the early evolution of life and events surrounding the transition from the ancient RNA-world to present cellular systems (Anantharaman et al. 2002).

Most eukaryotic RNAs are processed by other RNAs in one way or another (Figure1.2). rRNAs and tRNAs are released from larger precursors by the action of RNaseP, and in eukaryotes, transcribed mRNA (messenger RNA) must have any introns removed by one of the most intricate ribonuclearprotein complexes in the eukaryotic cell, the spliceosome.

Eukaryotic protein-coding genes usually contain one or more introns (depending on species) which need to be "spliced" out before the mature mRNA transcript can be translated into protein. A mass spectrometry study of the human spliceosome revealed a staggering ~300 proteins (Zhou et al. 2002). This large number of proteins was subsequently reduced (to ~200 proteins) core spliceosomal proteins (Jurica and Moore 2003), leaving still a large number of proteins to associate with five ncRNAs (snRNAs) during the splicing process.

Introns, snRNAs and spliceosome-associated proteins have been characterised in a number of eukaryotes (Archibald et al. 2000; Anantharaman et al. 2002; Nixon et al. 2002) suggesting that both introns and the spliceosomal machinery evolved very early in the eukaryotic lineage. A principle aim of this project is to examine what the distribution within eukaryotes of spliceosomal components, introns and splicing mechanisms can reveal about splicing in the eukaryotic ancestor, the last common ancestor of living eukaryotes.

1.1: Eukaryotic Phylogeny

In order to study ncRNA and RNA processing in eukaryotes a phylogenetic tree indicating relationships between the eukaryotic lineages must first be described. However, the phylogeny of eukaryotes is still plagued with uncertainties and is the source of much debate. Some unanswered questions include the root of the eukaryotic tree, the identity of early evolving lineages and the evolutionary relationships between many groups of unicellular eukaryotes (Dacks et al. 2002).

The eukaryotic phylogenetic tree shown in Figure 1.3 has been adapted from (Simpson and Roger 2002) (http://hdes,biochem.dat.ca/Rogerlab/) and indicates both branches that are relatively stable, and others that differ between hypotheses for eukaryotic relationships. This tree will be referred to throughout this project. Although there are still questions about the deeper relationships between the major groups of eukaryotes, the tree in Figure 1.3 can still be used to hypothesise fundamental characteristics of the "**eukaryotic ancestor**", defined here as the last common ancestor of all living eukaryotes (see later in this chapter).

Recently some extremely small eukaryotes have been discovered (Baldauf 2003; Stoeck and Epstein 2003). These are nano- (2 to 20 μm) and pico- (<2 μm) eukaryotes, that overlap bacteria (~0.5 to 2 μm) in size. For example, the smallest described eukaryote, *Ostreococcus tauri* (a phytoplankton belonging to the chlorophyte (green algae) lineage) is <1 μm in diameter but still has a nucleus, 14 linear chromosomes, one chloroplast and several mitochondria (Courties et al. 1998). The eukaryotic ancestor studied in this project can only represent eukaryotes that are known to date but there may well be other eukaryotes awaiting investigation.

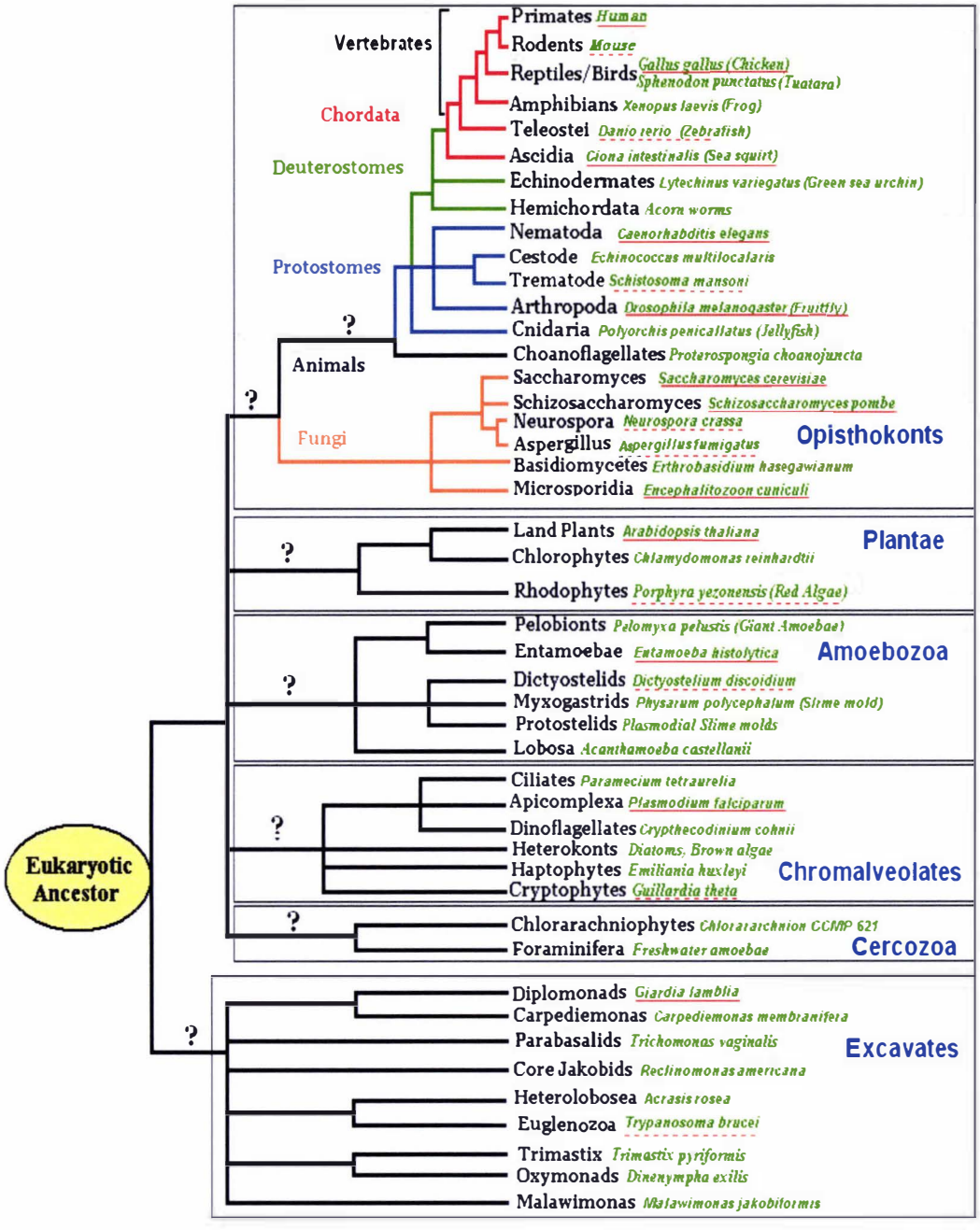


Figure 1.3: Eukaryotic phylogenetic tree that will be used throughout this project. This tree is based on the tree shown at http://hdes,biochem.dat.ca/Rogerlab/, including information from Simpson and Roger 2002. A representative species for each lineage are shown in green. Branches that join differently in alternative hypothesis are indicated with "?". The animal, fungal (Opisthokonts) and plant groups are often referred collectively as the 'Crown' eukaryotes, while the rest (Amoebozoa, Chromalveolates, Cercozoa and Excavates) are collectively called the basal eukaryotes. In animals, Protostomes (blue lineages) and Deuterostomes (green lineages) are grouped based on differences in embryo development. Within the Deuterostomes are the chordates (red lineage) which in turn contain the vertebrate lineages (as marked on the tree).

There have been some important developments in understanding the earliest divergences within eukaryotes. Some of the earliest eukaryotic phylogenetic trees, based on gene sequence data, were derived from small subunit ribosomal RNA (SSU-rRNA) sequences (Sogin 1991) (Figure 1.4). This tree splits into three parts. One contains the 'crown' eukaryotes namely animals, yeasts and plants. In addition the crown group included heterokonts (brown algae), alveolates (e.g. the malaria parasite, *Plasmodium falciparum*) and rhodophytes (red algae) appear to emerge almost simultaneously as the order of branching is irresolvable. The second part of the SSU-rRNA tree contains the basal eukaryotes, so named as they branch in stepwise emergence from the base of the eukaryotic tree.

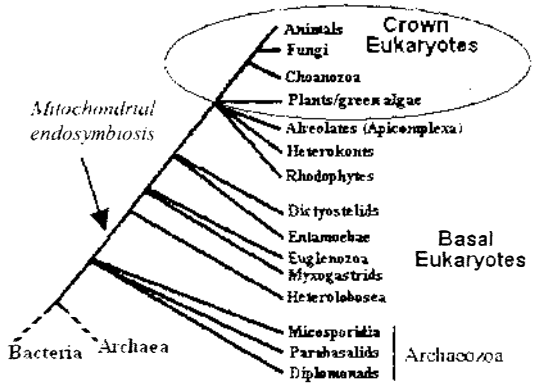


Figure 1.4: An early eukaryotic phylogenetic tree based on SSU rRNA sequences. Eukaryotes, traditionally referred to as the crown eukaryotes are indicated. The archaeozoa hypothesis and position of mitochondiral endoysymbiosis are no longer accepted and this tree is now not generally accepted.

The position of three most basal eukaryotic lineages on the SSU-rRNA tree (Figure 1.4) gave rise to the "Archezoa" hypothesis (Cavalier-Smith 1989). This is that that the three early-branching lineages which lack mitochondria (diplomonads, parabasalids and microsporidia), emerged before a mitochondrial endosymbiosis and would have been thus living relics from an amitochondrial period of eukaryotic evolution. Other eukaryotic phylogeny hypotheses include the 'eukaryotic big-bang" hypothesis that suggests that immediately after the endosymbiotic event leading to mitochondria, eukaryotes evolved in a massive radiation of 4-10 groups whose interrelationships are fundamentally irresolvable (Philippe and Germot 2000; Philippe et al. 2000).

Phylogenetic analysis of other genes including RNA polymerase II (Dacks and Doolittle 2001), dihydrofolate reductase (DHFR) and thymidylate synthase (TS) (Simpson and Roger 2002), completion of a number of basal eukaryotic genomes, and improved microbiological techniques has led to the SSU-rRNA tree being significantly changed. The main change being that there are no extant Archaeozoa, i.e. eukaryotes from before the origin of the mitochondrion. However, replacement trees retain some of the relational uncertainty between eukaryotic lineages due to two significant obstacles in reconstructing eukaryotic phylogeny, 1. long-branch attraction and 2. loss of information due to poor taxon sampling of free-living basal eukaryotes (Stiller and Hall 1999; Dacks and Doolittle 2001).

Long-branch attraction (also called long edge attraction) is an effect produced by phylogenetic tree-building programs where long branches tend to group together and cause misplacement of taxa on the tree (Hendy and Penny 1989; Stiller and Hall 1999), thus in

phylogenies rooted by a distant outgroup (such as the eukaryotes rooted by either bacteria or archaea) unrelated fast evolving ingroups will emerge independently as the deepest offshoots, being attracted by the long branch of the outgroup (Gribaldo and Philippe 2002). Many of the branches leading to the basal eukaryotic taxa are long and it has been argued that the tree constructed from SSU-rRNA analysis may be an artefact due to variation in the rate of eukaryotic molecular evolution (Stiller and Hall 1999). Furthermore, it has been shown mathematically that on current models it is impossible from sequence data itself to reconstruct ancestral data at the root of "deep" phylogenetic trees, even with normal mutation rates (Mossel 2003; Mossel and Steel 2004), and thus the 'correct' tree topology may not be able to be determined on sequence data alone. A major problem is that the majority of basal eukaryotes whose genomes have been sequenced or are in the process of being sequenced, are not free-living but parasitic (e.g. *Giardia lamblia* and *Entamoeba histolytic* are both human intestinal parasites), and thus are expected to have evolved in such a way as to enhance their parasitic lifestyle. Reductive evolution (Andersson and Kurland 1998) is one of the most common results, where parasites lose many features of their free-living relatives. In particular, parasites appear to have lost genes required by their free-living relatives (Bapteste et al. 2002). This raises questions about the relationships between different parasites as some genes may be similar not due to the phylogenetic relationship between their species but because there may be similar constraints on this gene due to its parasitism, although such factors are not well defined. Similar to the process shown in parasitic bacteria, horizontal transfer has occurred in the genomes of parasitic basal eukaryotes (Richards et al. 2003). This is unlikely to affect the results of this study as it has been found that genes involved in transcription, translation and relating processes (such as splicing and tRNA processing) are rarely horizontally transferred (Jain et al. 1999).

Figure 1.5 illustrates that the ancestral eukaryote is not the "first" eukaryote (the closest eukaryote to the prokaryotic lineages), and is not defined as such in this study. This study does not set out to "sort-out" or clarify any of the deeper relationships shown on the eukaryotic tree. Instead the focus is on identifying RNA processing characteristics which occur in several lineages of extant eukaryotes (Figure 1.3). Any similarities in these characteristics between say the crown eukaryotes

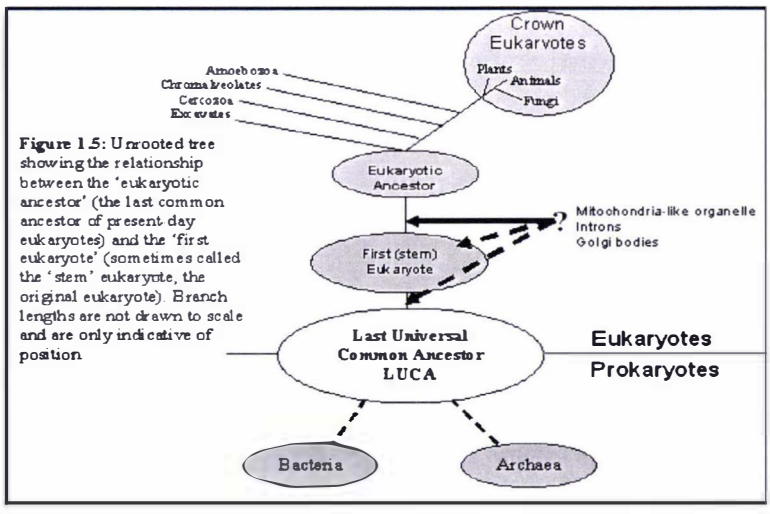


Figure 1.5: Unrooted tree showing the relationship between the 'eukaryotic ancestor' (the last common ancestor of present-day eukaryotes) and the 'first eukaryote' (sometimes called the 'stem' eukaryote, the original eukaryote). Branch lengths are not drawn to scale and are only indicative of position.

(animals, fungi and plants) and basal eukaryotes could be universal eukaryotic traits (Koonin et al. 2000; Anantharaman et al. 2002; Koonin et al. 2002; Baldauf 2003). Examination of some RNA processing systems, in particular tRNA processing (RNaseP) and mRNA splicing (the spliceosome) from both crown and basal eukaryotic lineages are used in this project to infer the nature of systems thought to have been present in the eukaryotic ancestor.

There are at present a number of eukaryotes whose genomes have either been completely sequenced, or nearly so (underlined in red in Figure 1.3). Animal genomes include human (*Homo sapiens*) (Lander et al. 2001), mouse (*Mus musculus*) (Bult et al. 2004), zebrafish (*Danio rerio*), nematode worm *Caenorhabditis elegans* (Wilson 1999), fruitfly (*Drosophila melanogaster*) (Adams et al. 2000) and sea-squirt (*Ciona intestinalis*) (Dehal et al. 2002). Sequenced yeast genomes include *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Schizosaccharomyces pombe* (Wood et al. 2002) and *Neurospora crassa* (Arnold and Hilton 2003; Galagan et al. 2003) and sequenced plant genomes include *Arabidopsis thaliana* (Schoof et al. 2002) and *Oryza sativa* (rice) (Yazaki et al. 2004). Some basal eukaryotic genomes are now available including *Giardia lamblia* (McArthur et al. 2000), *Entamoeba histolytica* (amoebic dysentery) (Mann 2002), the slime-mold *Dictyostelium discoideum* (Eichinger and Noegel 2003) and *P. falciparum* (malaria) (Gardner et al. 2002). This list is obviously incomplete and appears to increase weekly.

A number of types of ncRNAs have now been characterised in crown eukaryotic species (such as humans, *A. thaliana*, *S. cerevisiae* and *S. pombe*). However, few ncRNAs have been characterised from basal eukaryotes. Examining the differences between ncRNAs from both crown and basal eukaryotes it may be possible to reconstruct, ancestral aspects of their RNA processing systems. Because of the importance of identifying ncRNAs in basal eukaryotic lineages, it is important to describe these organisms in more detail

1.2: Basal Eukaryotes

1.2.1: Giardia lamblia

One basal eukaryote of great interest in this project is the pathogenic anaerobic diplomonad *G. lamblia* (also called *Giardia intestinalis* Figure 1.6). It was originally described as "ancient and primitive" on account of its many bacterial-like characteristics (including a pyrophosphate-based energy metabolism and an arginine dihydrolase pathway of energy production). It was also thought to be lacking organelles such as mitochondria and Golgi bodies, as well as introns, and was positioned as the most basal eukaryotic branch on the eukaryotic SSU-rRNA tree (Lloyd et al. 2002)

Figure 1.6: Photograph of *Giardia lamblia* trophozoite (mobile stage). Reproduced from Henze and Martin 2003.

Lately, advanced microbiological techniques have quashed *G. lamblia's* original amitochondrial status; *G. lamblia* has now been shown to contain both mitosomes (highly reduced mitochondria-like organelles (Tovar et al. 2003) and Golgi bodies[1] (Dacks et al. 2003). Although *G. lamblia*'s basal position on the eukaryotic tree has been modified (compare its position on the eukaryotic trees in Figure 1.3 and Figure 1.4), it is still considered to be one of the most 'primitive' (oldest-diverging) living eukaryotes (Li and Wang 2004). Recently published genome sequence data can now allow the putative identification of ncRNA and protein genes known to be part of RNA processing events. Comparisons of RNA processing genes between *G. lamblia* and other eukaryotes (especially crown eukaryotes), will give some insight into mechanisms that are likely to be found in all eukaryotes.

The mitosomes from *G. lamblia* are unlike classic mitochondria which use aerobic respiration to make ATP. Instead the giardial mitosomes instead synthesise iron-sulphur-based enzymes, which then are used to make ATP in the cytosol (Henze and Martin 2003; Tovar et al. 2003). Genes involved in the iron-sulphur cluster assembly pathway (*iscS* and *iscU*) have been cloned from *Trichomonas vaginalis* and *Giardia lamblia* and show mitochondrial ancestry by phylogenetic analysis (Tovar et al. 2003). Mitochondrially derived genes such as *cpn60* (Roger et al. 1998) and *hsp70* (Arisue et al. 2002) have been found in *G. lamblia* and in light of its "amitochondrial" status was once thought to have been the result of horizontal gene transfer (Horner and Embley 2001), either from a "lost" mitochondria or another prokaryote (Sogin 1997; Doolittle 1998; Roger et al. 1998). With the discovery of the mitosomes it is possible that these genes have been transferred to the nucleus from the ancestor of the reduced organelle in *G. lamblia* although the lack of association of the giardial Cpn60 protein with the mitosome cannot, for this gene rule out, horizontal gene transfer (Tovar et al. 2003). If all eukaryotes contain the Cpn60 protein then it is much simpler for this gene to have transferred from the mitochondrion to the nucleus.

Hydrogenosomes are another class of mitochondrial-like organelles that occur quite widely in anaerobic eukaryotic lineages, including ciliates and fungi (Embley et al. 2003). Hydrogenosomes use pyruvate:ferredoxin oxidoreductase (PFO) to carry out the metabolism of pyruvate, transferring energy via ferredoxin to hydrogenase and producing hydrogen as the end product. In contrast, aerobic eukaryotes use pyruvate dehydrogenase located in the mitochondrion to process pyruvate and hydrogen is not produced (Embley et al. 2003). Phylogenetic analysis suggests that ancient eukaryotes likely contained both PFO and hydrogenase, thus could produce hydrogen (Embley et al. 2003). The discovery of mitosomes and hydrogenosomes as mitochondrial 'relatives' has removed all of the "deeply-branching"

[1] Golgi bodies serve as the major sorting point in the secretory pathway, selectively targeting proteins and lipids to different organelles to prevent the inappropriate meeting of certain intracellular components (Dacks et al. 2003).

eukaryotes from the amitochondrial list leaving no 'known' eukaryote that have always been amitochondrial .

G. lamblia is a very distinctive organism among eukaryotes and has some unusual genetic features. Giardial trophozoites (the mobile stage (Figure 1.6) as opposed to the sessile cyst stage) contain two nuclei, each of which contain the same amount of DNA, contain equal copies of rRNA genes (Kabnick and Peattie 1990), lack nucleoli and are transcriptionally active (Adam 2000). During vegetative growth, both nuclei switch from a diploid genome (2 copies of genome – 2N ploidy) to a tetraploid genome with a resulting eight copies (8N ploidy) of the genome being present in the organism (in certain strains this number can increase to 10 – 12N ploidy) (Vanacova et al. 2003), creating interesting problems for whole genome assembly.

Giardial mRNAs contain extremely short 5' untranslated regions (UTRs) (0-14 nt compared to 90 in mammals and 52 in yeast) (Elmendorf et al. 2001; Li and Wang 2004). The 5'UTR is typically modified in crown eukaryotes by a methyl-guanosine cap and plays an important role in translation control by correctly positioning the first AUG codon for the ribosome (Li and Wang 2004). The majority of mRNAs in *G. lamblia* do not contain a cap, however at least one exception has been found (Vanacova et al. 2003), and the presence of a number of homologues of yeast and human translation initiation factors suggest that capped mRNA may be present and translated in *G. lamblia* (Li and Wang 2004).

The 3'UTR regions of eukaryotic mRNAs provide a site for polyadenylation (the formation of a poly(A) tail of ~200 adenines), a feature that plays a role in mRNA stability and export from the nucleus (Vanacova et al. 2003). 3'UTR regions in both *G. lamblia* and *Entamoeba sp.* are polyadenylated; most are shorter (lengths of 5-43 nt) than those found in crown eukaryotes (with a few exceptions in both species) but similar to the length of prokaryotic poly(A) tails (lengths of ~30 nt) (Anantharaman et al. 2002).

'Antisense' transcripts are mRNAs that are complementary to 'sense' mRNAs (i.e. code for proteins) (Elmendorf et al. 2001). Antisense transcripts are found in both eukaryotes and prokaryotes and are typically 'sterile' mRNAs, lacking an open reading frame (ORF) and thus are unable to code for any protein (thus, in this project will be classed as ncRNAs). There is evidence that these sterile mRNAs could sometimes have roles in splicing (Kramer 1996), rRNA maturation as well as roles in gene regulation. *G. lamblia* has a higher than expected level (~20%) of antisense mRNAs (Elmendorf et al. 2001). These transcripts come from low levels of expression of many loci and it is unclear as to whether they represent errors in transcription or they have in fact regulatory functions within the cell[2].

To date a single intron has been found in *G. lamblia* along with a number of proteins known to be involved in splicing (Nixon et al. 2002) demolishing *G. lamblia's* intron-less

[2] A study of some human sterile transcripts (Lehner et al. 2002) suggests regulatory functions for some of the human antisense RNAs.

status. From these investigations and similar studies in other basal eukaryotes such as *Ent. histolytica* and *Trichomonas sp.* (Williams and Keeling 2003) it is likely that the eukaryotic ancestor already contained features such as mitochondria, Golgi bodies and introns (Dacks and Doolittle 2001) but does not resolve the issue as to whether the "first" eukaryote (Figure 1.5) also had these features that are not found in prokaryotes.

1.2.2: Plasmodium, Entamoeba and Microsporidia

Other eukaryotic genomes used frequently in this project are the basal eukaryotes *Entamoeba histolytica* and *Plasmodium falciparum*, and the microsporidium *Encephalitozoon cuniculi*. *Plasmodium falciparum* is one of the four species of *Plasmodium* that infects humans to cause malaria and like other members of the apicomplexa phylum, harbour a relict plastid (the apicoplast) homologous to the chloroplasts of plants and algae. An international effort was launched in 1996 to sequence the *P. falciparum* genome to open new avenues of research (Gardner et al. 2002). Approximately 50% of the genome encodes proteins with few ncRNAs (mainly rRNA and tRNA) identified (Gardner et al. 2002). Its well annotated genome sequence is extremely useful to comparative genomic research as predicted open reading frames and known gene similarities are already indicated.

Entamoeba histolytica (abbreviated as *Ent. histolytica* during this project) is an amoebic parasite that causes amoebic dysentery in humans (Vanacova et al. 2003). Its genome is as yet unannotated in the public release (another release is available but has restricted access), but the data is in the advanced stages of assembly (Mann 2002). Like *G. lamblia*, *Ent. histolytica* contains mitosomes instead of mitochondria and genes of apparent mitochondrial origin (*cpn60* and *hsp70*) (Bakatselou et al. 2003). *Entamoeba* species contain slightly longer 5'UTR regions (5-20 nt) with a few longer (420, 126 and 265 nt) regions have been characterised but elements such as caps have not yet been described (Vanacova et al. 2003). The position of *Ent. histolytica* on the eukaryotic tree is quite distant from that of *G. lamblia* and although in both of these species ATP synthesis occurs in the cytosol (Henze and Martin 2003), no comparative studies between their mitosomes has yet been done to determine any features of an ancestral mitochondrial-like organelle.

Microsporidia are obligate intracellular parasites infesting many animal groups. They are responsible for various digestive and nervous clinical syndromes in immunocompromised humans (such as HIV-infected and transplant patients) (Vivares et al. 2002). *Encephalitozoon cuniculi* (abbreviated during this project as *Ecz. cuniculi* to avoid confusion with *Ent. histolytica*) is a microsporidium which was once thought to be a basal amitochondriate but is now recognised as being part of the fungal lineage. Its genome is incredibly reduced (only ~2.6 MBases) and like that of *P. falciparum*, well annotated with protein and some ncRNA information (Katinka et al. 2001). Similar to the other previously amitochondriate eukaryotes

(diplomonads (e.g. *G. lamblia*) and parabasalids (e.g. *Trichomonas vaginalis*)), *Ecz. cuniculi* also contains mitosomes and a "simplified" Golgi apparatus (Katinka et al. 2001). The *Ecz. cuniculi* genome, although greatly reduced, offers a lineage that branched before the split of two highly researched fungi (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*) and may indicate ancestral genes (both protein and ncRNA) that are essential for fungal viability.

1.3: Thesis Structure and Organisation

Genetic similarities between the highly-researched crown eukaryotes (vertebrates, yeasts and plants) and recently sequenced basal eukaryotes (such as *G. lamblia, Ent. histolytica* and *P. falciparum*) can indicate genetic mechanisms likely to have been present in their last common ancestor, the eukaryotic ancestor. This project compares ncRNA and protein sequences from crown and basal eukaryotes to examine splicing, and the spliceosome, in the eukaryotic ancestor. Investigation of the spliceosome in the eukaryotic ancestor involved a number of other issues (Figure 1.7), namely the evaluation and development of tools necessary for the search for relevant ncRNAs and their associated proteins, and the development of databases for effective genomic data management.

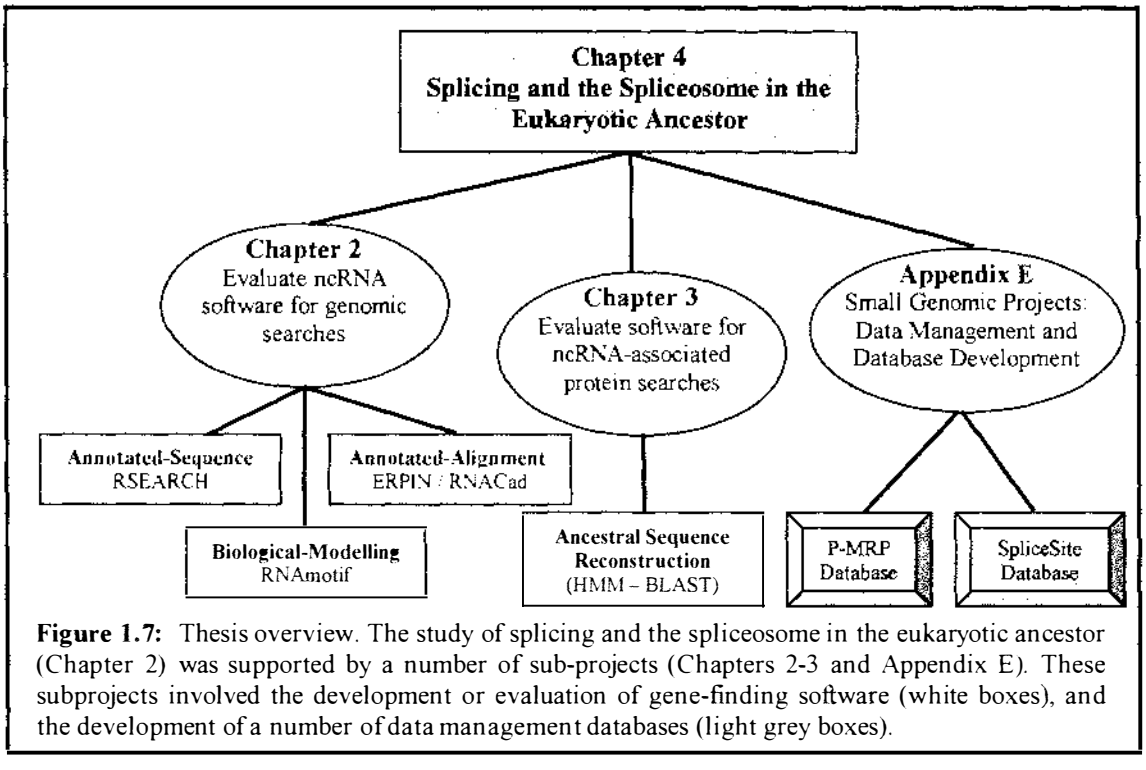


Figure 1.7: Thesis overview. The study of splicing and the spliceosome in the eukaryotic ancestor (Chapter 2) was supported by a number of sub-projects (Chapters 2-3 and Appendix E). These subprojects involved the development or evaluation of gene-finding software (white boxes), and the development of a number of data management databases (light grey boxes).

1.3.1: ncRNA Identification – Chapter 2

Searching for ncRNA genes, such as those involved in splicing, in genomic data is still a developing area. ncRNAs fold first into a two-dimensional "secondary structure" before folding into a three-dimensional structure (Rivas and Eddy 2000). Often ncRNAs from distantly related species retain the same secondary structure but have different nucleotide sequences

through co-variation, a process where a nucleotide may change so long as the pairing characteristics of the overall sequence is retained. This means that an ncRNA gene may have quite different nucleotide sequences in distant species, but still have all the required functional characteristics. This lack of sequence similarity is what makes ncRNA genes so hard to find and why secondary structure information is incorporated into both RNA-detection and RNA-analysis software (Eddy 2002).

Searching databases for homologues based on sequence similarity is only useful for the most slowly evolving or for large ncRNAs like ribosomal RNAs, and becomes much less reliable for other ncRNAs. If there is also a large evolutionary distance between the query species and the target species, sequence similarity methods often fail to uncover any potential gene candidates (Eddy 2002). A draft of the human genome showed that protein-coding sequences accounted for less than 2% of the total genome size leaving a lot of area uncovered by standard gene-finding techniques (Mattick 2001). This provides good reason for developing tools for finding and analysing ncRNA genes, which is why this project also investigates software for use in searching genome sequence databases for ncRNA genes (Chapter 2).

ncRNA-finding software was evaluated to find two ncRNAs, the U5snRNA and the RibonucleaseP RNA, in a number of protist species. The U5snRNP is one of the spliceosomal ribonucleoproteins involved in eukaryotic intron splicing, playing an important role in tethering the two exons to juxtapose them for catalysis (Newman 1997; Xu et al. 1997; Peng et al. 2002). The U5snRNA has been identified in many species including human, the yeast *S. cerevisiae* and the plant *A. thaliana*, but also from the microsporidian *Ecz .cuniculi*, and the basal eukaryote *P. falciparum*. Not only could the U5snRNA be used to test ncRNA-finding software by running searches of the *Ezh. cuniculi* and *P. falciparum* genomes, it was likely that the U5snRNA could also be found in other basal eukaryotes such as *G. lamblia*, *Ent. histolytica* and *Dictyostelium discoideum* (a soil-living amoeba, slime-mold).

The other ncRNA used for testing ncRNA-finding software was RibonucleaseP (RNaseP) RNA, which also forms a complex with specific proteins. RNaseP is a ribozyme (an RNA-based enzyme) that cleaves 5'-leader sequences from precursor-tRNA to leave a mature tRNA molecule (Xiao et al. 2002) (Khan and Lal 2003). Although this complex is found in all types of cells there are differences in both the RNA and protein RNaseP components from archaea, bacteria and eukaryotes (Collins et al. 2000). Evidence has suggested that the RNA component of RNaseP is a molecular fossil dating from the RNA world as the RNA is catalytic, ubiquitous and occupies a central position in metabolism (Jeffares et al. 1998). In bacteria, the RNaseP complex consists of one catalytically-active RNA and one protein, whereas in most archaea and eukaryotes the RNA has not been shown to be catalytic in the absence of its associated proteins (Xiao et al. 2002). RNaseP RNA contains both conserved and variable regions in its sequence and secondary structure, making it more of a challenge for ncRNA-

finding software. It has not been found previously in any basal eukaryote (or in microsporidia) although its proposed universal distribution makes it likely to be present.

Three different types of software were evaluated with the U5snRNA and the RNaseP RNA. The first type of software requires an alignment of the ncRNA sequences annotated with a consensus secondary structure (RNACad - http://www.cse.ucsc.edu/~mpbrown/rnacad/; and ERPIN- (Gautheret and Lambert 2001)); the second type uses a single sequence annotated with its secondary structure (RSEARCH - (Klein and Eddy 2003); and the third type uses a grammar (code) representation of the ncRNA which includes sequence and secondary structure motifs in a biological-modelling approach (RNAmotif - (Macke et al. 2001). Each of these software types has advantages and disadvantages that are discussed in Chapter 2. The most successful in searching for the U5snRNA and RNaseP RNA in this study was RNAmotif which used a "biological-modelling" approach. This approach models biological features such as conserved protein and RNA binding sites that cannot be obviously designated in the other two sequence-based approaches and offered the flexibility required to find ncRNA genes in basal eukaryotic genomes. The success of this program has resulted in the published manuscript "*Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif*" which is included in Chapter 2.

1.3.2: Identifying ncRNA-associated Proteins – Chapter 3

Sometimes it may not be possible to uncover particular ncRNA genes within a genome. However, if essential proteins that specifically bind to that particular ncRNA can be found then there is more confidence that the ncRNA is also there. One example is the U5snRNA specific protein Prp8 considered to be the most highly conserved protein in the spliceosome (Fast and Doolittle 1999). The mere presence of Prp8 has been used to argue for the presence of introns in *Trichomonas vaginalis*, a parabasalid protist, although as yet no introns have been identified in this species (Achsel et al. 1998; Fast and Doolittle 1999; Nixon et al. 2002). *G. lamblia* also contains Prp8 and to date one intron has been identified (Nixon et al. 2002).

Finding ncRNA-associated proteins is not always as straight-forward as throwing query sequences at the BLAST software. Some of these proteins may not be well conserved between species, nor contain known protein motifs, making identification with standard sequence similarity software difficult. As a result it is often difficult to identify these proteins in distantly related species. There is still a real need for accurate and fast tools to analyse sequences and, especially to find genes and determine their functions (Mathe et al. 2002) and thus a new technique called "Ancestral Sequence Reconstruction" (ASR) was developed in Chapter 3 to aid in finding proteins that have diverged greatly between distant species. The resulting published manuscript "Using ancestral sequences to uncover potential gene homologs" (Collins et al. 2003) is included in Chapter 3.

Some eukaryotic RNaseP-associated proteins (Pop1, Pop4, Pop5 and Rpp21) were used in this study as a prelude to searching for (and finding) the RNaseP RNA in some basal eukaryotes (Chapter 2), and to the larger spliceosomal-protein study (Chapter 4). A Pop4 protein candidate was found in *G. lamblia* using BLAST but candidates for the Pop5 and Pop1 proteins were found only with more extensive techniques such as HMMer (that uses HMM profiling), and ASR (both are explained more fully in Chapter 3) indicating that different techniques may be required to find different proteins.

1.3.3: Splicing and the Spliceosome in the Eukaryotic Ancestor – Chapter 4

Splicing in the crown eukaryotes is not by any means a 'simple' process. Three types of splicing have been shown to be present in eukaryotes (major, minor and trans-splicing). Each of these splicing mechanisms requires a different, but overlapping, set of snRNA and protein components. Recent studies from human and yeast spliceosomes (Rappsilber et al. 2002; Stevens et al. 2002; Jurica and Moore 2003) have characterised a large number of splicing-proteins that contribute to the spliceosomal complex. Basal eukaryotic spliceosomes have scarcely been studied but some snRNA and protein information is now available. Investigating the distribution of these spliceosome components among present eukaryotic lineages, including both crown and basal eukaryotes, can reveal how the spliceosome evolved within eukaryotes.

This study takes a parsimonious approach, in that it is more likely that complex features common to a number of eukaryotic lineages were present in the ancestor of those lineages (the alternative view is that common features arose independently in exactly the same way in different lineages). Here, computational searches of eukaryotic genomes are combined with literature and database information to determine the nature of the spliceosome and splicing present in the eukaryotic ancestor. Tools and techniques developed in the previous chapters are used in this study to aid in the identification of snRNAs and spliceosomal proteins from basal eukaryotes to enable comparisons with those already characterised from crown eukaryotes.

Some of the splicing proteins have been used in previously published surveys of eukaryotic genomes (Anantharaman et al. 2002; Koonin et al. 2004) but these surveys did not include any basal eukaryotic genome nor did they cover the range of spliceosomal proteins in depth. Previous studies at best could predict the presence of a few splicing proteins in the last common ancestor of animals, plants and fungi. This study incorporates information from these surveys but also expands the range of both the spliceosomal proteins and eukaryotic genomes included in the searches to look comprehensively at the spliceosome in the eukaryotic ancestor.

Results in Chapter 4 show that in the eukaryotic ancestor it was very likely that a spliceosome was present. Far from being a simplified molecule, the ancestral spliceosome may have contained many components (e.g. the core snRNPs, Sm-proteins and RNA helicases) that are found in today's eukaryotes. Of the three types of splicing found in today's eukaryotes, it is

possible that two of these mechanisms (i.e. major and trans splicing) were present in the eukaryotic ancestor as each contain conserved features between crown and basal eukaryotes.

Splicing can now be seen as a fundamental and ancestrally-derived aspect of eukaryotic life as it is likely to have evolved before the last ancestor of living eukaryotes. Contrary to the idea that splicing may have been a 'simplified' mechanism in this ancient organism it can now be suggested that this was not the case and that splicing and the spliceosome had already evolved in a sophisticated cellular process. Splicing has already been linked to other cellular processes such as mRNA export, transcriptional-elongation and polyadenylation (Lynch and Richardson 2002). Examples include the UAP^{56} protein which has multiple tasks in spliceosome assembly but is also a negative inhibitor of mRNA export (Luo et al. 2001), and Rds3 a spliceosome component which is also involved in mRNA export (Wang and Rymond 2003). Results in Chapter 4 show that these proteins were likely to have been present in the eukaryotic ancestor indicating that modern links between multiple RNA processing functions are may also be ancestrally derived.

1.3.4: Additional information

This project concludes with a summary and a look at future directions that this work could follow.

A significant issue that arose during the course of this project was the importance of including data management practices in small genomic analysis projects. It was found that with the large number of international databases holding very different types of genetic information, a very small number of proteins produced a large amount of recovered data, data that could very easily become uncontrollable without some type of management system. Technologies between databases differ and naming conventions can be inconsistent causing problems when managing data from a variety of databases. Appendix E summarises some of the major data management problems associated with small gene-finding projects which were encountered in this project. With the expansion of genomic sequencing there has been development of software for the handling and management of large genomic sequencing projects and there are many laboratory information management systems (LIMS) available but there is still a need for tools and software for small genomic project management. Appendix E looks at the problem of "data explosion" where a small number of query sequences can soon snowball into a large amount of data. Appendix E also summarises the development of two databases created to handle genomic information for this project, P-MRPbase (a database of RNaseP and RNaseMRP RNA and protein components) and SpliceSite (a database of spliceosomal components). These databases were created using a popular small database program (Microsoft Access), which perhaps would make diehard computer scientists cringe and professional data managers cry, but the ready

availability of both the software and training enabled small databases to be quickly developed and modified to handle the different types of data used.

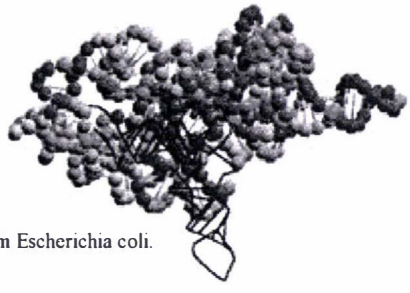
Additional information included in the appendices include an article written for NZBioscience (The journal of the New Zealand Society for Biochemistry and Molecular Biology) on RNA analysis, and a (2-page) poster abstract that was selected for a "Flash" presentation at the ECCB'2003 (European Conference on Computational Biology, Paris, France) in September 2003. Perl Scripts developed especially for this project are also included in Appendix D.

1.4: Summary

There are many more ncRNAs than was ever suspected (Storz 2002; Griffiths-Jones 2004; Herbert 2004; Miriami et al. 2004). A challenge for the future will be to identify the whole complement of ncRNAs in an organism and to investigate their functions. Techniques for achieving this are not as yet available though a good start has been made. Analysis of the spliceosome in present-day eukaryotes has shown that the eukaryotic ancestor was not by any means a "simple" organism and already had sophisticated mechanisms of gene regulation including gene splicing. Further study may show other mechanisms that are likely to be present in this ancient organism. This is only the first step towards looking at how RNA processing and other cellular processes have changed between prokaryotes and eukaryotes (or even between eukaryotes and prokaryotes if in fact eukaryotes are shown to have arisen earlier than prokaryotes (Jeffares et al. 1998).

The current situation in ncRNA analysis is reminiscent of the early days of protein sequence analysis. Not too long ago the few programs available for sequence searches were only known to the select few but were too impractical and expensive to run on the computers of the time (Eddy 2002). However, since then fast heuristic tools such as BLAST appeared to enable the wider community access to protein computational analysis. In order to allow effective evolutionary and functional analysis of ncRNAs, it is now time for such tools to be developed and expanded in this rapidly expanding area of genomics.

Welcome to the new RNA world!



Three dimensional structure of the RNaseP RNA from Escherichia coli.
Courtesy of the RNase P Database (Brown 1999).

Chapter 2: Zen and the art of finding non-coding RNA genes

"He who thinks everything is easy will end up finding everything is difficult. Therefore, the Sage, who regards everything as difficult, meets with no difficulties in the end" - Lao Tzu (Tao Teh Ching).

2.1: Introduction

In this post-genomic era, the emphasis is usually on the identification of protein-coding genes, but it is known that there are a large number of genes that produce RNA transcripts that do not code for proteins (ncRNAs). There is considerable interest in finding homologues of ncRNA genes, but almost all standard methods of gene identification assume that the gene encodes a protein, and thus many ncRNA genes remain invisible (Rivas and Eddy 2000; Eddy 2002). Moreover most biochemical analyses of cell fractions are not designed to detect ncRNAs (Mattick 2001). RNA is labile (i.e. easily degraded) and the main protocol for analysing ncRNA function is to use gene-knockouts, which are technically demanding, often ambiguous and not undertaken lightly (Mattick 2001). This provides good reason for the assessment of software to search for and investigate ncRNA candidate genes.

Functional ncRNAs have both a two-dimensional (secondary) and a three-dimensional (tertiary) component (Rivas and Eddy 2000). Structural focus is usually on the secondary-structure because it is easier to calculate efficiently. Some RNA tertiary structure information is available (Tamura et al. 2004) but there is still a long way to go before tertiary structure can be calculated usefully and applied to prediction software. Often ncRNAs may evolve different nucleotide sequences but still retain the same secondary-structure through covariation, a process where a nucleotide may change so long as the overall pairing characteristics of the sequence is retained. Thus the same ncRNA in two distantly related species may have a more conserved secondary-structure than sequence (Durbin 1998). This lack of sequence similarity is what makes ncRNA genes so hard to find and why it is desirable to incorporate secondary-structure information into both RNA-detection and RNA analysis software (Eddy 2002).

Software has been developed in order to identify ncRNA genes in genomic data. There are programs to detect specific ncRNAs such as tRNA (tRNAscan (Lowe and Eddy 1997)) and H/HCA-snoRNAs (Edvardsson et al. 2003), but the majority of ncRNA molecules do not have any specialised detection software available. The earliest types of ncRNA-searching programs were pattern matching programs (e.g. PatScan (Dsouza et al. 1997)), which could only look for small but highly conserved sequence motifs and could not include any secondary-structure information in their searches, thus their use was limited. Other programs have used predicted minimum-free-energy properties to screen for specific groups of ncRNAs (H/HCA-snoRNAs (Edvardsson et al.

2003) and tRNA (Tsui et al. 2003)). The few general RNA-detection programs available use a number of approaches, but all incorporate secondary-structure information, non-canonical base-pairing, as well as nucleotide sequence information of known ncRNAs. This chapter highlights some of the past and current software which can be used for ncRNA gene identification and analysis.

The majority of ncRNA-search software has been published with searches of specific ncRNAs (such as the Iron-Response-Element; (Macke et al. 2001)) that have highly conserved secondary-structure and some highly conserved sequence motifs. However, none of the published studies attempted to search for ncRNAs that contain highly variable as well as highly conserved regions. Two examples of these types of ncRNAs are the snRNAs and the eukaryotic RNaseP RNA. The U5snRNA is an ideal ncRNA to test ncRNA-search software. It is an essential component of the eukaryotic spliceosome (described in more detail in Chapter 4), has been extensively studied (Peng et al. 2002; Malca et al. 2003) and contains conserved secondary-structure and sequences involved in protein and nucleotide binding. U5snRNA genes have been characterised for a number of crown and basal eukaryotes including the 'excavates' *Trypanosoma brucei* and *Leptomonas collosoma*. The critical question is whether it is present in all eukaryotes, in which case it is expected to be present in the genomes of other basal eukaryotes such as *Giardia lamblia* and *Entamoeba histolytica*.

The other test ncRNA used in this study is the eukaryotic RNaseP RNA. RNaseP has been found in bacteria, archaea and eukaryotes, but it has not yet been characterised for **any** basal eukaryote although given its apparent universal distribution it is expected to be present. Apart from some short nucleotide motif sequences, the RNaseP RNAs have little nucleotide sequence homology making this gene difficult to find in more distant species using only sequence similarity-based software. Because secondary-structure characteristics are the conserved feature of the eukaryotic RNaseP RNA, being able to find it in any basal eukaryote is a real test of any search software's capacity to successfully incorporate secondary-structure and sequence information.

At present there are three main mechanisms used in ncRNA-search software:

1. Alignment of multiple sequences combined with secondary-structure annotation,
2. A single sequence with secondary-structure annotation,
3. Biological-modelling (using biological information by combining sequence and secondary-structure elements that represent protein or RNA binding sites).

In this study four programs, covering these different mechanisms, were evaluated for their use in finding ncRNA genes from eukaryotic genomes (including some basal eukaryotic genomes).

2.2:Results

2.2.1: Alignment with Secondary-structure Annotation

RNACad

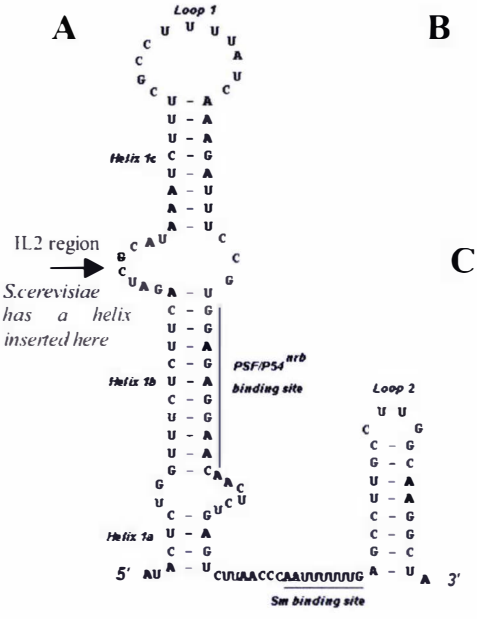
RNACad[3] uses stochastic context-free grammars (SCFGs) to model RNA sequence and secondary-structure information. It was designed to search databases for RNA secondary-structure and to produce structural multiple alignments for a set of structurally related sequences. Context-free grammars allow for nested long-distance pairwise correlations between terminal symbols, ideal for storing information about base-pairings involved in RNA secondary-structure. Whereas regular grammars such as those used in pattern search algorithms generate strings from left to right, context-free grammars can generate strings from outside-in (i.e. nested information). Stochastic (probabilistic) context-free grammars used in ncRNA searching can contain additional information about the likelihood of the structural event (i.e. how likely is it for this nucleotide to be paired in a helix or remain single-stranded), and the nucleotide composition of each position. Another SCFG-based program tRNAscan-SE (Lowe and Eddy 1997) is available but can only be used to search for tRNAs (Eddy 2002).

Unfortunately, although, the RNACad software offered great potential for ncRNA genomic searches, and despite help from departmental "experts", this software would not compile correctly on any Linux computer. Communication was tried to the programs' author and designated help e-mail address but no reply was ever received. Refereed publications that used RNACad could not be found although this program is listed at IMB-Jena (The RNA world site: http://www.imb-jena.de/RNA.html) under RNA software. After some time spent (with much appreciated assistance from Dr. Paul Gardner and other Linux gurus) it was decided to leave RNACad for now and concentrate on other ncRNA-search software.

Another program COVE (Eddy and Durbin 1994) also uses SCFGs (Stochastic-context-free-grammars) to model annotated RNA alignments. This software was not evaluated in this study due to time constraints but has been described as requiring heavy processing ability. This would be an ideal program for future testing of SCFG-based ncRNA software.

[3] RNACad does not have a published reference, but is available under the GNU public licence from http://www.cse.ucsc.edu/~mpbrown/rnacad/.

A



B

```
>Human U5 snRNA sequence
AUACUCUGGUUUCUCUUCAGAUCGCAUAAAUCUUUCGCCUU
..(((((..((((((((((((.........((((((((.....
UUAUCAAAGAUUUCCGUGGAGAGGAACAACUCUGAGUCUUA
.....))))))))...)))))))))))).....))))....
ACCCAAUUUUUUGAGCCUUGCCUUGGCAAGGCUA
............((((((((((....))))))))).
```

C

```
h5(tag = 'Helix1a', len = 4)
ss(len =2)
h5(tag = 'Helix1b' len = 11)
ss(len = 8)
h5(tag = 'Helix1c' len = 8)
ss(tag = 'Loop1', len = 11)
h3(tag = 'Helix1c')
ss(len = 3)
h3(tag = 'Helix1b' seq =~ "GGAGAGGAA")
ss(len = 6)
h3(tag = 'Helix1a')
ss(tag = 'Sm binding region', minlen = 14, maxlen = 20,
                                    seq =~ AAUUUUUNG")
h5(tag = Helix2', len = 8)
ss(tag = Loop2', len = 4)
h3(tag = 'Helix2)
```

D

```
# STOCKHOLM 1.0
#=GS HumanU5 DE HumanU5

HumanU5          AUACUCUGGUUUCUCUUCAGAUCGCAUAAAUCUUUCGCCUUUCAUCAAAGAUUUCCGU
#=GR HumanU5 SS  ..>>>>..>>>>>>>>>>>...>>...>>>>>>>>..........<<<<<<<<.<<<

HumanU5           GGAGAGGAACAACUCUGAGUCUUAACCCAAUUUUUUGAGCCUUGCCUUGGCAAGGCUA
#=GR HumanU5 SS  <<<<<<<<<<.....<<<<.................>>>>>>>>....<<<<<<<<<.
//
```

Figure 2.1: Human U5 snRNA sequence and secondary structure showing the different ways that ncRNA can be represented.
A: Human-readable format, a graphical representation but hard for computers to read.
B: "Bracket notation", a machine-readable format, good for computers but hard for humans to read. '(' indicates the 5' base of a basepair and ')' the 3' base of a basepair. '.' indicates an unpaired nucleotide.
C: RNAmotif 'descriptor' representing the U5snRNA. Lines beginning with "h5" represent the 5' side of a helix, "h3" represents the 3' side of the helix and "ss" represents a single-stranded region. "len" gives the length of this region; a variable length can be stated with "minlen" (minimum length) and "maxlen" (maximum length). "seq" can be used to state a compulsory motif within a region. An additional user-defined scoring section can also be added. This is only a representative descriptor to show how a model can be translated into this computer-readable format. A full U5snRNA descriptor is shown in the submitted manuscript
D: Stockholm format that is used for input into the RSEARCH program. '>' represent the 5' base of a base-pair and '<' the 3' base of a base-pair (Note that this is 'opposite to the bracket notation). '.' indicates an unpaired nucleotide.

ERPIN (Easy RNA Profile IdentificatioN)

The ERPIN (Easy RNA Profile IdentificatioN) (Gautheret and Lambert 2001) software uses multiple aligned RNA sequences in together with a consensus secondary-structure, to perform first a profile construction and then a database search. This software is useful if a number of sequences are known for an ncRNA family and are able to be reliably aligned. It has the advantage of being able to statistically capture biases in an alignment that could escape human inspection. It is written in the C programming language and has source code and Linux executable files available for download (http://tagc.univ-mrs.fr/pub/erpin).

The input of ERPIN is a multiple sequence alignment of RNA sequences annotated with secondary-structure information. A log-odds score (lod-score) profile is constructed for each helix and single strand in the alignment. Given aligned sequences, this is a score that gives a measure of the relative likelihood that the sequences are related (as opposed to being unrelated). This score is constructed from two parts; the first calculates the probability that a particular nucleotide in each position is independent of the others. The other part calculates the probability that the nucleotides *a* and *b* have each been independently derived from some unknown nucleotide *c* in their common ancestor – *c* might be the same as *a* and/or *b*). The ratio of these two probabilities is known as the odds ratio. In order to arrive at an additive scoring system, the logarithm of this ratio is taken and is known as the log-odds ratio (lod). Alignment gaps are not permitted in helical regions, but helical bulges may be accounted for by breaking a helix into two sections with a single-stranded region between them.

ERPIN installed without problem and was tested using an alignment of U5snRNA sequences. Alignments of known U5snRNA sequences were available from the Rfam database (Griffiths-Jones et al. 2003)-http://rfam.wustl.edu/). The U5snRNA has been characterised from the microsporidian *Encephalitozoon cuniculi* (Chromosome XI: position 114087-114198) thus searches against the *Ecz. cuniculi* genome would be expected to recover this sequence.

Alignments with different combinations of sequences were used to test the ERPIN program. The first alignment consisted of U5snRNA sequences from human, mouse, rat, *Xenopus laevis* (frog), *Drosophila melanogaster* (fruitfly) and *Caenorhabditis elegans* (nematode). A second alignment also included the U5snRNA from the yeast *Schizosaccharomyces pombe* which added more gaps to the alignment but included a sequence that was more closely related to the *Ecz. cuniculi* U5snRNA sequence than the animal sequences used in the first alignment. The *Saccharomyces cerevisiae* U5snRNA sequence is known to have an additional helix in the IL2 region (indicated in Figure 2.1A) which would have added an additional large insertion into the alignment so this sequence was omitted from the test alignment. A third (positive control)

alignment contained all the sequences in the first and second alignments but also included the *Ecz. cuniculi* U5snRNA sequence.

A

```
>Sec_structure
..(.((..((((((((((((---...((...((((((((.....)))))))).)))))))))))))).-.....)).)
>Human
AUACUCUGGUUUCUCUUCA---GAUCGCAUAAAUCUUUCGCCUUUCAUCAAAGAUUUCCGUGGAGAGGAACA-ACUCU-GAGU
>Mouse
AUACUCUGGUUUCUCUUCA---GAUCGUAUAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGGAGAGGAACA-ACUCU-GAGU
>Rattus
AUACUCUGGUUUCUCUUCA---GAUCGUAUAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGGAGAGGAACA-ACUCU-GAGU
>Xenopus
AUACUCUGGUUUCUCUUCA--AAUUCGAAUAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGGAGAGGAACG-ACCAU-GAGU
>Drosophila
AUACUCUGGUUUCUCUUCA--AUGUCGAAUAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGGAGAGGAACACUCUAA-GAGU
>Celegans
CAACUCUGGUUCCUCUGCAUUUAACCGUGAAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGCAAAGGAGCA-UUUACUGAGU
```

B

```
>Sec_structure
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
33030033000000000000003330033311111111444444444411111111400000000000004044444400 40
11234455666666666661117778899900000000111111111110000000038866666666665377777744 92
>Human
AUACUCUGGUUUCUCUUCA---GAUCGCAUAAAUCUUUCGCCUUUCAUCAAAGAUUUCCGUGGAGAGGAACA-ACUCU-GAGU
.
.
>Celegans
CAACUCUGGUUCCUCUGCAUUUAACCGUGAAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGCAAAGGAGCA-UUUACUGAGU
```

C

```
Training set:   "erpin_align1.epn":
                6 sequences of length 126
Database:       "enceph.genome2"
                2495538 nucleotides to be processed in 13 sequences
                ATGC ratios: 0.264  0.263  0.238  0.235
Cutoff:          30.26  126.16

>Human_U5
FW   1       3..113      141.65
A.C.TC.TG.GTTTCTCTTCA.---.GAT.CG.CAT.AAATCTTT.CGCCTTTCATC.AAAGATTT.C.CG.TGGAGAGGAAC.A.-
.ACTCT-.GA.G.T.CTTAACCC.-.AATTTTTTGA.G-.CCT.TGCC.-.TT.--.GGCA.AGG

4993596 bases processed
cutoff: 30.26
12 config. per site
1 hit
```

Figure 2.2: ERPIN ncRNA search software input and output example.
Input for ERPIN consists of an RNA sequence alignment annotated by secondary structure information. The alignments shown are truncated from those used in the study and do not contain the single-stranded "Sm-binding region" nor the second-loop structure
A: The secondary structure can be annotated using bracket notation where "." represents single-stranded nucleotides and "(" and ")" represent the opening and closing of a base-pair within a helix. Gaps are represented by "-" in both the sequence alignment and the secondary structure annotation.
B: The bracket notation is converted (using the supplied Perl script) to an ERPIN format with the numbers (read down each column) representing the open and closed pairings. These numbers can be adjusted to indicate non-canonical and/or pseudoknot pairing events.
C: Example output of ERPIN which gives information about the alignment used "Training set" and the database searched which in this case was the *Ecz. cuniculi* genome (with added positive controls).

The secondary-structure used to annotate the ERPIN-test alignments was based on vertebrate U5snRNA sequences (Frank et al. 1994; Dix et al. 1998; Peng et al. 2002). Areas where the structure varied between species were annotated as being single stranded. ERPIN does not allow any gaps in any helical position so these regions were annotated as being single stranded. An annotated alignment using the widely-used bracket notation (Figure 2.1B) can be converted into the ERPIN-readable format using a Perl script (provided with the ERPIN package). An example alignment with its secondary-structure in both bracket and ERPIN notation is shown in Figure 2.2.

The human and the *Ecz. cuniculi* U5snRNA sequences, were added to the end of the *Ecz. cuniculi* and *G. lamblia* genomic databases as positive controls. These controls were expected to be recovered with all alignments and indicated that the program was working correctly. Although the *Ecz. cuniculi* U5snRNA is expected to be recovered from within its genome, the known U5snRNA sequence was added to the end of its genome-file as a separate sequence. This ensured that this positive control would still be recovered even if surrounding sequences obscured the true sequence within the genome[4]. Prior to ERPIN testing, BLAST searches of the *Ecz. cuniculi* genome, with all other known U5snRNA sequences did not recover the *Ecz. cuniculi* U5snRNA sequence indicating that sequence alone could not recover the desired sequence from this genome.

ERPIN results (shown in Table 2.1) showed that the *Ecz. cuniculi* sequence was not recovered except with itself in the positive control alignment - Alignment 3. This indicates that the nucleotide sequence information is perhaps the major influence on the model with which ERPIN conducts its search.

Table 2.1: Erpin Results. Recovery of the positive control sequences from the *Ecz. cuniculi* and *G. lamblia* genomic databases. + indicates that the control was recovered, - indicates that the control was not recovered.

Table 2.1	*Ecz. cuniculi* Genome Search			*G. lamblia* Genome Search	
Alignment	Positive control 1	Positive control 2	Positive control 3	Positive control 1	*G. lamblia* sequences
1	+	-	+	+	-
2	+	-	-	+	+
3	+	+	-	+	-
4	-	-	+	+	-

Positive control 1: Human U5 snRNA sequence added to end of database
Positive control 2: *Ecz. cuniculi* U5 snRNA sequence added to end of database
Positive control 3: *Ecz. cuniculi* U5 snRNA sequence contained in genome (Chromosome XI :114087-114198)

Alignment 1 (human, mouse, rat, *Xenopus laevis* (frog), *Drosophila melanogaster* (fruitfly) and *C. elegans* (nematode)). This alignment did not detect the *Ecz. cuniculi* U5snRNA sequence (positive controls 2 and 3) in either genome.

Alignment 2 added the *S. pombe* U5 snRNA sequence into alignment1 to see if a fungal sequence could detect the *Ecz. cuniculi* U5snRNA but it made no difference to the results.

Alignment 3 added Ecz cuniculi to alignment 2 (Positive control alignment). All positive controls were expected to be recovered with this alignment which was the case.

Alignment 4 used the U5 seed alignment (minus the *S. cerevisiae* sequence) from the Rfam database (Griffiths-Jones 2003):198 sequences. This alignment contained both human and *Ecz. cuniculi* control sequences but sequence information could be diluted by the presence of the many other sequences. A large number of vertebrate sequences are present in this alignment, and thus the human U5snRNA control sequence was recovered. However, there were no other microsporidian sequences present and the Ecz. cuniculi control was not recovered with this alignment.

[4] This was not the case with any of the sequences tested in this study but could still be a factor in genomes that are either AT-rich or AT-poor.

Searches of the *G. lamblia* genome did not recover any U5snRNA candidate sequences with any of the alignments. Parameters were adjusted to extremely low levels to search both genomes but had the same negative results.

A chance meeting at the ECCB 2003 conference (Paris, France in late September 2003) with ERPIN's primary author (Dr. Daniel Gautheret) allowed further discussion of the use of this program for ncRNA gene searches. ERPIN is presently being developed for the computational detection of MicroRNAs (miRNA) in animal genomes (presented at ECCB'2003) and is very useful when conserved sequence alignments of short nucleotide sequences can be constructed. A later version of ERPIN (version 3.9.8) was sent to me to determine if the latest improvements to the algorithm and scoring mechanism would improve my results, however this was not the case as the results were identical to those produced with the earlier version.

It is likely that ERPIN's dependency on the nucleotide sequence information in annotated alignments might preclude using this program to search for ncRNAs with conserved secondary-structure, but little primary sequence conservation. It is unlikely that this method would work with ncRNAs that also contain areas of secondary-structure variability such as the eukaryotic RNaseP RNA (personal communication, D. Gautheret). Alignments of some RNaseP RNAs (both whole sequence and with the conserved sections) tested with ERPIN confirmed this (data not shown).

2.2.2: Biological-modelling software

RNAmotif

The RNAmotif program (Macke et al. 2001) models biological information, in addition to sequence and secondary-structure, to search for ncRNA candidate genes through the design of an appropriate descriptor. The RNAmotif program was developed from an earlier program RNAMOT (Laferriere et al. 1994), but uses an expanded syntax for describing motifs and implementation of nearest-neighbor rules, and other schemes, for ranking hits. RNAmotif is written in the C programming language and is freely downloadable from ftp.scripps.edu/pub/macke/rnamotif-version.tag.gz (where "version" is the version number, currently 3.3.0). An RNAmotif descriptor is a short piece of code using special grammar rules to model a particular ncRNA (Figure 2.1C). The descriptor incorporates secondary-structure and sequence characteristics into the search model without the requirement for RNA sequence alignments. Sequence length (including minimum and maximum allowable lengths) and any sequence motifs (representing protein/RNA binding sites) contained in single-stranded or helical regions can be incorporated into the descriptor allowing for maximum flexibility in designing the search model. Small protein or RNA-binding motifs are extremely useful in descriptor design and offer powerful selection criteria within the search with

RNAmotif. A downside of this flexibility is that in order to design an effective descriptor, it is necessary to research the desired ncRNA thoroughly for its biological characteristics, incorporating sequence and secondary-structure information from a full literature search.

The first version of RNAmotif that was trialled (Version 2.2.0) compiled easily but ran very slowly against the smallest genomic databases (e.g. the *Ecz. cuniculi* genome of 2.6 MBases). An additional problem was the massive amount of data retumed by RNAmotif for each sequence area; results files >1 gigabyte of information were often returned. In a later version (Version 3.0.0) these problems were overcome in by the use of parallel processing to increase speed, and a mechanism for filtering the results, thus RNAmotif became a practical program to use for ncRNA genomic searching.

The testing of RNAmotif for genome searching is detailed in the accompanying manuscript – "*Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif*". This study was done with the aid of one of the principle authors of the RNAmotif software, Dr. Thomas Macke (The Scripps Research Institute, La Jolla, CA, USA). In this manuscript my contribution was to supply the concepts and to conduct the genome searches, with Dr. Macke making valuable adjustments to the RNAmotif code to allow efficient parallel implementation and results-filtering. In this paper, descriptors were constructed for two ncRNAs, the U5snRNA and the eukaryotic RNaseP RNA, and then used in genomic searches of some eukaryotic genomes including those from the basal eukaryotes *Giardia lamblia* and *Entamoeba histolytica*. RNAmotif descriptors for the U5snRNA were successfully trialled against test data, and against the *Ecz. cuniculi* and *P. falciparum* genomes in which this ncRNA has already been identified. RNAmotif then recovered U5snRNA candidates from other basal eukaryotes (*G. lamblia*, *Dictyostelium discoideum* and *Ent. histolytica*) and the sea-squirt *Ciona intestinalis*. The success of this software with the U5snRNA led to descriptors being created for the eukaryotic RNaseP RNA. RNaseP descriptors were more challenging to construct, but were also successfully tested and recovered candidates from *Ecz. cuniculi*, and the basal eukaryotes *Ent. histolytica* and *G. lamblia*. These are the first RNaseP RNAs to be recovered from any basal eukaryote and although they contain all features shown to be essential for an RNaseP RNA (Frank et al. 2000) will require some biochemical analysis for validity. RT-PCR and sequencing of the *G. lamblia* U5snRNA and RNaseP candidates[5] confirmed that these sequences were expressed (contained in the RNA content of *G. lamblia*) and that the sequences shown in the manuscript are correct. Although it requires more non-computational input (i.e. literature searches for initial model creation), than other similar software, RNAmotif software has proved very successful to date, in ncRNA searches of basal eukaryotic genomes.

[5] Many thanks to Trish, Anu and Alica who fitted this in over a couple of summers.

A long (two page abstract (included in Appendix A.1) based on this research was accepted for a "flash" (five minutes- no questions) presentation at the European Conference on Computational Biology held in Paris, France (September 2003).

Further analysis of the candidate ncRNA sequences uncovered using RNAmotif is shown in Figures 2.3-2.5. RNAalifold (Hofacker 2003) is a program which, given a sequence alignment, will compute the likely consensus secondary-structure to fit that alignment. The five *Ciona intestinalis* candidate U5snRNAs were aligned and annotated with the RNAalifold consensus secondary-structure (Figure 2.3A – displayed using GeneDoc[6]). Results were visualised using RNAforester (Sczyrba et al. 2003) to show the differences between the *C. intestinalis* consensus U5snRNA and the human U5snRNA (Figure 2.3B).

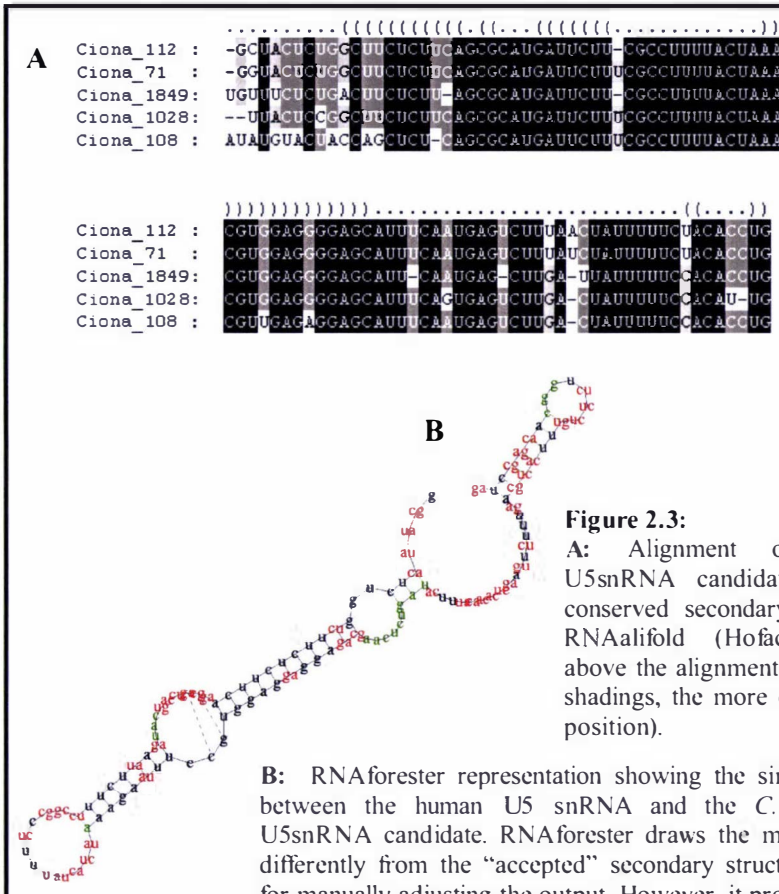


Figure 2.3:
A: Alignment of *Ciona intestinalis* U5snRNA candidate sequences with the conserved secondary structure predicted by RNAalifold (Hofacker 2003) annotated above the alignment. (The darker the column shadings, the more conserved the nucleotide position).

B: RNAforester representation showing the similarities and differences between the human U5 snRNA and the *C. intestinalis* 112-region U5snRNA candidate. RNAforester draws the modelled sequence a little differently from the "accepted" secondary structure with no mechanism for manually adjusting the output. However, it provides a clear comparison of the two secondary structures. Nucleotides in black are conserved in both sequence and secondary structure. Nucleotides in red have different nucleotides but maintain the same secondary structure. Nucleotides found in the *C. intestinalis* sequence only are in blue and those found only in the human sequence are in green.

[6] Nicholas Karl B., and Nicholas Hugh B. Jr., 1997. GeneDoc, a tool for editing and annotating multiple sequence alignments. Distributed by the author. http://www.cris.com/~ketchup/genedoc.shtml.

Comparison of the human and *C. intestinalis* U5snRNA sequences indicate that the *C. intestinalis* sequences contain essential U5snRNA elements such as the consensus loop1 and Sm-protein binding site sequences as well as conserved structural elements such as helix1c. Thus there is high confidence that the *C. intestinalis* U5snRNA candidate sequences are genuine. Similar analysis of the candidates from *G. lamblia*, *Ent. histolytica* and *D. discoideum*, with the human U5 sequence (Figure 2.4A-C) also show considerable conserved secondary-structure between these RNAs although the sequences are very different. The candidate sequences from *Ent. histolytica* and *D. discoideum* contain the consensus loop1 and Sm-protein binding site sequences but the *D. discoideum* U5snRNA has only 5 base-pairs in helic1c as opposed to the consensus 8 base-pairs found in other eukaryotes.

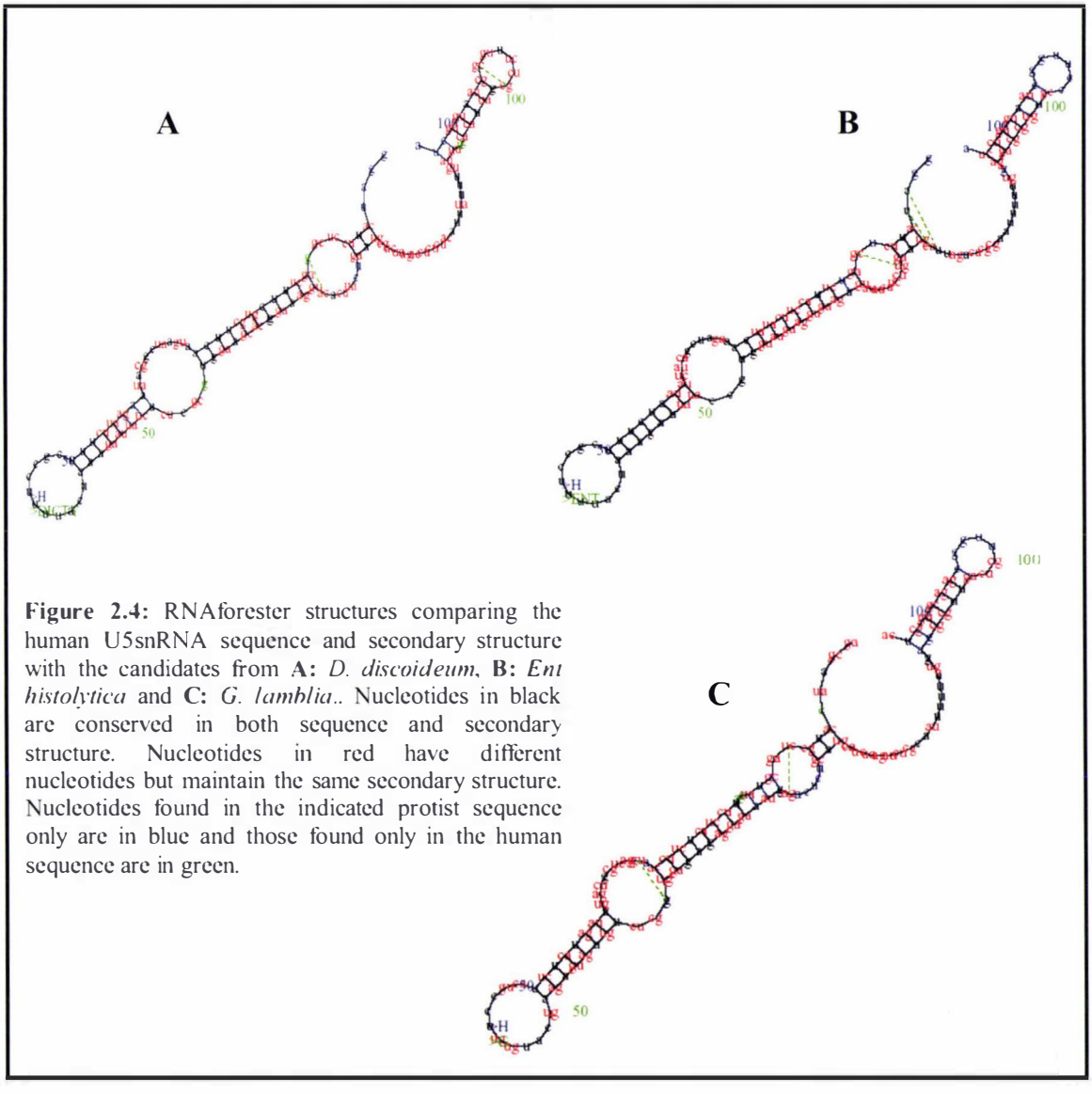


Figure 2.4: RNAforester structures comparing the human U5snRNA sequence and secondary structure with the candidates from **A:** *D. discoideum*, **B:** *Ent histolytica* and **C:** *G. lamblia*.. Nucleotides in black are conserved in both sequence and secondary structure. Nucleotides in red have different nucleotides but maintain the same secondary structure. Nucleotides found in the indicated protist sequence only are in blue and those found only in the human sequence are in green.

The *G. lamblia* U5snRNA contains some differences in its loop1 sequence (AUCCUGGUACG as opposed to the consensus CGCCUUUUACU) but maintains other essential features such as the Sm-protein binding site and helix1c of 8 base-pairs, as well as the optional PSF-binding site. The loop1 sequence has an important function during splicing and binds to the 5' splice site of the pre-mRNA (McConnell et al. 2003). The one intron published for *G. lamblia* (Nixon et al. 2002) is short (35bp) containing a canonical 3' splice site (AG) and a non-canonical 5' splice site (CT). Thus the differences in the loop1 sequence found in the *G. lamblia* U5snRNA candidate may be a reflection of subtle differences in the splicing mechanism of this organism. In yeasts there is some variability permitted in loop1 sequences (O'Keefe 2002) indicating that the highly conserved sequence found in most organisms may represent the optimal sequence for efficient splicing of the full complement of pre-mRNAs in a cell. The low (to date: 1) number of introns found in *G. lamblia* may have allowed the loop1 sequence of its U5snRNA to evolve differently from U5snRNAs found in other organisms.

RNaseP RNA is a difficult ncRNA to draw using RNA-drawing software, thus RNAforester results (Figure 2.5C) look somewhat different to the consensus eukaryotic RNaseP structures (Figures 2.5A and B). RNAforester results comparing the *G. lamblia* candidate RNaseP RNA with the human RNaseP RNA (Figure 2.5C) indicate that there is a high level of secondary-structure conservation although their nucleotide sequences are quite different (positions shown in red). RNaseP sequences from crown eukaryotes contain a distinctive P3-region containing a large bulge in the middle of both the top and bottom helical strands (Figure 2.5), whereas the P3-region in bacteria is shorter and does not contain this bulge. The RNaseP candidates found in *G. lamblia*, *Ent. histolytica* and *Ecz. cuniculi* contain the "bacterial-like" P3 helix region instead of the expected crown-eukaryotic-like P3-region. Some archaeal RNaseP RNAs also contain the bacterial-type of P3-region (Harris et al. 2001) indicating that either the bacterial and crown-eukaryotic-type of P3-region, or an intermediate-type, may have been present in the ancestral RNaseP. Biochemical analysis including protein-binding studies will be required to confirm that the candidate RNaseP sequences found during this study are genuine but this lies outside the scope of this project. A summary of the U5snRNA and RNaseP candidates found with RNAmotif is shown in Table 2.2.

RNAmotif Results	U5snRNA	RNaseP RNA
Ecz. cuniculi	K	+
P. falciparum	K	-
G. lamblia	+	+
Ent. histolytica	+	+
D. discoideum	+	-
C. intestinalis	+	-

Table 2.2: Summary of results from the U5snRNA and RNaseP descriptors used with the RNAmotif software.
K Sequence was known prior to this study.
\+ Candidate sequence was recovered.
\- No candidate sequence was recovered.

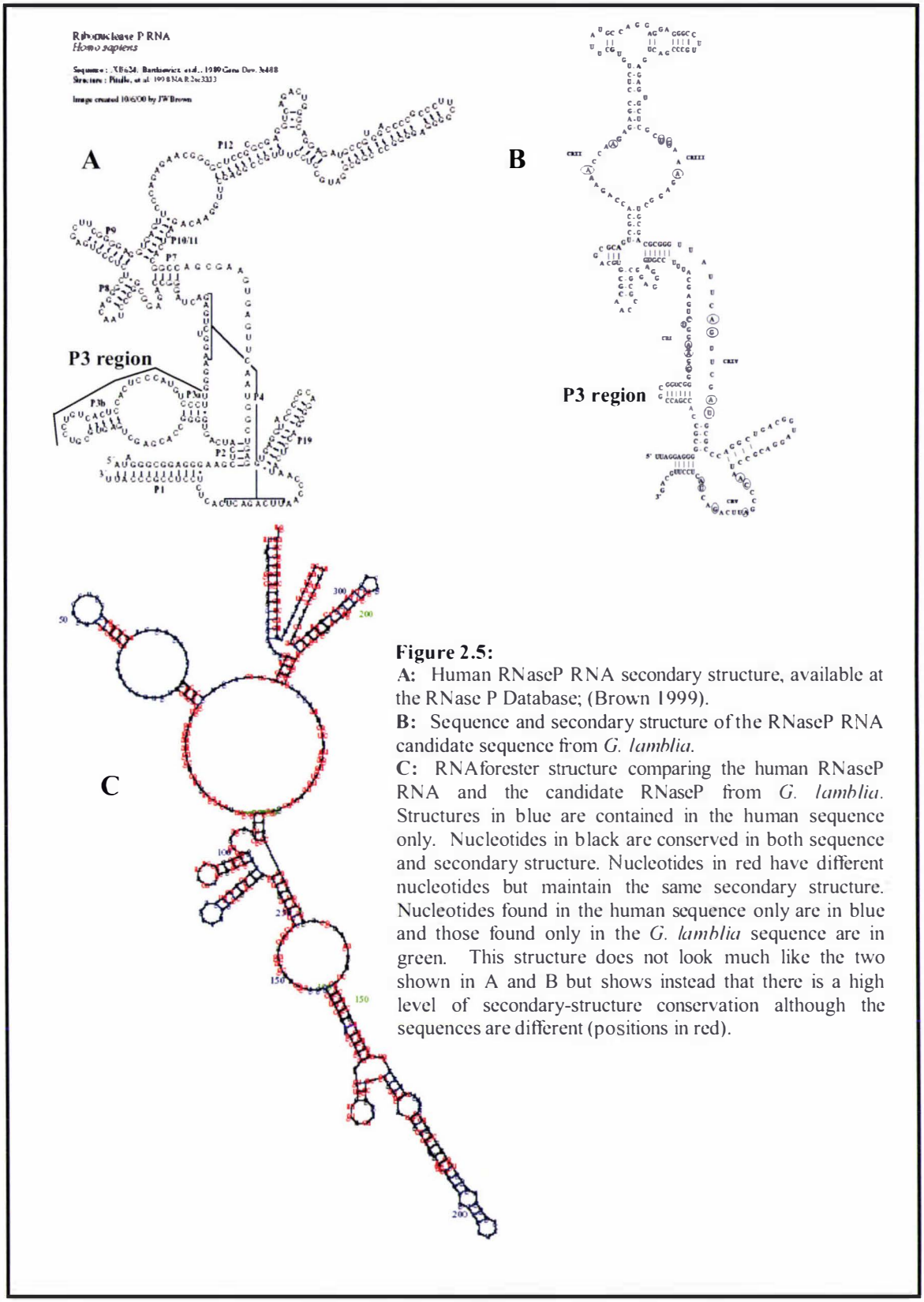


Figure 2.5:
A: Human RNaseP RNA secondary structure, available at the RNase P Database; (Brown 1999).
B: Sequence and secondary structure of the RNaseP RNA candidate sequence from *G. lamblia*.
C: RNAforester structure comparing the human RNaseP RNA and the candidate RNaseP from *G. lamblia*. Structures in blue are contained in the human sequence only. Nucleotides in black are conserved in both sequence and secondary structure. Nucleotides in red have different nucleotides but maintain the same secondary structure. Nucleotides found in the human sequence only are in blue and those found only in the *G. lamblia* sequence are in green. This structure does not look much like the two shown in A and B but shows instead that there is a high level of secondary-structure conservation although the sequences are different (positions in red).

2.2.3: Sequence with Secondary-structure Annotation

RSEARCH

RSEARCH (Klein and Eddy 2003) is a program that takes a single RNA sequence and its known secondary-structure and utilises a local alignment algorithm for database searches. It also reports back the statistical confidence and the structural alignment for each hit. The power of this program comes when only a single member of an ncRNA family is known. Uncertainty in secondary-structure can have a significant effect on the results, thus the secondary-structure of the ncRNA family must be well established. This program is slow but like RNAmotif, is aided by parallel-implementation using a clustered computing environment. RSEARCH offers a "BLAST-like" approach to ncRNA-searching with a simple input of a single sequence and its secondary-structure in "Stockholm" format (Figure 2.1D page 20).

This program has the advantage that neither a comprehensive model of the desired ncRNA, nor an RNA sequence alignment need be constructed, and that an independent statistical confidence score can be calculated. The disadvantage to this approach is that, like BLAST, a sequence from one species may not have enough sequence and secondary-structure similarity to find a candidate sequence in a distant genome (however, this has not yet been fully tested).

RSEARCH (version 1.1) was tested with some U5snRNA sequences (human, *S. pombe* and *Ecz. cuniculi*) against the *Ecz. cuniculi* and *G. lamblia* genomes, both of which were appended with the human U5snRNA sequence. The human and *S. pombe* U5snRNAs successfully recovered the *Ecz. cuniculi* U5snRNA from its genome but the *S. pombe* query did so only just. None of the above sequences recovered any viable U5snRNA candidates from the *G. lamblia* genome. Some *G. lamblia* sequences were recovered but all either had large gaps in the alignment with the query sequence/structures and did not contain essential features such as the Sm-protein binding site or anything resembling a loop1 consensus sequence. The candidate U5snRNA recovered from the *G. lamblia* genome with RNAmotif was not recovered with RSEARCH. Similarly, a search with the *G. lamblia* U5snRNA candidate sequence found by RNAmotif, did not recover either the human nor *Ecz. cuniculi* U5snRNA sequences during testing. Thus the RSEARCH testing could not confirm the *G. lamblia* U5snRNA candidate but could not offer any alternative candidates.

Some RNaseP RNA sequences were used during the original testing phase of RSEARCH (Klein and Eddy 2003) on a number of archaeal genomes. This testing showed that RSEARCH outperformed sequence similarity programs (such as BLAST and FASTA) in recovering the appropriate RNaseP RNA sequences. Searches of the *Ecz. cuniculi* and *G. lamblia* genomes (appended with the human RNaseP sequence) and a test database consisting of some eukaryotic, archaeal and bacterial RNaseP sequences, with the human RNaseP sequence were attempted but ran

into processing problems causing the program to prematurely abort. The nature of this problem is not known at this stage. This effect was not seen during the U5snRNA tests and time constraints have precluded attempting to solve the programming problems. Once the problems have been solved then it will be interesting to see if the RNaseP candidate sequences recovered with RNAmotif can also be recovered with RSEARCH and if RNaseP candidates can be recovered from other eukaryotic genomes with this method.

2.3: Concluding remarks

There is still much work to be done in the area of ncRNA genomics. RNA-detection software still requires considerable development to generate robust searching techniques and for the large part, requires a sizable amount of computer processing power. The current situation in RNA analysis is reminiscent of the early days of protein sequence analysis. As Eddy (2002) commented, it is not too long ago that the few programs available for sequence searches were too impractical and expensive to run on early computers. However, since these early days fast heuristic tools such as BLAST enabled the wider community access to protein computational analysis. In order to allow effective evolutionary and functional analysis of ncRNAs, it is now time for equivalent tools to be developed and expanded for RNA analysis in this rapidly expanding area of genomics.

Each of the approaches taken by the present ncRNA-search software may be better for finding specific types of ncRNA. The advantages of the annotated secondary-structure approach are that it is comparatively quick to set up the search, will run on a single processor (parallel-processing is not required but could, in theory be applied if necessary), and that the annotation can represent pseudoknots and other RNA-characteristic features. The model is constructed automatically from the sequence alignment annotated with bracket notation (the most common format for representing secondary-structure). The downside to this approach is that it relies on the ncRNA being alignable, with no gaps in any of the helices and little variability in the consensus secondary-structure used to annotate the alignment. Areas of secondary-structure variability can cause a large number of gaps in the alignment which leads to an inaccurate search model. This approach may be more appropriate for finding closely related ncRNA families where sequence and/or secondary-structure has little variability.

The "biological-modelling" approach taken by RNAmotif requires much more background work in constructing a descriptor using information largely taken from the literature. However, this approach has had good success when searching basal eukaryotic genomes whereas the "secondary-structure annotated alignment" approach did not. At present this was the most successful approach with ncRNAs that contain variable sequence and secondary-structure. The current drawback to the

RNAmotif software is its lack of any statistical significance calculation attached to any returned sequences. With a sequence and/or secondary-structure input, a value indicating the amount of similarity between the query and the recovered sequence can be calculated but this is not an easy issue when the input is based on a 'biological' model (as it is with RNAmotif). A calculation of some value of statistical significance associated with RNAmotif results should be possible. Future development of RNAmotif may include a statistical value calculation based on randomised representative RNA, randomised sections of a genome-database or even randomised sections of the descriptor itself. Another future option may be to use a algebraic dynamic programming (ADP) approach (Meyer and Giegerich 2002) to compare (in a similar way to RNAmotif) two ncRNAs. This technique requires presently, the ncRNA model to be described in a grammar based on the "Haskell" programming language; a specialist language not for the faint-hearted. It may be possible to design some type of "translation" software to enable an ncRNA to be described in both the RNAmotif and ADP grammars enabling a non-programmer to use these potentially powerful methods.

RSEARCH offers a faster approach for ncRNA searches (in a BLAST-like way) and would be very useful as a first step in finding a particular ncRNA from a sequenced genome, before alignments and/or descriptors are created. This software has only been recently published and is likely to become more efficient in later versions.

Although there are now some ncRNA-finding programs available, at present all require a measure of computing skills not normally associated with molecular biology. ncRNA models have to be translated in a number of machine-readable formats that are often not very human-readable and the programs tested in this study required the knowledge of Linux, parallel-processing and some programming language skills. It is hoped that future versions of ncRNA-finding and analysis software will take a user-friendly approach, incorporating common input/output languages and graphical interfaces. In this way, the tools that will be of great use to ncRNA researchers can actually be used by them.

Journal of Integrative Bioinformatics, 0001, 2003.
Online Journal: http://journal.imbio.de/index.php?paper_id=6

Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif*

Lesley J. Collins[1*], Thomas J. Macke[2] and David Penny[1]

[1] Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 11222, Palmerston North, New Zealand.

[2] Department of Molecular Biology,
The Scripps Research Institute, La Jolla, CA 92037, USA

[*] Corresponding author: L.J.Collins@massey.ac.nz

Summary

Non-coding RNAs (ncRNAs) contain both characteristic secondary-structure and short sequence motifs. However, "complex" ncRNAs (RNA bound to proteins in ribonucleoprotein complexes) can be hard to identify in genomic sequence data. Programs able to search for ncRNAs were previously limited to ncRNA molecules that either align very well or have highly conserved secondary-structure. The RNAmotif program uses additional information to find ncRNA gene candidates through the design of an appropriate "descriptor" to model sequence motifs, secondary-structure and protein/RNA binding information. This enables searches of those ncRNAs that contain variable secondary-structure and limited sequence motif information. Applying the biologically-based concept of "positive and negative controls" to the RNAmotif search technique, we can now go beyond the testing phase to successfully search real genomes, complete with their background noise and related molecules. Descriptors are designed for two "complex" ncRNAs, the U5snRNA (from the spliceosome) and RNaseP RNA, which successfully uncover these sequences from some eukaryotic genomes. We include explanations about the construction of the input "descriptors" from known biological information, to allow searches for other ncRNAs. RNAmotif maximizes the input of biological knowledge into a search for an ncRNA gene and now allows the investigation of some of the hardest-to-find, yet important, genes in some very interesting eukaryotic organisms.

1 Introduction

Non-coding RNAs (ncRNAs) make transcripts that function as RNA, rather than encoding proteins, the best-known examples being ribosomal-RNA (rRNA) and transfer-RNA (tRNA) [1]. Many ncRNAs form part of RNA-protein complexes (Ribonucleoproteins, RNPs) and play roles in cellular processes such as RNA processing and splicing. Some ncRNAs have catalytic functions e.g. RNaseP RNA, whereas others serve key structural roles in ribonucleoprotein complexes e.g. snRNAs [2]. Searching databases for homologues based on sequence similarity is only useful for the larger, more slowly evolving ncRNAs (such as ribosomal RNAs) and is less reliable for other ncRNAs. Sequence similarity methods may fail to find ncRNA gene candidates when there is a large evolutionary distance between the query species and the target genome being searched [3].

---

[*] Formatting has been changed from the published manuscript to enable integration into this thesis.

In the past, programs such as RNAMOT [4] and PatScan [5] were developed to define and search for RNA structures and these led to programs such as tRNAscan-SE [6], which were designed to look for specific kinds of structural RNA. Recent ncRNA search techniques (e.g. ERPIN [7] and RSEARCH [8]) take both sequence and structure into account but are unable to model small sequence and secondary-structure motifs that correspond to protein or RNA binding sites in the ncRNA. These programs rely on sequence and secondary-structure alignment, either between multiple ncRNA sequences (ERPIN) or between query and subject sequences (RSEARCH). Alignment is difficult for such ncRNAs as there is often little sequence homology between distantly related species. Although these RNAs have both a conserved secondary-structure and some highly conserved sequence motifs, they also contain some secondary-structure differences [9, 10].

The RNAmotif program [11] was developed from RNAMOT [4] and uses an expanded syntax for describing motifs along with an implementation of nearest-neighbor rules and other schemes for ranking hits. RNAmotif has previously been used to find two groups of ncRNAs; tRNA [12] and the Iron Response Element (IRE) [11, 13], both of which contain highly conserved secondary-structures. This program uses a user-defined "descriptor" as input, modelling allowable secondary-structure and sequence motifs. It also has a scoring section that assesses the different features of the match [11]. A criticism of RNAmotif software is the lack of any value of statistical significant attached to any returned sequences. This value can be easily calculated based on sequence and/or secondary-structure similarity but is difficult to compute based on a biologically-derived model. To overcome this hurdle, and until more sophisticated RNA-model comparison techniques become available, we introduce "positive and negative controls", a fundamental concept of molecular biology, to provide significance to the RNAmotif results. First a test database is constructed consisting of positive controls (sequences we expect to be returned with a descriptor) and negative controls (sequences we do not expect to be returned). A second testing phase tests the performance of a descriptor against genomic background noise and a third testing phase was to search a genome for its known ncRNA sequence, testing a descriptor against similar ncRNAs found in that genome.

This study also shows how the use of a user-defined scoring section, results filtering and parallel implementation reduce the problems associated with searches of both crown (animal, yeast and plants) and basal (protist) eukaryotic genomes. This resulted in the identification of candidates for both the U5snRNA and the RNaseP RNA, from *Giardia lamblia, Entamoeba histolytica* (*Ent. histolytica*) and the microsporidian, *Encephalitozoon cuniculi (*abbreviated here as *Ecz. cuniculi* to avoid confusion with *Ent. histolytica*), and the U5snRNAs from *Dictyostelium discoideum* and *Ciona intestinalis.*

The U5 snRNA molecule is part of the U5 snRNP ribonucleoprotein complex that is involved in the splicing of nuclear pre-messenger RNA [14]. U5snRNA has already been identified from a number of completely sequenced genomes including *Ecz. cuniculi* and *Plasmodium falciparum* making them ideal test subjects for this study. After the testing stage, the U5 descriptors were used to search other small eukaryotic genomes such as *G. lamblia*[15], *Dictyostelium discoideum,* [16], *Entamoeba histolytica* [17] and *Ciona intestinalis* [18].

The other ncRNA investigated here is Ribonuclease P (RNaseP) RNA, part of the ribonucleoprotein complex that cleaves 5'-leader sequences from precursor-tRNA to leave a mature tRNA molecule [19]. Apart from some short nucleotide motif sequences, eukaryotic RNaseP RNAs have little nucleotide sequence homology (except between

closely related species) making this gene difficult to find in more distant species. RNaseP RNA contains features that make it more challenging to write an effective descriptor. We used an improved version of RNAmotif implementing parallel processing to search for the RNaseP RNA in the genomes mentioned above. A common criticism of software descriptions is often there is not enough detail on parameter-tuning to enable a researcher in the biological field to effectively use the program [20]. To this end, we provide a comprehensive explanation of the construction of the U5snRNA and RNaseP descriptors from known biological information (i.e. RNA and protein binding sites), to enable researchers in the ncRNA field to design descriptors for their molecules of choice.

2 Methods

RNAmotif [11] is written in ANSI C and available as source code via 'anonymous ftp' from (ftp.scripps.edu/pub/macke/rnamotif-version.tar.gz where "version" is the version number, currently 3.0.0). RNAmotif supports parallel searches via an MPI based driver, called mrnamotif, which is included in the RNAmotif distribution. Parallel processing was done on the Helix Cluster, a distributed-memory Beowulf cluster with 65 nodes (128 processors) running the Linux RedHat (version 7.3) operating system and communicating with the MPI protocol (http://helix.massey.ac.nz). All nodes used in testing and searching with RNAmotif had AMD Athlon MP-2100 processors running at 1733.335 MHz. A Perl script used to split large databases into smaller units suitable for parallel processing is available from the corresponding author upon request.

The program "Getbest" (available from the corresponding author upon request) was incorporated into the RNAmotif searching technique filtering the results from each worker node to give a condensed results file. Getbest works by selecting only the best solution found at each position of the sequence being searched, in this case, the position with the lowest free energy (ΔG). As expected thermodynamic stabilities improve with length [21], the sequence with the lowest free energy will tend to have the longest sequence which is retained using Getbest.

2.1 Sequences and Genomes

U5snRNA sequences were downloaded from the Rfam database [22] and the databases at NCBI (http://www.ncbi.nlm.nih.gov/). The genomes of *Encephalitozoon cuniculi* [23], (AL391737 and AL590442-AL590451), *Ciona intestinalis* [18] (AABS00000000) and *Pyrococcus abyssi* (AL096836) were also downloaded from NCBI. The *Plasmodium falciparum* genome was downloaded from PlasmoDB [24, 25] http://plamodb.org). *Dictyostelium discoideum* (soil-living amoeba) [16] preliminary sequence data was obtained from The Wellcome Trust Sanger Institute (http://www.sanger.ac.uk). The *Entamoeba histolytica* genome sequencing data [17] was produced by the Sanger Institute Pathogen Sequencing Unit at the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/E-histolytica).

Early releases of the *Giardia lamblia* genome (WB strain, clone C6) was kindly provided by the *Giardia lamblia* Genome Project [15] is based at the Marine Biological Laboratory at Woods Hole, Massachusetts, U.S.A. (http://jbpc.mbl.edu/Giardia-HTML/index2.html). This "Whole Genome Shotgun" sequencing project has now been completed and deposited at DDBJ/EMBL/GenBank under the project accession AACB01000000. RNaseP RNA sequences were downloaded from the RNaseP Database ([26], http://www.mbio.ncsu.edu/RNaseP/main.html) and NCBI. The RNaseP eukaryotic consensus secondary-structure was taken from Frank et al. 2000 [27].

2.2 RNAmotif Descriptors

2.2.1 Descriptor Design – U5snRNA

A descriptor is read by the RNAmotif program from the 5' end of the model to the 3' end, so both 'sides' of a helix must be represented in the code. For example, *h5 (tag = 'helix1', len = 4)* opens a helix of 4 base-pairs, and *h3(tag = 'helix1')* closes the helix. Single-stranded regions are represented by *ss(tag = 'single_stranded 1')*. Both helices and single-stranded regions may contain length (*minlen* - minimum length; *maxlen* - maximum length), sequence, mismatch and mispairing parameters to allow for the small differences that are found in ncRNAs from different species. For all descriptors in this study, parameters were set to allow G:U pairing, and the folding of the structure to have a user-defined maximum energy level (emax).

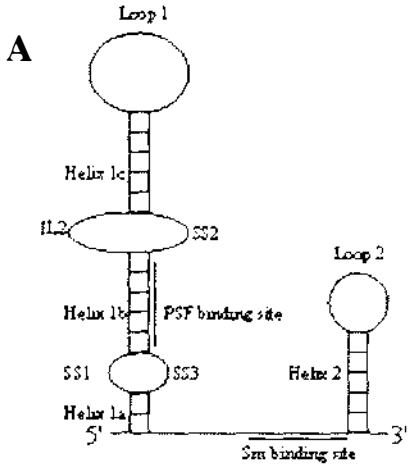
The scoring section was designed to allow the user to see at a glance the different type of motifs that have been added together to produce the final score. The absence of a motif recorded a '*' in the motif position in a string of motif characters. The presence of a motif changed this '*' into a letter designating the selected motif. This motif scoring visualization is useful when looking for an ncRNA that contains some, but not other, elements, yet can still be a legitimate candidate.

Three descriptors were constructed for the U5snRNA based on features found in different combinations of species. The U5snRNA consensus secondary-structure contains a Sm protein-binding site, a highly conserved loop of eleven nucleotides next to a helix of 6-8 base-pairs [9, 14] (Figure1A). Features that are not present in U5snRNA sequences from some species include a second helix-loop structure and a PSF/p54^{nrb} protein-binding site. Figure 1 shows the U5_A descriptor and lists the differences between that descriptor and the other two U5snRNA descriptors used in this study, U5_B and U5_C. Secondary structure regions either absent or extremely variable between species (e.g. Helix 1a) were not included in the descriptors or converted to single-stranded regions. Mispairing events (i.e. *mispair = 1*) were permitted in some of the helices to improve the range of sequences recovered during testing, however including these events increased processing time. Highly variable single-stranded regions such as "IL2" were given a wide length range (in this case, between 3 and 18 nucleotides to allow for an extra helix that is present in some yeast species).

Helix1c, Loop1 and the Sm-binding site are important biological features of the U5snRNA [9]. Helix1c in some species has an internal mispairing event (a G:A pairing) which was modelled differently in descriptors U5_A and U5_B. U5_A allowed a mispairing on either end of the helix as well as internally (mispair = 1, ends = 'mm') whereas U5_B used stricter settings with mispairing only permitted on the distal (farthermost from the loop) end (mispair = 1, ends = 'mp'). Loop1 consists of eleven nucleotides containing a highly conserved sequence motif [28]. Loop1 was modelled differently in the three descriptors as shown in Figure 1B to allow for differences from the consensus model shown in the few basal eukaryotic U5snRNAs available (e.g. *P. falciparum* and *L. collosoma*). U5_A allowed for proximal mispairing whereas the U5_B descriptor was again stricter. The sequence within loop1 was scored the same between U5_A and U5_B but an alternative "less-strict" scoring scheme was used in U5_C. The Sm-protein binding site provides an example of how a sequence can be used for selection (in the "descr" section) then have viable sequence alternatives scored against in the scoring section. The Sm-binding sequence was also anchored to the end of the single-stranded region (using $). This greatly improved processing speed (non-

anchored sequence, "ayuuuung" = 4 minutes, anchored sequence "ayuuuung$" = 33 seconds) and lowered the number of redundant hits and the output file-size.

Optional sequence motifs (e.g. The PSF-binding site on the 3' side of helix 1b) can be scored against in the scoring section but not included in the "descr" section as this would make this motif inclusion compulsory.



B

```
# U5snRNA Descriptor (U5_A)
#Maximum score is 6
params
    wc += gu;
    emax = -5;
descr
  h5(tag='1b', minlen=3, maxlen=10,mispair = 1)
  ss(tag='IL2',minlen=3, maxlen=18)
  h5(tag='1c',minlen=6, maxlen=9, mispair=1, ends='mm')
3 { ss(tag='loop1', minlen=7,maxlen=12)
  }h3(tag='1c')
  ss(tag='ss2',minlen=0, maxlen=9)
  h3(tag='1b')
  ss(tag='Sm',minlen=15, maxlen= 34,mismatch = 1,seq = "ayuuuung$")
  h5(tag='2',minlen = 0,maxlen=9)
  ss(tag='loop2', minlen=3, maxlen=7)
  h3(tag='2')
score
 {
 s1 = 0.0;
 x = '', y = '', t = '', r = '', q = '';
 if(length(h5(tag='1c')) == 8)  {s1 += 1.0; x = 'A';}
 if(length(ss(tag='loop1')) == 11)  {s1 += 1.0; y = 'B';}
4 if(ss(tag='loop1') =~ "ggccuuuuacu") {s1 += 2.0; t = 'a';}
  else if(ss(tag='loop1') =~ "ggccuuuuac") {s1 += 1.5; t = 'b';}
  else if(ss(tag='loop1') =~ "ccuuuuac") {s1 += 0.5; t = 'c';}
  else if(ss(tag='loop1') =~ "cuuuu") {s1 += 0.5; t = 'd';}
 if(h3(tag='1b') =~ "gragaggaac") {s1 += 1.0; r = 'a';}
 else if(h3(tag='1b') =~ "gaggaa") {s1 += 1.0; r = 'b';}
 else if(h3(tag='1b') =~ "ggaga") {s1 += 0.5; r = 'c';}
 if(ss(tag='Sm') =~ "aauuuung$") {s1 += 1.0; q = 'a';}
 else if(ss(tag='Sm') =~ "ayuuuung$") {s1 += 0.5; q = 'b';}
 else if(ss(tag='Sm') =~ "ayuuuuung$") {s1 += 0.6; q = 'c';}
 else if(ss(tag='Sm') =~ "ayuuuuuung$") {s1 += 0.5; q = 'd';}
 if(s1 < 2.5) REJECT;
 s2 = sfn( h5[1], h3[NSE] );
 if( s2 > emax )
   #REJECT;
   s1 -=0.01;
 SCORE = sprintf( "%s,%s,%s,%s,%s,%s,%1.2f,%1.3f", x, y, t, r, q, s1, s2 );
}
```

Figure 1: U5 snRNA model and RNAmotif descriptor. **A**: Consensus model of the U5snRNA secondary-structure based on the model of Peng et al. 2002. Loop2 is not present in the U5snRNAs of some species. **B**: U5_A descriptor used to search for potential U5snRNA genes. Other U5snRNA descriptors were based on this descriptor but contained the following differences:

U5_B descriptor

Point 1, h5(tag='1c',minlen=6, maxlen=8, mispair=1, ends='mp');

Point 2, ss(tag='loop1', len = 11) ;

Point 3, deleting these two lines thus not scoring on the length of helix 1c and loop1.

U5_C descriptor

Point 4, the sequences scored upon were as follows:

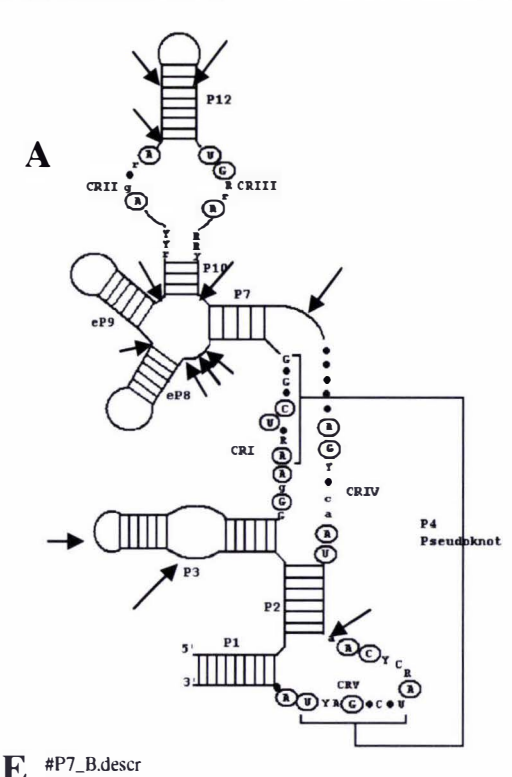
```
if(ss(tag='loop1') =~ "ygccuuuuacu") {s1 += 1.0;t = 'a';}
else if(ss(tag='loop1') =~ "ccuuuuac") {s1 += 0.7;t = 'b';}
else if(ss(tag='loop1') =~ "ccuuuu") {s1 += 0.5;t = 'c';}
```

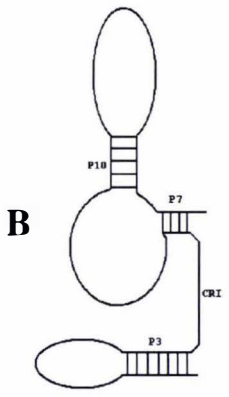
The scoring for option d is the same as in U5_A.

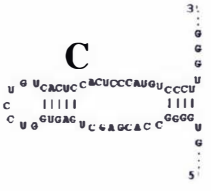
In the scoring section for each descriptor the individual letters i.e. 'A', 'B', 'x', 'y', 't', 'r' and 'q' represent the option scored for each sub-element allowing the user to see in the output the makeup of the overall score.

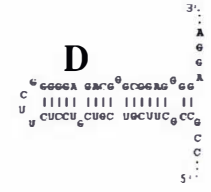
2.2.2 Descriptor Design – RNase P

The eukaryotic RNaseP RNA has a generally conserved secondary-structure (Figure 2A) with parts of the structure highly conserved while other parts contain variability in both sequence and secondary-structure [10, 29]. Thus the RNaseP RNA descriptor had to take into account both conserved and variable features.









E

```
#P7_B.descr
parms
wc += gu;
emax = -15;
descr
h5(tag='P3', minlen = 3, maxlen = 12)
ss(tag='P3loop', minlen = 3, maxlen = 45)
h3(tag='P3')
ss(tag='pre-CRI', minlen = 0, maxlen = 3)
ss(tag='CRI', minlen = 12, maxlen = 25, seq = "^gnaannuc")
h5(tag='P7', minlen = 3, maxlen = 6)
ss(minlen = 2, maxlen=50)
h5(tag='P10', minlen = 5, maxlen = 10, mispair = 1)
ss(tag = 'CRII', minlen = 40, maxlen = 150, seq = "ugnna")
h3(tag = 'P10')
ss(minlen = 0, maxlen = 5)
h3(tag = 'P7')
```

```
score
{
s1 = 0.0;
t = '*';
if(ss(tag='CRI') =~ "ggaarnucngng") {s1 += 2.5; t = 'a';}
else if(ss(tag='CRI') =~ "gaarnucngng") {s1 += 2.0; t = 'b';}
else if(ss(tag='CRI') =~ "ggaaanucc") {s1 += 1.7;t = 'c';}
else if(ss(tag='CRI') =~ "ggaannuc") {s1 += 1.7;t = 'd';}
else if(ss(tag='CRI') =~ "gnaannuc") {s1 += 1.0;t = 'e';}
s2 = efn( h5[1], h3[NSE] );
if( s2 > emax ) REJECT;
if(s1 < 1.7) REJECT;
SCORE = sprintf( '%s,%1.2f,%1.3f',t,s1,s2 );
}
```

Figure 2: A. Consensus secondary-structure model for the eukaryotic RNaseP RNA taken from Frank et al. (2000). Named nucleotides are conserved within Eukaryotes and circled nucleotides are conserved within all three kingdoms. Arrows indicate where optional helices may be inserted showing areas of variability within the RNaseP RNA secondary-structure model. **B.** General secondary-structure of the P7 descriptors. **C.** Example of a "eukaryotic-type" P3-region from the human RNaseP RNA. **D.** Example of a "bacterial-type" P3-region from the *Escherichia coli* RNaseP RNA. **E.** P7_B descriptor. P7_A descriptor has the two lines at point 2 replaced by a single line: ss(tag='CRI', minlen = 12, maxlen = 25, seq = "gnaannuc"). The P7_C descriptor has an additional helix added to simulate the conventional P3-region structure of eukaryotic RNaseP RNA. The lines at Point 1 have been replaced by the following lines:

```
h5(tag='P3', minlen = 3, maxlen = 12)
ss(tag='P3loop1', minlen = 3, maxlen = 20)
h5(tag = 'P3b', minlen = 3, maxlen = 12)
ss(tag='P3loop2', minlen = 3, maxlen = 10)
h3(tag = 'P3b')
ss(tag = 'P3loop3', minlen = 3, maxlen = 20)
h3(tag='P3')
```

It was found that descriptors designed for the full eukaryotic secondary-structure were computationally-prohibitive, requiring weeks to search the simplest databases and often failed to return accurate, if any, results (data not shown). Using a descriptor for part of the RNaseP RNA structure allowed sequences to be returned that were then

analyzed further for essential downstream elements (e.g. CRV pseudoknot pairing). Descriptors were designed (Figure 2E) for the P3-CRI-P7-P10 section of the complete secondary-structure (Figure 2B) and designated P7_A, P7_B and P7_C.

The P3-region consists in archaea and bacteria of one helix-loop structure, but in eukaryotes has a large bulge in the middle of both the 5' and 3' strands of the helix forming two stacked helices (shown in Figures 2C and D) thus it was decided to allow for both types of structures in the RNaseP descriptors. Descriptors P7_A and P7_B code for a single helix-loop structure but allow a large loop length to compensate for any second helix. Descriptor P7_C codes for a second helix with a minimum length of 3 base-pairs and single-stranded regions on either side. Adding this second helix both dramatically increased the processing time (from 3 minutes to 35mins) and lowered the specificity of the descriptor; however, most folding energies were improved with P7_C for each sequence region recovered. Designating the second helix optional (i.e. setting the minimum length of helix 1b to 0) recovered those sequences not detected with the P7_C descriptor, but increased the processing time tenfold. The lesson learnt here was to keep the number of helices to the minimum required for the desired species specificity.

The CRI region is the most highly conserved sequence motif in the RNaseP RNA and is critical for selection as an RNaseP RNA gene candidate. The CRI-minimal sequence motif (found in all RNaseP RNA sequences from all three kingdoms) was made mandatory for selection by including it in the CRI-motif parameter settings. This is the only sequence motif scored in the RNaseP descriptors as other CR-regions in the descriptor area (CR-II and CR-III) were too general to be useful. The CRI-minimal sequence motif was anchored in descriptor P7_B by adding a separate single-stranded region (*tag* – *'preCRI'*) before this sub-element (*tag* *'CRI'*).

The P10 region contains allowances for the characteristic mispairing event that occurs in bacteria and some archaea but does not occur in sequences from crown eukaryotes (the situation in basal eukaryotes is unknown). The CRII-P12-CRIII single-stranded region (tag = 'CRII') is highly variable in length, and the number and position of helices between species and kingdoms [27]. This region was set to a single-stranded loop with a large range in length (between 40 and 150 nucleotides) to allow for this variability.

2.2.3 Descriptor Testing

Known U5snRNA and RNase P sequences were downloaded from the Rfam database [22] and from the NCBI databases. Test databases, TestDatabaseA (Table1) and Pdatabase (Table 2) were constructed containing sequences that were expected to be returned with the descriptors (positive controls), and sequences from other ncRNAs that were not expected to be returned (negative controls). Descriptors were tested against test databases to understand the performance of different variations of the descriptor.

Other ncRNAs contain some of the motifs found in our descriptors and thus representative sequences from all these ncRNAs were included in these test databases. It is unnecessary to construct overly large test databases (such as downloading the complete Rfam database) as long as appropriate positive and negative controls, usually determined with the biological knowledge of the ncRNA, have been included. If appropriate controls cannot be selected for a desired ncRNA then the Rfam database, although large, will be a reasonable alternative. Scoring cutoffs for each descriptor are selected after the analysis of positive and negative control results and determine the selectivity of the descriptor.

In order to test the level of background from other ncRNAs, three U5snRNA sequences, human (M23822) *Schizosaccharomyces pombe* (X15504) and *Caenorhabditis elegans* (Z69665) and the human RNaseP RNA (X15624) were randomly inserted into the *Pyrococcus abyssi* genome. To date there have been no U5snRNA sequences described for any archaeal species so it was expected that the inserted human, *C. elegans* and *S. pombe* sequences would be recovered with higher scores than any 'native' *P. abyssi* sequences. *P. abyssi* contains its own RNaseP but as the RNaseP descriptor was designed primarily for eukaryotic RNaseP, it was expected that the human RNaseP would be recovered with higher scores than the native *P. abyssi* RNaseP.

Table1	Species	U5	U1	U2	U4	U6	U11	U12	RNaseP
Plants	Arabidopsis thaliana	X13012	X53175	X06474	X67146	X52527			
	Rice	AC104179	AC025783	AF106845	AB026295	AC079128			
	Pea	X15934	X15926	X15936	X15931				
Fungi	Aspergillus nidulans	AC004395		AL683874		AY136823			
	Saccharomyces cerevisiae	M16510	M17411	M14625	U18778	Z73279			M27035
	Schizosaccharomyces pombe	X15310	X55773	X55772	X15491	M55650			X04013
	Encephalitozoon cuniculi	AL590450				AL590448			
Animal	Human	M77840	AC097369	X59360	X59361	AC114982	X13707	L43846	X15624
	Mouse	M10336	M14121	K00027	M10328	AC116657			
	Zebrafish	AL591593	AL929029	AL92108		AL929029			
	Frog	X06020	K02698	K02457					AF044330
	Caenorhabditis elegans	Z68215	Z81556	X51372	X51382	Z22178			
	Drosophila melanogaster	AC099022	X02136	X04241	K03095	M24605			
Basal	Entosiphon sulcatum	AF09539+		AF095839		AF095841			
	Leptomonas collosoma	AF006632		X56453	AF204671	X79014			
	Plasmodium falciparum	AE014823		AE014841	Z98547				
	Tetrahymena thermophila	X63789	X58845	X63786	X58844	X63790			
	Chlamydomonas reinhardtii	X67000	X70869	X71483		X71486			
	Trypanosoma brucei			X04678	M25777	X13017			

Table 1: Accession numbers of the sequences contained in the test database "TestDatabaseA" used in the evaluation of the U5snRNA descriptors. An empty cell indicates that this sequence was not available for inclusion in this database. Theoretically all U5snRNAs should be returned with the U5snRNA descriptors and thus be positive controls and all non-U5snRNAs be negative controls. For practical reasons only some sequences were selected to be specific positive and negative controls in the testing of the U5snRNA descriptors. Blue indicates a positive control and red indicates a negative control.

Another testing stage is to run a descriptor against a genome in which the ncRNA has already been characterized. This could be done easily for the U5snRNA descriptors as the *Ecz. cuniculi* and *P. falciparum* U5snRNAs have already been identified and are available from the Rfam database (AL590450 and AE014823 respectively). This could not be done, however, for the RNaseP descriptors as to date there have been no RNaseP RNAs characterized in any of the small eukaryotic genomes that were available.

All genomes used in this study were appended to contain positive controls (a file of U5snRNA and RNaseP sequences attached to the end of the genome file), so that if there were no sequences returned with a search, it could be determined that the program had run to completion successfully.

Table 2: Sequences in the "Pdatabase" used for evaluation of the RNaseP descriptors. Included in this database are RNaseP RNA sequences from Eukaryotic, Archaeal and Bacterial species and RNase MRP sequences from Eukaryotic species. Note that RNase MRP has not been found in any Archaeal or bacterial species to date. An empty cell indicates that this sequence was not available for inclusion in this database. Theoretically all RNaseP RNAs should be returned with the RNaseP descriptors; however certain sequences are selected to be specific positive and negative controls. *Blue* indicates a positive control and red indicates a negative control. K - From Kiss and Filipowicz (1992)

Table 2	Species	RNaseP	RNase MRP
Eukaryote	Saccharomyces cerevisiae	M27035	Z14231
	Schizosaccharomyces pombe	X04013	AL009197 (31216-31615)
	Homo sapiens	X15624	
	Mus musculus	L08802	
	Danio rerio	U50408	
	Xenopus laevis	AF044330	
	Drosophila melanogaster	AF434763	
	Arabidopsis thaliana		X65942 (34)
	Bos Taurus (Bovine)		Z25280
	Nicotiana tabacum (Tobacco)		K
	Rattus norvegicus (Rat)		J05014
Archaea	Pyrococcus abyssi	AJ248283	
	Sulfolobus acidocaldarius	L13597	
	Methanobacterium thermoautotrophicum	U42986	
	Methanococcus vannielii	AF192357	
	Archaeoglobus fulgidus	AE000782	
	Halobacterium cutirubum	U42983	
	Aeropyrum pernix	AP000060	
Bacteria	Escherichia coli	V00338	
	Bacillus subtilis	M13175	
	Thermus aquaticus	Z15006	
	Streptomyces lividans	M64552	
	Agrobacterium tumefaciens	M59352	

3 Results

3.1 U5snRNA

3.1.1 Descriptor Testing Results

RNAmotif searches against TestDatabaseA with each U5snRNA descriptor indicated their sensitivity (U5snRNA sequences from which species were returned) and their specificity (which ncRNAs other than the U5snRNA were returned). Results are shown in Table 3. All three descriptors returned all the designated positive controls (U5snRNAs from human, *Ecz. cuniculi*, *P. falciparum* and *Entosiphon sulcatum*). The looser U5_A and U5_C descriptors detected other ncRNAs with scores lower than 3.0, determining this number as the minimum score cutoff for subsequent genomic searches. The tighter U5_B descriptor did not return any other ncRNAs from TestDatabaseA and also failed to detect some of the known basal eukaryotic U5snRNAs.

RNAmotif searches of the 'doctored' *P. abyssi* genome with the U5 descriptors recovered all three inserted U5snRNA sequences above the cutoff score. The *S. pombe* U5snRNA was recovered with a lower score than some native sequences, indicating that with the parameters set in these descriptors, yeast-like U5snRNA sequences could not be expected to be recovered reliably above background noise.

In genomic testing of the U5snRNA descriptors, the known *Ecz. cuniculi* U5snRNA sequence (AL590450) was successfully recovered from its genome as the only top scoring hit with all three descriptors. The known *P. falciparum* U5snRNA (AE014823) was also easily recovered from the *P. falciparum* genome by all three U5 descriptors with the highest score. Recovery of their known U5snRNA sequences from the *Ecz. cuniculi* and *P. falciparum* genomes indicated that it was possible with the U5snRNA descriptors to distinguish between the U5snRNA and other closely related ncRNAs in their own genomes.

Table 3	U5_A	U5_B	U5_C
Test Database Testing			
Processing Time	39.5 seconds	5.9 seconds	31.9 seconds
Output File Size	69 KB	33 KB (40 KB)	56KB
	Highest Scores	**Highest Scores**	**Highest Scores**
Human U5	4.49	1.50 (3.5)	3.99
Drosophila melanogaster U5	3.99	-	2.99
Caenorhabditis elegans U5	4.49	2.49 (4.49)	3.49
Arabidopsis thaliana U5	3.49	2.49 (3.49)	2.49
Oryza sativa U5	4.50	2.50 (4.50)	3.50
Schizosaccharomyces pombe U5	3.49	-	2.69
Tetrahymena themophila U5	4.00	-	3.00
Encephalitozoon cuniculi U5	4.00	2.00 (4.00)	3.50
Plasmodium falciparum U	3.99	2.00 (4.00)	3.00
Pysarum polycephalum U5	4.99	3.00 (5.00)	3.99
Entosiphon sulcatum U5	4.49	2.50 (4.50)	3.49
Human U11	2.99	-	2.99
Mouse U12	2.49	-	2.49
Caenorhabditis elegans U4	2.49	-	2.49
Genome Searches	**Processing Time**	**Processing Time**	**Processing Time**
Ecz. cuniculi genome	26 min 49 sec	5 min 35 sec	26 min 50 sec
"P. abyssi" genome	19 min 43 sec	3 min 12 sec	19 min 46 sec
G. lamblia genome	17 min 33sec	3 min 19 sec	17 min 41 sec
P. falciparum genome	738min 41 sec	132min 14sec	746min 58sec
Ent. histolytica genome	640 min 0 sec	106min 39sec	Not run
D. discoideum genome	1014min 52sec	366min 30sec	Not run
C. intestinalis genome	495min 35sec	83min 14sec	Not run

Table 3: Evaluation results for the U5snRNA descriptors. Representative results from searches of TestDatabaseA are shown although other similar sequences that were in this database were also returned. All descriptors had "emax = 5". Scores below this threshold were not rejected during descriptor testing but had "0.01" subtracted from their overall score for indicative purposes. Descriptors U5_A and U5_C had score cutoffs set to 2.5. U5_B was run with two variations, the first containing scoring for the length of helix1c and loop1, and the second without this scoring. The scoring cutoff was set lower at 1.5 for latter run to compensate for the lessened maximum possible score. Timing differences between the two runs were the same so only one set of timing results are given. The *P. abyssi* genome has a number of U5snRNA sequences inserted for testing purposes. '=' indicates that this sequence was not detected with this descriptor. Species in **bold** were positive control for testing the U5snRNA descriptors.

Other small eukaryotic genomes (*C. intestinalis*, *G. lamblia*, *Ent. histolytica* and *D. discoideum*) were then searched with the U5snRNA descriptors. A prior BLAST search of these genomes with all known U5snRNA sequences returned no significant results. With RNAmotif and the U5 descriptors, five candidate sequences were returned from the *C. intestinalis* genome, all of them contained the consensus loop1 sequence and could be folded into the consensus U5snRNA secondary-structure (Scaffold112:58432-58331;Scaffold71:40056-39955;Scaffold1849:5070-5172;Scaffold1028:15981-5885;Scaffold108:18565-18656). An alignment of these candidate sequences show that they are extremely similar to each other, with only a few nucleotide differences between them. Subsequent analysis showed that the *C. intestinalis* U5snRNA candidates showed similarity to other vertebrate U5snRNAs including human and mouse U5snRNAs. The proposed secondary structure for one of these sequences is shown in Figure 3E

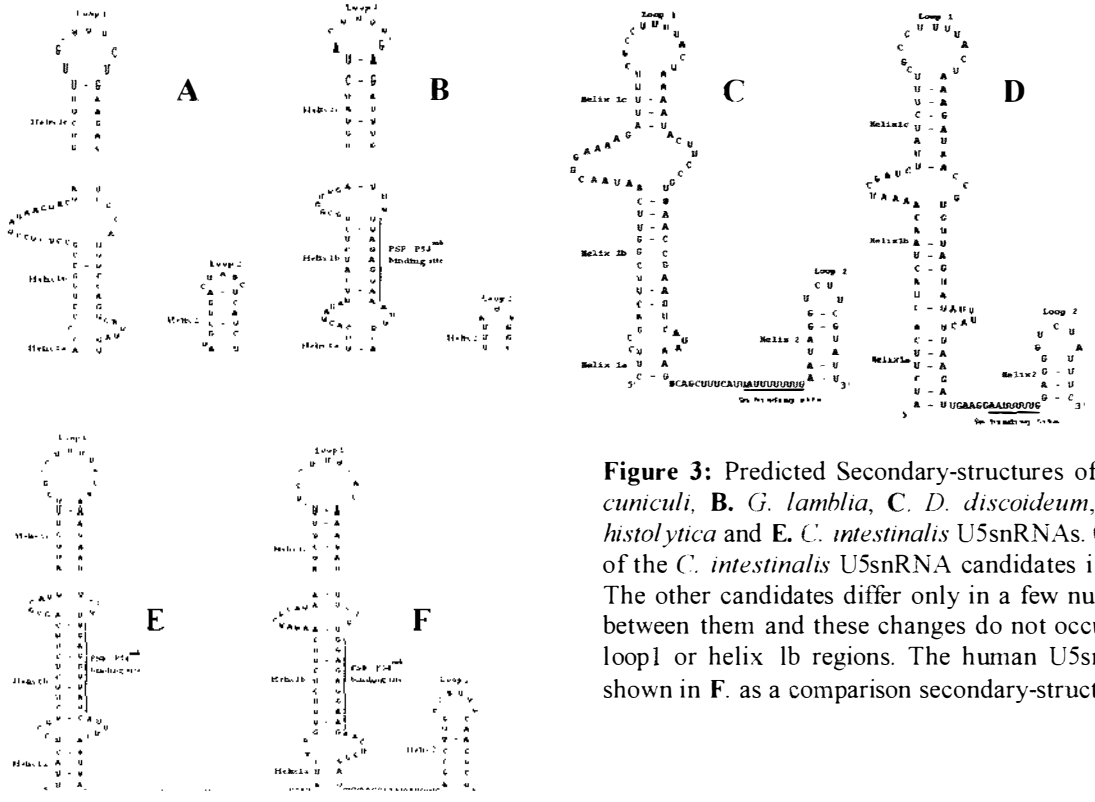


Figure 3: Predicted Secondary-structures of **A.** *Ecz. cuniculi*, **B.** *G. lamblia*, **C**. *D. discoideum*, **D.** *Ent. histolytica* and **E.** *C. intestinalis* U5snRNAs. Only one of the *C. intestinalis* U5snRNA candidates is shown. The other candidates differ only in a few nucleotides between them and these changes do not occur in the loop1 or helix 1b regions. The human U5snRNA is shown in **F**. as a comparison secondary-structure [9].

RNAmotif searches against the *G. lamblia* genome with the U5_A and U5_B descriptors (the U5_C descriptor failed to recover any clear candidate sequence) recovered a candidate sequence (AACB01000156: 17548-17456) that could be folded into the consensus U5snRNA secondary-structure. The sequence in loop1 does not conform entirely to the consensus loop1 motif (constructed from crown and basal eukaryotic U5snRNA loop1 sequences, [28]. The loop1 sequence in particular, affects splice site selection, particularly for introns with non-ideal 5' splice sites [30], and thus the differences in the loop1 sequences of the *G. lamblia* U5snRNA candidate may be a reflection of a difference in the splicing mechanism of this organism. The *G. lamblia* U5snRNA candidate has been shown to be expressed using RT-PCR (data not shown) and its sequence has been confirmed. The proposed secondary-structure of the *G. lamblia* U5snRNA is shown in Figure 3B.

An U5snRNA candidate was also recovered from the *D. discoideum* genome (Figure 3C). Helix1c is shorter than those found in other species, having only 5bp but allowing some non-canonical base-pairing i.e. G-A pairing could lengthen this helix. As the candidate *D. discoideum* U5snRNA sequence was recovered from only preliminary contig data, more work needs to be done to establish its viability when the genome has been more fully sequenced and assembled. However, a search of the Rfam database with the candidate *D. discoideum* U5snRNA sequence returned the U5snRNA alignment increasing the validity of this candidate.

Searches of the *Ent. histolytica* genome also produced a candidate U5snRNA sequence (Figure 3D). Again this candidate U5snRNA sequence was recovered from

preliminary sequencing data and will require more investigation once the complete genome has been sequenced. As with the *D. discoideum* candidate, the *Ent. histolytica* U5snRNA returned the U5snRNA alignment with a search against the Rfam database.

3.2 RNaseP RNA

Testing against the Pdatabase (results shown in Table 4) showed that although the RNaseP descriptors covered only part of the total RNaseP RNA secondary-structure, they were still able to detect RNaseP RNA sequences from all three kingdoms. Descriptor P7_A showed the greatest ability not only to recover RNaseP RNA sequences from all three kingdoms, but to distinguish between them using CRI-motif scoring. RNaseMRP sequences were selected as negative controls during RNaseP descriptor testing because the RNaseMRP CRI-regions are similar to the RNaseP CRI-regions, with the expectation that the RNaseP descriptors should distinguish against the two different ncRNAs. RNaseMRP sequences were detected with the RNaseP descriptors at the lowest level (CRI-motif = 'e'; the consensus CRI motif common to all three kingdoms), indicating that results returned with this motif may not be specific to RNaseP.

The RNaseP descriptors were also tested against TestDatabaseA to determine if other ncRNAs were also detected. Some U4 snRNA sequences were returned with the lowest scoring motif (CRI-motif = 'e') but no other snRNAs were detected above this level. Thus, future genomic searches used score cutoffs above this level. The P7_B returned the same sequences from both the Pdatabase and TestDatabaseA with less than half the processing time but did not distinguish clearly between the RNase MRP and the archaeal and bacterial RNaseP sequences. The P7_C descriptor took longer than the other two descriptors and did not return some of the archaeal and bacterial sequences, but gave much higher folding scores for the eukaryotic sequences.

To date there have been no RNaseP RNA sequences described for *G. lamblia, Ecz. cuniculi, Ent. histolytica, D. discoideum, P. falciparum* and *C. intestinalis*. BLAST searches of these genomes with all known RNaseP RNA sequences failed to find any significant sequences. Candidate sequences were required to contain the CRIV, CRII and CRIII consensus regions in expected places and the CRI-region had to contain conserved nucleotides present in RNaseP RNAs from all three kingdoms.

RNAmotif searches against the *Ecz. cuniculi* genome recovered an RNase P candidate (Figure 4C) (Chromosome VII – start position 87184) which also contains some sequence similarity to RNaseP RNA sequences from rat and mouse. Another candidate was recovered from the *Ecz. cuniculi* genome (Figure 4A) with a proposed secondary structure that fits the general eukaryotic consensus secondary-structure; except for the P3-region which is more bacterial-like (compared with the example structures shown in Figures 2D and E).

Table 4	P7_A	P7_B	P7_C
Database Testing			
Processing Time	7 min 54 sec	3 min 12 sec	35 min 28 sec
Output File Size	47 KB	36 KB	39KB
	CRI motif	**CRI motif**	**CRI motif**
Human RNaseP	a (-15.03)	a (-13.81)	a (-20.43)
S. cerevisiae RNaseP	b (-15.05)	a (-15.05)	a (-19.75)
S. pombe RNaseP	e (-14.16)	e (-15.05)	-
P. abyssi RNaseP	d (-22.43)	d (-24.81)	c (-27.41)
S. acidocaldarius RNaseP	d (-13.52)	c (-16.25)	-
M. thermoautotrophicum RNaseP	d (-17.15)	c (-24.08)	d (-20.38)
A. fulgidus RNaseP	d (-26.70)	-	-
H. cutribricum RNaseP	c (-25.53)	c (-29.33)	-
E. coli RNaseP	b (-21.46)	a (-21.46)	a (-26.46)
B. subtilis RNaseP	c (-15.27)	c (-15.27)	-
A. tumefaciens RNaseP	c (-24.26)	a (-26.89)	a (-25.87)
S. lividans RNaseP	c (-31.73)	d (-25.63)	c (-27.93)
A. thaliana RNase MRP	e (-19.95)	c (-19.95)	c (-26.15)
N. tabacum RNase MRP	e (-18.92)	c (-18.92)	c (-22.96)
Human RNase MRP	e (-14.74)	c (-13.43)	c (-21.74)
S. pombe RNase MRP	e (-12.10)	-	c (-16.51)
S. cerevisiae RNase MRP	-	-	-
Other snRNAs that occur with descriptor:	L.collosoma U4 (e) T. brucei U4 (e)	L.collosoma U4 (e)	L.collosoma U4 (d) T. brucei U4 (d)
Processing Time			
TestDatabaseA (Used for U5 testing)	2 min 54 sec	1 m 18 sec	8 min 23 sec
Ecz. cuniculi genome	866 min 25 sec	290 min 26 sec	1619 min 9 sec
P. falciparum genome	Not run	617min 27sec	Not run
G. lamblia genome	508min 52 sec	174 min 48 sec	Not run
Ent. histolytica genome	Not run	1627 min 37sec	Not run
D. discoideum genome	Not run	1093 min 26 sec	Not run

Table 4: Evaluation results from the RNaseP descriptors. Indicative results are shown. For processing times all descriptors had the following settings "emax = 15" with scores below this threshold being rejected and score cutoffs at 1.7. Detection of sequences with an "e" motif was achieved by setting the emax to -12 and the score cutoff to 1.0. The best folding energy scores are shown in brackets next to the CRI motif. Searches of the *P. falciparum*, *G. lamblia*, and *D, discoideum* genomes used parallel processing with emax at -20 and the score cutoff at 2.0. Searches of the *G. lamblia* genome were run with 8 nodes. Searches of the *P. falciparum*, *D. discoideum* and *Ent. histolytica* genomes were run with 16 nodes. '-' indicates that this sequence was not detected with this descriptor.

CRI motif a = "ggaarnucngng" (Eukaryotic consensus CRI with first "g");
CRI motif b = "gaarnucngng" (Eukaryotic consensus CRI without first "g");
CRI motif c = "ggaaanucc" (Bacterial consensus CRI);
CRI motif d = "ggaannuc" (Archaeal consensus CRI);
CRI motif e = "gnaannuc" (Universal consensus CRI)

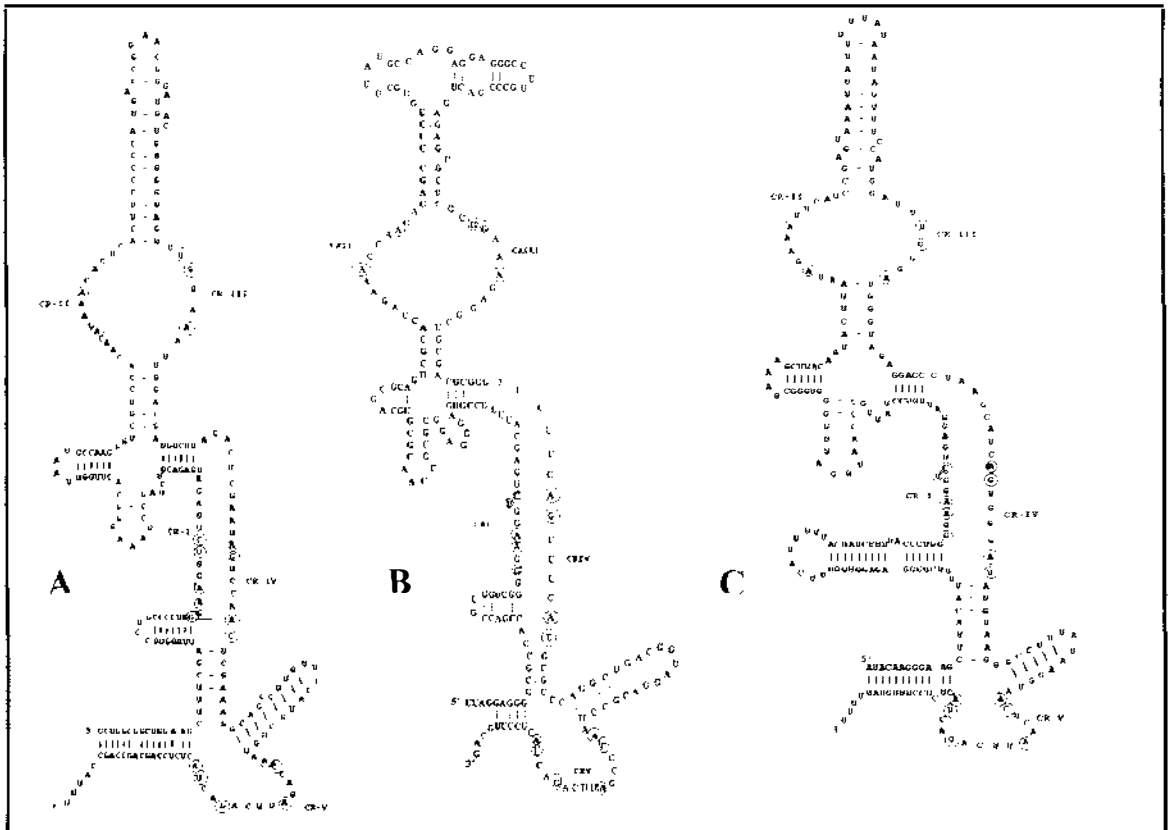


Figure 4: Putative RNaseP RNA secondary-structures of **A:** *Ecz. cuniculi*, **B:** *G. lamblia* and **C:** *Ent. histolytica* RNaseP RNA candidates. Circled nucleotides are those that are conserved between all three kingdoms (refer Figure 2) and must be present in these locations for the candidate to be considered a potential RNaseP RNA.

RNAmotif searches against *G. lamblia* using P7_A and P7_B descriptors (P7_C was not used due to its much longer processing time, even with parallel processing) returned a candidate RNase P sequence (Figure 4B)(AACB01000012: 65152-64918). As with the *Ecz. cuniculi* candidate, the P3-region was bacterial-like but the rest of the structure fitted the eukaryotic model. RT-PCR has shown that this sequence is expressed in *G. lamblia* and the sequence has been confirmed by sequencing.

RNAmotif searches against the *Ent. histolytica* genome also recovered a candidate RNaseP RNA sequence (Figure 4C) (*Ent1359g08.p1k*:355-440). The P3-region was longer than that found in the *Ecz. cuniculi* and *G. lamblia* RNaseP candidates, and although there was a two-nucleotide bulge in the 3' side of the helix, the *Ent. histolytica* P3-region still resembled the bacterial-like P3-region as opposed to the eukaryotic model. As this sequence was obtained from preliminary sequence data, further investigation will be required when the genome sequencing has been completed.

RNAmotif searches against *P. falciparum* failed to find any viable RNaseP RNA candidates, which may be due in part to its high A+T content. RNAmotif searches against the *D. discoideum* genome also failed to find any RNaseP RNA candidate. The genomic data for this genome is still preliminary and it is possible that the region containing an RNaseP RNA has not yet been sequenced. New releases of this genome will be screened in the future for possible RNaseP RNA candidates.

The RNaseP RNA candidates found in *Ecz. cuniculi*, *Ent. histolytica* and *G. lamblia* contain all the features that are expected in an RNaseP RNA, including nucleotides that have been shown to be conserved between RNaseP RNAs from all three kingdoms [27]. The Rfam database was searched with the three RNaseP candidate sequences from *G. lamblia*, *Ent. histolytica* and *Ecz. cuniculi* but returned no hits. As a test, the RNaseP sequence from the fruitfly, *D. melanogaster* (the only eukaryotic RNaseP RNA not found in Rfam, AF434763), also failed to return any hits from Rfam. The eukaryotic RNaseP RNA is notoriously difficult to align which may account to some extent to the lack of any positive results from Rfam for any of our RNaseP candidates.

4 Discussion

Non-coding RNA genes are hard to find in genomic data. Previously, RNAmotif has been used to find ncRNAs with highly conserved secondary-structure, using very "tight" and efficient descriptors [11, 12]. This technique has now been taken to the next level, integrating sequence and secondary-structure sub-elements (representing protein and RNA binding sites) to search for ncRNA genes in eukaryotic genomic data.

A potential criticism of the RNAmotif software is that it cannot yet give a value of statistical significance on each returned sequence. Here positive and negative controls are used instead, because they can compensate at least partially. The concept of positive and negative controls is fundamental in molecular biology (as many of its techniques also lack statistical analysis). However, a measure of statistical significance would be desirable to compare whole ncRNA sequences. Sequence and secondary-structure alone may not be enough to produce a true representation of the features required by an ncRNA molecule in order to retain biological function. Algebraic dynamic programming techniques that can model the complete structure in a similar way to RNAmotif, are being developed [31]. Such methods at present are mathematically challenging but in future may result in a statistical evaluation method that will integrate well with RNAmotif and other ncRNA-associated software.

The design of appropriate descriptors is presently not a simple task. Testing has shown as the "descr" section of the descriptor becomes more complicated (i.e. more helices, sequences and parameters), processing time increases dramatically. Ideally the simplest possible descriptor should be designed to search for a particular ncRNA. However, this may not be appropriate when searching for ncRNAs for which little information is available. In these cases, 'atypical' elements found in some species but not in others may have to be included to facilitate sequence recovery. An example of this is the specific descriptor U5_A which was more accurate in finding candidates in "typical" eukaryotic organisms such as *C. intestinalis*, but did not work as well as the less-specific descriptor (U5_B) for searching basal eukaryotic genomes. By having descriptors with differing stringency and related in different ways, candidates could be recovered from distant genomes. At present, there is still a balance that must be achieved between a descriptor that can run in a realistic time-frame and one that allows enough variability for a search of a distantly-related genome.

Testing of the RNaseP descriptors showed that the complete ncRNA need not be modelled, and areas of high variability between species can be described as single-stranded regions, aiding both processing performance and species detection. Analysis of regions "downstream" of the descriptor region was done manually in this study, but could also be automatically incorporated into future RNAmotif releases.

The RNAmotif program is still a program under development. The largest genome size that we used here was that of the sea-squirt *C. intestinalis* (155Mbases). It is feasible to search larger genomes using parallel processing and descriptors, "fine-tuned" for performance efficiency. However, when searching into a little known genome such as those basal eukaryotes, many descriptor parameters cannot be completely optimized for performance without running the risk of not recovering the ncRNA of choice. Future RNAmotif releases may include the incorporation of other biological information, such as whether a resulting sequence is within an open reading frame, or ways to compensate for AT-rich target genomes. However, RNAmotif now offers a realistic and biologically-orientated way to search for other non-coding RNAs within the increasingly wide range of sequenced genomes.

5 Acknowledgements

Many thanks to the administrators of the Helix parallel processing facility at Massey University, Albany, New Zealand for their help and advice. Many thanks to Mitchell L. Sogin, and Andrew G. McArthur and their teams at the *Giardia lamblia* Genome Project, (funded by the NIAID/NIH under cooperative agreement AI 043273), Marine Biological Laboratory at Woods Hole, for access to non-public data. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for the RT-PCR and sequencing of some of the ncRNA gene candidates. This work was supported by the NZ Marsden Fund.

6 References

[1] S. R. Eddy. Non-coding RNA genes and the modern RNA world. Nat Rev Genet, 2(12): 919-29., 2001.

[2] G. Storz. An expanding universe of noncoding RNAs. Science, 296(5571): 1260-3, 2002.

[3] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2(1): 8, 2001.

[4] A. Laferriere, D. Gautheret and R. Cedergren. An RNA pattern matching program with enhanced performance and portability. Comput Appl Biosci, 10(2): 211-2, 1994.

[5] M. Dsouza, N. Larsen and R. Overbeek. Searching for patterns in genomic data. Trends Genet, 13(12): 497-8, 1997.

[6] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 25(5): 955-64, 1997.

[7] D. Gautheret and A. Lambert. Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. J Mol Biol, 313(5): 1003-11., 2001.

[8] R. J. Klein and S. R. Eddy. RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics, 4(1): 44, 2003.

[9] R. Peng, B. T. Dye, I. Perez, D. C. Barnard, A. B. Thompson and J. G. Patton. PSF and p54nrb bind a conserved stem in U5 snRNA. Rna, 8(10): 1334-47., 2002.

[10] L. J. Collins, V. Moulton and D. Penny. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. J Mol Evol, 51(3): 194-204., 2000.

[11] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res, 29(22): 4724-35., 2001.

[12] V. Tsui, T. Macke and D. A. Case. A novel method for finding tRNA genes. Rna, 9(5): 507-17, 2003.

[13] G. B. Fogel, V. W. Porto, D. G. Weekes, D. B. Fogel, R. H. Griffey, J. A. McNeil, E. Lesnik, D. J. Ecker and R. Sampath. Discovery of RNA structural elements using evolutionary computation. Nucleic Acids Res, 30(23): 5310-7., 2002.

[14] A. J. Newman. The role of U5 snRNP in pre-mRNA splicing. Embo J, 16(19): 5797-800., 1997.

[15] A. G. McArthur, et al. The Giardia genome project database. FEMS Microbiology Letters, 189(2): 271-273, 2000.

[16] L. Eichinger and A. A. Noegel. Crawling into a new era-the Dictyostelium genome project. Embo J, 22(9): 1941-6, 2003.

[17] B. J. Mann. Entamoeba histolytica Genome Project: an update. Trends Parasitol, 18(4): 147-8, 2002.

[18] P. Dehal, et al. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science, 298(5601): 2157-67, 2002.

[19] S. Xiao, F. Scott, C. A. Fierke and D. R. Engelke. Eukaryotic Ribonuclease P: A Plurality of Ribonucleoprotein Enzymes. Annu Rev Biochem, 71: 165-89, 2002.

[20] C. Mathe, M. F. Sagot, T. Schiex and P. Rouze. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res, 30(19): 4103-17, 2002.

[21] E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics, 16(7): 583-605., 2000.

[22] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna and S. R. Eddy. Rfam: an RNA family database. Nucleic Acids Res, 31(1): 439-41, 2003.

[23] M. D. Katinka, et al. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature, 414(6862): 450-3., 2001.

[24] M. J. Gardner, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature, 419(6906): 498-511, 2002.

[25] A. Bahl, et al. PlasmoDB: the Plasmodium genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). Nucleic Acids Res, 30(1): 87-90., 2002.

[26] J. W. Brown. The Ribonuclease P Database. Nucleic Acids Res, 27(1): 314, 1999.

[27] D. N. Frank, C. Adamidi, M. A. Ehringer, C. Pitulle and N. R. Pace. Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. RNA, 6(12): 1895-904., 2000.

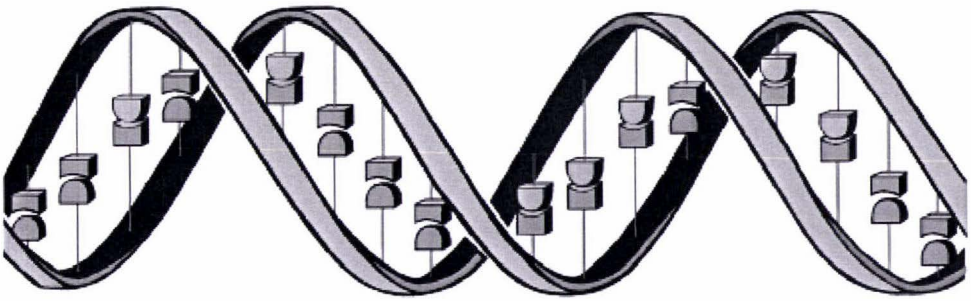
[28] A. Szkukalek, E. Myslinski, A. Mougin, R. Luhrmann and C. Branlant. Phylogenetic conservation of modified nucleotides in the terminal loop 1 of the spliceosomal U5 snRNA. Biochimie, 77(1-2): 16-21., 1995.

[29] S. Xiao, F. Houser-Scott and D. R. Engelke. Eukaryotic ribonuclease P: increased complexity to cope with the nuclear pre-tRNA pathway. J Cell Physiol, 187(1): 11-20., 2001.

[30] T. S. McConnell and J. A. Steitz. Proximity of the invariant loop of U5 snRNA to the second intron residue during pre-mRNA splicing. Embo J, 20(13): 3577-86., 2001.

[31] C. Meyer and R. Giegerich. Matching and Significance Evaluation of Combined Sequence-Structure Motifs in RNA. Z. Phys. Chem., 216: 193-216, 2002.

[32] D. Penny and A. Poole. The nature of the last universal common ancestor. Curr Opin Genet Dev, 9(6): 672-7., 1999.



Chapter 3 Having a BLAST with Ancestral Sequences

"A wise man gathers from the past what is to come"– Sophocles

One of the fundamental paradigms in computational biology is function prediction by homology. A gene is compared against others in a database and if a second sequence is detected whose similarity to the first gene is statistically significant; then the function of the unknown protein is inferred based on the function of the known protein (Sjolander 2004). A number of computationally efficient methods for pairwise sequence comparisons have been developed including BLAST (Altschul et al. 1990), FASTA (Pearson and Lipman 1988) and PSI-BLAST (Altschul et al. 1997), which have been used in high-throughput, homology-based prediction (Sjolander 2004).

However there are three evolutionary factors that can cause 'interpretation errors' with these programs (Sjolander 2004). The first factor is gene duplication where genes related by duplication events (paralogous genes) can diverge in function, but still contain statistically similar sequences. A homologous sequence from a distantly related species may be more similar to the sequence from the duplicate copy and an incorrect function can be inferred. The second evolutionary factor is the shuffling of protein domains where in two functionally different proteins, motifs may be highly conserved but in different positions. Standard methods typically ignore whether two proteins align globally or only locally, and thus again the wrong protein may be used to assign function. The third factor that can cause interpretation errors is speciation, where orthologous proteins share a common ancestor but have different functions if the species are distantly related. The correct function may be assigned if the species that are being compared are close but not if the species are evolutionarily distant. Automated prediction programs may predict incorrect coding region limits and/or split long genes into several coding regions (or alternatively merge several regions into one) (Mathe et al. 2002). Current programs aim at optimising performance on the majority of data, usually at the expense of the 'outliers', however these 'outliers' may, in some organisms be more frequent than suspected. The prediction of protein encoding genes is still in need of improvement (Mathe et al. 2002).

Within this project, spliceosomal protein studies depend on the accurate prediction of protein function, thus techniques were developed to test this accuracy. From this, a new method was developed for protein prediction using inferred ancestral sequences (Ancestral Sequence Reconstruction - ASR). This chapter will examine protein prediction and the ASR technique using a number of RNaseP-associated proteins that may have been present in the eukaryotic ancestor. It contains a published manuscript, "Using ancestral sequences to uncover potential gene homologues" (Collins et al. 2003). This was presented initially at the New Zealand Bioinformatics Conference 2003, and shows how inferred ancestral sequences can be used in

combination with standard search algorithms to identify candidate gene homologues from newly sequenced genomes (such as *Giardia lamblia* and *Entamoeba histolytica*).

In genome comparisons of distantly related species, it is often difficult to identify highly diverged protein-coding regions (Collins et al. 2003). This is especially true when searching for either protein or ncRNA genes from any of the basal eukaryotes. Theoretically identifying an ncRNA from a genome increases the chance of finding suspected associated-proteins, and vice versa. Thus finding the RNaseP RNA from *G. lamblia* (Chapter 2) makes it possible that proteins (such as Pop1, Pop4, Pop5 and Rpp21) known to be part of the RNaseP from other eukaryotes, are also present in *G. lamblia*. This study used these proteins to test this premise, and evaluated techniques that would be used later for studying the large group of proteins associated with the spliceosome (Chapter 4).

RNaseP processes the 5' ends of precursor tRNA molecules in all organisms (Altman 1989). In bacteria RNaseP may also processes other RNA substrates as several polycistronic mRNAs have recently been found to be cleaved by RNaseP in *Escherichia coli* intergenic regions (Li and Altman 2003). Cellular processes other than tRNA processing may also be affected by bacterial RNaseP including the maturation of SRP-RNA (a major component of bacterial protein secretion) and protein translocation (Li and Altman 2003).

In eukaryotes, RNaseP contains a single catalytic RNA and multiple protein subunits (possible exceptions include the spinach chloroplast which may not contain a functional RNA component) (Xiao et al. 2002; de la Cruz and Vioque 2003). Because the RNA provides the catalytic function, evolution maintains higher-order protein structure rather than conservation of amino-acid sequences. As a result, it is often difficult to identify homologous RNaseP proteins in distantly related lineages. There are 9-10 proteins (depending on eukaryote) and all are essential to the function of the ribonucleoprotein (Xiao et al. 2002). Only some of these proteins (Pop4 – also called Rpp29; Rpp21, Rpp30 and Rpp38 in humans) bind to the RNA while others act as stabilisers or aid in transport throughout the cell (Xiao et al. 2002). Recent studies show that the human RNaseP RNA (H1 RNA), Rpp21 and Rpp29 are sufficient for the 5' cleavage of precursor tRNA (Mann et al. 2003). Finding RNaseP proteins in other eukaryotes should help in understanding the nature of RNaseP in the ancestral eukaryote (Penny and Poole 1999).

The four RNaseP proteins chosen for this study are Pop1, Pop4, Pop5 and Rpp21, which were used to search the *Encephalitozoon cuniculi* (*Ecz. cuniculi*), *Giardia lamblia* and *Entamoeba histolytica* (*Ent. histolytica*) genomes. The first step in phylogenomic analysis involves the identification of homologous proteins from different species (Sjolander 2004). Homologue identification can then lead to sequence alignments, that in turn are used for input into HMM-based programs, or as is seen later in this chapter, ancestral sequence reconstruction.

Searches of the protein databases at NCBI (http://ncbi.nlm.nih.gov) found a number of sequences for each of the four proteins (Pop1, Pop4, Pop5 and Rpp21). Proteins, accession

numbers and references are shown in Table ASR-3.1, page 71). Downloaded data was stored in a custom designed database "P-MRPbase", named because the majority of proteins that bind to RNaseP also bind to a closely related ncRNA, RNaseMRP (Mann et al. 2003). This database was central to managing a wide range of information about RNaseP and is discussed in more detail in Appendix E. Database management techniques developed and applied in studying RNaseP proteins were also applied to the larger spliceosomal project (Chapter 4).

BLAST involves taking a known sequence (the 'query' sequence) and attempting to align it to any part of any sequences in a designed database (the 'subject' database) and thus recover a homologous ('candidate') sequence from the subject database. By BLASTing in turn, the query sequences from different species against a genome, consistency indicates if there are any query sequences that have been clustered in error, as well as recovering sequence regions corresponding to the desired candidate protein. For example a BLAST search with the human Pop4 as query sequence will recover a Pop4 candidate protein from the *G. lamblia* genome - but homology is scattered throughout the alignment between the two proteins. The *S. pombe* Pop4 also recovers the same *G. lamblia* candidate but there are differences in the aligned sequences (due to sequence differences between the human and *S. pombe* query sequences). If, theoretically, a kiwi Pop4 sequence (only used here as example) recovered a different sequence from *G. lamblia* with very different scores and alignment then the kiwi sequence would have to be examined to see if it really belonged in the Pop4 sequence cluster. This checking for anomalies is extremely useful when multiple alignments are being created for any protein.

Other factors such as length and physiochemical properties can also be used to check both the query and the recovered candidate sequences (Han et al. 2004). Amino-acid composition, charge, polarity and hydrophobicity play prominent roles in protein classification. Amino-acid composition and hydrophobicity are important factors for the interaction of a protein with other biomolecules, as well as for protein folding. Similarly, charge and polarity are important for electrostatic interactions and hydrogen-bonding to RNA (Han et al. 2004), and thus it is expected that candidate sequences will contain similar physiochemical properties to those of the known protein sequences. Such features can be predicted using software such as WinPep (Hennig 1999), however predictions based on amino-acid sequence alone may differ from biochemically determined features.

Using this procedure of 'comparative' BLASTing, the Pop4 protein was identified in *G. lamblia*, Pop1 in *Ent. histolytica* and *Ecz. cuniculi*, Rpp21 in *G. lamblia* and Pop5 in *Ecz. cuniculi* and *Ent. histolytica*. Results and candidates are indicated in the included manuscript. 'Back-BLASTing' (BLASTing the candidate sequence against the NCBI protein databases) checks to see if a candidate sequence is similar only to it's original query protein or is similar to multiple proteins (e.g. the candidate protein could belong to a protein family). This tests the reciprocity of the BLAST search and was used here to check the validity of protein candidates.

Some proteins were not found within the basal eukaryotic genomes using BLAST, so other techniques were investigated that could recover potential sequences. BLAST is a useful quick method but more powerful methods are sometimes required.

One such program is the HMMer software (Eddy 1998) that builds a hidden markov model profile from an input alignment then searches a database to find candidate sequences to fit this profile. Hidden Markov models (HMMs) are statistical models of the consensus of sequences from a protein family (Durbin 1998); Eddy - HMMer users guide). They develop scores for amino-acids (or opening and extending gaps) that are specific for each amino-acid (or nucleotide) position. In contrast, traditional pairwise alignment (for example, BLAST or FASTA) uses position-*independent* scoring parameters (a match or mismatch is scored equally anywhere in the sequence). This position-dependent property of HMM-profiles captures important information about the degree of conservation at various positions in multiple alignments, and the varying degree to which gaps and insertions are permitted. HMM- or profile-based methods typically outperform pairwise methods, such as BLAST, in both alignment accuracy and database search sensitivity and specificity, but do not capture any higher-order correlations such as sequence folding (Eddy – HMMer Users guide). This is why HMM-profiling makes poor models of RNAs, because an HMM profile cannot presently, describe base-pairing events (Durbin 1998); Eddy - HMMer users guide).

HMMer recovered a candidate for the Pop5 protein from the *G. lamblia* genome that was not found using any other method in this study (results in the included manuscript). This Pop5 candidate gene was detected on a 9 amino-acid conserved sequence at the N-terminal end (beginning) of the protein. HMMer was the only software tested that was sensitive enough to recover this small region. Other attributes of the *G. lamblia* Pop5 sequence such as length, predicted molecular weight and predicted isoelectric point confirm that this is a likely Pop5 candidate.

3.1 ASR - Ancestral Sequence Reconstruction

A new method was developed in this project to aid in finding proteins from distant genomes. This method used 'ancestral sequences' inferred using a sequence alignment and a phylogenetic tree of the sequences. Previously, reconstruction of inferred ancestral proteins has been used to infer functional properties of current proteins and to look at how they may have evolved (Chandrasekharan et al. 1996). Ancestral sequences have also been used to develop vaccine antigens against HIV (Nickle et al. 2003). A new application of ancestral sequence reconstruction (ASR) for homologue identification is reported here. ASR involves aligning sequences from extant (living) species and calculating the most likely sequence (i.e. the maximum likelihood estimate; Steel and Penny 2000) of an 'ancestor' at a node on an evolutionary tree (Collins et al. 2003). The accuracy of a predicted ancestral sequence is

reduced sharply with increasing evolutionary distance (Koshi and Goldstein 1996) and it has been shown under current models of sequence evolution, to be "impossible" to reconstruct ancestral data at the root of "deep" phylogenetic trees, such as those representing the eukaryotic phylogeny (Mossel 2003). However, in applying ASR as a tool in gene finding, the reconstruction of the sequence does not have to be perfect; only close enough to advance the construction of the next node, or to find a candidate homologue. Basically the ancestral sequence, even if not perfect is closer to the target than any extant sequence. The ASR technique, its development and application to finding proteins within protist genomes, are covered in the manuscript included in this chapter. Supplementary data to this manuscript is given in Appendix B. RNaseP-associated protein candidate sequences are given in AppendixC.3.

There is some information additional to that in the included manuscript. In the manuscript summary it is stated that the effect of different substitution matrices with BLAST (BLOSUM62 was used throughout) was not tested. Since this publication, the PAM250 substitution matrix was tested with the Pop1 proteins and inferred ASR sequences; results showing a decrease in scores and E-values from all searches (Table 3.1). For this reason the default substitution matrix BLOSUM62 was used for subsequent ASR testing.

G. lamblia genome	**BLOSUM62**		**PAM250**	
Pop1	**Score**	**E-value**	**Score**	**E-value**
Human	35	0.09	32	1.0
Mouse	37	0.03	35	0.2
D. melanogaster	36	0.033	38	0.02
C. elegans	39	0.006	33	0.63
C. briggsae	38	0.009	34	0.37
S. cerevisiae	-	-	-	-
S. pombe	32*	0.52	-	-
A. nidulans	-	-	28*	1.3
C. albicans	-	-	-	-
P. falciparum	38	0.007	-	-
C. parvum	39	0.001	-	-
E. histolytica	33	0.084	-	-
Node A	43	7e-04	40	0.007
Node B	45	1e-04	41	0.003
Node C	44	2e-04	41	0.003
Node D	44	2e-04 +	42	0.001
Node E	38	0.022 +	35	0.27
Node F	35	0.11 +	-	-
Node G	41	0.003	-	-

Table 3.1: Comparison of different substitution matrices used with the ASR technique. Results from the search of the *G. lamblia* genome with the Pop1 protein sequences and inferred ancestral sequences. The full length ancestral sequences for the Pop1 proteins were used here. * indicates that the *G. lamblia* candidate sequence was not the top scoring hit but within the first 3 hits. + indicates that another sequence was equal at the top with this hit. In general the score value reflects the number of similar or identical residues (i.e. the higher the better). The E-value (Expectation value) is the number of different alignments with scores equivalent to or better than the one recovered, that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

During this study two programs were used to infer ancestral sequences but there are other software that can also do so. One such program is PAUP[8] (Phylogenetic Analysis Using Parsimony) which is a widely used software package for the inference of evolutionary trees. This package is not free and requires some time for the non-experienced user to understand command and parameter settings but offers a different option for the calculation of ancestral sequences. The AllAll program of the Darwin package (Le Bouder-Langevin et al. 2002) can also calculate ancestral sequences, but has a disadvantage in that it does not allow the input of a sequence alignment, instead using a file of sequences to create an alignment. This is not a problem if the sequences can be easily aligned, but poorly constructed alignments between sequences from diverse species may result in unhelpful ancestral sequences.

A disadvantage with the more 'intense' search techniques such as HMMer and ASR compared with BLAST is that they require much more time, both in the construction of confident sequence alignments and in the case of HMMer, in the running time of the software. Pre-assembled alignments and HMM models are now available from the Pfam database (Bateman et al. 2004); http://www.sanger.ac.uk/Software/Pfam/) and can be downloaded for searching local databases. Pop4 belongs to a Pfam alignment (UPF0086 – domain of unknown function) which contained some of the sequences that were used in this study but none from any basal eukaryote. Pop1 had a Pfam alignment (PF06978 -Ribonucleases P/MRP protein subunit POP1) which includes sequences from *G. lamblia* and *Ecz. cuniculi* but no sequence from *Ent. histolytica*. Rpp21 also had a Pfam alignment (PF04032 - RNaseP Rpr2/Rpp21 subunit domain) which now includes many more sequences than were used during this study including sequences from *Cryptosporidium parvum*, *Plasmodium falciparum*, *G. lamblia* and some archaeal sequences. There is currently no Pfam alignment for the Pop5 protein.

The availability of these alignments speeds up the search process when techniques other than BLAST are necessary. ASR uses one other factor, the construction of a phylogenetic tree appropriate for the sequences and species contained in the sequence alignment. SSU rRNA trees are readily available from the Ribosomal Database Project (http://rdp.cme.msu.edu/html), bearing in mind that branch positions of the basal eukaryotes, microsporidia and sometimes other eukaryotic species, on these trees may not be representative of the true relationship between organisms.

For a significant percentage of proteins encoded in a 'typical' genome, no amount of BLAST will identify homologues of known function (Sjolander 2004). In prokaryotic organisms >35% of genes are annotated as "function unknown" (Karaoz et al. 2004). In

[8] Swofford, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0 Sinauer Associates, Sunderland, Massachusetts. Information at http://paup.csit.fsu.edu/about.html. Currently version 4.0 is offered but earlier versions (such as version 3.1) may also be used for inferring ancestral sequences.

eukaryotes, functional annotation is especially daunting, for example >60% of the genes in *Plasmodium falciparum* are "hypothetical" proteins (Gardner et al. 2002). Further annotation of these genomes will probably come mostly from individual genes analysis (in contrast to high-throughput annotation). BLAST will probably remain the most popular gene-finding tool due to its ease of use and that it has been largely successful to date but its limitations are becoming all too clear with the large number of 'hypothetical proteins' listed in every genome sequence. ASR becomes another tool to join HMMer and other protein software in the protein-finding 'toolbox'.

Finding RNaseP proteins, Pop1, Pop4, Pop5 and Rpp21 in *G. lamblia* increases the expectation of also finding an RNaseP RNA (and vice versa). A candidate RNaseP RNA was found in *G. lamblia* (Chapter 2) suggesting that the RNaseP complex in *G. lamblia* is eukaryotic-like (i.e. many proteins and a single RNA) rather than bacterial-like (one protein and one RNA). Although there is significant similarity between the RNA subunits of bacterial and archaeal RNaseP (Brown 1999), database mining of the available archaeal genomes has failed to identify a protein homologous to the bacterial RNaseP protein, however, proteins homologous to some of the eukaryotic proteins have instead been found (Boomershine et al. 2003).

Pop4, Rpp21, Rpp30 and Pop5 homologues have been experimentally identified in the archaeal species, *Methanothermobacter thermoautotrophicus* (Hall and Brown 2002) and *Pyrococcus horikoshii* (Kouzuma et al. 2003). It was shown that in *P. horikoshii*, the Pop5, Rpp21 and Pop4 homologues are minimal components of RNaseP activity, that is greatly increased by the inclusion of the Rpp30 protein (Kouzuma et al. 2003). All four of the *P. horikoshii* proteins bind to the RNaseP RNA from this species. Other tests showed that the reconstituted RNaseP complex may have a slightly different structure than the native complex, suggesting that there may be other RNaseP proteins that have not yet been tested (Kouzuma et al. 2003). With the finding of Pop4, Rpp21 and Pop5 homologues in *G. lamblia*, it may be likely that at least these three proteins were present in the last common ancestor between eukaryotes and archaea. Pop1 has not yet been identified in any archaeal species but its presence in *G. lamblia* indicates that it is likely to have been present at least, in the eukaryotic ancestor. RNaseP proteins Rpp14 and Rpp30 have been characterised in archaea (Jarrous et al. 1999; Koonin et al. 2001; Jiang and Altman 2002) and belong to alignments in Pfam (Rpp14 family and RNase P p30 alignments respectively). No Rpp14 or Rpp30 protein candidate genes were found in *G. lamblia* or any other basal eukaryote, although Rpp14 and Rpp30 have been identified in the microsporidian *Ecz. cuniculi* (Pfam sequences: Q8STU1 and Q8SUN8 respectively). An Rpp25 protein homologue has been isolated from the ciliate *Stylonychia lemnae* (called Mdp2) (Fetzer et al. 2002; Aravind et al. 2003). Rpp25 and Pop7 (Rpp20 in humans) proteins belong to a protein superfamily of which members have also been identified in some archaeal genomes (Aravind et al. 2003). BLAST results of Rpp25 and Pop7 sequences returned no significant hits in either *G. lamblia*, *Ent. histolytica* or *Ecz. cuniculi*.

From the results in the accompanying paper and later publications it can be suggested that at least Pop1, Pop4, Pop5, Rpp14, Rpp21, Rpp30 and perhaps Rpp25 and Pop7, were present in the RNaseP complex of the eukaryotic ancestor. The identification of the RNaseP RNA, Pop4 and Rpp21 from *G. lamblia* opens the way for biochemical analysis to determine if like humans (Mann et al. 2003), these three components are enough for 5' cleavage of the precursor tRNA. It will also be interesting to determine in the future if any of the other more species specific proteins found in humans (Rpp38 and Rpp40) and *S. cerevisiae* (Pop3, Pop6 and Pop8) can also be found in *G. lamblia* or any other basal eukaryote, determining the likely composition of the complete RNaseP complex in the eukaryotic ancestor and perhaps, the last common ancestor between eukaryotes and archaea. From here comparisons between the single protein and the multi-protein complexes from bacterial and archaeal RNasePs respectively, should shed some light on the evolution of this 'ancient' ribozyme.

ORIGINAL RESEARCH

Using ancestral sequences to uncover potential gene homologues

Lesley J Collins,[1] Anthony M Poole,[1,2] David Penny[1]

[1]Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand; [2]Department of Molecular Biology and Functional Genomics, Stockholm University, Stockholm, Sweden

Abstract: Gene homologues between distantly related species can be difficult to identify. We test the idea that inferred ancestral sequences could aid in finding gene homologues. Ancestral sequences are inferred by aligning gene homologues on a known tree and estimating the most likely amino acid for each position at each node in that tree. BLAST and HMMER are used separately and together with ancestral sequences to search the genome sequence databases of *Encephalitozoon cuniculi*, *Entamoeba histolytica* and *Giardia lamblia* for RNase P protein homologues. RNase P proteins Pop4, Pop1, Pop5 and Rpp21 have been reported in humans and at least two other eukaryotic species but have yet to be identified in the above genomes. Using ancestral sequences reconstruction (ASR) for these proteins, we successfully identified putative homologues from *E. cuniculi*, *Ent. histolytica* and *G. lamblia*. In some cases, the use of ASR outperformed BLAST and HMMER. Overall, including ancestral sequences in searches with BLAST and or HMMER was the most successful approach in the recovery of potential RNase P protein gene homologues, making this a useful technique in early homologue identification.

Keywords: ancestral sequence reconstruction, *Giardia lamblia*, *Encephalitozoon cuniculi*, *Entamoeba histolytica*, RNase P

Introduction

In genome comparisons of distantly related species, it is often difficult to identify homologous protein coding regions that are highly diverged. One example is RNase P, a ribonucleoprotein that is responsible for processing tRNA transcripts into mature tRNAs (Altman 1989). In eukaryotes, RNase P contains multiple protein subunits and a single catalytic RNA. Because the RNA provides the catalytic function, evolution maintains higher-order protein structure rather than conservation of amino acid sequences. As a result, it is often difficult to identify homologous RNase P proteins in distantly related lineages.

Eukaryotic RNase P protein sequences vary greatly between species. Few have been identified in species other than human, mouse, *Drosophila melanogaster*, *Caenorhabditis elegans* and the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. There are 9–10 proteins (depending on species) plus one catalytic RNA present in the eukaryotic RNase P complex, and all are essential to the function of the ribonucleoprotein (Xiao et al 2002). Only some of these proteins bind to the RNA species, while others act as stabilisers and aid in transport of the complex and or its components (Xiao et al 2002). Many of the proteins are also found in the RNase MRP complex. The distribution of these RNase P and RNase MRP proteins (Figure 1) may also aid our understanding of how these ribonucleoprotein complexes are related (Collins et al 2000). Finding candidate RNase P proteins in protist

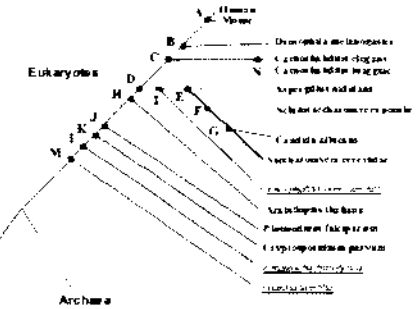


Figure 1 Tree (based on Embley and Hirt 1998) showing the relationship between the genomes of the eukaryotic species used in this study. Branch lengths are not to scale but represent an approximate distance between species. Underlined species are genomes that are being searched in this study. Ancestral sequences are represented by letters of the alphabet that are used throughout this study.

Figure 3.1:

Correspondence: Lesley J Collins, AllanWilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 11222, Palmerston North, New Zealand, tel +64 6 356 9099, fax +64 6 350 5626, email l.j.collins@massey.ac.nz

Applied Bioinformatics 2003 2(3 Suppl) S85–S95

genomes may help us to understand the ancestral condition of RNase P in eukaryotes and may add to the understanding of the first ancestral eukaryotes (Penny and Poole 1999).

The amitochondrial protist genomes *Giardia lamblia* and *Entamoeba histolytica* are highly diverged from each other and from species in which RNase P proteins have been characterised (eg humans and *S. cerevisiae*). The microsporidian protist, *Encephalitozoon cuniculi*, was once thought to be part of a deeply branching protist lineage, but microsporidia are now considered atypical fungi (Katinka et al 2001) so should provide an ideal 'stepping stone' in the path from the yeasts to the other protist genomes. To date, no RNase P components have been identified from any of these genomes. Figure 2 shows the putative relationship between the species whose sequences and genomes are used in this study. Protein names and accession numbers are given in Table 1.

In the past, reconstruction of inferred ancestral proteins has been used to test functional properties of current proteins and to look at how they may have evolved (Chandrasekharan et al 1996). Some predicted sequences have been synthesised, cloned and expressed in cell cultures. The expressed proteins have then been purified and tested in functional assays. In this way, the activity of the inferred ancestral protein can be compared to that of a present-day protein (eg with serine proteases as in Chandrasekharan et al 1996).

Another potential application of ancestral sequence reconstruction (ASR) is homologue identification. When BLAST* (Altschul et al 1997) fails to pick up a distant homologue, another solution is to use a hidden Markov model (HMM) profiling program such as HMMER (Eddy 1998). While HMM profiling makes use of a training dataset (a multiple sequence alignment), and is therefore potentially more accurate than pairwise BLAST, the relationship between the species represented in the input alignment is not taken into account, and often there are many sequences from closely related species available and only a few from species closely related to the organism of interest. This may have the effect that the resulting profile is well trained for sequences closely related to the majority in the initial alignment and not so well trained for under-represented (or unrepresented) sequences that are more divergent. Finally, HMM profiling can be difficult in the absence of information about annotated open reading frames from incomplete or newly sequenced genomes.

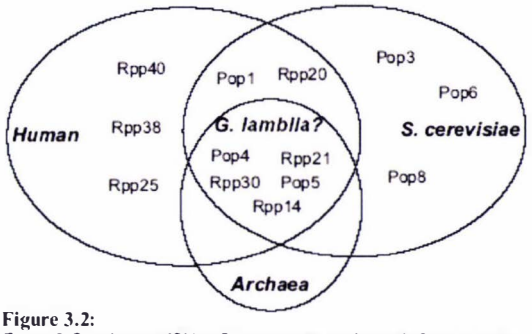


Figure 3.2:
Figure 2 Distribution of RNase P proteins prior to this study. Some proteins are only found in humans and closely related species, and some are only found in the yeast *Saccharomyces cerevisiae* and its close relatives. Proteins such as Pop1 and Rpp20 are found in both, whereas there is a group of RNase P proteins found also in Archaea (Archaebacteria). A search of the *Giardia lamblia* genome might expect to find the proteins that are found in the intersection of humans, yeast and the Archaea, namely Pop4, Pop5, Rpp21, Rpp30 and Rpp14 and perhaps Pop1 and Rpp20, but may not find proteins such as Rpp40 or Pop3.

ASR involves aligning sequences from extant species and calculating the most likely sequence of an 'ancestor' at nodes on an evolutionary tree. Because the output is a sequence, this can, in principle, be used in combination with both BLAST and HMMER. Figure 1 depicts such a strategy where a homologous sequence may be present in each of the represented species and the lettered nodes A–M correspond to the most likely ancestral sequences. Searching with the known species, eg human or mouse, may not uncover any homologues in the very distant genomes of *Giardia lamblia* and *Entamoeba histolytica*, but they may be found by searching with the ancestral sequences at nodes A or B.

As expected, the accuracy of the reconstruction of an ancestral sequence is reduced sharply with increasing evolutionary distance (Koshi and Goldstein 1996). Often, the most interesting evolutionary changes do not occur at the low levels of sequence divergence where the ancestral rates are more easily inferred (Chang and Donoghue 2000). However, in applying ASR as a tool in gene finding, the accuracy of the sequence does not have to be perfect; only close enough to advance the construction of the next node or to find a candidate homologue. There are a number of methods available for calculating ancestral sequences (as reviewed in Chang and Donoghue 2000). Likelihood methods such as PAML and FastML identify the most likely ancestral state according to a specified model of evolution. Their scores reflect the probability of observing the sequence data, given a particular tree and model of evolution.

In this study, we show how inferred ancestral sequences can be used in combination with standard search algorithms

Table 1 Sequences used in this study

Protein name	*Species*	*Accession number*	*Reference*
Pop1	Human	X99302	Lygerou et al 1996
	Mouse	BAB30296.1	Direct sub[a]
	Caenorhabditis elegans	U00048	Lygerou et al 1996
	Caenorhabditis briggsae	CBRG18D16	NCBI[b]
	Drosophila melanogaster	AAF46049.1	Direct sub[a]
	Aspergillus nidulans	gnl|TIGR_5085|afumi_1341[c]	TIGR,[d] this study[e]
	Saccharomyces cerevisiae	X80358	Lygerou et al 1996
	Schizosaccharomyces pombe	T50203	Direct sub[a]
	Candida albicans	gnl|SDSTC_5476|C.albicans_Contig6_1614[c]	NCBI,[b] this study[e]
	Arabidopsis thaliana	AAB63829	Direct sub[a]
	Cryptosporidium parvum	gnl|CVMUMN_5807|cparvum_Contig598	NCBI[b]
	Entamoeba histolytica	gnl|TIGR_5759|ehistolyt_ENTLB32TF	TIGR,[d] this study[e]
	Plasmodium falciparum	gnl|Sanger_36329|Sanger_BLOB_002759	Sanger,[f] this study[e]
	Giardia lamblia	00060.Contig670	GGSP,[g] this study[e]
Pop4	Human (Rpp29)	AF001176	Jarrous et al 1999
	Mouse	From reference	van Eenannaam et al 1999
	Caenorhabditis elegans	T19305	Direct sub[a]
	Arabidopsis thaliana	AAC64308	Lin et al 1999
	Drosophila melanogaster	AAF50498	Direct sub[a]
	Saccharomyces cerevisiae	CAB66445	Direct sub[a]
	Schizosaccharomyces pombe	NP_009816	van Eenannaam et al 1999
	Entamoeba histolytica	gnl|TIGR_5759|ehistolyt_ENTLX67TR	TIGR,[d] this study[e]
	Giardia lamblia	00060.Contig602	GGSP,[g] this study[e]
	Methanobacterium thermoautotrophicum (MTH11)	O26119	Smith et al 1997
	Halobacterium halobium	O24785	Miyokawa et al 1996
	Haloarcula marismortui	P22527	Arndt et al 1990
	Pyrococcus abyssi	H75146	Genoscope
	Archaeoglobus fulgidus	O28362	Klenk et al 1997
	Methanococcus vannielii	P14022	Auer et al 1989
	Methanococcus jannaschii (MJ0464)	Q57903	Direct sub[a]
Rpp21	Human	AAK39955	Jarrous et al 1999
	Mouse	NP_080584	Shibata et al 2001
	Saccharomyces cerevisiae (Rpr2p)	AF055991	Chamberlain et al 1998
	Schizosaccharomyces pombe	T39293	Direct sub[a]
	Encephalitozoon cuniculi	>gi|17158052|ref|NC_003236.1	Katinka et al 2001
	Giardia lamblia	00060.Contig293	GGSP,[g] this study[e]
Pop5	Human	AAF17213	van Eenannaam et al 1999
	Mouse	BAB23935	Direct sub[a]
	Drosophila melanogaster	AAF49372	Direct sub[a]
	Saccharomyces cerevisiae	NP_009369	Chamberlain et al 1998
	Schizosaccharomyces pombe	T41635	Direct sub[a]
	Arabidopsis thaliana	AAB80636	Direct sub[a]
	Giardia lamblia	00060.Contig490	GGSP,[g] this study[e]
	Methanobacterium thermoautotrophicum (MTH687)	AAB85512	Andrews et al 2001
	Pyrococcus abyssi	AJ248285[e]	Direct sub[a]
	Pyrococcus horikoshii	AP000006[e]	Kawarabayasi et al 1998
	Archaeoglobus fulgidus	AE001070[e]	Klenk et al 1997
	Methanococcus jannaschii	U67499[e]	Direct sub[a]

NOTE. Alternative sequence names are given next to the species name in parentheses. Archaeal sequence names are in italics.

[a] Sequences were directly submitted to GenBank and do not have a published reference listed

[b] Sequences were recovered from the NCBI unfinished genome database

[c] Homologue within mentioned sequence was recovered with BLAST

[d] Preliminary genome sequence data was obtained from *The Institute For Genomic Research* website at http://www.tigr.org

[e] Although the genome sequence data has been obtained from the source listed, its identification has been determined in this study

[f] Preliminary genome sequence data was obtained from the Sanger Institute Pathogen Sequencing Unit: http://www.sanger.ac.uk/Projects/Protozoa

[g] Giardia Genome Sequencing Project: http://jbpc.mbl.edu/Giardia-html/index2.html

Table ASR-3.1:

to identify potential gene homologues from newly sequenced genomes, where homologues of a gene are only known from species that are phylogenetically distant from that of the sequenced organism. These ancestral sequences can also be incorporated into HMM profiles or into other algorithms to aid in other genomic homologue search strategies.

The target genomes used in this study were *E. cuniculi*, *Ent. histolytica* and *G. lamblia*. The last two genomes consist of preliminary contig data and represent new information coming from recent sequencing projects. However, preliminary contig data may also be incomplete and can test the robustness of any search strategy. We show here that the use of ancestral sequences can in some cases improve on BLAST and HMMER in the recovery of potential gene homologues, complementing existing search strategies.

Materials and methods

Ancestral sequence reconstruction

Ancestral sequences were calculated for each RNase P protein using the PAML and FastML programs. All sequences, initial and calculated, were then used in BLAST searches of the target genomes. Not all nodes (as shown in Figure 2) could be calculated for each protein owing to limited sequence availability for different proteins. As genes for many RNase P proteins have been identified in *S. pombe*, we used this genome as a control for testing the utility of ASR in homologue identification.

Trees were calculated for each set of sequences based on 18S rRNA sequence trees (sequences downloaded from the Ribosomal Database Project, http://rdp.cme.msu.edu/html/). The sequences were aligned with ClustalX (Thompson et al 1997), and the resulting tree file used as the tree-input file for the ancestral sequence calculation. Trees were constructed for each set of sequences used to calculate ancestral sequences.

Ancestral sequences were calculated using the Codeml program in PAML v3.1 July 2001 (Yang et al 1995). This program uses the marginal reconstruction approach, using most-parsimonious likelihood (Steel and Penny 2000), that compares the probabilities of different amino acids for an interior node at a site and then selects the amino acid that has the highest posterior probability. The WAG (Whelan and Goldman 2001) model and the fixed branch length option (to retain the relationship information inherent in the SSU rRNA trees) were used throughout the PAML calculations.

FastML was also used to reconstruct ancestral nodes from the sequences available for each organism (see Figure 2). FastML is a fast dynamic programming algorithm developed for maximum-likelihood joint reconstruction of the set of all ancestral amino acid sequences in a phylogenetic tree (Pupko et al 2000) (http://kimura.tau.ac.il). The same trees that were used for PAML were used with FastML. Default parameters were used throughout.

BLAST

BLAST searches of the *Giardia lamblia*, *Entamoeba histolytica* and *Encephalitozoon cuniculi* genomes were done using the BioEdit2 program (Hall 1999). The newest *Giardia lamblia* contigs were available through an online BLAST facility at http://jbpc.mbl.edu/Giardia-html/index2.html. Searches of GenBank® and other unfinished eukaryotic genome sequences were available at the NCBI (National Center for Biotechnology Information) BLAST site (http://www.ncbi.nlm.nih.gov/BLAST/). For all BLAST searches the default parameters were used. 'Back-BLASTing' takes a candidate protein sequence from the above genomes and does a BLAST search against GenBank and other genome databases available at the NCBI BLAST site listed above. This establishes whether a candidate sequence is primarily similar to other homologues of the original test sequence, or whether it just carries motifs that are similar to many other proteins. This tests the reciprocity of the BLAST search.

Genome databases

The *Giardia lamblia* Genome Project (McArthur et al 2000) is based at the Marine Biological Laboratory at Woods Hole, Massachusetts, USA (http://jbpc.mbl.edu/Giardia-html/index2.html).[1] The *Entamoeba histolytica* genome sequence data was obtained from the Sanger Institute Pathogen Sequencing Unit (http://www.sanger.ac.uk/Projects/Protozoa/ and ftp://ftp.sanger.ac.uk/pub/pathogens/E_histolytica/).

The *Encephalitozoon cuniculi* genome (Katinka et al 2001) and the *S. pombe* genome (Wood et al 2002) were available through the NCBI site (http://www.ncbi.nlm.nih.gov).

Graphs

The ln (natural log) of the resulting E values was plotted against the sequences or HMM model. For graphing

purposes, the highest value (usually the *S. pombe* protein sequence against the *S. pombe* database) was set to –150 or –200 and a 'no result' was set to 0. A 'good return' threshold for the E value was set at 1×10^{-3} (ln=–6.90).

Profile hidden Markov models

Profile hidden Markov models (Durbin et al 1998) were produced using HMMER2 (Eddy 1998) available as part of the Wise2 package (http://www.sanger.ac.uk/Software/Wise2/). Input protein alignments were constructed using ClustalX (Thompson et al 1997). To build the profile HMM from the alignment, HMMbuild was set to –f to configure the model for finding multiple domains per sequence, where each domain can be a local alignment. The HMMcalibrate program, which calibrates the profile, used default settings.

Genewisedb, available with the Wise2 package (Birney and Durban 1997), enabled searches of nucleotide databases with protein HMM profiles. Default parameters were used, except the splice parameter was set to flat and the bits cutoff set to 10 to allow for more distant sequences to be recovered.

Results

Search results for four RNase P proteins are reported for a number of genomes, with *S. pombe* serving as a control to evaluate the alternative search methods (BLAST, profile HMMs and ASR).

Pop1

Pop1 is the largest protein in the eukaryotic RNase P complex. In humans it is 115 kDa (1024 amino acids) and in the yeast *S. cerevisiae*, 100.5 kDa (875 amino acids). To date, it has been found in some eukaryotic species but not in any Archaea (see Table 1). It is thought to interact directly with the RNA component of RNase P in *S. cerevisiae* (Ziehler et al 2001) but not in humans (Jiang et al 2001). BLAST searches of the NCBI unfinished genomes database recovered full-length candidate Pop1 sequences from *Candida albicans*, *Aspergillus nidulans* and *C. briggsae* with fragments from *Cryptosporidium parvum*, *Arabidopsis thaliana*, *Plasmodium falciparum* and *Entamoeba histolytica* (sequence identifications are given in Table 1).

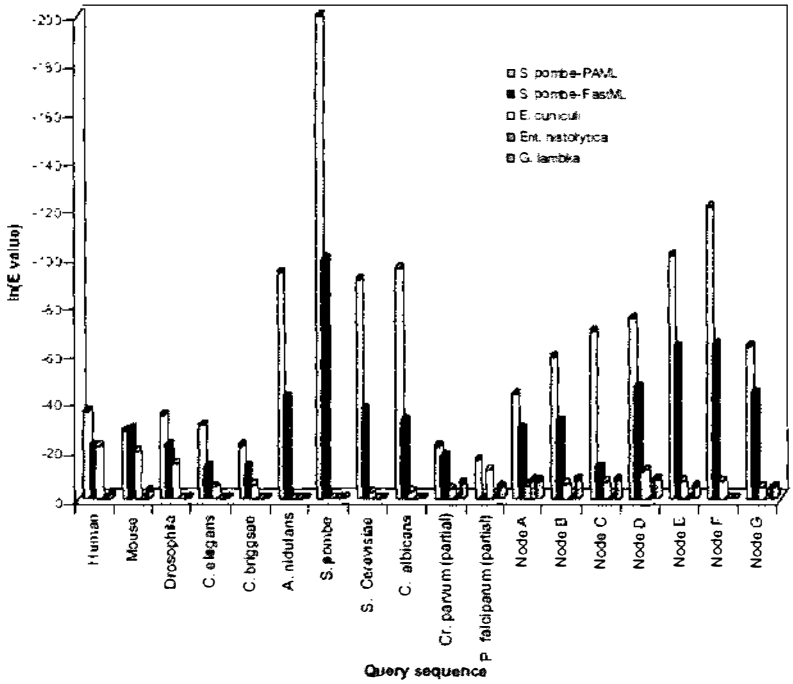
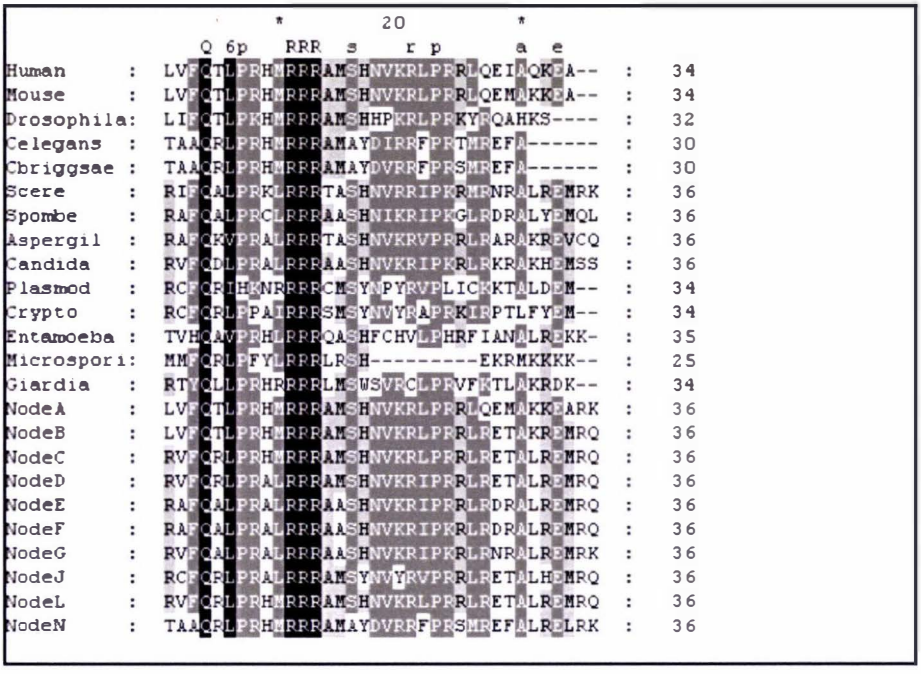


Figure 3.3:

Figure 3 Graph showing the results of BLAST searches of the *Schizosaccharomyces pombe*, *Encephalitozoon cuniculi*, *Entamoeba histolytica* and *Giardia lamblia* genomes with Pop1 known and ancestral sequences (nodes A–H). Each sequence is plotted against the natural log (ln) of the E value returned with the search of the genome. For graphing purposes, the highest result was set to –200 and a 'no result' was set to 0. The higher bars indicate a better match. Searches of the *S. pombe* genome used sequences calculated with both PAML and FastML, while searches of the other genomes used PAML only. Results indicate that ancestral sequences recovered sequences at nodes E and F with a higher probability than any of the known sequences (except for the *S. pombe* Pop1 sequence) finding itself in the *S. pombe* genome. The *G. lamblia* Pop1 homologue is recovered above the threshold (ln (1×10^{-3}) = –6.90) only with ancestral sequences.

A



B

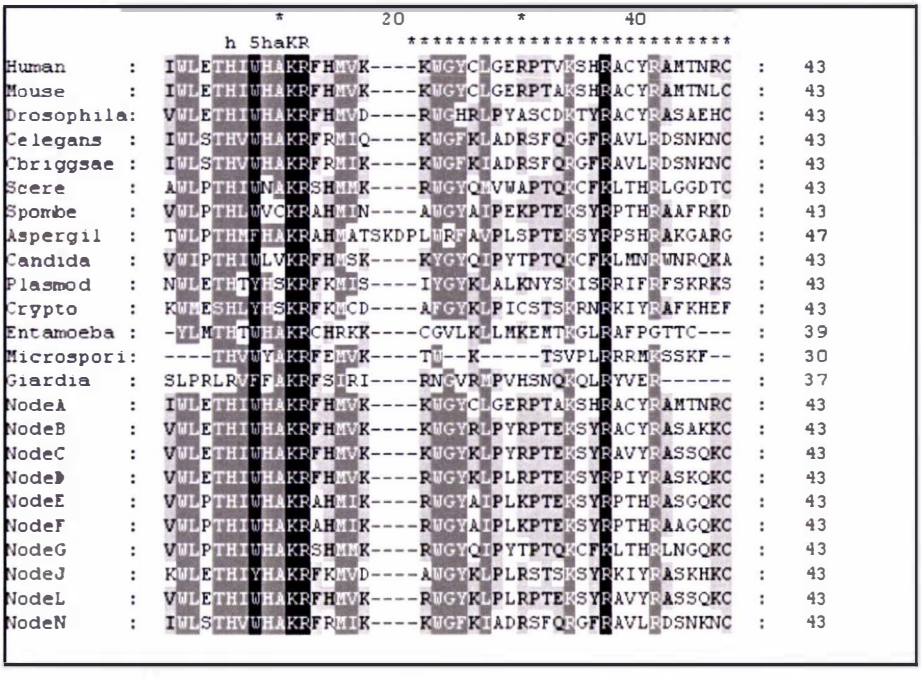


Figure 3.4:

Figure 4 Partial alignments of eukaryotic Pop1 protein sequence and the candidate Pop1 sequences from *Giardia lamblia*, *Entamoeba histolytica* and *Encephalitozoon cuniculi*. **(A)** RRR motif region. **(B)** W-Box motif region. Darker shadings represent the increased amount of homology of the amino acid position in the alignment. Key: Scere – *S. cerevisiae*; Aspergillu – *A. nidulans*; Plasmod – *P. falciparum*; Crypto – *Cr. parvum*; Microspori – *E. cuniculi*

The full-length candidate homologues were included in the ancestral sequence calculations, but the partial sequences were not.

Results of BLAST searches of the target genomes with Pop1 known and ancestral sequences are shown in Figure 3. This shows that the recovery of the *S. pombe* Pop1 sequence is successfully achieved with sequences calculated for ancestral nodes. The ancestral sequences generally recover the expected sequence with a higher E value than obtained by BLASTing the known sequences, although in some cases the actual scores can be marginally lower.[2] The highest score (other than with the *S. pombe* Pop1 sequence itself) is that from the closest node to the *S. pombe* sequence, node F. The next highest is node E, before the other fungi, *A. nidulans*, *S. cerevisiae* and *C. albicans*. When the *S. pombe* sequence is not included in the alignments used to calculate ancestral sequences, node E again recovers the expected sequence more strongly than the other yeast sequences (data not shown).

BLASTing the *G. lamblia* genome database with any of the known Pop1 sequences (including those found above) recovered no candidate sequences. However, a potential candidate was found using the ancestral sequence at node A. The *G. lamblia* Pop1 candidate was examined experimentally by RT-PCR (real-time polymerase chain reaction) and showed that this gene was expressed (data not shown), and the sequence was confirmed.

The ancestral sequences calculated with FastML were truncated to the most conserved region near the N-terminal end of the proteins, as were the original known sequences in the output alignment. These truncated ancestral sequences generally scored lower in BLAST searches than the equivalent full (PAML) ancestral sequence.

A partial *Ent. histolytica* Pop1 candidate recovered with BLAST from the NCBI unfinished genome database was also recovered from the *Ent. histolytica* database using ASR. An *E. cuniculi* Pop1 candidate sequence was recovered using BLAST, both with the known Pop1 and with inferred ancestral sequences.

Figure 4 shows the most conserved part of the Pop1 protein alignment, including the partial candidate sequences recovered from the NCBI unfinished genomes database and the candidate Pop1 sequences from *E. cuniculi* and *G. lamblia*.[3]

The Pop1 candidate sequences were 'back-BLASTed' against the NCBI databases to check for any previous designated function, and as a way of checking similarity with other Pop1 sequences. The Pop1 candidates from *G. lamblia* and *Ent. histolytica* did not recover any known Pop1 sequences nor were there any significant results with any other proteins having homology only to small lysine-rich regions. These results show that the *G. lamblia* and *Ent. histolytica* Pop1 candidates do not show any significant homology to any other sequence in the large NCBI databases. The *E. cuniculi* Pop1 candidate recovered the human and mouse Pop1 proteins from GenBank with respective E values of 0.001 and 0.004 showing more homology to other Pop1 homologues than to any other protein in the larger databases.

Pop4

Pop4 is the central protein in the eukaryotic RNase P complex, binding to many other proteins and the RNase P RNA molecule. It also plays an important part in the RNase MRP ribonucleoprotein complex (Xiao et al 2002). Ancestral nodes were calculated for Pop4, with and without archaeal sequences. Because our target protist genomes, *G. lamblia* and *Ent. histolytica*, are basal on the eukaryotic tree, archaeal sequences may share greater identity to the candidate sequences from these protist genomes than the known 'crown group' eukaryotic sequences. BLAST results for Pop4 are shown in Figure 5.

The *S. pombe* Pop4 sequence was again readily recovered; the ancestral nodes generally scoring higher than the known sequences. In contrast to Pop1, node A did not score higher than the human or mouse Pop4 sequences. A Pop4 candidate was recovered in the *G. lamblia* genome using BLAST, but only ancestral sequences successfully recovered a potential homologue in *Ent. histolytica*, although only a partial sequence was recovered; however, the genome is at this stage incomplete. Again the ancestral sequences gave better scores than the known Pop4 sequences.[4] Results from FastML were very similar to those calculated with PAML, although the sequences were again slightly truncated (results not shown).

The archaeal sequences and their respective ancestral nodes (not shown in Figure 2) did not recover any candidates from any of the target genomes. This indicates that the protist Pop4 protein candidates share a greater sequence similarity with the eukaryotic Pop4 proteins than with the archaeal sequences used in this study.

Again, the Pop4 candidate from *G. lamblia* was shown by RT-PCR to be expressed, and the sequence was confirmed (data not shown). No Pop4 candidate was recovered from *E. cuniculi* with any known or ancestral Pop4 sequence. Because this genome is considered complete, this is unexpected. As Pop4 candidates were recovered from the

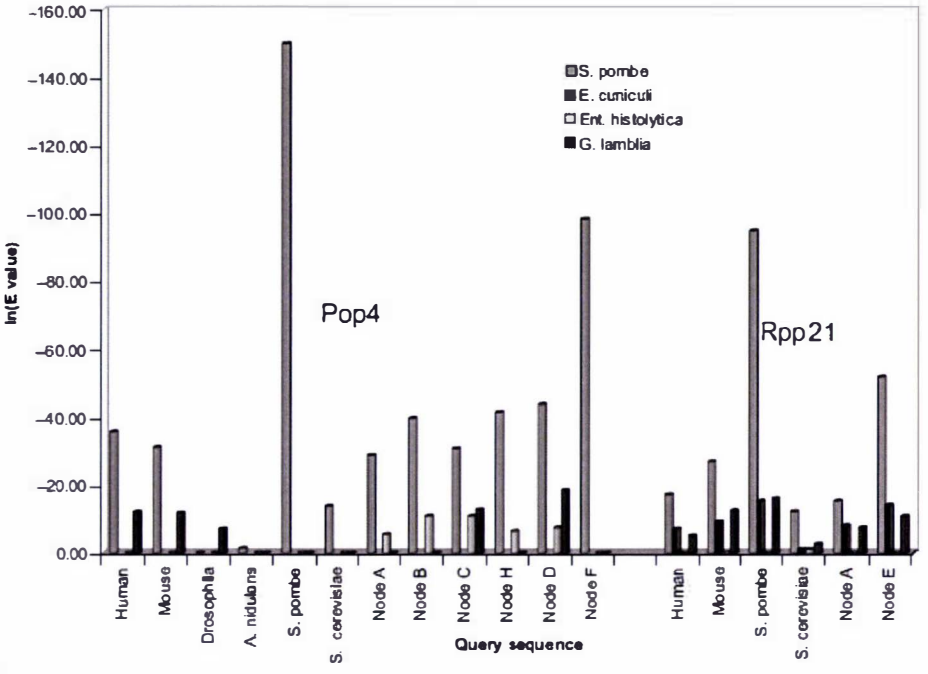


Figure 3.5:

Figure 5 Graph showing the results of BLAST searches of the *Schizosaccharomyces pombe*, *Encephalitozoon cuniculi*, *Entamoeba histolytica* and *Giardia lamblia* genomes with known and ancestral Pop4 and Rpp21 sequences. Each sequence is plotted against the natural log (ln) of the E value returned with the search of the genome. For graphing purposes, the highest result was set to −150 and a 'no result' was set to 0. The higher bars indicate a better match. The Pop4 homologue is only recovered from *Ent. histolytica* using ancestral sequences. No Pop4 homologue was recovered from *E. cuniculi*. There were no results from any genome with any of the archaeal Pop4 sequences or ancestral sequences calculated for nodes I, J, K, L, M and O. The candidate Pop4 and Rpp21 homologues from *G. lamblia* and the potential *E. cuniculi* Rpp21 homologue were recovered using known sequences. No Rpp21 homologue was identified in *Ent. histolytica*.

more basally branching protists (*Ent. histolytica* and *G. lamblia*), this suggests that Pop4 is either absent or highly diverged in the microsporidian *E. cuniculi*. The partial sequence from *Ent. histolytica* aligned only in lysine-rich regions and was not considered a strong candidate.[5] Back-BLASTing of the *G. lamblia* Pop4 candidate recovered the human, mouse, *S. pombe* and *Drosophila* Pop4 genes from GenBank. This showed that the *G. lamblia* Pop4 candidate has more similarity to Pop4 proteins from other species than to any other protein in the larger databases.

Rpp21

The Rpp21 protein has only been characterised to date in four eukaryotes (human, mouse, *S. cerevisiae* and *S. pombe*). In humans, Rpp21 binds directly to the RNase P RNA (Jiang et al 2001) but does not interact strongly with any other proteins in the RNase P complex (Jiang and Altman 2001). BLAST searching (results in Figure 3) uncovered potential homologues in *G. lamblia* and *E. cuniculi*. In both cases, the ancestral sequences readily recovered the same sequences.[6] There were no candidate homologues found within the *Ent. histolytica* genome. As there are only four known sequences, only two ancestral sequences could be calculated.[7] Back-BLASTing of the *G. lamblia* and *E. cuniculi* Rpp21 candidates returned both the human and mouse Rpp21 protein from GenBank.

Pop5

Pop5 was originally found in *S. cerevisiae* and was thought until recently to be absent in humans (van Eenennaam, Lugtenberg et al 2001). However, it has now also been identified in a number of eukaryotic and archaeal species (Auer et al 1989; van Eenannaam, Lugtenberg et al 2001; Hall and Brown 2002) (sequences in Table 1).

With the genome data available, BLAST searches with known eukaryotic and archaeal Pop5 protein sequences were unable to pick up any Pop5 candidates from the *G. lamblia* genome. However, a profile HMM model of the eukaryotic sequences recovered the first 9 amino acids of a potential candidate right on the end of a contig sequence. This was not recovered when ancestral sequences were added to the model or used in individual BLAST searches of the genome.

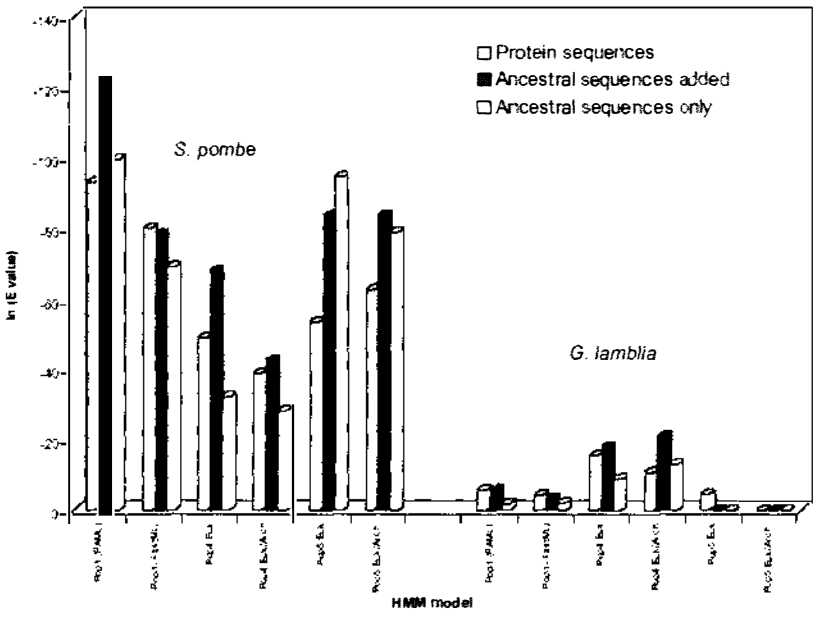


Figure 3.6:

Figure 6 Graph indicating how well each profile HMM returned the correct homologue from *Schizosaccharomyces pombe* and *Giardia lamblia*. Each HMM model is plotted against the natural log (ln) of the E value returned with the search of the genome. The higher bars indicate a better match. In most cases, the addition of ancestral sequences increased the ln(E value) thus indicating a greater possibility of finding the correct homologue.

However, when these 9 amino acids were BLASTed against a later release of the *G. lamblia* contigs (currently available at http://jbpc.mbl.edu/Giardia-html/index2.html), a full Pop5 candidate sequence was uncovered.

A BLAST search of this release of the *G. lamblia* contigs with known Pop5 sequences and ancestral sequences failed to pick up the same motif or contig. This is a case where the HMM profile outperformed a BLAST search using ancestral sequences, demonstrating that ASR does not always outperform HMM-based searching. BLAST searches of *Ent. histolytica* and *E. cuniculi* also failed to find any potential candidates.[8] As expected, back-BLASTing with the *G. lamblia* Pop5 candidate did not recover any proteins from GenBank.

Ancestral sequences and HMM

Profile HMM models were made from RNase P protein homologues, either with or without ancestral sequences and both with and without archaeal homologue sequences. The HMM models were then used to search the *S. pombe* and *G. lamblia* genomes to test if the addition of ancestral sequences helped or hindered the search for candidate homologues. As is seen in Figure 6, the profile HMMs of Pop1 sequences recovered the Pop1 sequence from the *S. pombe* genome. The truncated sequences generated by FastML gave lower scores than the full-length PAML sequences. This may be a reflection of the scoring system, where a hit over a longer region of sequence will score higher than a hit over a small but more conserved region.[9]

In all cases studied here, the addition of ancestral sequences increased the chance of finding the candidate sequence (Figure 6). When the original sequences were omitted from the profile, using only ancestral sequences for the profile HMM model the results were sometimes better and sometimes worse than when only the known sequences were used.

Testing of Pop4 and Pop5 profile HMM models showed that for these proteins the addition of archaeal sequences had little effect (Figure 6). Profile HMMs of Pop4 proteins also failed to recover a Pop4 homologue from *E. cuniculi*. Further work is required to see if this protein is present in this species or just not contained in the data.

By increasing the scores in all cases tested here, the results show that ancestral sequences can improve the sensitivity of homologue identification using profile HMMs.

Summary

We have demonstrated that ancestral sequences can be useful in searching genomic data for gene homologues when protein sequences are not well conserved or only contain

limited sequence homology. This includes cases where the genomes being searched are phylogenetically distant from known protein sequences. Ancestral sequences can also be used in combination with known protein sequences to build profile HMM models, another powerful tool in searching genome databases for gene homologues. Models should, however, be built with and without the ancestral sequences to check that the addition of ancestral sequences does not bias the model in the 'wrong' direction.

The method of calculation of ancestral sequences may not be critical for their use. We found that PAML gives ancestral sequences over the full length of the alignment, whereas FastML produces truncated sequences. Both sets of ancestral sequences gave similar results as far as finding the potential gene candidates, but the longer sequences generated by PAML provide greater confidence that one has recovered the candidate sequence, reflected in higher similarity scores. In homologue identification, the accuracy of ASR is not critical as for other applications since it is a means to an end; its utility is to improve homologue identification. The BLAST and HMMER applications also allow for mismatches in their searches, which is perhaps analogous.

For the RNase P proteins studied here, the addition of archaeal sequences did little to help find candidate eukaryotic sequences. For other proteins from basal eukaryotic lineages, archaeal sequences may nevertheless aid identification of gene homologues and should not be excluded solely on the basis of the results reported here.

In genome sequencing projects, it is desirable to identify potential gene homologues early after sequencing, often on preliminary contig data. Searching the preliminary contig data of the *E. cuniculi*, *Ent. histolytica* and *G. lamblia* genomes, as well as the complete *S. pombe* genome, provided a useful test of the utility of ASR in homologue identification. In some cases, inclusion of ASR in BLAST and HMMER searches outperformed equivalent searches with extant sequences making this a viable supplementary technique in homologue identification.

In the present study, we have not examined the effect of using different substitution matrices with BLAST; BLOSUM62 was used throughout. Using a matrix more suited to distant relationships (eg PAM250) might improve the results of BLASTing with extant sequences and may likewise improve the results using ancestral sequences. This could, for instance, be helpful in searching *E. cuniculi* for candidate Pop4 sequences.

It would also be interesting to compare BLAST using ancestral sequences with the performance of PSI-BLAST (Altschul et al 1997). PSI-BLAST produces a profile through iterative searches of a database. At each step, the generated profile is used in subsequent search iterations so can pick up new sequences not found in previous iterations. ASR could be used in combination with PSI-BLAST (at the first iteration), though it is unclear whether this would significantly improve the quality of the hits returned.

In conclusion, our results demonstrate the utility of ASR in detailed homologue identification where BLAST and profile HMMs using extant sequences have failed to recover any homologues. Furthermore, there is scope to fine-tune the approach, when needed, by making full use of the parameters available in currently available search strategies.

Acknowledgements

Many thanks to Mitchell L Sogin and Andrew G McArthur for hosting AM Poole in their lab and the laboratory team at the *Giardia lamblia* Genome Project (funded by the NIAID/NIH under cooperative agreement AI 043273), Marine Biological Laboratory at Woods Hole for access to non-public data. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for performing the RT-PCR work. This study was supported by the NZ Marsden Fund.

Notes

[1] Since the submission of this paper, genomic sequence data from the whole genome shotgun sequencing *Giardia lamblia* sequencing project has been lodged in GenBank under the accession number AACB00000000. The candidate sequences can be found as follows: Pop1 (gi|29250263), Pop4 (gi|29250923), Rpp21 (gi|29248569) and Pop5 (gi|29250880).

[2] See Supplementary Table 1, available at http://awcmee.massey.ac.nz/downloads.htm

[3] A full alignment is available as Supplementary Figure 1, available at http://awcmee.massey.ac.nz/downloads.htm. The two sequence motifs, the R-box (Supplementary Figure 1A) and the W-box (Supplementary Figure 1B) (van Eenennaam, van der Heijden et al 2001) are part of the Pop1 conserved region and are present in the candidate sequences.

[4] See Supplementary Table 2, available at http://awcmee.massey.ac.nz/downloads.htm

[5] An alignment of known Pop4 proteins and the candidate Pop4 protein from *G. lamblia* is available as Supplementary Figure 2 at http://awcmee.massey.ac.nz/downloads.htm. The *Ent. histolytica* partial sequence does not align well so was omitted from this alignment. Back-BLASTing of the candidate Pop4 protein from *Ent. histolytica* returned no significant result.

[6] See Supplementary Table 3, available at http://awcmee.massey.ac.nz/downloads.htm

[7] An alignment of known Rpp21 protein sequences and the candidate sequences from *G. lamblia* and *E. cuniculi* is available as Supplementary Figure 3, available at http://awcmee.massey.ac.nz/downloads.htm

Supplementary Figures and Tables are shown in Appendix B.

[8] An alignment of known Pop5 proteins and the candidate Pop5 protein from *G. lamblia* is presented in Supplementary Figure 4, available at http://awcmee.massey.ac.nz/downloads.htm

[9] See Supplementary Table 4, available at http://awcmee.massey.ac.nz/downloads.htm

Supplementary Figures and Tables are shown in Appendix B.

References

Altman S. 1989. Ribonuclease P: an enzyme with a catalytic RNA subunit. *Adv Enzymol Relat Areas Mol Biol*. 62:1–36.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25: 3389–402.

Andrews AJ, Hall TA, Brown JW. 2001. Characterization of RNase P holoenzymes from *Methanococcus jannaschii* and *Methanothermobacter thermoautotrophicus*. *Biol Chem*. 382: 1171–7

Arndt E, Kromer W, Hatakeyama T. 1990. Organization and nucleotide sequence of a gene cluster coding for eight ribosomal proteins in the archaebacterium *Halobacterium marismortui*. *J Biol Chem*. 265: 3034–9

Auer J, Lechner K, Bock A. 1989. Gene organization and structure of two transcriptional units from *Methanococcus* coding for ribosomal proteins and elongation factors. *Can J Microbiol*. 35:260–4

Birney E, Durbin R. 1997. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc Int Conf Intell Syst Mol Biol*. 5:56–64.

Chamberlain JR, Lee Y, Lane WS, Engelke DR. 1998. Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. *Genes Dev*. 12:1678–90

Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, Husain A. 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science*. 271:502–5

Chang BS, Donoghue MJ. 2000. Recreating ancestral proteins. *Trends Ecol Evol*. 15:109–14

Collins LJ, Moulton V, Penny D. 2000. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J Mol Evol*. 51:194–204

Durbin R, Krogh SEA, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK: Cambridge Univ Pr

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics*. 14: 755–63

Embley TM, Hirt RP. 1998. Early branching eukaryotes? *Curr Opin Genet Dev*. 8:624–9

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser*. 41:95–8

Hall TA, Brown JW. 2002. Archaeal RNase P has multiple protein subunits homologous to eukaryotic nuclear RNase P proteins. *RNA*. 8: 296–306

Jarrous N, Eder PS, Wesolowski D, Altman S. 1999. Rpp14 and Rpp29, two protein subunits of human ribonuclease P. *RNA*. 5:153–7

Jiang T, Altman S. 2001. Protein–protein interactions with subunits of human nuclear RNase P. *Proc Natl Acad Sci USA*. 98:920–5.

Jiang T, Guerrier-Takada C, Altman S. 2001. Protein–RNA interactions in the subunits of human nuclear RNase P. *RNA*. 7:937–41

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*. 414:450–3.

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*. 390:364–70.

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res*. 5:55–76.

Koshi JM, Goldstein RA. 1996. Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol*. 42:313–20.

Lin X, Kaul S, Rounsley SD, Shea TP, Benito M-I, Town CD, Fujii CY, Mason TM, Bowman CL, Barnstead ME et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*. 402:761–8.

Lygerou Z, Pluk H, van Venrooij WJ, Seraphin B. 1996. hPop1: an autoantigenic protein subunit shared by the human RNase P and RNase MRP ribonucleoproteins. *EmboJ*. 15:5936–48.

McArthur AG, Morrison HG, Nixon JE, Passamaneck NQ, Kim U, Hinkle G, Crocker MK, Holder ME, Farr R, Reich CI et al. 2000. The Giardia Genome Project database. *FEMS Microbiol Lett*. 189:271–3

Miyokawa T, Urayama T, Shimooka K, Itoh T, Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA et al. 1996. Organization and nucleotide sequences of ten ribosomal protein genes from the region equivalent to the S10 operon in the archaebacterium *Halobacterium halobium*. *Biochem Mol Biol Int*. 39:1209–20

Penny D, Poole A. 1999. The nature of the universal ancestor. *Curr Opin Genet Dev*. 9:672–7

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*. 17:890–6.

Shibata K, Itoh M, Aizawa K, Nagaoka S, Sasaki N, Carninci P, Konno H, Akiyama J, Nishi K, Kitsunai T et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature*. 409:685–90

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol*. 179: 7135–55

Steel M, Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol*. 17:839–50

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 25:4876–82

van Eenennaam H, Pruijn GJ, van Venrooij WJ. 1999. hPop4, a new protein subunit of the human RNase MRP and RNase P ribonucleoprotein complexes. *Nucleic Acids Res*. 27:2465–72

van Eenennaam H, Lugtenberg D, Vogelzangs JH, van Venrooij WJ, Pruijn G. 2001. hPop5, a protein subunit of the human RNase MRP and RNase P endoribonucleases. *J Biol Chem*. 276:31635–41

van Eenennaam H, van Der Heijden A, Janssen RJ, van Venrooij WJ, Pruijn GJ. 2001. Basic domains target protein subunits of the RNase MRP complex to the nucleolus independently of complex association. *Mol Biol Cell*. 12:3680–9

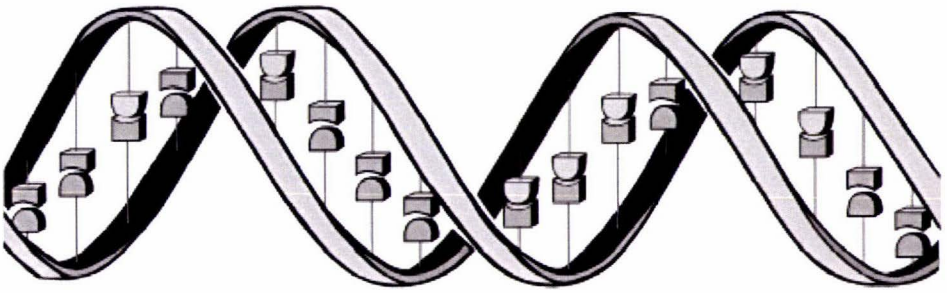
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. 18:691–9

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature*. 415:871–80

Xiao S, Scott F, Fierke CA, Engelke DR. 2002. Eukaryotic ribonuclease P: a plurality of ribonucleoprotein enzymes. *Annu Rev Biochem*. 71:165–89.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*. 141:1641–50.

Ziehler WA, Morris J, Scott FH, Millikin C, Engelke DR. 2001. A essential protein-binding domain of nuclear RNase P RNA. *RNA*. 7:565–75



Chapter 4 Splicing and the Spliceosome in the Eukaryotic Ancestor

"Everything should be as simple as possible, but no simpler"- Albert Einstein

4.1: Introduction

Most genes in 'higher' eukaryotes such as plants and animals, are interrupted by non-coding sequences (introns) that must be excised precisely from precursor mRNA (pre-mRNA) to yield mature functional mRNAs (Patel and Steitz 2003). Intron removal and the ligation of the coding sequences (exons) occurs through two sequential trans-esterification reactions carried out by a massive ribonucleoprotein complex known as the spliceosome (Nilsen 2003). The standard spliceosome (Figure 4.1[1]) is made up of five snRNPs (U1, U2, U4, U5 and U6 snRNPs) each containing a small stable RNA bound by several proteins, together with >150 less-stably associated proteins (Jurica and Moore 2003). In contrast eukaryotic ribosomes contain only 82 integral proteins (Jurica and Moore 2003), thus the spliceosome certainly deserves its appellation of "massive". In addition, the spliceosomal complex has been implicated in other cellular functions such as mRNA capping and the addition of the polyA tail (Lynch and Richardson 2002). Consequently these processes will be referred to frequently.

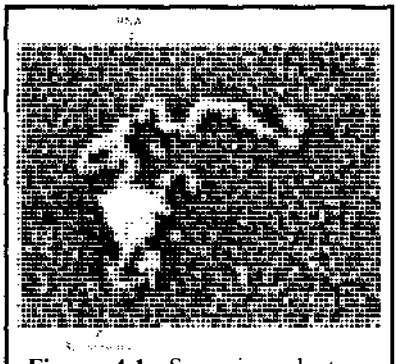


Figure 4.1: Scanning electron micrograph of a spliceosome processing an mRNA.

Introns, snRNAs and splicing-associated proteins have now been characterised in a number of basal eukaryotes (Wilihoeft et al. 2001; Archibald et al. 2002; Nixon et al. 2002) suggesting that introns and the spliceosomal machinery evolved very early in the eukaryotic lineage, and likely occur in the last common ancestor of living eukaryotes, the "Eukaryotic Ancestor". It is outside the scope of the work reported here to consider the origin of the eukaryotic ancestor or as to how ncRNAs and proteins evolved to this point; the question of interest concerns just which ncRNAs and proteins were likely to have been present.

This study takes a parsimonious approach, in that the larger the number of deep eukaryotic lineages that contained a feature, the more likely it was that the feature was present in the ancestor of those lineages. The alternative view is that a common feature arose independently in each lineage. By identifying spliceosomal features in many eukaryotic lineages, it is possible to infer the properties of their ancestor. This study looks at a range of eukaryotic species, especially the basal eukaryotes, and notes intron and splicing characteristics.

[1] The origin of this micrograph is unknown. It was found (unreferenced) on a number of internet sites including the following: http://oregonstate.edu/instruction/bb492/fignumbers/figL12-25.html.

4.1.1: Major (U2-dependent) splicing

The major-spliceosome is the predominant splicing mechanism in vertebrates, yeasts and plants and splices introns containing 'canonical' splice site characteristics (i.e. 5'splice-sites with the "GT" motif and 3'splice-sites with the "AG" motif; often referred to as having GT-AG boundaries). This type of spliceosome contains the U1, U2, U4, U5 and U6 snRNPs and also numerous associated protein factors mentioned previously. Each snRNP consists of a specific snRNA, several snRNP-specific proteins and the Sm core proteins (B/B', D1, D2, D3, E, F and G) (Labourier and Rio 2001). A summary diagram of the major splicing cycle (Gesteland et al. 1999; Nagai et al. 2001; Valadkhan and Manley 2001) is shown in Figure 4.2 and is briefly described next.

The first step in major splicing is the formation of the pre-spliceosome complex. This involves the binding of the U1snRNP to the 5'splice-site of the intron, which then promotes the binding of the U2snRNP with the branch site positioned at the other end of the intron. This binding to the branch-site results in the bulging out of an adenosine residue (often referred to as the branch-site-adenosine) from the mRNA required for the first step of catalysis (Step A). Independently, the U4snRNP binds to the U6snRNP then binds with the U5snRNP before this U4/U6.U5tri-snRNP joins the pre-spliceosome complex, thus activating it to become the 'spliceosome' (StepB1). During spliceosome activation base-pairing between the U4 and U6snRNAs is disrupted and a new base-pairing between the U2 and U6snRNAs occurs. Also the base-pairing of the U1snRNA with the 5'splice-site is exchanged for base-pairing between U6snRNA and the 5'splice-site.

After these rearrangements the U1 and U4snRNPs are released from the spliceosome prior to the first transesterification step of splicing (Step B2). In the first catalytic splicing step the 2'OH group of the bulged adenosine is activated to attack the phosphodiester bond at the 5'splice-site, resulting in the formation of a branched (lariat) intron containing a 2' to 5' phosphodiester bond and release of the exon (Step C1). The second catalytic step (Step C2) involves a nucleophilic attack by the 3' OH of the 5'exon on the phosphodiester bond at the 3'splice-site, resulting in ligation of the two exons and excision of the intron lariat.

Splicing Complex	Summary of Stage
Pre-spliceosome	U1snRNP attaches to the pre-mRNA.
A Complex	U2snRNP binds to the complex at the branch point adenosine.
B1 Complex	The U4/U6.U5 tri-snRNP complex binds to the pre-mRNA.
B2 Complex	U4snRNP leaves the complex allowing the U6snRNP to bind with the U2snRNA.
C1 Complex	Step I catalysis: separation of the mRNA at the 5' exon/intron boundary and formation of the lariat RNA.
C2 Complex	Step II catalysis: joining of the exons and excision of the lariat RNA.
I Complex	Joined exons leave the complex leaving the lariat RNA behind.
Completion	Spliceosome dissociates, snRNPs recycle and lariat RNA moves to the RNA degradation pathway or acts as a regulatory molecule.

Table 4.1: Summary of events taking place during different stages of splicing. Complex names refer to the names of stages indicated in Figure 4.2.

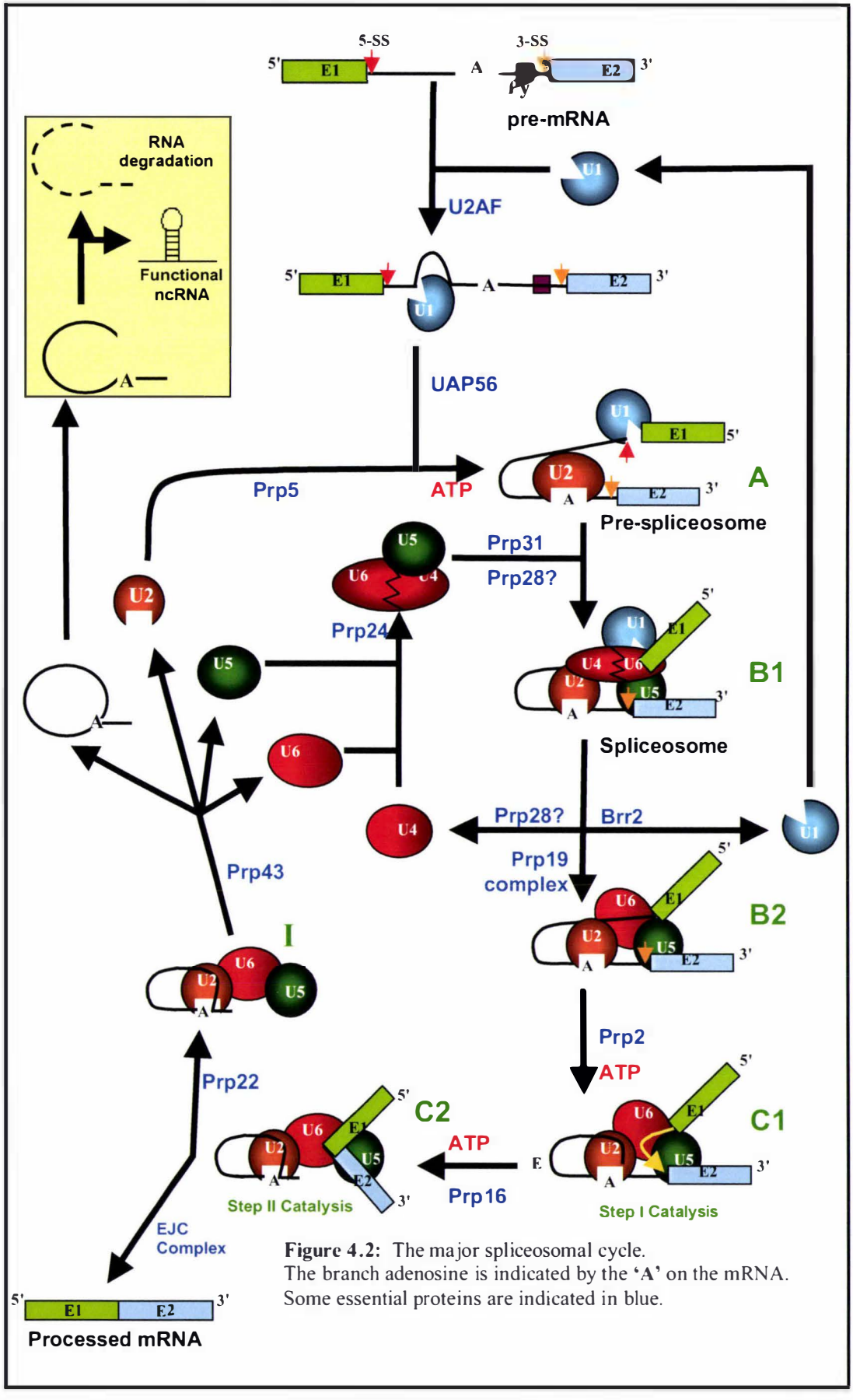


Figure 4.2: The major spliceosomal cycle.
The branch adenosine is indicated by the '**A**' on the mRNA.
Some essential proteins are indicated in blue.

Of the five snRNAs, U1 and U4snRNA leave the spliceosome before the catalytic steps occur. U2 and U6snRNAs are both highly conserved and in addition to their role in positioning the 5'splice-site and the branch site, studies suggest a crucial role for two invariant regions in the U6snRNA in splicing catalysis (Valadkhan and Manley 2000). The current data also provides evidence for participation of the U2 and U6snRNAs in both catalytic steps of splicing (Valadkhan and Manley 2003). The possibility of an RNA catalytic site in the spliceosome is strengthened further by the mechanistic and structural similarities between the spliceosome and the self-splicing groupII introns (these ribozymes are found in bacteria, basal eukaryotes and organelles of higher eukaryotes). Although catalysis in the spliceosome appears to be RNA-based, both protein-protein and RNA-protein interactions contribute to the formation to its catalytic core (Valadkhan and Manley 2003).

While the accepted view of ordered assembly has been supported by numerous studies, a number of reports have identified a variety of interactions that contradict the proposed chronology of events (Malca et al. 2003). One such study indicates that a complex of all five major snRNAs (U1, U2, U4, U6 and U5), a 'penta-snRNP' can specifically bind to the 5'splice-site through base-pairing of the 5' end of the U1snRNA. Similarly an early functional interaction between the U4/U6.U5 tri-snRNP complex and the 5'splice-site occurs independently of prior binding of U2snRNA to the branch site (Malca et al. 2003). Although the above interactions may be valid, for the purposes of this study the "accepted" spliceosomal cycle will be used as it tidily separates the different stages of splicing to enable uncomplicated analysis and it doesn't affect the analysis of splicing requirements.

4.1.2: Minor (U12-dependent) splicing

Another class of introns containing non-canonical boundary sequences has been found in jellyfish, insects, animals and plants, and is spliced by a distinct splicing machinery (Patel and Steitz 2003). The excision of these "minor" class introns is dependent on the U12 snRNP and is known as U-12 type or U12-dependent splicing (whereas the canonical "major" class introns require U2snRNA thus the process is also known as U2-type or U2-dependent splicing).

Minor spliceosomes contain a different set of snRNPs to that used in major splicing (Figure 4.3B). The U11snRNP replaces the U1snRNP, the U12snRNP replaces the U2snRNP, and the U4atac and U6atac snRNPs replace the U4 and U6snRNPs respectively. Of the snRNPs only the U5snRNP is shared between the two spliceosomes. In contrast the majority of spliceosomal proteins appear to be shared (Lynch and Richardson 2002). Although the first U12-type introns characterised had AT-AC boundaries (hence the naming of the U4atac and U6atac snRNAs) GT-AG boundaries appear to be more common (Burge et al. 1999).

In general, the minor and major snRNAs are engaged in analogous snRNA/snRNA and snRNA/pre-mRNA interactions such that a similar dynamic network is formed (Schneider et al. 2002) (Figure 4.3). The U4atac/U6atac snRNPs undergo base-pairing like that seen with the major spliceosomal U4 and U6snRNPs forming very similar secondary structures. The main difference is that unlike the separate binding of the U1 and U2snRNPs to the pre-mRNA, in the minor spliceosome the U11 and U12 snRNPs form a stable complex and interact with the pre-mRNA as such. This mechanism is suggested to prevent the formation of mixed spliceosomes (Patel and Steitz 2003).

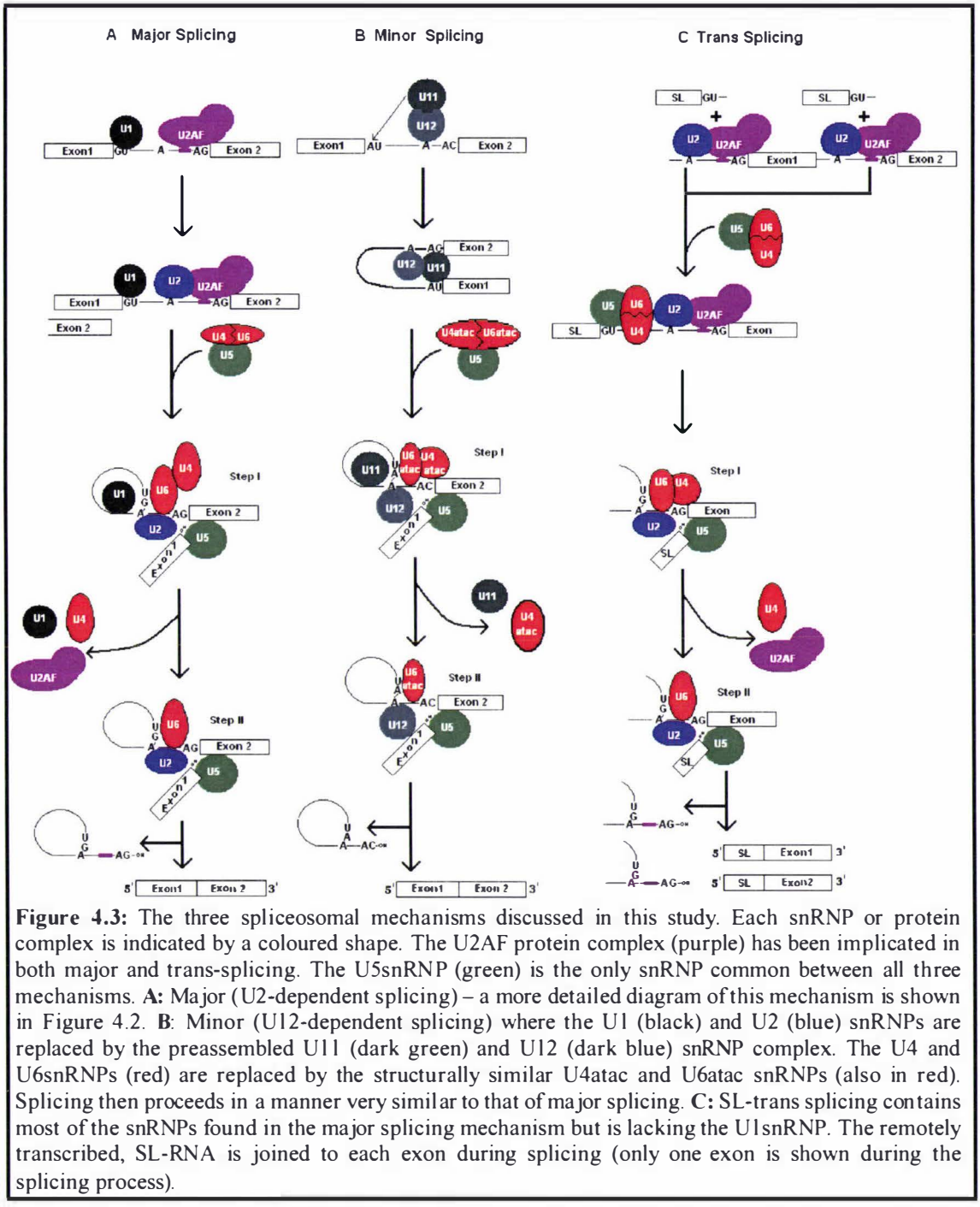


Figure 4.3: The three spliceosomal mechanisms discussed in this study. Each snRNP or protein complex is indicated by a coloured shape. The U2AF protein complex (purple) has been implicated in both major and trans-splicing. The U5snRNP (green) is the only snRNP common between all three mechanisms. **A:** Major (U2-dependent splicing) – a more detailed diagram of this mechanism is shown in Figure 4.2. **B**: Minor (U12-dependent splicing) where the U1 (black) and U2 (blue) snRNPs are replaced by the preassembled U11 (dark green) and U12 (dark blue) snRNP complex. The U4 and U6snRNPs (red) are replaced by the structurally similar U4atac and U6atac snRNPs (also in red). Splicing then proceeds in a manner very similar to that of major splicing. **C:** SL-trans splicing contains most of the snRNPs found in the major splicing mechanism but is lacking the U1snRNP. The remotely transcribed, SL-RNA is joined to each exon during splicing (only one exon is shown during the splicing process).

Recently a number of minor splicing-specific proteins have been identified from analysis of the human U11/U12snRNP (Will et al. 2004) and the fruitfly U11 snRNP (Schneider et al. 2004). U11snRNP-specific proteins, U11-25, U11-35, U11-48 and U11-59 and the U11/U12-specific proteins U11/12-20, U11/12-31 and U11/12-65 have similar sequences in the mouse and zebrafish genomes and some of these are also found in the fruitfly, mosquito and some plant genomes (Schneider et al. 2004).

A computational scan of the human genome found 404 U12 introns (Levine and Durbin 2001) and indicate that 0.34% of human introns are spliced by the minor spliceosome, a number of which occur within the larger gene families (e.g. calcium and sodium voltage-gated cation channels and the protein kinase superfamily). Although there are few U12-type introns in the genome of any given species, their presence in insects, metazoa and plants (though not *C. elegans*, *S. cerevisiae* nor another yeast *Schizosaccharomyces pombe*) indicate that the minor spliceosome must have been present in the common ancestor to plants and animals and has been lost from some lineages (Lynch and Richardson 2002; Zhu and Brendel 2003).

4.1.3: Trans-splicing

A type of splicing that may contribute to generate protein diversity is trans-splicing. This process joins exons from two independently transcribed pre-mRNAs to form a single mature transcript. The most common form of trans-splicing is where a special spliced leader (SL)-RNA is added to the 5' end of an exon. SL-trans-splicing is found in trypanosomes (e.g. *Trypanosoma brucei*, *Euglena gracialis*), cestodes (flatworms e.g. *Echinococcus multilocalaris*), nematodes (e.g. *C. elegans*) and the sea squirt *Ciona intestinalis* and is used to process a polycistronic (multi-gene) pre-mRNA to form multiple mature transcripts (Tschudi and Ullu 2002). Trans-splicing requires the U2, U4/U6 and U5snRNA as well as the SL-RNA, joining a small non-coding "mini-exon" derived from the SL-RNA to each protein-coding exon present in the pre-mRNA (Figure 4.3C). The SL-RNA secondary structure (containing three stem-loops) is similar between all organisms carrying out trans-splicing, although the nucleotide sequence is not conserved (Liang et al. 2003). The U1snRNP has been shown not to be required for trans-splicing but it is present in trypanosomes, though shorter than the equivalent snRNA in animals. (Liang et al. 2003). In *C. elegans* polycistronic pre-mRNAs are common, as opposed to the monocistronic (one transcript = one mRNA) system predominant in mammals. *C. elegans* divides its polycistronic mRNAs into discrete monocistronic mRNAs by trans-splicing SL1 (SL-like RNA) to the 5'splice-site for the first exon in the mRNA then trans-splices SL2 (similar to SL1) to the 5'splice-sites for the subsequent exons (Pirrotta 2002).

Vertebrates have been shown able to trans-splice *C. elegans* SL-RNAs (Bruzik and

Maniatis 1992; Garcia-Blanco 2003) but SL-trans-splicing in vertebrates although reported (Sullivan et al. 1991; Eul et al. 1995; Caudevilla et al. 2001; Tasic et al. 2002), is also questioned (Maniatis and Tasic 2002). Some other forms of trans-splicing have been reported, such as the "SMaRT" RNA-reprogramming system (Garcia-Blanco 2003; Mansfield et al. 2003) and "discontinuous groupII" splicing (Bonen 1993). Spliceosome-mediated-RNA-trans-splicing (SMaRT) uses the spliceosome to carry out RNA recombination using trans-splicing to replace exons within a protein (Garcia-Blanco 2003; Mansfield et al. 2003). Discontinuous groupII trans-splicing has been found in plant and algal chloroplasts and some plant mitochondria (Bonen 1993) where two exons transcribed at a distance from each other are spliced together. Alternative trans-splicing occurs in mammalian cells and involves the joining of exons from independent mRNAs without using an SL-RNA or equivalent leader sequence (Labrador and Corces 2003). This process is essentially a form of alternative splicing which is discussed later in this introduction.

The basal eukaryotic lineage, Euglenozoa (this lineage is shown on the tree in Figure 1.3 (page 4) and Figure 4.6) contains a group of eukaryotes collectively referred to in the literature as the trypanosomes (includes *Trypanosoma brucei*, *Euglena gracialis* and *Entosiphon sulcatum*), whose mRNAs are spliced predominantly by the trans-splicing mechanism (Nilsen 1995; Frantz et al. 2000; Liang et al. 2003). However a single cis-spliced intron was discovered (Mair et al. 2000; Liang et al. 2003) suggesting that these two splicing processes coexist in trypanosomes, as in all other organisms capable of SL-trans-splicing. In trypanosomes the SL-addition serves two purposes, splicing together with polyadenylation to first dissect the polycistronic mRNA, and then provide the cap to the mature mRNA (Liang et al. 2003).

There are mechanistic parallels between trans-splicing and cis-splicing including the use of the same set of nucleotide sequence features to mark splice-sites and there is structural similarity between the SL-RNAs and the spliceosomal snRNAs (Vandenberghe et al. 2001). These similarities imply an evolutionary relationship between cis-splicing and trans-splicing (Bonen 1993). However, the nature of this relationship is unclear because the phylogenetic distribution of trans-splicing has not yet been fully determined (Vandenberghe et al. 2001). The presence of trans-splicing in both crown and basal eukaryotes suggests that either it is an ancient eukaryotic characteristic or it has arisen independently in a number of lineages. As the number of lineages in which trans-splicing is known increases, the independent-origin hypothesis becomes less likely (Vandenberghe et al. 2001).

4.1.4: Exon /Intron Recognition and Alternative splicing

No matter which type of splicing is used to excise introns from an mRNA, the intron/exon boundaries of the mRNA must be recognised by the spliceosome. In vertebrates,

exons are short (average 150nt) and the introns are long (average 2300 nt) requiring that the splicing region is recognised by molecules bridging across the exon; the exon-definition model of splicing (Figure 4.4) (Berget 1995; Lam and Hertel 2002). These exons contain recognition motifs (exonic splicing enhancers -ESEs) that facilitate exon recognition by the spliceosome. Studies suggest that a group of proteins called "SR proteins" tether the U1, U2snRNPs, U2AF and other splicing factors, recruiting the pre-spliceosome to the pre-mRNA.

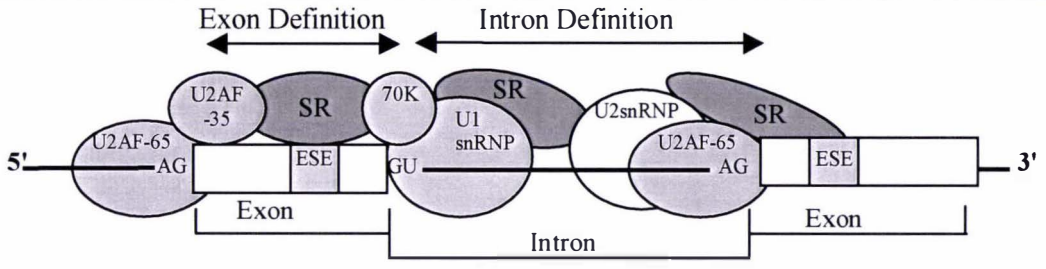


Figure 4.4: Exon and intron definition models for splice site recognition (Based on (Lam and Hertel 2002; Romfo et al. 2000 and Lorkovic et al. 2000). The boundaries between the introns and exons are recognised through the binding of multiple proteins either across the exon (exon definition) or across the intron (intron definition).

Very small exons (<30nt) can cause recognition problems for the vertebrate splicing machinery and lead to exons being spliced out with the intron (a process called exon skipping). Exons larger than 300 nt can also cause problems and also leads to exon skipping due to a lack of protein interaction across the exon (Romfo et al. 2000). Exon size is not as constrained, however, in yeasts, and unicellular eukaryotes where small introns predominate. In these organisms a mechanism of "Intron-definition" is present where introns rather than exons serve as the initial unit of recognition during splicing (Figure 4.4) (Romfo et al. 2000). Both exon and intron definition mechanisms are present in mammals, the fruitfly, plants and the yeast *S. pombe* (Lorkovic et al. 2000; Romfo et al. 2000) but it is not known yet which mechanism(s) are present in other eukaryotes.

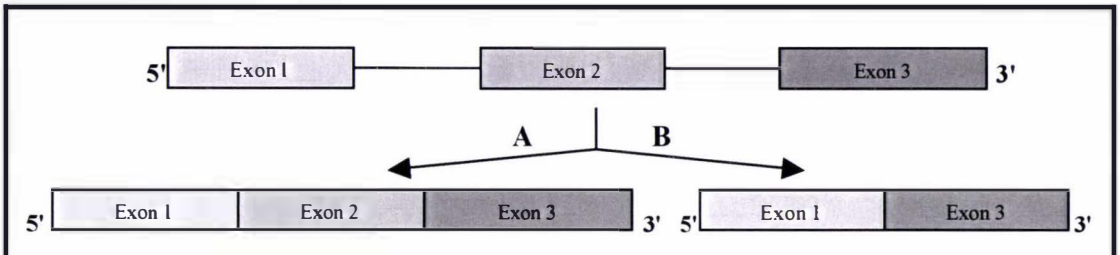


Figure 4.5: Simplified diagram describing alternative splicing. In situation A the exons are spliced together in the usual fashion but in situation B exon 2 is removed with the introns, resulting in a much shorter mature mRNA transcript.

Alternative splicing is the process by which more than one mRNA can be generated from the same pre-mRNA by the differential joining of 5' and 3' splice-sites (Figure 4.5). It is a major source of protein diversity (Maniatis and Tasic 2002; Sorek et al. 2004). The result of

this is that exons can be extended, shortened, skipped or included; and conversely introns can be removed or retained in the mRNA. The process of alternative splicing results in mature mRNAs that encode proteins with functional differences, often in a tissue-specific or developmental-stage manner. Regulatory proteins interact with specific sequences (including ESEs) within pre-mRNAs stimulating or repressing exon recognition. A large percentage of human genes appear to undergo alternative splicing (between 33 – 60%) with considerable amounts also in the fruitfly and nematode genomes (Kondrashov and Koonin 2003). Alternative splicing has been shown to be associated with all three types of splicing; major, minor (13/404 human U12 introns) (Boue et al. 2003), and trans-splicing (in the fruitfly (Maniatis and Tasic 2002; Labrador and Corces 2003) and trypanosomes (Manning-Cela et al. 2002)), and is thus considered an important mechanism in gene expression (Sorek et al. 2004).

4.1.5: Splicing and the Spliceosome in the Eukaryotic Ancestor

Investigating the distribution of splicing mechanisms and spliceosome components among present eukaryotic lineages (Results shown in Figure 4.6) can reveal how splicing and the spliceosome evolved within eukaryotes. There have been recently a number of trees representing eukaryotic evolution (Cavalier-Smith and Chao 1996; Embley and Hirt 1998; Dacks and Doolittle 2001; Simpson and Roger 2002) but there is still debate as to the placement of some lineages on these trees. Excavates are one such lineage which includes jakobids, heteroloboseans, diplomonads, retortamonads, *Trimastix* and *Carpediemonas*. While cytological data suggests that these organisms have a common excavate ancestor, there is no consensus view on the phylogenetic relationship between these species, or their relationship to other basal eukaryotes (Archibald and Keeling 2002). There is also the problem produced by present tree-building programs with deep phylogeny long-branch attraction where disproportionably long branches tend to be placed together. This has raised doubts about the position on the eukaryotic tree of some of the most "deeply-branched" or basal eukaryotes such as *Giardia lamblia* and *Entamoeba histolytica*. This creates problems when investigating RNA and protein sequences across a wide range of eukaryotes, of how to show the relationship between investigated sequences. The eukaryotic tree used during this study (Figure 4.6) is based on (Simpson and Roger 2002) and indicates which branches are affected by different hypotheses, so that any uncertainty can be taken into account when drawing any conclusions.

Recent studies from human and yeast spliceosomes (Jurica and Moore 2003) have characterised the large number of proteins that contribute to the spliceosomal complex. To determine whether it was likely that any or some of these proteins were present in the eukaryotic ancestor three basal eukaryotic genomes (*Plasmodium falciparum*, *Entamoeba histolytica* and *Giardia lamblia*) were computationally searched for protein splicing factors.

The genome of the microsporidian, *Encephalitozoon cuniculi* (*Ecz. cuniculi*) was also used for searches as it represented a highly reduced genome between the animals and yeast, as microsporidia are thought to have branched early within the fungi (Vivares et al. 2002).

4.2: Materials and Methods

Information about the taxonomic distribution of the snRNAs was gained from the Rfam database (Griffiths-Jones 2003; http://rfam.wustl.edu/), Genbank and other NCBI databases (http://www.ncbi.nlm.nih.gov) and the literature. The detection of the U5snRNA from some basal eukaryotes is described in Chapter 2. Intron taxonomic information was taken from the literature. Intron length was classified into three levels; high (>1000 nt), medium (50-1000 nt) and short (<50 nt). The presence of minor (U12-type), alternative and trans-splicing was extracted from the literature. snRNA candidate sequences are given in Appendix C.1.

The genomes of *Giardia lamblia* and *Encephalitozoon cuniculi* were downloaded from NCBI (http://www.ncbi.nlm.nih.gov). The *Plasmodium falciparum* genome was downloaded from PlasmoDB (Bahl et al. 2002) as was the *Entamoeba histolytica* genome produced by the Pathogen Sequencing Unit at the Sanger Institute (ftp.sanger.ac.uk/pub/pathogens/E_histolytica/). Protein database searches at NCBI started with the known proteins from human, *S. cerevisiae* and *S. pombe* and used the associated BLink function which displays the graphical output of pre-computed BLASTP results against the protein non-redundant (nr) database (http://www.ncbi.nlm.nih.gov/sutils/static/blinkhelp.html). Protein homologues were also recovered from 'KOG' (eukaryotic orthologous groups ftp:ftp.ncbi.nih.gov/pub/COG/KOG) (Koonin et al. 2004), a subset of the 'Clusters of Orthologous Groups' (COG) database (Tatusov et al. 2001).

Protein homologues were selected with the following criteria; proteins either had to have been experimentally confirmed as being the homologue of a query protein as determined either by the associated literature or within the GenPept file itself (designated "E" in the results tables); annotated as being similar in sequence to the query protein (designated "S" in the results tables); or a hypothetical open reading frame (ORF) with a BLink score greater than 300 and with a length no less than 75% (or no more than 25% greater) of the query protein (designated "H" in the results tables). Protein (sequence and annotation) and genomic search results data was managed using the "SpliceSite" database developed for this study (described in detail in Appendix E).

Genomic searches of many splicing proteins used the BLAST program (Altschul et al. 1997). Results were ranked (1-4; 1 having the highest confidence of validity) on the following system: A 'candidate' sequence of similar length (within 100 amino acids) to the query protein sequence and containing greater than 65% amino acid similarity was ranked highest as '1'. A candidate sequence of similar length to the query protein sequence and containing 50% to

65% amino acid similarity was ranked '2'. A candidate sequence (which may be of a different length to the query protein) but containing a protein motif present in the query sequence was ranked '3'. Candidates that displayed low sequence homology but across the whole protein length were ranked '4'. If no significant results were returned for a query protein against a genome the result was designated "-". In the situation where a query protein from different species returned different sequences from the target genome (e.g. human proteinA returned sequence1, but the homologous proteinA sequence from *C. elegans* returns sequence2), then the result was designated "?" for indicating that the result was unclear. All candidate sequences were "Back-BLASTed" against the protein databases at NCBI, the genomes from which they were recovered, and from a database of proteins contained in the SpliceSite database (Appendix E). Back-BLASTing could confirm a sequence's candidacy but also reveal any other closely related protein. Protein candidate sequences are given in AppendixC2.

Table 4.2: Letter codes used for the eukaryotic species used in this study. These codes are used in tables shown throughout this chapter.

	Species	Common Name	Euk lineage
Hu	Homo sapiens	Human	Primates
Mu	Mus musculus	House mouse	Rodents
X	Xenopus laevis	African clawed frog	Amphibians
Z	Danio rerio	Zebrafish	Teleostei (fish)
Ci	Ciona intestinalis	Sea squirt	Ascidia
Ce	Caenorhabditis elegans	Nematode worm	Nematoda
Dr	Drosophila melanogaster	Fruitfly	Arthropods
Sc	Saccharomyces cerevisiae	Bakers' yeast	Fungi
Sp	Schizosaccharomyces pombe	Fission yeast	Fungi
Nc	Neurospora crassa	Bread mould	Fungi
Ec	Encephalitozoon cuniculi		Microsporidia
Ar	Arabidopsis thaliana	Thale cress	Land plants
Os	Oryza stativa	Rice	Land plants
Pf	Plasmodium falciparum	Malaria parasite	Apicomplexa
Py	Plasmodium yoelii yoelii	Mouse parasite	Apicomplexa
Tp	Trypanosoma brucei		Euglenozoa
Lm	Leishmania major		Euglenozoa
D	Dictyostelium discoideum	Slime mold	Myxogastrids
Gt	Guillardia theta nucleomorph		Cryptophytes
Eh	Entamoeba histolytica		Entamoebae
Gl	Giardia lamblia		Diplomonads

The Ancestral Sequence Reconstruction technique (Collins et al. 2003) was used on a selected number of proteins that could be reliably aligned. Ancestral sequences were predicted using PAML (Yang 1997) then combined with BLAST to search genomic databases. This technique is described in more detail in the published manuscript included in Chapter 3. Results of snRNA and protein searches are shown in tables throughout this study. Information from published comparative genomic studies, that included some splicing proteins (Anantharaman et al. 2002; Koonin et al. 2004), has been included in the results tables. The

seven eukaryotic genomes searched in Koonin et al. 2004 are human, *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *S. pombe*, *A. thaliana* and *Ecz. cuniculi*.

Protein presence was traced to the eukaryotic ancestor using MacClade version 4.0 (http://macclade.org/). The phylogenetic tree from Figure 4.6 was used for all three runs. The following settings were used: Run A: {E, S, H, 1, 2, 3} = 1; {-, 4} = 0; {?} = ?. Run B (slightly stricter): {E, S, 1, 2, 3} = 1; {-, 4} = 0; {?, H} = ?. Run C (strict): {E, S, 1, 2} = 1; {-, 4, H, 3, ?} = 0. The likely presence of a protein in the eukaryotic ancestor was scored as follows: 3 = protein highly likely present in eukaryotic ancestor (Ancestor positive), 2 = protein likely present in eukaryotic ancestor (Ancestor equivocal), 1 = protein low likelihood of being in eukaryotic ancestor (Ancestor negative but protein present in at least 2 basal eukaryotic lineages (i.e. lineages outside animals, yeast or plants). MacClade results are shown in each of the results tables.

4.3: Results and Discussion

4.3.1: Intron presence and length in the eukaryotic ancestor

The presence of spliceosomal introns has been previously described for a number of basal eukaryotic lineages and are collated and summarised in Table 4.3. This further supports the premise (Lynch and Richardson 2002) that spliceosomal introns and some form of spliceosomal splicing were present in the eukaryotic ancestor.

Another characteristic of introns is their length. Intron length is divided into three groups in this study (short (<50 nt), medium (50nt to 1kb) range and long introns (> 1kb)) with the distribution of these groups shown on the tree in Figure 4.6. Although short introns are present in mammals, data from the Xpro database (Gopalan et al. 2004) indicate that they are not common. Humans have a mean intron length of 2.3 kb, Mouse 1.1kb and Rat 733nt, but *C. elegans* has a mean length of only 300 nt, while in the plant *Arabidopsis thaliana* genes have a mean intron length of 171nt (Table 4.3). These statistics are dependent on the distribution of the data that has been lodged in Genbank and may not reflect yet, the true distribution of intron sizes in the actual genomes (Gopalan et al. 2004). The yeasts *S. cerevisiae* and *S. pombe* also have shorter introns than mammals with the average lengths 270 nt and 57 nt respectively.

Many of the introns described for the basal eukaryotes are classed in our system as being short. Only one short intron (35bp) has been found to date in the Diplomonad *Giardia lamblia* (Nixon et al. 2002) containing a canonical 3'splice-site but a non-canonical 5'splice-site. Short canonical introns have been found in *Carpediemonas mambranifera*, a free-living relative of *G. lamblia* (Simpson et al. 2002). Introns that are shorter still have been found in the ciliate *Paramecium tetraurelia* (18-35 nt) (Dessen et al. 2001) and the Chlorarachniophyte algae CCMP621 nucleomorph (18-20 nt) (Gilson and McFadden 1996). Our study of intron

length depends largely on computer prediction, which in turn is based on extensive experimental intron research. This research is not yet to a stage with the basal eukaryotic species that splice-site prediction software can be applied accurately and thus introns may be either missed or incorrectly annotated in basal genomic annotation. From the analysis shown here it is feasible that the intron length in the eukaryotic ancestor was within the short (<50 nt) and medium (50nt to 1kb) range, with long introns (> 1kb) arising later in some lineages.

Intron frequency does not appear to be conserved across eukaryotic lineages. Introns have been predicted in 54% of *Plasmodium falciparum* genes, a proportion roughly similar to that in *S. pombe* and *Dictyostelium discoideum* but much higher than observed in *S. cerevisiae* and *Cryptosporidium parvum* where only 5% of the genes contain introns (Anantharaman et al. 2002; Gardner et al. 2002). Elevated AT% (% of nucleotides A or T) content is a characteristic of introns (when compared to exonic sequences) from a number of basal eukaryotes including *Ent. histolytica* (Wilihoeft et al. 2001) and *D. discoideum* (Rivero 2002).

Species	Ref	Group	Intron Length	5' splice-site	3' splice-site	Branch Point	Py Tract
Section A							
Human	X	Vertebrate	2300av	AG/GTAAGT	CAG/GT	CTGAC	Py tract
C. elegans	X	Vertebrate	300av	GT	TTTCAG	TTT(C/G)AA	Py rich
Drosophila	X	Vertebrate	826av	GTATGT	YAG	CTAAT	Py rich
Arabidopsis	X	Land Plant	171av	AAG/GTAAGT	TTGCAG/GT	CTGAT	UA-rich
S. cerevisiae	X	Yeast	270av	TGT	YAG	TACTAAC	None
S. pombe	X	Yeast	57av	GT	AG	CTRAY	Py rich
Ecz. cuniculi	m	Microsporidia	129av	NS	NS	NS	NS
Section B							
Chlamydomonas reinhardii	j	Chlorophytes	79-297	GT	AG	CTCAC	NS
Entamoeba histolytica	c	Entamoebae	46-115	GTTTGTT	TAG	YNYYRAY	NS
Entamoeba dispar	c	Entamoebae	46-84	GTTTGTT	TAG	YNYYRAY	NS
Dictyostelium discoideum	d	Dictyostellids	70-150	AG/GTAAGT	ATAG/	TACTAAY	weak Py
Paramecium tetraurelia	k	Ciliates	18-35	NS	NS	NS	NS
Cryptosporidium parvum	l	Apicomplexa	566av	NS	NS	NS	NS
Plasmodium falciparum	h	Apicomplexa	NS	GT	AG	NS	NS
Giardia lamblia E	b	Diplomonad	35	CT	AG	AACTAAC	NS
Carpediemonas E membranifera	i	Carpediemonas	31-33	GT	AG	TYCTTAT	NS
Reclinomonas E americana	a	Core Jakobid	67-145	GT	AG	NS	NS
Acrasis rosea E	a	Heterolobosea	NS	NS	NS	NS	NS
Euglena gracialis E	e	Euglenozoa	50-9200	GT	AG	NS	NS
Trypanosoma brucei E	g	Euglenozoa	302-653	GT	AG	NS	NS
Entosiphonsulcatum E	f	Euglenozoa	102	GT	AG	CGTCGAT	Py tract
Malawinomonas E jacobiformis	a	Malawinomonas	58-127	GT CT	AG	NS	NS

Table 4.3: Intron characteristics from eukaryotes. **Section A:** Characteristics of introns from the 'crown group' of eukaryotes (animals, plants and yeast). **Section B:** Characteristics of introns from basal eukaryotes. **Key:** "NS" indicates that although at least one intron has been described for this species, this particular characteristic has not been determined and is not stated in any literature to date. Intron length is given in number of nucleotides. "av" is the average intron size as determined in the Xpro database **X** (Gopalan et al. 2004) or literature. Intron lengths are given in the number of nucleotides (nt). Splice-site characteristics divided by '/' indicate nucleotides on either side of splice-site boundaries; 5' splice-site -exon/intron boundary and 3' splice-site: intron/exon boundary. **E** -indicates that this species is part of the Excavate basal eukaryotic lineage. Data for this table has been collated from the following literature: **a**-(Archibald et al. 2002), **b**-(Nixon et al. 2002), **c**-(Wilihoeft et al. 2001), **d**-(Rivero 2002), **e**-(Breckenridge et al. 1999; Canaday et al. 2001), **f**-(Ebel et al. 1999), **g**-(Djikeng et al. 2001), **h**-(Huestis and Fischer 2001), **i**-(Simpson et al. 2002), **j**-(Watanabe and Ohama 2001), **k**-(Dessen et al. 2001), **l** -(Abrahamsen et al. 2004) ,**m**-(Katinka et al. 2001). Crown eukaryotic information came mostly from the Xpro database (Gopalan et al. 2004) **X**.

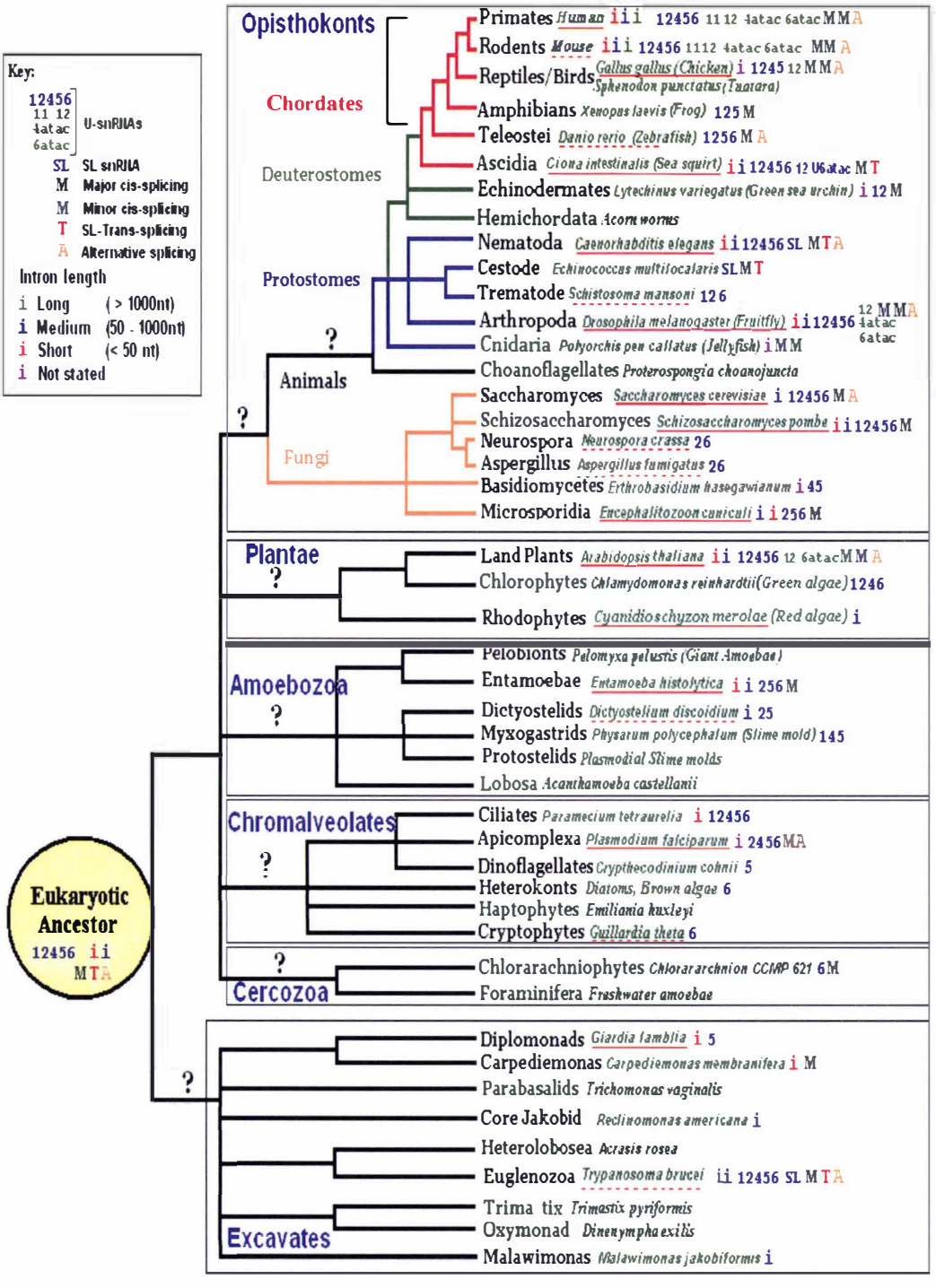


Figure 4.6: Distribution of introns, snRNPs and modes of splicing in eukaryotes. This tree is based on the tree shown at http://hdes,biochem.dat.ca/Rogerlab/, including information from Simpson and Roger 2002. A representative species for each lineage are shown in green. Branches that join differently in alternative hypothesis are indicated with "?". The animal, fungal (Opisthokonts) and plant groups are often referred collectively as the 'Crown' eukaryotes, while the rest (Amoebozoa, Chromalveolates, Cercozoa and Excavates) are collectively called the basal eukaryotes. Protostomes (blue lineages) and Deuterostomes (green lineages) are grouped based on differences in embryo development. Within the Deuterostomes are the chordates (red lineage) which in turn contain the vertebrate lineages (as marked on the tree).
Intron, snRNA and splicing characteristics present in the eukaryotic ancestor are determined from their distribution over eukaryotic lineages. References additional to those in Table 4.3 and the text are (Takahashi et al. 1996; Spafford et al. 1999; Dodgson 2003; Tombes et al. 2003; Matsuzaki et al. 2004). This tree will be referred to frequently throughout this chapter as it is central to much of the reasoning.

AT% is only slightly elevated in the introns from other species including *C. parvum*, *P. falciparum*, *S. pombe*, *S. cerevisiae* and *Ecz. cuniculi* (Anantharaman et al. 2002) and may only be a consequence of containing a pyrimidine-tract (Py-tract) or Uracil-enriched region. At this stage, elevated AT% cannot be attributed to introns in the eukaryotic ancestor.

The presence of short introns in basal eukaryotes indicates that the intron definition mechanism of splice site recognition may be predominant for these eukaryotes. However, some introns described from Euglenozoa (e.g. *Euglena gracialis*) are long (> 1kb), one as long as 9.2kb (Canaday et al. 2001), which may cause problems for the intron-definition mechanism requiring instead some form of exon-definition. Since a number of eukaryotes (fruitfly, *S. pombe* and plants) contain both an intron- and an exon-definition system of splice site recognition, it is not impossible that both mechanisms exist in basal eukaryotes and that both systems were present in the eukaryotic ancestor.

4.3.2: snRNAs in the eukaryotic ancestor

Computational searches (including Rfam (Griffiths-Jones et al. 2003), NCBI databases; and BLAST searches of basal eukaryotic genomes) combined with an intense literature search indicated the distribution of snRNAs and their associated proteins throughout the eukaryotic tree. A conclusion from the present work is that all five major spliceosome snRNAs (U1, U2, U4, U5 and U6 snRNA) are found throughout the eukaryotic tree in both crown and basal eukaryotes (Table 4.4 and Figure 4.6) and thus were likely to have been present in the eukaryotic ancestor. However some of the snRNAs found in basal eukaryotes (e.g. U1 and U5) are shorter in length than their crown eukaryote equivalents, and missing helices, suggesting that the structure of the snRNAs within the eukaryotic ancestor may not be the same as is found in living eukaryotes.

MacClade A	B	C	EA	snRNA	Hu	Mu	X	Z	Ci	Cc	Dr	Sc	Sp	Nc	Ec	Ar	Os	Pf	Tp	Lm	D	Eh	Gl
2	2	2	***	U1snRNA	E	E	E	S		E	E	E	E			E	S		E	E			
3	3	3	***	U2snRNA	E	E	E	E	2	E	E	E	E	S	2	E	E	2	E	E	S	2	
2	2	2	***	U4snRNA	E	E			2	E	E	E	E			E	S	S	E	E			
2	2	2	***	U5snRNA	E	E	E	E	2	E	E	E	E		S	E	S	S	E		2	2	1
3	3	3	***	U6snRNA	E	E		S	2	E	E	E	E	S	S	E	S	H	E	E		E	
-	-	-		U11 snRNA	E	E			2		E					E							
-	-	-		U12 snRNA	E	E			2		E					E							
-	-	-		U4atac snRNA	E						E												
-	-	-		U6atac snRNA	E				2		E					E							
2	2	2	***	SL RNA					E	E									E	E			

Table 4.4: Eukaryotic distribution of snRNAs. **Key:** please refer to the key on page 93. An empty cell indicates that this type of snRNA has not to date been found in this species. Species names are shown in full in Table 4.2. Animals are shown in blue (vertebrates in darker blue), yeasts are in red, microsporidia is in brown, plants are in green and the basal eukaryotes are in purple. The U5snRNA identified from *G. lamblia* is described in Chapter 2.

Trans-splicing has been characterised in nematodes, cestodes, the sea-squirt *Ciona intestinalis* (crown eukaryotes) and in trypanosomes (excavate lineage, basal eukaryotes) indicating that trans-splicing was likely to be present in the eukaryotic ancestor. The present definition of SL-trans-splicing requires the presence of an SL-RNA, thus the presence of trans-splicing in the eukaryotic ancestor requires also some form of SL-RNA to have been present.

The minor-spliceosome snRNAs (U4atac, U6atac, U11 and U12) have been found in humans and the fruitfly *Drosophila melanogaster* (U12 and U6atac snRNAs only have been characterised from the plant *Arabidopsis thaliana*). BLAST searches of the *P. falciparum, Ent. histolytica, G. lamblia* and *Ecz. cuniculi* genomes with known minor-spliceosomal snRNA sequences failed to recover any potential candidates. However, since known minor-snRNA sequences are evolutionarily distant from the basal eukaryotes used in this study, searches may require more powerful techniques. Potential candidates for the U6atac and U12 snRNAs were recovered from the sea-squirt genome, which, besides having major splicing has also been shown to contain trans-splicing (Vandenberghe et al. 2001). If these candidate minor-snRNA sequences are genuine then the sea-squirt would be the first species to contain all three types (major, minor and trans) of splicing.

Finding snRNAs in genomic sequences is not easy; often there is little or no sequence similarity between snRNAs from distantly related sequences. For example, in Chapter 2 U5snRNA candidates were recovered from the *G. lamblia, Ent. histolytica and D. discoideum* genomes only with the use of specialised RNA-finding software. It is likely that such software will also be required to find other snRNAs (major, minor and SL) in basal eukaryotes before we can resolve their true distribution and consequently their condition in the eukaryotic ancestor.

4.3.3: Splicing mechanisms in the Eukaryotic Ancestor

Both the major and minor spliceosomal splicing predate the separation of animals and plants (Lynch and Richardson 2002), however at present, we are unable to determine if minor splicing evolved any earlier than the animal-plant ancestor. Most of the proteins associated with the major spliceosome snRNAs appear to be shared between both types of cis-splicing (Lynch and Richardson 2002). The major and minor spliceosomes exhibit similar salt stability, suggesting that the structural organisation of these particles may be similar (Schneider et al. 2002). However, one protein (U11-35kD) has been characterised that appears to be part of the minor spliceosome but not present in the major spliceosome (Will et al. 1999).

One theory for having more than one type of cis-splicing (i.e. both major and minor splicing) put forward by Patel and Steitz (2003), is that the removal of U12-type (i.e. minor) introns may be a rate-limiting step in pre-mRNA processing. The conversion of a U12-type

intron to a U2-type (spliced by the major-spliceosome) markedly increased (~sixfold) the amount of mature mRNA and protein expression (Patel and Steitz 2003). Another potential role of U12-type introns may be to change splicing speed in response to stresses in the cellular environment (Patel and Steitz 2003). Gene regulation can also take advantage of competition between major and minor splicing. Alternative splicing in the fruitfly *prospero* pre-mRNA produces one form (*pros-L*) with major-splicing in early embryogenesis and another (*pros-S*) with minor-splicing later in embryo development (Scamborova et al. 2004). The discovery of this type of regulation is the first step in elucidating the mechanism involving competition between the major and minor spliceosomes (Scamborova et al. 2004).

There are three models proposed for the evolution of U12-type introns and the minor-splicing mechanism (Schneider et al. 2002); co-divergence, fission-fusion and parasitic invasion. In the first two models the U2 and U12 systems diverged from an ancestral spliceosome possibly following duplication. In the convergence model, each of the major and minor splicing mechanisms evolved separately in the same nucleus. With the fission-fusion model each system developed further in separate lineages but was then reunited through a merging of genetic material before the ancestor of animals and plants (Burge et al. 1998). In the third model -parasitic invasion, where a GroupII intron invaded a number of genes of a common ancestor of metazoa and plants which had a pre-existing spliceosome. Subsequent fragmentation of the new intron gave rise to novel snRNAs and the utilisation of many of the proteins of the pre-existing spliceosome. The distribution of minor splicing among eukaryotic lineages cannot indicate which of these evolutionary processes is more likely to have occurred. U12-type introns have not yet been characterised in any basal eukaryotes but this may not mean that it is not present, just not yet detected. At this stage minor splicing is not thought to have been present in the eukaryotic ancestor but evolved before the ancestor of animals and plants. Any discovery of any U12-type introns in any basal eukaryotic lineage will of course allow this view to change.

Trans-splicing in the wider sense (i.e. the joining of two independently transcribed exons) is considered to be in the eukaryotic ancestor due to it's presence in both basal and crown eukaryotic lineages. SL-trans-splicing is predominant in the excavate trypanosomes (see Figure 4.6 for its position on the eukaryotic tree) but is also found in some crown eukaryotes (discussed previously) indicating that this type of trans-splicing may also be present in the eukaryotic ancestor. Although an SL-RNA has not been identified in humans it is possible to induce SL-trans-splicing if the appropriate leader sequence is added (Vandenberghe et al. 2001; Garcia-Blanco 2003), indicating that the necessary machinery may be present in humans and other crown eukaryotes. The true phylogenetic range of SL-trans-splicing presently remains unknown as each discovery of SL-trans-splicing was a fortuitous

result of a detailed study of particular genes (Vandenberghe et al. 2001), and because these extensive studies have only been carried out in a small number of eukaryotes.

Alternative splicing is associated with all three types of splicing (major, minor and trans-splicing) (Maniatis and Tasic 2002; Manning-Cela et al. 2002; Boue et al. 2003). It's presence in both crown and basal eukaryotic lineages suggest that alternative splicing may have been present in the eukaryotic ancestor. As alternative splicing appears to increase protein diversity (Boue et al. 2003) this mechanism may have been important in the diversification of eukaryotes subsequent to the eukaryotic ancestor.

The distribution of major, minor and trans-splicing across the eukaryotic lineages as shown on Figure 4.6 indicates that multiple splicing mechanisms are common in today's eukaryotes. It is possible that multiple splicing mechanisms also existed in the eukaryotic ancestor. Information is emerging about the interaction of splicing mechanisms and how the different spliceosomes interact (Scamborova et al. 2004) and future studies may indicate how this interaction may have evolved but this is beyond the scope of the present work.

4.3.4: Spliceosomal proteins in the eukaryotic ancestor

The next step is to search for specific proteins known to be associated with the spliceosome. A search of protein and nucleotide databases with known human, *S. cerevisiae* and *S. pombe* spliceosomal proteins (Kaufer and Potashkin 2000; Lorkovic et al. 2000; Zhou et al. 2002; Jurica and Moore 2003) found likely homologues in other eukaryotes. More detailed searches of three basal eukaryotic genomes (*P. falciparum*, *G. lamblia* and *Ent. histolytica*) and the microsporidian *Ecz. cuniculi* recovered potential homologues of the most conserved spliceosomal proteins. The object of these searches was, firstly to determine if enough spliceosomal components could be found in both crown and basal eukaryotic lineages to indicate that a spliceosome was present in the eukaryotic ancestor; and secondly, if a spliceosome was present determine its complexity (i.e. was it a simplified version of today's spliceosomes or just as complex). Over 150 of the most conserved spliceosomal proteins were examined in this study and were grouped based on common snRNA-binding properties (e.g. the U1snRNA-specific proteins) or containing common distinguishing sequence motifs (e.g. Sm/Lsm proteins). Results are summarised under the different protein group headings on the following pages. As each protein group is required to be examined in full the reporting of this information can become somewhat repetitive. This unfortunately[2] is necessary to unravel the complexity of today's spliceosomes and thus comprehend ancestral spliceosomal characteristics. Each group of proteins is reported separately, beginning with the groups of snRNP-associated proteins then onto proteins that have other functions in the spliceosome.

[2] Please refer to Figure E.4 and perhaps feel some sympathy for the researcher who had to evaluate this information.

Table 4.5A: U1snRNP associated proteins

MacClade A	B	C	U1snRNP			Protein Name		Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Nc	Ar	Os	Pf	Py	Tp	Ec*	Pf*	Eh	Gl
			EA	K	An	Human	Scere																			
2	2	1	**	K-E		U1-70	SNP1	E	E	E		S	S	H	E	E	H	E	S	S	S	E	?	2	?	?
2	2	2	***	K	Euk-S	U1-A	MUD1	E	S	S		H			E	E	H	E	S	S	S	E	1	2	2	4
2	2	1	**	K		U1-C	YHC1	E	E	E	E	H	H		E	E	S	S		H	H	E	1	2	3	-
1	1	-	*	K		FBP11	Prp40	E					H		E	E	H						2	2	3	-
-	-	-					Snu56								E								-	-	-	-
1	1	1	*		Euk-C		Nam8								E		H						2	2	2	4
-	-	-					Snu65								E								-	-	-	-
-	-	-		K-E	Sc		Snu71								E								?	4	-	-
-	-	-					Usa1								E								-	-	-	-
1	1	1	*		Euk		Prp39								E	E		S	S				3	2	?	1

Table 4.5B: U2snRNP associated proteins

MacClade A	B	C	U2snRNP			Protein Name			Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Os	Pf	Py	Tp	D	Ec*	Pf*	Eh	Gl
			EA	K	An	Human	Scere	Spombe																					
3	3	1	**	K		SAP155	HSH155	Prp10	E	E	E		H	H	H	E	E	H	H	H		H	H	H	H	2	1	3	2
2	1	1	**	K	Euk	SAP145	CUS1	Sap145p	E	S			S	H	H	E	E	S	S	H		H	H			1	-	2	1
1	1	-	*	K		SAP130	RSE1	Prp12p	E	H		S	H	H	H	E	E	H		S	H				H	4	1	3	-
2	1	1	**	K-E		SAP114	Prp21	Sap114p	E	S		S		H	H	E	E	H		H		H	H			2	2	2	?
3	3	1	***	K		SAP62	Prp11	Sap62p	E	H	S	S	H	H	H	E	E	H	S	S	H	S	S	H		1	1	1	1
2	1	1	**	K		SAP61	Prp9	Sap61	E	H			H	H	H	E	E	S		E	H	S	S	H		2	2	1	-
3	3	2	***	K		SAP49	HSH49	Sap49p	E	S	S	S	S	H	H	E	E	S		S	H	S	S	H		2	2	2	1
2	1	1	**	K-E		U2-A'	LEA1	U2-A'	E	E		H	H	H	H	H	E	H		E		H	H	E		-	4	?	2
1	1	1	**	K		U2-B''	MSL1	U2-B'	E	E	H				H	E	E			E						2	3	2	2
2	2	1	**			p14	SNU17		E				E	H	H	E	H		H	H	H	H	H	S		2	2	?	2
-	-	-					Ist3									E										?	3	4	-
1	1	1	**			SF3b14b	Rds3		E				H	H	H	E	H	H		H	H					2	2	4	2

Table 4.5C: U5snRNP associated proteins

MacClade A	B	C	U5snRNP			Protein Name		Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Os	Pf	Py	Tp	Lm	D	Gt	Ec*	Pf*	Eh	Gl
			EA	K	A	Human	Scere																							
3	3	3	***	K	Euk	U5-220	Prp8	E	S	S	H	H	H	H	E	E	H	S	S	S	S	H	S	S	S	S	1	1	2	1
2	2	1	**	K	Euk	U5-200	Brr2	E	E			H	H	H	E	E	H		S	H	S	H					1	1	1	1
2	1	1	**	K	Euk	U5-116	Snu114	E	E	S	S		H		E	E	H		S		S	H					?	1	2	2
2	2	2	***		Euk	U5-102	Prp6	E	H				H	H	E	E	H		S	S	H	S			S		?	2	1	?
3	2	2	***H	K-E	Euk	U5-100	Prp28	E			H	H	H	H	E	E	H		S	S	S	S	H		S		2	2	2	?
2	1	1	*			U5-52	Snu40	E	S	S		H	H	H	H	H	H		S			H					-	2	2	-
2	1	-	*			U5-40		E	S	S	S	H	H	H		E	H		S		S	H			H		?	1	?	?
3	3	3	***	K-C		U5-15	DIB1	E				H	H	H	E	E	H	S	S	S	S	H	S	S			2	1	1	1
-	-	-					Aar2								E												-	-	-	-
1	-	-				PSF		E	S		S																3	3	3	3
1	-	-				P54nrb		E																			3	3	3	3

Table 4.5D: U4/U6.U5 tri-snRNP associated proteins

MacClade A	B	C	U4/U6.U5 tri-snRNP		Protein Name			Hu	Mu	X	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Pf	Py	Ec*	Pf*	Eh	Gl
			EA	K	Human	Scere	Spombe																	
2	1	1	**	K-E	SART-1	Snu66	Snu66p	E	E	S	H	H	H	E	E	H		S	H	S	-	2	3	-
2	1	1	**		Tri--65	SAD1	Sad1p	E	H		H	H	H	E	E	H		H	H	H	4	2	1	?,3[a]
1	-	-			Tri--27			E	S		H	H	H		S	H			H	H	-	2	-	-
1	-	-		K-S			Prp38	S	S		S	S	S		E	S	S	S	H	H				

Table 4.5E: U4/ U6snRNP associated proteins

MacClade A	B	C	U4/U6snRNP			Protein Name			Hu	Mu	X	Ce	Dr	An	Sc	Sp	Nc	Ar	Pf	Py	Tp	Lm	D	Gt	Ec*	Pf*	Eh	Gl
			EA	K	A	Human	Scere	Spombe																				
3	2	1	**	K-E	Euk	U4/U6-90	Prp3	Prp3	E		H	H	H	H	E	E	S	H	H	S		H	S		-	2	2	-
2	1	1	**	K-E		U4/U6-60	Prp4	Prp4	E		S	H	H	H	E	S		S			H				?	3	2	?
1	1	1	**	K-E	Euk/Arch	U4/U6-61	Prp31	Spp13p	E	E		S			E	E			H		H				2	?	2	2
1	1	1	**C			U4/U6-20	Cpr1	USA-Cyp1	E						E	E									2	2	2	2
1	1	1	**			U4/U6-15.5	Snu13	Snu13p	E						E	E	S							S	2	?	2	2
-	-	-			Euk	Prp24	Prp24		E						E										?	2	-	-
1	1	-				RY-1			E	S						S									-	3	3	-
-	-	-					Spp41								E										-	-	4	-
-	-	-					Snu23								E										-	4	-	-

Table 4.5F: Sm/ Lsm proteins

MacClade A	B	C	Sm/ LSm			Protein Name		Hu	Mu	X	?	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Os	Pf	Py	Tp	Lm	D	Gt	Ec*	Pf*	Eh	Gl
			EA	K	A	Human	Scere																							
3	2	2	***			Sm B/B'	SmB1	E	S	S	E	S	E	S	E	E	H		S	S	H	H	E		S		-	-	2	1
3	3	3	***	K	Euk	Sm D1	SMD1	E	E	H	E	S	S	H	E	E	H	S	S	S	S	S	E	S		S	2	1	-	1
3	3	3	***	K	Euk	Sm D2	SMD2	E	H			S	S	H	E	E	S	S	S	H		S	E			S	2	1	2	1
3	3	2	***		Euk	Sm D3	SMD3	E	E	H	S	H	S	H	E	E	S		S	H	S	H	E		H		2	1	2	2
3	3	3	***	K	Euk	Sm E	SME1	E	E		S		H	H	E	E	H		S		S	S	E	E		S	1	-	-	2
2	2	1	**	K		Sm F	SMX3	E	E			S	H	H	E	E	S	S	S		S	S	E			S	2	1	-	?
1	1	1	*	K	Euk	Sm G	SMX2	E	E			S	H	H	E	E	H		S	S	S	S	E				2	1	-	-
3	3	3	***	K		LSm2	LSM2		E			S	H	H	E	E	H	S	S	S	S	S	S	E			2	2	2	1
2	1	1	*	K	Euk-C	LSm3	LSM3	E					H	H	E	E	H			H	S	S	E				2	-	2	-
2	2	1	**	K	Euk-A	LSm4	LSM4	E	E			S	H	H	E	E	H	S		S	S		E		S		2	2	-	2
1	1	1	*		Euk	LSm6	LSM6	E	E			S	H	H	E	E					S	S	E				-	-	-	2
1	1	1	*	K-E	Euk	LSm7	LSM7		H				H	H	E	E	H		S			S	E				-	4	2	-
1	-	-	*	K	Euk	LSm1		E	H			S	H	H	E	E	S		S								2	-	-	-
1	1	1	*	K-E	Euk	Lsm5		E				S	H	H	E	E	S				S		E				-	-	2	4
1	1	1	*	K-E	Euk-C	LSm8		E	S			S	H	H	E	E			S		S	S	E				-	1	-	-

Table 4.5G: Catalytic Step II and late acting proteins

MacClade A	B	C	Catalytic step II and late acting proteins			Protein Names		Hu	Mu	Ce	Dr	An	Sc	Sp	Nc	Ar	Os	Pf	Py	Tp	Pf*	Eh	Gl
			EA	K	A	Human	Scere																
1	1	1	*H	K	Euk	Prp16	Prp16	E	S		H	H	E	E	S			S	S		3	2	3
1	1	1	*H		Euk	Prp22	Prp22	E				H	E	E	H	S		S		H	2	2	3
1	1	1	*H		Euk	Prp43	Prp43	E	E	H	H	H	E	E						H	2	2	3
1	1	1	*H	K-E	Euk	Slu7	Slu7p	E	S	H	H	H	E	E	H	S					2	2	-
2	1	1	*H	K		Prp17	Prp17	E	S	H	H	H	E	E	H	S	S			H	2	3	?
-	-	-				Prp18	Prp18	E	S		S		E	E	H	S	H	S			2	-	4

Table 4.5H: Proteins associated with the Minor spliceosome

MacClade A	B	C	Minor-Splicing	Hu	Mu	Z	Ci	Ce	Dr	An	Ar	Os	Ec	Pf	Eh	Gl
-	-	-	U11-25	E	S	S	2				S	S	-	-	-	-
-	-	-	U11-35	E	S	S	2			S	S	S	4	4	-	-
-	-	-	U11-48	E	S	S	3						-	-	-	-
-	-	-	U11-59	E	S	S	-						-	-	-	-
-	-	-	U11/U12-20	E	S	S	2		S	S			-	-	-	-
-	-	-	U11/U12-31	E	S	S	3			S	S	S	-	-	-	-
-	-	-	U11/U12-65	E	H	S	2		S	S	S	S	-	-	-	-
-	-	-	C114	E	S		2	H	S	S			-	-	-	-
-	-	-	YB1	E	E		2						-	-	-	-
-	-	-	Toe-1	E	E		2	H					-	-	-	-

Table 4.5I: SR proteins

MacClade A	B	C	SR Proteins EA	K	A	Protein Names Human	Scere	Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Nc	Ar	Os	Pf	Py	Tp	Pf*	Eh	Gl
-	-	-		K-E		SRp75		E	H		S	S	H	H						S	S		3	-	-
-	-	-				SRp54	SFRS11	E	H		S	H	S	H				S	S			H	-	-	-
1	1	-		K-E	Euk-A-S	SRp55		E	H	S			S	H									3	3	3
-	-	-				SRp40	Srp40p	E	E	S					E	E	S						-	-	4
-	-	-				SF2	ASF	E	H		S	E	E	H				E	S	S	S		2	-	4,-[a]
-	-	-			C,D,H	9G8		E	H									E	S				2,2[a]	-	4,-[a]
1	1	-			Euk-S-Sp	SC35		E	S	H	S		E	H		E		E	S	S	S	H	3	4	-
1	1	-				SRp30c		E	H									E	S				3	3	3
-	-	-			C,D,H	hTra2		E	S	S	H		E	H									3	?	-
-	-	-				SRp20		E	E	H	S												3	-	-
1	1	-		K-E		SRm300		E	S														3	3	-
1	1	-				SRm160		E	E				H	H				S					3	3	-

Table 4.5J: Spliceosomal proteins also associated with other cellular events

MacClade A	B	C	Other cellular events EA	K	Protein Names Human	Scere	Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Ec	Ar	Os	Pf	Py	Lm	D	Pf*	Eh	Gl
1	1	1	**H		UAP56	SUB2	E	S	H	H		H	H	E	E		H		S	S			2	1	2
2	-	-			TAT-SF1	CUS2	E	H			H	H	H	E	E	H	S	H	H	H			?	3	-
1	1	1	*		SKIP	Prp45	E	S						E	E		S						2	2	-
1	1	-			THO2	Rlr1	E	S		S	E	H	H	E	S	H	H	H					3	3	-
-	-	-			HPR1		E	S						E									-	-	-
1	1	1	*		hPrp4 kinase		E	E		S	S	H	S								S		H	?	3	2
1	1	1	*		TEX1		E	H			H	H	H	E			H						2	2	?
2	1	-			XAB2	SYF1	E	H			H	H	H	E	S	H	H	H	H	H			3	3	-
1	1	-			CA150		E	S				H	H				H	H					3	3	3
-	-	-			CF I-68		E	S			H	H	H										-	-	-
1	1	1	*		CF I-25		E	S		H	H	H	H		H		H		S				-	2	-
-	-	-			ASR2B		E	E		E	H	H	H										-	3	-
-	-	-	C	K-E	Aly	YRA1/Rai	E							E	E								-	4	2
2	2	2	***		PABP	PAB1	E		E		E	E		E	E							E	2	3	2

Table 4.5K: Proteins associated with the Prp19 complex

MacClade A	B	C	Prp19 complex (NTC) EA	K	Protein Names Human	Scere	Hu	Mu	Z	Ce	Dr	An	Sc	Sp	Nc	Ar	Os	Pf	Py	D	Pf*	Eh	Gl
1	1	1	**	K	CDC5L	CEF1	E	S	S	H	H	H	E	E	H						2	2	3
2	1	1	**H	K-E	Prp5	Prl1p	E	H		H	H	H	E	E	H	H	H	H	H	H	2	2	3
2	-	-	*		fSAP33	ISY1	E	H		H	H	H	E	E	H	H		H	H		?	2	-
1	1	1	*		hCrn	CLF1	E	H	S	H	E	H	E	S	H	S					1	2	-
1	1	1	*		PLRG1	Prp46	E	S		H	H	H	E	E	S	E					1	3	2
2	2	-	**	K-E	Prp19	Prp19p	E	S	H		H	H	E	E	H	S	S	H	H	H	2,2[a]	3,3[a]	3,3[a]

Table 4.5L: Proteins associated with the Exon-Junction complex

MacClade A	B	C	EJC-associated EA	K	A	Protein Names Human	Scere	Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Os	Pf	Py	D	Pf*	Eh	Gl
1	1	-				Y14		E	S										S	E				3,3[a]	3,4[a]	-,-[a]
2	2	2	***			Magoh		E	H	E		S	E	H		S	H		S	E	S	S		2	2	-
-	-	-		K		RNPS1		E	S		H	S	H	H		E	H							?	4	-
-	-	-		K-E	Euk	CBP80	Sto1	E				H	H	H	E	S	S		E	S				-	-	-
2	2	2	***	K	Euk	CBP20	Cbc2	E	S	E	S	S	S	H	E	S	E	S	E	S	S	S	S	2	?	-

Table 4.5M: Other DExD/DExH proteins

MacClade A	B	C	Other DExD/H proteins		Protein Names		Hu	Mu	X	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Os	Pf	Py	D	Pf*	Eh	Gl
			EA	A	Human	Scere																		
1	1	-	*H	Euk		Prp2	E	E			H		E						H			?	2	3
1	1	1	*H		Abstrakt		E	S		S	E	H		S			E	S	H			1	2	2
1	1	-	*H		p68	DPB2	E	S	E		H	H	E	E	H	S	H	H				?	2	3

Table 4.5N: Other splicing proteins

MacClade A	B	C	Other Splicing Proteins			Protein Names		Hu	Mu	X	Z	Ce	Dr	An	Sc	Sp	Nc	Ec	Ar	Os	Pf	Py	Tp	D	Pf*	Eh	Gl
			EA	K	A	Human	Scere																				
1	1	1	**	K	Euk	U2AF65	MUD2	E	E	S		E	E	H	E	E	H		E	E	S	S			?	?	2
2	2	2	***	K-S	Euk	U2AF35		E	S		E	H	S	H	E	E	E	S	E	E	S	S	E		2	1	-
2	2	1	**	K		SF1	Msl5	E	H	S		E	E	H	E	E	H		S		S	S			2	2	L
1	1	1	**			fSAP118		E	H				H	H											2	2	1
1	-	-		K-E		TIP39	Ylr424w	E	S			H	H	H	E	H	H		S	H	H			S	-	-	-
-	-	-			Euk-C-S	SPF45		E							E										?	-	-
-	-	-		K		LUC7A	Luc7p	E							E	E									-	-	3
2	1	1	**			MFAP1		E	S	S	S	S	H	H		S	H		S		S	H			1	2	-
2	2	1	**H			IFN4	FAL1	E	H		H	H	H	H	E	S	S		S	S	S	S			2	2	2
1	1	1	**			RHA		E				S	H	H					S	S					2	2	3
1	1	1	**			CCAP2		E				H	H	H	S	S	H		H		S			S	2	2	-
-	-	-				SPF31		E	S			S	H	H		E		H	H		S	S			2	4	?
-	-	-				RED		E	H		S	S	H	H					H						2	-	-
1	1	-				PUF60		E	H				E	H					S						?	3	3
-	-	-				DGSI		E	S			H	H	H					H						3	-	-
1	1	1	**			fSAP15	YCR903	E		E	S	H	H	H	E	S	H	S	S	S	S	S	S	S	2	-	4
-	-	-				fSAP29		E	S	S		H	H	H		H			H		H	H			2	4	4
1	1	-				OTT		E	S			H	H	H											-	3	3
-	-	-				IMP3		E	E	S	S	S	E	H		S			H	S					2	-	-
-	-	-				fSAP94		E		H	S	H	H												2	-	-
1	1	-				fSAP59		E	S		S	H	H	H											2	3	3
2	-	-				GCFC		E	S			H	H	H		S	H		H	H	H	H			-	3	-
1	1	-				fSAP57	PFS2	E							E										2	3	3
1	1	-				fSAP164		E	E				H	H	S	H					H	H			2	3	?
-	-	-				fSAP11		E	S			H	H						H						-	2	-
1	1	-				fSAPa		E	H			H	H	H					H						2	-	3
1	1	-				fSAP113		E																	2	3	3
1	1	-				SPF38		E																	2	3	3
-	-	-				SPF27		E								E									2	-	-
1	1	-				CrkRS	CTK1	E	H			H	H	H	E										3	3	3
-	-	-				fSAP18		E	H				H	H											-	-	-
-	-	-				fSAP105		E	H			H	H	H					S	H					-	-	-
-	-	-				fSAP121		E					H			E			H						-	-	-
-	-	-				fSAP79		E	S	S	S	H	H	H					H						-	-	-
-	-	-				fSAP24		E	H	H			H	H		H									-	-	-
-	-	-				SPF30	Spf30	E																	-	-	-
-	-	-				SNP70		E	H				H	H											-	-	-
-	-	-				NAP		E	S			S	H	H		S	H		H		H	H			-	-	-
1	1	-				ZNF207		E				H	H	H					H			H			3	-	3
-	-	-				fSAP71		E	H	S			S	H					H						3	-	-
-	-	-				WTAP		E																	3	-	-
-	-	-				fSAP152		E																	-	3	-
1	1	-				SHARP		E	S				S												3	3	3
-	-	-				CIRP		E	E	S															?	4	4
-	-	-				FBP3		E																	3	-	-
1	1	-				fSAP35		E					H	H											3	3	4

Key:

EA The number of * indicates the increased likelihood of the protein being present in the eukaryotic ancestor (* - low likelihood to *** - very high likelihood).
H indicates that this protein is a member of the RNA-Helicase protein family.
C indicates that this protein is a member of the cyclophilin protein family.

K: Data from Koonin et al. 2004. (genomes searched are indicated in methods section)
- K Protein present in all 7 genomes.
- K-E All genomes except *Ecz. cuniculi*.
- K-S All genomes except *S. cerevisiae*.
- K-C All genomes except *C. elegans*.

An
- Euk Protein present in all eukaryotic genomes tested in (Anantharaman et al. 2002).
- Euk-S Protein not found in *S. cerevisiae*.
- Euk-C Protein not found in *C. elegans*.
- Euk-A Protein not found in *A. thaliana*.
- Euk-A-S Protein not found in *A. thaliana* or *S. cerevisiae*.
- Euk-S-S Protein not found in *S. cerevisiae* or *S. pombe*.
- C,D,H Protein found *C. elegans, D. melanogaster* and humans only.
- Arch Protein also found in Archaea

This study:
- E Experimental evidence.
- S Sequence similarity
- H Hypothetical sequence.

An empty cell indicates that this protein has not been identified in this species nor could a homologous sequence be recovered.

Candidate sequences are ranked 1-4 with 1 having the highest confidence (Rankings are explained in full in the methods section).

a indicates that the results were found using the ASR technique.

? indicates that the results were uncertain.

- indicates that a candidate sequence could not be found.

Species names are shown in full in Table 4.2.

Colour coding:

Animals	blue (vertebrates in darker blue)
Yeasts	red
Microsporidia	brown,
Plants	green
Basal eukaryotes	purple.

Results from BLAST searches of basal eukaryotes are shown in black.

Ec* and Pf* separate these candidate sequences obtained using local BLAST searches from sequences in from these organisms in purple that were obtained from the literature and other sources.

A column with no entries is not shown.

MacClade Results

Run A: {E, S, H, 1, 2, 3} = 1; {-, 4} = 0; {?} = ?.
Run B: {E, S, 1, 2, 3} = 1; {-, 4} = 0; {?, H} = ?.
Run C: {E, S, 1, 2} = 1; {-, 4, H, 3, ?} = 0.

Scores:
- 3 = protein highly likely present in eukaryotic ancestor (Ancestor positive).
- 2 = protein likely present in eukaryotic ancestor (Ancestor equivocal).
- 1 = protein low likelihood of being in eukaryotic ancestor (Ancestor negative but protein present in at least 2 basal eukaryotic lineages.

U1snRNP-specific proteins: *(Table 4.5A)*

U1-A, Nam8, U1-C, U1-70, Prp40, Snu56, Snu71, Usa1, Prp39 and U11-35

The U1snRNP binds to the mRNA in the pre-spliceosome and leaves the spliceosome before the first step of catalysis (see Figure 4.2). It's function in the early stages of splicing is in the recognition of the 5'splice-site on the pre-mRNA (Labourier and Rio 2001). Although the Sm-core group of proteins (B/B', D1, D2, D3, E, F and G) are associated with the U1snRNP, these proteins also bind to the other snRNPs and will be covered in the Sm/Lsm core proteins section. Results for the U1snRNA-specific proteins are shown in Table 4.5A and are described below.

The U1-70 and U1-C proteins interact with the Sm-core proteins during U1snRNP assembly (Nelissen et al. 1994) whereas U1-70 and U1A interact with the U1snRNA (Labourier and Rio 2001). The U1-A protein also plays a role in cap modification in trypanosomes linking splicing and polyadenylation in these excavates (Tschudi and Ullu 2002). U1-A candidates were found in *G. lamblia, Ent. histolytica* and *P. falciparum*, having previously been found in animals, yeast and plants and thus was likely present in the eukaryotic ancestor. U1-C has also been found in crown eukaryotes and candidate sequences were recovered from *Ecz. cuniculi, P. falciparum* and *Ent. histolytica* but a candidate could not be found in *G. lamblia*. It's presence in crown and basal eukaryotes suggest that U1-C was also present in the eukaryotic ancestor. The U1-70 protein binds to the AFS/SF2 splicing factor to promote binding to the 5'splice-site (Forch et al. 2003) and contributes greatly to the exon-definition mechanism of splice site recognition. It has been found across crown eukaryotes and a candidate was found in *P. falciparum*. However, the results from *Ent. histolytica* and *G. lamblia* were unclear. This protein has been characterised from trypanosomes (Tschudi and Ullu 2002). U1-70 was also placed in the eukaryotic ancestor.

Yeasts contain a number of additional U1snRNP-specific proteins. Prp40 aids in the addition of the U2 snRNP to the pre-spliceosomal complex but is also involved in the export of proteins out of the nucleus (Murphy et al. 2004). There is some similarity between the yeast Prp40 protein and FBP11 protein from humans (Allen et al. 2002) but FBP11 has not yet been implicated in splicing. Prp40 has been found *S. cerevisiae*, *S. pombe* and *N. crassa* and a candidate sequence was found in *Ecz. cuniculi* indicating that Prp40 was likely present in the fungal ancestor. However, candidates were also found in *P. falciparum* and *Ent. histolytica* indicating a possibility that Prp40 could have been present in the eukaryotic ancestor.

The Prp39 protein (Lockhart and Rymond 1994) and is required to stabilise the U1snRNP complex to the 5'splicesite. It has been found in yeasts and plants but not in animals and candidate sequences were found in *Ecz. cuniculi*, *P. falciparum* and *Ent. histolytica*. This suggests that Prp39 may have been present in the eukaryotic ancestor and have either been lost from animals, or no longer contain enough sequence similarity to be considered protein

homologues. The *S. cerevisiae*-specific proteins (Snu56, Snu65, Snu71 and Usa1) have not been found to date in any other eukaryote (Gottschalk et al. 1998) and were not found in any basal eukaryotes during this study. Such findings are reassuring in that these proteins act as negative controls (i.e. to ensure that not every protein was found in basal eukaryotes).

U2snRNP-specific proteins: *(Table 4.5B)*
Sap155, Sap145, Sap130, Sap114, Sap62, Sap61, Sap49, U2-A', U2-B'', Ist3, Rds3, SF3b14b and SF3b10b

The U2 snRNP binds to the branch-site of the pre-mRNA early in splicing resulting in the bulging out of the branch-site-adenosine and completing the pre-spliceosome (see Figure 4.2). Results for the U2-specific proteins are shown in Table 4.5B. The majority of the U2snRNP-specific proteins belong to two U2snRNP-specific protein complexes (SF3a and SF3b). The first protein complex, SF3a consists of Sap61, Sap62 and Sap114 (Will et al. 2001). None of the SF3a proteins are found in the minor spliceosome, indicating that either SF3a is not a component of the minor spliceosome (or conversely, have been disassociated during minor spliceosome preparation) (Will et al. 1999). All of the three SF3a proteins (Sap61, Sap62 and Sap114) have been characterised from crown eukaryotes and candidate sequences were found in basal eukaryotic genomes (Sap62 was recovered from *P. falciparum*, *Ent. histolytica* and *G. lamblia* whereas Sap61 and Sap114 were recovered from *P. falciparum* and *Ent. histolytica*).

The other complex, SF3b (containing the P14, Sap49, Sap130, Sap145 and Sap155, Rds3/SF3b14b and SF3b10 proteins) has been shown to be present in both the major and minor spliceosome (Golas et al. 2003). Two proteins of the SF3b complex, P14 (interacts directly with Sap155 and also interacts directly with the branch site adenosine of the intron) and Rds3 (required for stable U2snRNP recruitment to the spliceosome) (Wang and Rymond 2003) are well conserved across eukaryotic species and are good candidates for being present in the eukaryotic ancestor. Other SF3b proteins, Sap155, Sap145 and Sap49 are also conserved across eukaryotes and good candidates for being present in the eukaryotic ancestor. Sap130 was only recovered confidently from *P. falciparum* in addition to sequences found in crown eukaryotes indicating a lower likelihood of being present in the eukaryotic ancestor. SF3b10 protein sequences have been difficult to locate and thus not been analysed in this study

The U2-A' and U2-B'' proteins associate stably with U2snRNA (Will et al. 2001) and are found throughout the crown eukaryotes. Candidate sequences recovered from basal eukaryotes makes them likely to have been present in the eukaryotic ancestor. The *S. cerevisiae*-specific protein Ist3 did recover some candidate sequences from *P. falciparum* and *Ent. histolytica* but since this protein has not yet been identified from other crown eukaryotes it has not been placed in the eukaryotic ancestor.

Overall, most of the U2snRNP-associated proteins may have been present in the eukaryotic ancestor. The SF3a and SF3b complexes may have been similar to what is seen in extant eukaryotes. Since U2snRNA is thought to be part of the spliceosome catalytic core then this is evidence that the entire U2snRNP evolved into a sophisticated complex before, or within, the eukaryotic ancestor.

U5snRNP-specific proteins: *(Table 4.5C)*

Prp8, Brr2, Snu114, Prp6, Prp28, Snu40, U5-40, U5-15, PSF and P54nrb

The U5snRNP is required for both steps of splicing interacting with both the 5' and 3'splice-sites of the mRNA (Dix et al. 1998) and is the only snRNP found in all three types of splicing. The yeast U5snRNP has fewer proteins than its mammalian equivalent and contains Prp8, Brr2, Snu114, Prp28, Snu40 and the Sm proteins (Stevens et al. 2001) while the human U5snRNP additionally contains Prp6, the U5-40 protein and the U5-15 protein (Zhou et al. 2002). The *S. cerevisiae* Dib1 (U5-15 homologue) has been found not in the U5snRNP but in the U4/U6.U5 tri-snRNP (Stevens et al. 2001) but for convenience is dealt with in this section. Results of the U5snRNP-specific proteins are shown in Table 4.5C.

U5snRNA associated proteins such as Prp8 and Brr2 are found throughout crown eukaryotes and also in a number of basal eukaryotes including *G. lamblia* (Nixon et al. 2002) *Trypanosoma brucei* (Lucke et al. 1997) and Trichomonas vaginalis (Fast and Doolittle 1999). The Prp8, Snu114 and U5-40 proteins interact with each other forming an RNA-free complex which then interacts with the U5snRNA (Dix et al. 1998). The Prp8 protein is the largest and most highly conserved protein in the spliceosome (Kuhn et al. 2002). It spans the entire 5' stem loop of U5snRNA in both yeasts and humans indicating that not only is the protein highly conserved but also is conserved in its interactions with the U5snRNA and interactions within the U5snRNP (Urlaub et al. 2000). This protein may stabilise the fragile interactions between the U5snRNA and the non-conserved exon sequences at the splices sites, anchoring them in the catalytic centre of the spliceosome (Dix et al. 1998).

Early steps in splicing catalysis include the unwinding of the U1snRNA/5'splice-site helix and unwinding of the U4/U6 helices. These unwinding events are thought to be catalysed by two DExD/H-box RNA helicases, Prp28 and Brr2 respectively (Kuhn et al. 2002). Recent studies indicate that Prp8 coordinates the functioning of Prp28 and Brr2 by inhibiting their activity until spliceosome assembly is complete and correct (Kuhn et al. 2002).

Prp8 is also involved in the recognition of the pyrimidine tract and the 3'splice-site suggesting a role in the regulation of the second splicing step (Dix et al. 1998). However, its highly conserved sequence contains no distinct RNA binding or other recognisable motifs that may give clues to its function (Dix et al. 1998). With such a variety of important interactions

within the spliceosome it is easy to see why Prp8 is so highly conserved and it is not hard to place it within the eukaryotic ancestor.

Brr2 has been shown to interact with a number of proteins (including Prp2, Prp8, Slu7, U1-C and Snu66) and also mediates the recruitment of Prp16 to the spliceosome (van Nues and Beggs 2001). It also interacts with a number of non-splicing proteins involved in signal transduction and transcription in yeast (van Nues and Beggs 2001). Prp28 has an important role by displacing U1snRNP from the 5'splice-site (Chan et al. 2003). The Brr2 and Prp28 DExD/H RNA helicases are highly conserved in crown and basal eukaryotes but have the added complication that DExD/H RNA helicases share contain highly conserved motifs making positive identification from Blast results difficult; without care a wrong identification could be made. Brr2 is extremely conserved and the basal eukaryotic candidate sequences could be treated with a higher confidence than for the other DExD/H RNA helicases recovered during this study. Both Prp8 and Brr2 are components of both the major and minor spliceosomes (Luo et al. 1999) and can be placed with confidence in the eukaryotic ancestor. Prp28 is a member of the DExD group of proteins is placed in the eukaryotic ancestor and is reviewed in that section.

Snu114 plays an important role in the dissociation of the U4snRNA from the U6snRNA during spliceosomal activation by triggering the function of Prp8 and/or Brr2 (Bartels et al. 2003). It is also required for the stable formation of the U4/U6.U5tri-snRNP and may be required later in catalysis to locate the 3'splice-site so that both exons can be precisely aligned by the U5snRNA (van Nues and Beggs 2001). Snu114 shares strong homology to the ribosome translocating factor EF-2 (Dix et al. 1998), thus it is difficult for sequence similarity programs such as BLAST to distinguish between these two proteins. Snu114 from vertebrates and yeasts has been biochemically analysed and has possible homologues in plants. Some candidate sequences were recovered from basal eukaryotes which upon domain analysis are more likely to be Snu114 than EF-2 proteins. Thus Snu114 was likely to be present in the eukaryotic ancestor.

The U5-15 protein (Dib1p in *S. cerevisiae* and Dim1p in *S. pombe*) has a highly conserved sequence across crown eukaryotes (Zhang et al. 2000) and has roles in splicing and cell cycle progression (Berry and Gould 1997). Candidate sequences were found in a number of basal eukaryotes, and thus U5-15 was also placed in the eukaryotic ancestor.

Snu40 has been shown to be important in U5snRNP assembly but is not present in the U4/U6.U5.tri-snRNP complex (Stevens et al. 2001) meaning that this protein leaves the tri-snRNP before spliceosome activation. Candidate sequences were found in *P. falciparum* and *Ent. histolytica* and since Snu40 is found in both crown and basal eukaryotes it was placed in the eukaryotic ancestor.

Prp6 has been shown to be important for the U4/U6.U5 tri-snRNP complex formation and may act as a bridging factor between the U5 and U4/U6 snRNPs (Makarov et al. 2000). It has been found in crown eukaryotes and there is a possible candidate in *P. falciparum* but results were unclear in two of the other basal eukaryotic genomes (*G. lamblia* and *Ent. histolytica)*. Given this uncertainty Prp6 was given a lower likelihood of being present in the eukaryotic ancestor.

One U5-snRNP-associated protein that cannot, as yet, be placed in the eukaryotic ancestor is PSF and a closely related sequence $P54^{nrb}$ (Shav-Tal and Zipori 2002). PSF binds to the polypyrimidine tract (Py tract) of vertebrate mRNAs and is an essential splicing factor during the second splicing step. In vertebrates, both PSF and $P54^{nrb}$ form a complex with the U5snRNA and associate with the 5'splice-site throughout splicing (Peng et al. 2002). Despite these proteins being essential for human splicing reactions, there have been no homologues found in yeast suggesting that either the yeast U5snRNP functions differently from humans, or that yeast contain an as yet identified functional homologue of PSF and $P54^{nrb}$. Searches of basal eukaryotic genomes also failed to find any viable homologous sequences although some small "motif" areas shared some similarity.

The *S. cerevisiae*-specific protein Aar2 (associated with the yeast U5snRNP but not with the U4/U6.U5tri-snRNP and may affect snRNP recycling) was not found in any other species searched and is thus highly unlikely to be present in the eukaryotic ancestor. Thus nearly all of the U5snRNA associated proteins can be placed in the eukaryotic ancestor indicating that this snRNP that is required throughout splicing was already well established within the eukaryotic ancestor.

U4/U6snRNA-specific proteins: *(Table 4.5E)*

Prp24, Prp3, Prp4, Snu13, Prp31, Cpr1, RY-1, Spp41 and Snu23

The U4 and U6 snRNPs exist as separate entities but form a complex (U4/U6snRNP complex) prior to binding to the U5snRNP to form the U4/U6.U5tri-snRNP complex that then attaches to the spliceosome. The assembly of the U4/U6snRNP, its recruitment into the pre-spliceosome and numerous conformational changes of its snRNA components are not well understood (Gonzalez-Santos et al. 2002). U4-specific, U6-specific and proteins specific to the U4/U6 complex will be discussed in this section. Results from the U4/U6-specific proteins are shown in Table 4.5E.

The discrete U4snRNP (when not coupled with the U6snRNP) is not an abundant species in yeast and may behave as a limiting factor in U4/U6.U5 tri-snRNP and spliceosome assembly (Stevens et al. 2001). Snu13 and Prp4 are associated with the 5' stem-loop of the U4snRNA. Both Prp3 and Prp4 are required for spliceosome activation (Gonzalez-Santos et al. 2002) as Prp3 interacts with Prp4 and binds to the paired U4/U6snRNAs. The Snu13

protein, as well as being an essential splicing factor, is also a component of yeast C/D-snoRNAs and its function may extend to other cellular processes. It is present in the yeast U4/U4.U5 tri-snRNP but has not been found in the purified human tri-snRNP (Stevens et al. 2001). All three of the above proteins (Prp3, Prp4 and Snu13) were detected in at least two basal eukaryotic genomes and indicating that these U4snRNP-associated proteins were likely to have been present in the eukaryotic ancestor.

Yeast U6snRNP contains Prp24 and the Lsm proteins (Stevens et al. 2001). The Lsm proteins are discussed in a later section. Prp24 was only detected in *P. falciparum* and thus has only a low likelihood of being present in the eukaryotic ancestor.

Prp31 is required for the U4/U6.U5 tri-snRNP assembly (Makarova et al. 2002). It remains bound to the U4/U6snRNP complex while binding to the U5snRNP-protein Prp6 tethering the complexes together. Prp31 has been characterised in animals and yeast with candidate sequences found in some basal eukaryotes including *G. lamblia* and *Ent. histolytica*. This protein has also been reported in some archaeal genomes (Anantharaman et al. 2002) and is very likely to have been present in the eukaryotic ancestor ;and in this case even in the first eukaryote (Figure 1.5 page 6).

Cpr1 (also called USA-CypP) is a member of the cyclophilin protein family that is thought to have roles in protein folding or conformational changes but can also act as chaperones (Horowitz et al. 2002). During splicing Cpr1 binds to Prp18 and to the Prp3/Prp4 complex but its function is still not clear. Although Cpr1 candidates were found in *Ecz. cuniculi*, *P. falciparum*, *Ent. histolytica* and *G. lamblia*, it cannot be ruled out that these candidates may in fact be other closely related cyclophilins. These basal eukaryotic candidates do however, suggest that at least one cyclophilin (either Cpr1 or related to Cpr1), may have been present in the eukaryotic ancestor.

The *S. cerevisiae*-specific proteins, Spp41 and Snu23, and the RY-1 protein recovered some sequences similarity *P. falciparum* and *Ent. histolytica*. However, these candidates contained domain motifs and low overall levels of sequence similarity and these proteins were not placed in the eukaryotic ancestor.

U4/U6.tri snRNA-specific proteins: *(Table 4.5D)*

Snu66, Sad and, Tri-snRNP27

Three U4/U6.U5tri-snRNA specific SR related proteins (Tri-snRNP27, Sad1 and Snu66) mediate the recruitment of the tri-snRNP to the pre-spliceosome during spliceosome formation (Gottschalk et al. 1999; Makarova et al. 2001). Results for these three proteins are shown in Table 4.5D.

All three tri-snRNP-specific proteins (Sad1, Snu66 and Tri-27) recovered possible candidates from *P. falciparum* but only Sad1 recovered a candidate from *Ent. histolytica*.

BLAST results from *G. lamblia* were less successful with unclear results with Sad1 and no sequences recovered with any Snu66 and Tri-snRNP27K proteins.

The ancestral sequence reconstruction (ASR) was applied with Snu66 and Sad1 to *G. lamblia* and recovered a motif-associated area with Sad1 indicating a possibility that Sad1 is present in *G. lamblia*. ASR with Snu66 did not recover any significant hits against *G. lamblia* but was present in other basal eukaryotes. The absence of a candidate recovered from *G. lamblia* is not evidence that these proteins were not present in the eukaryotic ancestor, it is just that the approach used in this study is to only accept positive evidence. At this stage Sad1 and Snu66 were given a low possibility of being present in the eukaryotic ancestor.

Although the U4/U6snRNP complex and the U5snRNP ancestor and some U4/U6-specific proteins (Prp3, Prp4, Prp31, Cpr1, Snu13 and Prp24) were likely to have been present in the eukaryotic ancestor it is uncertain at present as to whether the U4/U6.U5 tri-snRNP was present as a single complex or as individual components.

The SPF30 protein has been shown in humans to tether the U4/U6.U5tri-snRNP to the pre-spliceosome, apparently via interactions with the Prp3 protein (Schneider et al. 2002). Sequences for this protein have been difficult to locate and thus SPF30 has not been analysed in this study. The Prp38 protein from *S. pombe* (thought to release the U4snRNP from the spliceosome) (Lybarger et al. 1999) recovered some animal, plant and *Plasmodium* sp. sequences during a search of protein databases there is still some ambiguity with these sequences that will need to be resolved before they are used for searching basal eukaryotes.

Sm and Lsm proteins: *(Table 4.5F)*

SmB'/B, SmD1, SmD2, SmD3, SmE, SmF, SmG, Lsm1, Lsm2, Lsm3, Lsm4, Lsm5, Lsm6, Lsm7 and Lsm8

Sm-core proteins (B'/B, D3, D2, D1, E, F, and G) are found in both the major and the minor spliceosomes (Hastings and Krainer 2001) and are involved in the biogenesis (assembly) of the snRNPs. Sm-core proteins bind to a conserved Sm-binding site situated in a single-stranded region of the U1, U2, U4 and U5 snRNAs and are present in all major and minor snRNPs (Vidal et al. 1999; Donahue and Jarrell 2002). Despite structural similarities, Lsm proteins play distinct roles from Sm proteins (Chan et al. 2003). Lsm proteins (Lms1, Lsm2, Lsm3, Lsm4, Lsm5, Lsm6, Lsm7 and Lsm8) play roles in the rearrangement of U6snRNP during splicing, and in promoting U4/U6 formation during recycling of the spliceosome (Chan et al. 2003; Liu et al. 2004). Results from searches of the Sm/Lsm protein group are shown in Table 4.5F. Some of the Sm/Lsm proteins (SmE, SmF, SmG, Lsm1, Lsm3 and Lsm5) have been found in both eukaryotic and archaeal genomes (Anantharaman et al. 2002) and thus are good candidates for also being present in the eukaryotic ancestor as well as in the first eukaryote. Other Sm/Lsm proteins (SmB/B'[3], SmD1, SmD2, SmD3, Lsm2 and

[3] Sm B and B' are alternatively spliced products of the same gene (Vidal et al. 1999).

Lsm4) recovered candidate sequences in at least two basal eukaryotic species and are also likely to be present in the eukaryotic ancestor. Of the last three Lsm proteins (Lsm6, Lsm7 and Lsm8), Lsm8 recovered a good candidate from *P. falciparum*; Lsm6 a possible candidate from *G. lamblia* and Lsm7 a possible candidate in *Ent. histolytica.* Recently, Lsm2 to Lsm8 have been experimentally identified in *Trypanosoma brucei* (Liu et al. 2004) increasing the chances that these proteins were present in the eukaryotic ancestor. The Sm-core and the Lsm proteins have all been placed in the eukaryotic ancestor indicating by their presence that their key roles in splicing today may be ancestrally derived.

U11/U12 snRNP-specific proteins *(Table 4.5H)*

Recently a number of minor splicing-specific proteins have been identified from analysis of the human U11/U12snRNP (Will et al. 2004) and the fruitfly U11 snRNP (Schneider et al. 2004). U11snRNP-specific proteins, U11-25, U11-35, U11-48 and U11-59 and the U11/U12-specific proteins U11/12-20, U11/12-31 and U11/12-65 have similar sequences in the mouse and zebrafish genomes and some of these are also found in the fruitfly, mosquito and some plant genomes (Schneider et al. 2004) (Table 4.5H). Searches against the *C. intestinalis* (sea-squirt) genome recovered candidates for most of these proteins (the U11-59 being the exception). Candidates for three other U11/U12-associated proteins, YB1, Toe-1 and C114 were also recovered from the sea-squirt. These protein candidates, together with the presence of candidate sequences for the U11, U12 and U6atac snRNAs, strongly suggest the presence of minor splicing (as well as major and trans-splicing) in the sea-squirt. Searches of basal eukaryotic genomes failed to find any clear candidates for any of the minor-splicing proteins used in this study although small low homology sequences could be recovered by U11-35 from *P. falciparum* and *Ecz. cuniculi.* At this time there is no evidence to suggest that minor-splicing was present in the eukaryotic ancestor.

Catalytic Step II and late acting proteins : *(Table 4.5G)*

Prp16, Prp22, Prp43, Slu7, Prp17 and Prp18

Protein interaction in the second catalytic splicing step can be divided into two stages; Prp16 and Prp17 activate the first stage, then Prp18 and Slu7 activate the second stage (Chawla et al. 2003). Prp18 weakly associates with the U5snRNP in yeast but not in animals and is involved only in the second step of splicing (Dix et al. 1998). Slu7, Prp18 and Prp22 are essential *in vitro* (*in vivo* not yet determined) for the removal of introns with long distances between their branch site and their 3'splice-site (van Nues and Beggs 2001). After the second splicing reaction, ATP-hydrolysis by Prp22 releases the mature mRNA from the spliceosome (Gesteland et al. 1999). Prp17 has extensive interactions with a number of other splicing proteins (Prp18, Prp16, Prp8, Slu7, Prp22) and also with the U2 and U5snRNAs. Prp17 has an additional role in cell division (Chawla et al. 2003). Prp17 and Prp18 have candidate

sequences in *P. falciparum* but only low homology or small motif areas could be found in *G. lamblia and Ent. histolytica.* Prp16, Prp22 and Prp43 are DExD box RNA helicases and are dealt with in the section below. Slu7 recovered candidate sequences from *P. falciparum* and *Ent. histolytica* and has potential homologues in crown eukaryotes, and thus was placed in the eukaryotic ancestor.

Other DExD/H Proteins: *(Table 4.5M)*

Prp2, Abstrakt, P68, P72, Prp16, Prp22 and Prp43

In *S. cerevisiae*, eight DExD/H proteins (Prp2p, Prp16p, Prp22p, Prp43p, Brr2p, Prp5p, Prp28p and Sub2p/UAP56) have been identified as being required for pre-mRNA splicing (Jurica and Moore 2003). Seven additional proteins emerging have been found in mammalian spliceosomes (DICE1, Abstrakt, eIF4a3, DDX35, DDX9, KIAA0052, p72) (Jurica and Moore 2003). These DExD/H motif containing proteins are classed as RNA helicases and are essential to change the mRNA structural conformation at most stages of the splicing cycle. Some of these proteins are covered under different protein groups (Brr2, Prp28-U5specific proteins; Prp5 – NTC proteins; UAP56 –Other important splicing proteins).

Problems arise when searching for DExD/H proteins as they contain large conserved sequence motifs and often searching with one DExD/H protein will find many proteins of the same family. By comparing the length of the candidate sequence and the position of a motif to that of the query proteins, it is possible *sometimes* to narrow down the choices of proteins that are the most similar for that open reading frame, but this was not often possible with proteins containing the DExD/H motif. For example the *G. lamblia* Contig 41 (9432-13868) is 1478 amino acids in length and was recovered highly with searches of Prp16, Prp22, Prp43 and Prp2 (also recovered at a lower level with other DExD/H proteins). From the lengths of known proteins used in this study, this protein could fit Prp16, Prp22 or Prp43.

Because not all the splicing factors that have been identified were able to be screened in this study there may be other DExD/H proteins that may also fit these candidate sequences. Searches of NCBI databases with DExD/H protein candidates recover many DExD proteins with very similar scores. For this reason only the DExD/H proteins that showed outstanding homology, with no conflict with other DExD/H proteins (e.g. Brr2 and UAP56), were placed in the eukaryotic ancestor although it is likely that other DExD/H proteins were also present.

SR proteins: *(Table 4.5I)*

ASF/SF2, 9G8, Srm160, Srm300, Srp20, Srp30, Srp40, Srp54, Srp55, Srp75, SC35 and Tra2

SR (Ser-Arg rich) proteins are required for splice site recognition in all three types of splicing (major, minor and trans) (Hastings and Krainer 2001; Furuyama and Bruzik 2002;

Graveley 2004). They have also been shown to stimulate exon inclusion in alternative splicing They contain a characteristic C-terminal 'RS' domain of variable length, rich in serine-arginine repeats that can be extensively phosphorylated (Portal et al. 2003). This phosphorylation can mediate regulatory interactions with other proteins. These proteins are commonly found in mammalian splicing (Hastings and Krainer 2001) but are absent in yeast (Zhou et al. 2002). Some novel SR proteins have been found in *Trypanosoma cruzi* (Ismaili et al. 1999; Ismaili et al. 2000; Portal et al. 2003), evidence that SR proteins may have been present in early eukaryotes but may have been lost in some later lineages. Results from searches with SR proteins are shown in Table 4.5I.

Candidate sequences for the proteins ASF/SF2 and 9G8 were recovered from *P. falciparum,* but BLAST searches with other SR proteins of *P. falciparum, G. lamblia* and *Ent. histolytica* returned at best, only motif-associated areas. ASR searching with 9G8 recovered a possible candidate in *P. falciparum* but again only motif-associated areas in the other two genomes.

The RS motif (the predominant feature of SR proteins), however, has been found almost exclusively in splicing related proteins (Portal et al. 2003) indicating that the motif-associated areas found in *P. falciparum, G. lamblia* and *Ent. histolytica* may be part of novel SR proteins. The presence of the RS-domain in these three basal eukaryotes as well as the novel *T. cruzi* SR proteins (protein sequences were not found) indicates that SR proteins as a group may have been in the eukaryotic ancestor but no specific protein as yet can be ancestral to living eukaryotes

Prp19 associated complex (NTC): *(Table 4.5K)*

Prp19, Cdc5 Cef, Snt309, Plrg1, Clf1

The Prp19 associated complex (NTC or nineteen complex) is required for the stable association of U5 and U6 snRNPs with the spliceosome after U4snRNP dissociation and for the dissociation of Lsm3 from the spliceosome during spliceosome activation (Chan et al. 2003). The NTC has been isolated as a distinct unit indicating that its constituents bind directly with one another (Ohi and Gould 2002). The yeast NTC consists of Cef1p (CDC5L homologue), Snt309, Ntc31, Isy1, Ntc20 and at least another six uncharacterised proteins. Another 30 uncharacterised proteins have been copurified with the human Cdc5/Cef1 protein (Ohi and Gould 2002). Results of searches with NTC proteins are shown in Table 4.5K.

Prp19 itself is required to maintain the organisation of the NTC complex (Ohi and Gould 2002) and is associated with the spliceosome either after or simultaneously with U4snRNP dissociation. A Prp19 candidate was found in *P. falciparum* but only motif areas could be determined from *Ent. histolytica* and *G. lamblia* using both BLAST and ASR.

However, Prp19 was recovered from *D. discoideum* and thus, Prp19 was placed in the eukaryotic ancestor.

Prp5 and PLRG1/Prp46 are core components of the mammalian NTC (Ohi and Gould 2002). Although Prp5 is a DExD/H protein, clear candidate sequences from *P. falciparum* and *Ent. histolytica* were recovered but only a motif-associated region was recovered from *G. lamblia*. PLRG1 also recovered candidates from *P. falciparum* and *G. lamblia* with a motif-associated region recovered from *Ent. histolytica*. Cdc5l/Cef1 is suggested to associate the whole NTC complex to the spliceosome and may also be involved in transcription and the cell division cycle (Ajuh et al. 2001). Cdc5l/Cef1 is also associated with PLRG1 which as well as being an NTC-associated splicing factor is involved in cell shape maintenance and/or regulation of the cell cycle (Ajuh et al. 2001). Cdc5l/Cef1 candidate sequences were both found in *P. falciparum* and *Ent. histolytica* and a motif-associated area recovered from *G. lamblia*. As all three of these proteins (Prp5, PLRG1 and Cdc5l) are found also throughout crown eukaryotes they are likely to have been present in the eukaryotic ancestor.

Clf1 is part of the NTC and promotes the functional integration of the U4/U6.U5trisnRNP into the pre-spliceosome (Wang et al. 2003). Clf1 recovered candidate sequences from *P. falciparum* and *Ent. histolytica* but no candidate could be recovered from *G. lamblia* at this stage. This protein is also found throughout crown eukaryotes and was placed in the eukaryotic ancestor.

Some NTC-associated proteins have been isolated during the large spliceosomal studies (e.g. Jurica and Moore 2003), but there is little information about how these proteins function in the NTC-complex or splicing. fSap33 is one of these proteins, and is found throughout crown eukaryotes with candidate sequences recovered from *P. falciparum* and *Ent. histolytica*. There is as yet no known function for fSap33 but its distribution is indicative that it may have been present in the eukaryotic ancestor.

Results here indicate that a number of NTC-associated proteins, as well as Prp19 itself, were present in the eukaryotic ancestor, indicating that the NTC-complex as a whole is of ancestral origin.

Coupling of splicing with other major cellular events: *(Table 4.5J)*

There are still several more complexes to consider. In today's eukaryotes, almost all of the major events in the production of mature mRNAs are highly coupled with splicing (Lynch and Richardson 2002) and there are many interactions between various splicing factors and elongation factors to promote transcription elongation, mRNA export, transcriptional termination and polyadenylation. Some of the complexity of the spliceosome may be accounted for by proteins that are not essential for catalysis but instead play important post-splicing roles (Nilsen 2003). Results from the search of these proteins are shown in Table 4.5J.

Prp4Kinase (not to be mistaken with Prp4) is present in the yeast S. *pombe* and mammals but has not been found in *S. cerevisiae* (Kuhn and Kaufer 2003). It plays a key role in regulating splicing and in connecting this process with the cell cycle. A candidate Prp4Kinase sequence was found in *G. lamblia* but results were uncertain in *P. falciparum*. A hypothetical sequence was found in *Trypanosoma brucei* (through a search of this genome's annotation) thus this protein was placed in the eukaryotic ancestor.

The transcription cofactor Tat-SF1 also occurs in the (major) spliceosome (Zhou et al. 2002), interacts with snRNPs, and is thought to reciprocally activate transcription elongation and splicing. Additional transcription factors (e.g. CA150, Xab2 and Skip) as well as polyadenylation factors (e.g. CF1-68, CF1-25) are also found in the spliceosome (Zhou et al. 2002). The proteins CA150, Tat-SF1 and Xab2 recovered only at best motif sequence areas from *P. falciparum*, *Ent. histolytica* and *G. lamblia* and were not placed in the eukaryotic ancestor. The Skip/Prp45 protein, recovered candidate sequences in *P. falciparum* and *Ent. histolytica* and was likely to have been present in the eukaryotic ancestor.

Every protein that is in the TREX-complex (involved in transcription elongation) may also be present in the spliceosome (Zhou et al. 2002), suggesting that transcription; splicing and export may all be coupled via this complex. Both Aly (covered earlier and placed in the eukaryotic ancestor) and UAP^{56} are components of the TREX-complex. Aly is recruited to the mRNA during splicing and specifically interacts with UAP^{56}. UAP^{56} has multiple tasks in spliceosome assembly, including dissociation of $U2AF^{65}$ from the spliceosome (Luo et al. 2001). Excess UAP^{56} is a dominant negative inhibitor of mRNA export and prevents the recruitment of Aly to the spliced mRNP (Luo et al. 2001) UAP^{56} protein candidates were found in *P. falciparum*, *Ent. histolytica* and *G. lamblia* as well as being present in crown eukaryotes and thus likely to have been present in the eukaryotic ancestor. The Aly protein recovered a candidate sequence in *G. lamblia*, however Aly is a member of the closely related cyclophilin protein family and thus there is a chance that candidate sequences may in fact be other related cyclophilin proteins. Aly could not then be placed in the eukaryotic ancestor. Another TREX complex-protein Tex1 recovered possible candidate sequences in *P. falciparum* and *Ent. histolytica* and thus was placed in the eukaryotic ancestor. Other components Tho2, Hpr1 (Reed 2003), Asr2B, CF1-68kD and CF1-25kD recovered at best motif-associated areas and were not placed in the eukaryotic ancestor.

The PolyA-binding protein (PabP) has been characterised in vertebrates, yeast and the basal eukaryote *Leishmania major* and has important roles in translation initiation and mRNA biogenesis, export and degradation (Chekanova and Belostotsky 2003). Candidates for PabP were found also found in *P. falciparum* and *G. lamblia* and thus this protein was placed in the eukaryotic ancestor.

The U2-SF3b protein Rds3 (which has already been placed in the eukaryotic ancestor), as well as being an essential component of the U2snRNP, also interacts with the Yra1p export factor in yeast (this protein is not actually part of the spliceosome), showing another link between splicing and RNA export (Wang and Rymond 2003). The results shown here indicate that within the eukaryotic ancestor there may have already been strong links between pre-mRNA splicing and other cellular processes such as RNA export and transcription.

Post- transcriptional EJC proteins: *(Table 4.5L)*

Y14, Magoh, Rnps1, Cbp80 and Cbp20

The exon-junction complex (EJC) consists of several proteins that, upon the completion of intron excision, are deposited on the mRNA product at a conserved position, 20-24 nt upstream of exon-exon junctions (Nott et al. 2004). Core components include Y14, Magoh and the Aly proteins, of which only Y14 and Magoh remain stably associated with mRNA after nuclear export. The EJC contains several proteins involved in nonsense-mediated-decay (the degradation of mRNAs containing premature stop codons) and in the cytoplasmic localisation of mRNAs. The splicing proteins Aly and Srp20 (an SR protein) join three other proteins (p170, p95 and p57) at the mRNA contact region in the EJC. Once formed, the EJC is associated with mRNAs that are bound in the nucleus by the cap-binding protein Cbp80 but not with the mRNAs bound by the cytoplasmic eIF4E cap-binding protein. It is suggested that the entire EJC is exported, then dissociates from the mRNA in the cytoplasm. This mechanism has also been analysed in yeast suggesting conservation between yeast and animal systems (Reed 2003). In *C. elegans* and *D. melanogaster*, Y14 and Magoh are required for late embryogenesis and proper germline sexual differentiation indicating that the conserved interaction between Y14 and Magoh proteins is important for multiple developmental processes in various organisms (Kawano et al. 2004). Spliced mRNAs exhibit increased translational yield as compared with no-intron mRNAs in mammalian tissue culture cells, much of which can be attributed to EJC deposition (Nott et al. 2004). Results of searches for EJC-associated proteins are shown in Table 4.5L.

Of the proteins associated with the EJC complex, proteins that have already been placed in the eukaryotic ancestor are UAP^{56} and Aly. The SR proteins Srp20 and SRm160 recovered at best motif-associated areas and there was some uncertainty with Rnps1 so these proteins were not placed in the eukaryotic ancestor. Candidates for the Magoh protein were recovered from *P. falciparum* and *Ent. histolytica* but there were no significant hits in *G. lamblia*. Magoh is also found in vertebrates, yeasts and plants and was placed in the eukaryotic ancestor. However, Magoh's partner Y14 recovered at best motif-associated areas with both BLAST and ASR from *P. falciparum* and *G. lamblia* but ASR searches recovered a

candidate with some low sequence homology from *Ent. histolytica*. As no clear candidate was found in any basal eukaryotes, Y14 was not placed in the eukaryotic ancestor. If Y14 was not present in the eukaryotic ancestor then it is possible that the ancestral Magoh protein either worked on its own or with another protein which contained some of the properties of Y14 but not sequence similarity.

Other Essential Splicing proteins: *(Table 4.5N)*

U2AF65, U2AF35, UAP56, fSap118, SF1

Not all splicing factors can be conveniently grouped into any of the previous sections and are thus placed here. Results from searches with the proteins in this group are shown in Table 4.5N. Essential splicing factors SF1, Luc7a and U2AF (U2AF^{65} and U2AF^{35} subunits) play important roles in splice-site recognition during early spliceosome assembly (Fortes et al. 1999; Selenko et al. 2003). During this assembly phase the U1snRNP binds to the 5'splice-site, U2AF^{65} binds to the Py-tract and U2AF^{35} binds to the downstream AG dinucleotide respectively. Specific recognition of the branch point sequence is mediated by SF1 (Selenko et al. 2003). The Py-tract is present in most animal introns but is absent from *S. cerevisiae* introns, however U2AF subunits are still found in *S. cerevisiae*. Other roles of U2AF are to promote U1snRNP recruitment to the 5'splice-site and U2snRNP recruitment by association to the RNA helicase UAP56 (Forch et al. 2003). A U2AF^{65} was recovered from *G. lamblia* but results were uncertain in *P. falciparum* and *Ent. histolytica*. U2AF^{35} has been characterised in *T .brucei* (known as U2AF^{23}) and candidate sequences were recovered from *P. falciparum* and *Ent. histolytica*. Candidate SF1 sequences were recovered from *P. falciparum*, *Ent. histolytica* and *G. lamblia*. Both U2AF subunits and SF1 have been characterised throughout crown eukaryotes and in basal eukaryotes, and thus were likely present in the eukaryotic ancestor.

Possible fSap118 candidates were found in *P. falciparum*, *G. lamblia* and *Ent. histolytica* but will require further analysis to determine their validity because of the close sequence similarity between members of the RNA helicase protein family. Little is known about fSap118 except that it is a helicase of the SKI2 subfamily and has been implicated in mammalian splicing.

The IFN4/Fal1p protein is involved in 40S ribosomal subunit biogenesis in *S. cerevisiae* (Forch et al. 2002) but it is also found in vertebrates and plants. IFN4 recovered candidates from *P. falciparum*, *Ent. histolytica* and *G. lamblia* and although it is a DExD/H protein was placed in the eukaryotic ancestor. Little is known as to how this protein is involved in splicing other than it was isolated with spliceosomal complexes (Zhou et al. 2002) and thus it may not even be a splicing protein at all but merely caught up in the biochemical extraction of the spliceosomal complex. This situation is the same for other proteins such as

Mfap1, Ccap2 and Rha that were also placed in the eukaryotic ancestor based on protein distribution data only.

The Tip39 protein was originally identified as a tooth enamel protein (Paine et al. 2000) but was later suggested to be a splicing factor due to its presence in a number of spliceosomal complexes (Jurica and Moore 2003). This protein recovered no candidates in *P. falciparum*, *Ent. histolytica* and *G. lamblia* and thus was not placed in the eukaryotic ancestor. There were a number of other splicing proteins that did not recover any candidates in either *P. falciparum*, *Ent. histolytica* and *G. lamblia*. This does not imply that these proteins are not present in these genomes, but merely that they were not found with any technique used in this study.

4.4: Summary

This study set out to determine if a spliceosome existed in the eukaryotic ancestor and if so, whether it was a simplified version of today's spliceosomes or just as complex. What was found was that snRNAs and splicing-specific proteins are found conserved throughout crown and basal eukaryotes indicating a probable ancestral presence. Conclusions drawn from the work shown in this chapter confirm the premise (Lynch and Richardson 2002) that introns and the spliceosomal machinery to process them, were present in the eukaryotic ancestor.

Another major conclusion of this work is that the splicing process in the eukaryotic ancestor may have been very similar to that seen today in living eukaryotes, i.e. not simplified but just as complex. All five major-splicing snRNPs (U1, U2, U4, U5 and U6) are likely to have been present in the eukaryotic ancestor. These ancestral snRNPs, far from being simplified versions, may have contained most of the U-snRNP-specific proteins found in today's eukaryotes (Table 4.6). Other groups of proteins such as the Sm-core proteins (bound within each snRNP) and the Lsm proteins have also remained highly conserved throughout the eukaryotic lineage and have likely ancestral eukaryotic origins (Table 4.6).

Some protein groups, however, have not been easy to characterise across eukaryotic lineages. Proteins that belong to highly conserved protein families (e.g. DExD/H, cyclophilin and SR proteins) may be very similar in sequence to other members of the same family. This creates problems, both with sequence annotation in general, and in determining if a particular protein was likely to have been present in the eukaryotic ancestor. Sequence-linked properties such as length and predicted physiochemical properties (e.g. isoelectric point and amino-acid composition) are of limited use in this situation because they are often shared by the other members of the family. Thus biochemical analysis including protein/RNA binding studies will be required for true identification for many of these spliceosomal proteins. Candidate sequences found using queries from members of a protein family may indicate however, the likely presence of that protein family and its distribution rather than the distribution of

individual members. For example, the presence of RS-motif sequences throughout crown and basal eukaryotes indicates that SR proteins that contain this motif are likely to have been present in the eukaryotic ancestor.

Proteins in the Eukaryotic Ancestor			
Table	**Page**	**Group**	**Proteins**
4.5A	89	U1-specific	U1-70, U1-A, U1-C, Prp40, Nam8, Prp39
4.5B	89	U2-specific	Sap155, Sap145, Sap130, Sap114, Sap62, Sap61, Sap49, U1-A, U1-B, P14, Rds3
4.5C	89	U5-specific	Prp8, Brr2, Snu114, Prp6, Prp28, Snu40, U5-40, U5-15
4.5E	90	U4/U6-specific	Prp3, Prp4, Prp31, Cpr1, Snu13, Prp24
4.5D	89	U4/U6.U5-specific	Snu66, Sad1
4.5F	90	Sm/Lsm core	SmB/B', SmD1, SmD2, SmD3, SmE, SmF, SmG, Lms1, Lsm2, Lsm3, Lsm4, Lsm5, Lsm6, Lsm7, Lsm8
4.5G	90	Catalytic Step II	Prp16, Prp22, Prp43
4.5M	92	DExD-motif	Abstrakt, P68
4.5I	91	SR proteins	SF2, 9G8
4.5K	91	Prp19-complex	Cdc5l, Prp5, fSap33, Crn, Plrg1, Prp19
4.5J	91	Other cellular events	UAP56, SKIP, Tex1, PabP
4.5L	91	Exon-junction proteins	Magoh, Cbp20
4.5N	92	Other splicing proteins	U2AF65, U2AF23, SF1, fSap118, Mfap1, Ifn4, Rha, Ccap2

Table 4.6: Summary of the 75 spliceosomal proteins likely to be present in the eukaryotic ancestor. Results tables and their respective page numbers are also listed.

Some proteins however are species (or lineage specific), i.e. yeast-specific proteins not found in animals and vice-versa. Intron size, splice-site recognition and processes linked to the splicing mechanisms can differ between lineages. Lineage-specific splicing proteins are evidence that changes in splicing mechanisms were accompanied by compensatory changes in spliceosomal protein composition.

Due to time limitations and the sheer amount of data that had to be obtained and managed, not all proteins identified as belonging to spliceosomes (Kaufer and Potashkin 2000; Lorkovic et al. 2000; Zhou et al. 2002; Jurica and Moore 2003) were used in searches during this study. The "parts-list" (Nilsen 2003) of the spliceosome may still not be complete as many other splicing-associated proteins are constantly being identified and new functions applied to those already identified. There are still relatively few splicing-associated proteins biochemically characterised from any of the basal eukaryotes (compared with the numbers characterised from yeasts and vertebrates), and as yet no complete spliceosomes isolated. Until then a complete picture of eukaryotic splicing (for all mechanisms in all eukaryotic lineages) cannot realistically be constructed but comparative studies can aid in sorting out the most likely scenarios for splicing in the eukaryotic ancestor.

The distribution of major and trans-splicing indicate that both splicing mechanisms may have been present in the eukaryotic ancestor. Both major and trans-splicing mechanisms contain many similarities even in highly diverse eukaryotic lineages, thus it is more likely that

they were separate entities already in the eukaryotic ancestor than the converse view that each instance evolved separately in each lineage. This is especially true because the present work leads to the inference that the original spliceosome was complicated. Another option is that splicing mechanism similarities may be the result of horizontal transfer. This is unlikely as it has been found that genes involved in transcription, translation and relating processes, such as splicing, are rarely horizontally transferred (Jain et al. 1999). The use of SL-trans-splicing to process polycistronic (many genes) mRNAs may have been lost or "downgraded" in some lineages (i.e. mammals) with the advent of monocistronic (single gene) mRNAs. However, the ability to join two independently produced pre-mRNAs in a trans-splicing reaction has remained in lineages that do not contain SL-trans-splicing (Garcia-Blanco 2003). It is possible that multiple splicing mechanisms may have allowed more diversity in mRNA processing in early eukaryotes. Minor splicing has not yet been demonstrated in any basal eukaryotes and at this stage is not seen likely to have been present in the eukaryotic ancestor, but evolved sometime before the separation of plants and animals. Alternative splicing is found associated with all three types of splicing (major, minor and trans) (Boue et al. 2003), present in many diverse eukaryotic lineages and is an important mechanism in gene expression and regulation.

Alternative splicing is thought to have played a major role in genome evolution (Boue et al. 2003). It is suggested that by allowing new exon inclusion or exclusion in one transcript, yet having the original transcript still present, alternative transcripts can be "trialled" within a cellular environment to allow a 'trial and error' approach for the evolution of gene structure (Boue et al. 2003; Modrek and Lee 2003). It is therefore possible that alternative splicing was present in the eukaryotic ancestor and played an important role in the evolution of eukaryotic lineages.

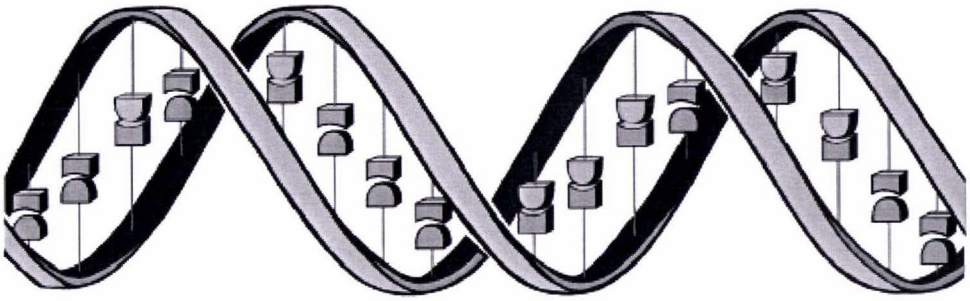
The distribution of intron characteristics throughout crown and basal eukaryotes was also examined during this study. Intron characteristics can reveal a number of things about how they are managed. Intron size (i.e. length) is constrained by the mechanism by which the boundaries between the introns and exons are recognised. The distribution of intron length over crown and basal eukaryotes indicates that in the eukaryotic ancestor, the introns were likely less than 1000 nucleotides in length and may even have been shorter due to the predominance of short introns in the basal eukaryotes. Short introns use the intron-definition mechanism of splice-site recognition to recognise boundaries across the intron, compared with the exon-definition mechanism described in animals to recognise splice-site boundaries across exons. It is likely that the eukaryotic ancestor contained a mechanism to recognise splice-site boundaries but as it is still not known what mechanism(s) is used in the basal eukaryotes; neither intron nor exon-definition can be determined for the eukaryotic ancestor.

This study did not investigate the introns-early (introns evolved early and lost from lineages) versus the intron-late (introns gained in lineages) theories of intron evolution as it

did not examine whether a particular intron has been gained or lost from a lineage. Rather it characterised introns in general from each lineage. Due to the presence of alternative splicing in many lineages (a mechanism by which it is suggested an intron could be permanently inserted or removed (Kondrashov and Koonin 2003)), detailed analysis of the loss and gain of introns from specific proteins may not in fact reveal the nature of the ancestral protein.

A number of studies in the past (Anantharaman et al. 2002; Koonin et al. 2004) have used computational surveys of eukaryotic (and archaeal) genomes to uncover proteins that may have been present in "ancestral" organisms. However, until recently only a small number of crown eukaryotic genomes have been available for analysis. With genomic sequencing of species from a number of basal eukaryotic lineages, it is now possible, as was done in this study, to look closer at how these species are related and how their cellular mechanisms evolved. The biochemical analysis of complete spliceosomes and snRNPs (Jurica and Moore 2003) combined with computational comparative genomics offers a powerful tool to study the evolution of splicing mechanisms.

Splicing can now be seen as a fundamental aspect of eukaryotic life and appears to have evolved before the last ancestor of living eukaryotes. Contrary to the idea that splicing may have been a 'simplified' mechanism in this ancient organism it can now be suggested that this was not the case and that splicing and the spliceosome had already evolved in a sophisticated cellular process, already linked to other cellular processes such as transcription, capping, mRNA export and polyadenylation. This may not have been the case with the much earlier 'first' eukaryote (Figure 1.6) and much study will be required to compare the spliceosomal process found in eukaryotes and the self-splicing mechanism of prokaryotes. An interesting prospect for the future



Chapter 5 Conclusions and Future Work

"I have often tried to picture you lying on a beach with absolutely nothing to do...and the picture always ends with your head imploding." - Babylon5

Although conclusions and future directions have been dealt with, in part, at the end of each chapter, it is appropriate to sum up here the overall project. RNA processing systems are complex in eukaryotic organisms, but most of the research has been carried using data from the crown eukaryotes (i.e. animals, fungi and plants). With the completion (or near completion) of a number of basal eukaryotic genomes, we can now use comparative genomics to analyse the same systems in basal eukaryotes, and from there look at ancestral features conserved between the two eukaryotic groups. However, previously there have been relatively few ncRNAs and their associated proteins characterised from basal eukaryotes, resulting in the need for the extensive genomic searches undertaken as part of this project. Gene-finding software (that can be used with eukaryotic genomes) was designed initially for the crown eukaryotes, and is based solely on sequence-similarity between the query gene and the target genome. Searching for sequence similarity between distantly related eukaryotes (such as between crown and basal eukaryotes) is often not successful, and resulted in the need to evaluate software parameters for searching basal eukaryotes, for both ncRNA and proteins. Collecting, storing and managing large volumes of diverse data indicated that data management systems were essential for comparative genomic projects. RNaseP was chosen to test search and data management strategies as it contained both ncRNA and protein components and, because of its ubiquitous nature, is expected to be found in basal eukaryotes. The principle study in this project, the nature of the spliceosome in the eukaryotic ancestor involved all of the issues mentioned above, and as a result characterised a large number of spliceosomal components conserved between crown and basal eukaryotes, leading to the perhaps unexpected inference that a complex spliceosome existed in the eukaryotic ancestor.

5.1: ncRNA identification

ncRNA-search software is still at a relatively early stage (when compared to protein-search software) and it is likely that no one program alone will find every ncRNA within a genome. ncRNA-search software tested in this project may suit different types of ncRNA. Future versions of RSEARCH (that are less processor-intensive than the present version) could become a first-pass option to quickly scan for an ncRNA before constructing alignments and/or descriptors. ERPIN may be useful in finding closely related RNAs (e.g. MicroRNAs) within a genome and RNAmotif used for the hard-to-find cases.

The RNAmotif software was modified during this study[14] to effectively search eukaryotic genomes. Future improvements could include a graphical interface both for the building of RNAmotif descriptors and for running the RNAmotif software itself. Graphical interfaces would make this program more accessible to the non-programmer, by helping to avoid annoying and time-consuming syntax errors in descriptor construction. An extension idea could lead to a researcher using a graphical interface to input biological data, check a graphically-constructed model, then search a designated genome, without even having to view the descriptor code. For now, more immediate improvements to the RNAmotif software include; changes in the parallel implementation to allow automatic splitting of genomic databases (as is seen already in RSEARCH), multiple RNAmotif searches (a primary scan for essential motif regions followed by an more detailed RNAmotif search on sequences found in the primary scan), and downstream post-processing for essential motifs.

Work with the RNAmotif software highlighted the usefulness of a parallel computing facility such as Helix (http://helix.massey.ac.nz). The Helix cluster is a distributed-memory Beowulf cluster with 65 nodes (128 processors) running the Linux RedHat (version 7.3) operating system and communicating with the MPI protocol. The ability to divide a genome into smaller pieces allows processor-intensive programs (such as RNAmotif) to be used practically in genomic searches, and is a key tool in the development of software that can then one day be run on smaller computer systems.

In general, ncRNA software is at present very programming-orientated (i.e. written for someone who understands computer programming). However, this is mostly because these programs are still in development. Once the development phase is completed, internet and graphic interfaces are likely to be included in the software to enable non-programming researchers to translate the biological knowledge into the required grammar, then to run their queries.

The ncRNAs recovered from basal eukaryotic genomes during this project (U5snRNA and RNase P sequences) are being published, and also submitted to Genbank and Rfam databases. It is hoped that other ncRNA-researchers may be able to use this information profitably.

5.2: ncRNA-associated protein identification

Spliceosomal protein studies depended on the accurate prediction of protein homology, thus protein-search strategies were evaluated to test their effectiveness against basal eukaryotic genomes. BLAST and FASTA are the most popular and widely used sequence-similarity based software for protein searches. BLAST is incorporated into the NCBI database search facility and thus, is a very

[14] With the help of Dr. T. Macke, the principle programmer of the RNAmotif software.

useful tool for finding homologous proteins from many species in a timely manner. However, sequence-similarity methods may be limited when searching basal eukaryotic genomes because the majority of the known protein homologues are from crown eukaryotes. The large evolutionary distance between crown and basal eukaryotes means that during a BLAST search of a basal eukaryote with a query crown eukaryotic sequence, candidate sequences are often recovered with low scores of statistical significance (i.e. scores that are too low to be confident of the identity). During this project, methods were developed to evaluate these low scoring candidate sequences to determine their validity. The first method was the use of comparative-blasting, where homologues of the same protein from different species are blasted against a genome and results compared for consistency. A candidate basal eukaryotic sequence had increased validity if it was recovered with different crown eukaryotic query sequences (i.e. with animal, yeast and plant homologues). The second method developed to validate candidate sequences was back-BLASTing; where a candidate was used as the query against a number of test databases (i.e. NCBI databases, databases of similar proteins and against the genome from whence it came). Back-BLASTing revealed if the candidate sequence was the member of a protein family (grouping of proteins with similar sequences, motifs and function). These validation methods showed that BLAST could still be used for searching basal eukaryotic genomes, even when query proteins were only available from crown eukaryotic species.

However, even with added validity checks, BLAST could not be expected to find candidates for some of the less conserved proteins associated with the spliceosome, so other methods were evaluated using RNaseP proteins. The ancestral sequence reconstruction (ASR) technique was developed during this project to aid in finding protein sequences from basal eukaryotes. This technique was successful in recovering candidate sequences from some basal eukaryotes that were not found using any other technique. Although the prediction of ancestral sequences has been used for many years in evolutionary and functional studies, its application to protein-searches is new. ASR was also combined with HMMer, another protein-search technique that constructs HMM-profiles from sequence alignments. It was found that ancestral sequences could fill-out an alignment, increasing values of statistical significance for any candidates. Two software packages (PAML and FastML) were used to infer the ancestral sequences, but there are other programs available. A comparison of ASR using a range of prediction software would be useful in developing this technique further. Development of ASR and ASR-HMMer techniques should also include testing with other protein datasets to understand the limitations of these techniques.

As well as sequence-based software there are a number of programs under development that increase the use of a protein's folding features in sequence search and analysis. These

programs, when they become available will offer advantages to searches that no longer rely only on sequence-similarity, which is often not conserved between the crown and basal eukaryotic sequences.

5.3: Data management in the post-genomic era

Data management is becoming more important now in the post-genomic era, than it has ever been in the biological sciences. Management systems are not only required to store the vast amounts of data being produced as increasing numbers of genomes are being sequenced, but also to pull together this information from different sources, to form a coherent view of a cellular environment. There are now many jigsaw pieces scattered over remote databases that can be easily accessed; the trick is finding each piece out of different databases, and then putting the puzzle together. Retrieving data from remote databases is often not straightforward and much information is missed due to inadequately annotated and referenced data. Natural-language keyword searches often fail to recover all the relevant data from a database as often genes have been given different names when they have been found in a number of species. For example, the RNaseMRP RNA gene in rat is not known as such but as 7-2 RNA. Multiple keywords and/or sequences have to be submitted in order to gain the maximum amount of relevant information from a database. Using the above example, RNaseMRP sequences from a number of vertebrates (human, cow and frog) can be obtained from the NCBI databases using the keyworks "RNase" and "MRP". However, the RNaseMRP sequence from rat can only be found using "7-2" and "RNA". Luckily in this case, a BLAST search of the NCBI databases with the human RNaseMRP can recover the rat homologue. This may not always be the case and data can be missed merely because of different naming conventions for different species.

Once data has been retrieved, it needs to be effectively stored and managed. As was found with P-MRPbase and SpliceSite, the development of personal databases for genomic project management does not have to be complicated. With a small carefully constructed database, loss of data and time (due to having to repeatedly find specific data) can be minimised, and data-mining supported to allow quick retrieval of required information. Designing, constructing and using a small database (e.g. P-MRP) was extremely useful in sorting out issues before working with a larger database (e.g. SpliceSite). There is a need for database construction and data management instruction at many levels. A study could be done to construct a relevant template database (one that can be downloaded and modified to suit requirements) so that users do not have to start from scratch each time. There is no doubt that the amount of genomic information available will continue

to grow, but there is now a great need for genomic management as well as genomic analysis skills in bioinformatics.

5.4: RNaseP in the Eukaryotic Ancestor

RNaseP was ideal as a test ncRNA-protein complex during this study before the larger spliceosomal complex was examined, but it is highly interesting in its own right. RNaseP activity has been found in every cell examined and is encoded in bacteria, archaea, crown eukaryotes and some organelles (Xiao et al. 2002). To date, RNaseP has not been biochemically isolated from any basal eukaryote (an extremely complex procedure) and thus, conclusions about RNaseP in the eukaryotic ancestor can only be preliminary. By comparing components of RNaseP between crown and basal eukaryotes, it is possible to examine the nature of RNaseP in the eukaryotic ancestor. Crown eukaryotes are known to have 9-10 proteins (depending on the species) associated with their RNaseP RNA but until now, nothing was known about RNaseP in basal eukaryotes.

The RNaseP RNA candidates recovered from the basal eukaryotic genomes used in this study (*G. lamblia* and *Ent. histolytica*) mostly resemble their crown eukaryotic counterparts. An exception is in the P3-region of the RNaseP RNA, where a bacterial-like structure is formed instead of the traditional eukaryotic-like structure. Secondary-structures were calculated by folding local regions with computer software, using the consensus RNaseP structures (from all three kingdoms) as guides. The true secondary-structure can only be determined using sophisticated biochemical analysis that tests each nucleotide for binding. The RNaseP RNA candidate sequences from *G. lamblia* and *Ent. histolytica* contain essential motif regions that are present in RNaseP RNAs from all three kingdoms (Frank et al. 2000) and there is confidence in their validity. It is hoped that once published these sequences will be examined further by researchers who specialise in eukaryotic RNaseP RNA.

RNaseP from basal eukaryotes, as expected, seems to be comprised of a single RNA and multiple protein components, like that of crown eukaryotes and some archaea. Since a complete RNaseP complex has never been characterised from any basal eukaryote, it is unknown as to its total protein content. Four proteins (Pop1, Pop4, Pop5 and Rpp21) were characterised from *G. lamblia* but, there are other proteins (Rpp14, Rpp25 and Rpp30) that have been characterised in archaea that may also be present. There are differences in the total protein complement between crown eukaryotes. Humans and the yeast *S. cerevisiae* each contain an acidic protein (Rpp40 and Pop8 respectively) which bear no sequence homology to each other and are not found in any other species (Jarrous et al. 1998). It is therefore possible that *G. lamblia* also contains an acidic protein which could not be detected by sequence-similarity but would only be detected upon the isolation of

a complete RNaseP complex. For pre-tRNA catalysis, human RNaseP requires only the RNA and the Pop4 and Rpp21 proteins (Mann et al. 2003), all of which have been identified in *G. lamblia*. Thus it is possible that a similar experiment could be done with *G. lamblia* to investigate the catalytic properties of a basal eukaryotic RNaseP.

With a number of proteins found conserved between the crown and basal eukaryotes (Pop1, Pop4, Pop5, Rpp21), or present in crown eukaryotes and archaea (Rpp14, Rpp30, Rpp25, Rpp20/Pop7), it is likely that the RNaseP present in the eukaryotic ancestor contained (as well as the RNaseP RNA) at least seven out of the 9 proteins found in humans.

RNaseP in bacteria contains only one protein which has not to date, been found in any archaeal or eukaryotic genomes, but the archaeal RNaseP contains multiple proteins (Kouzuma et al. 2003). It is likely that the ancestor of eukaryotes and archaea had an RNaseP that contained multiple proteins. Further analysis of archaeal and basal eukaryotic RNaseP components could shed some light on RNaseP evolution and especially the ultimate RNaseP ancestor present in LUCA (the last common ancestor between eukaryotes, archaea and bacteria). Because RNaseP has been found in all organisms that have been studied, it is certainly expected to be found in basal eukaryotes. Because of this, RNaseP serves as an important control for the spliceosomal results that come next, as it was found using methods to find an RNP (ribonucleoprotein) complex that is expected to be present. That the same methods find another RNP complex, the spliceosome whose presence and complexity is unknown, gives additional confidence in the spliceosomal results.

5.5: Splicing and the Spliceosome in the Eukaryotic Ancestor

A major aim of this study was to determine if a spliceosome existed in the eukaryotic ancestor, and if so, investigate its complexity. The triple combination of intron analysis, the presence of both snRNA and of spliceosomal proteins is convincing evidence for the presence of a spliceosome in the eukaryotic ancestor. A surprising outcome was the number of spliceosomal components that are conserved between crown and basal eukaryotes. All of the major snRNAs, together with most of their specific proteins are conserved throughout eukaryotes and were thus, likely to have been in the eukaryotic ancestor. Other splicing proteins, including those known to be connected to other cellular processes, such as transcription and capping were also conserved. Of the 153 proteins examined during this study, 75 were determined to have been present in the eukaryotic ancestor, indicating that the spliceosome present in the eukaryotic ancestor was not a simplified version of spliceosomes found in extant species, but of similar complexity. There is evidence to suggest that the present major splicing cycle (Figure 4.2) is conserved across eukaryotes and was present, in a general sense, in the eukaryotic ancestor.

It is also likely that multiple splicing mechanisms may have been present in the eukaryotic ancestor. Major and trans-splicing have been characterised in diverse lineages in crown and basal eukaryotes. Although trans-splicing has not been described for a large number of lineages (nematodes, cestodes, sea-squirt and trypanosomes), its distribution and similarities between the systems found in these eukaryotes suggest a common ancestral mechanism. However, minor splicing (found in animals and plants) could not be determined to be present in any basal eukaryotic lineage and could not be placed in the eukaryotic ancestor at this time. Alternative splicing has been found associated with all three splicing mechanisms (major, minor and trans) and was found throughout both crown and basal eukaryotic lineages. By allowing different mature transcripts from the same mRNA, alternative splicing is thought to have played a role in diversifying protein function throughout eukaryotic evolution (Boue et al. 2003). Given its distribution (as determined in this study) it is likely that the process of alternative splicing was present in the eukaryotic ancestor.

This study could not screen all of the proteins known to be associated with splicing (~200 Jurica and Moore 2003) but concentrated on the most conserved, and hence proteins more likely to be found in basal eukaryotes. The list of splicing-associated proteins is by no means complete. During this study, new proteins were being characterised in the literature on a monthly basis and some proteins were reclassified into different groups. Proteins that were not included in this study included the H-Complex proteins (proteins that bind to pre-mRNA under conditions where splicing is not supported; Jurica and Moore 2003). Similarly many spliceosomal proteins found only in humans (Jurica and Moore 2003) were not tested, although some species specific proteins (e.g. from human and *S. cerevisiae*) were included as a negative control for the searches of basal eukaryotes. Although it was possible that some species-specific proteins may be detected in basal eukaryotes (and hence, would no longer be species specific), we would not expect all to be found. Some proteins that had already been characterised in basal eukaryotic species (e.g. Prp8 and Brr2 from *G. lamblia*; Nixon et al. 2002) were used as positive controls.

This study focused mainly on three basal eukaryotic genomes, *G. lamblia*, *P. falciparum* and *Ent. histolytica* representing three diverse eukaryotic lineages (Figure 1.3, page 4). However, all three of these basal eukaryotes are also parasitic and it is unknown as to the effect that a parasitic life-cycle may have had on splicing and spliceosome composition. Many parasites show a loss of genes compared to their free-living relatives in a process called reductive evolution (Andersson and Kurland 1998). As more basal eukaryotic genomes are completed, especially free-living relatives of the above species, it will be interesting to gauge any consequences of a parasitic lifestyle on intron and splicing characteristics.

There were 72 spliceosomal-protein sequences identified in *G. lamblia* which will be used to further annotate the *G. lamblia* genome. Biochemical analysis may be done on some of these proteins to determine if they are expressed in *G. lamblia* and to confirm their sequences. Although the presence of a number of spliceosomal proteins indicates the presence of a spliceosome, *G. lamblia* to date has still only one characterised intron (A. McArthur, personal communication). It is highly unlikely that a complex spliceosomal process exists (or remains) in *G. lamblia* to process a single intron, thus it is likely that other introns are present. Most small introns characterised in basal eukaryotes have been found by chance when genes of interest have been experimentally analysed (an exception being *Entamoeba* species, Wilihoeft et al. 2001). As the *G. lamblia* genome is further analysed, both computationally and biochemically, it is expected that more introns will be characterised and a clearer picture will emerge of spliceosomal processing in this organism.

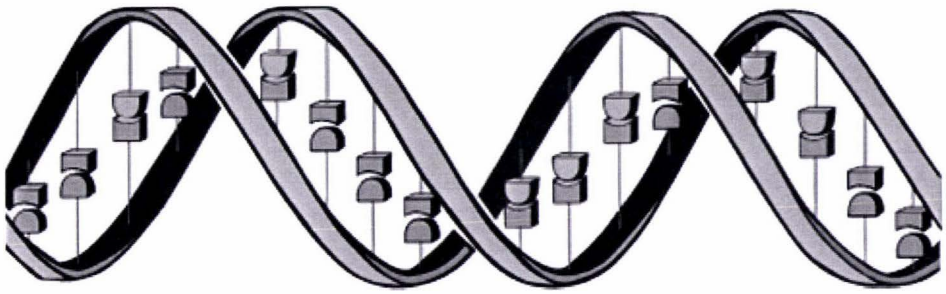
At present, the U5snRNA is the only snRNA characterised from *G. lamblia* (Chapter 2) although the other snRNAs have been found throughout basal eukaryotes and were likely present in the eukaryotic ancestor. It is desirable to construct more snRNA RNAmotif-descriptors to search for other snRNAs (especially the U2 and U6 snRNAs that are thought to form the spliceosomal catalytic core region) in *G. lamblia* and other basal eukaryotes.

The complexity of extant spliceosomes, especially the large number of associated proteins, meant that a large amount of sequence, literature and results data had to be collated, analysed and managed. SpliceSite was developed as a personal database to store and manage the massive amount of information accumulated about the spliceosomal components examined in this project. With some development (with the aid of a professional software developer) this database could become available over the internet to other researchers. However, more immediate aims would be to allow local (research team) access, and to use SpliceSite to develop a genomic-project database template that could then be used on other projects.

It is beyond the scope of this project to examine how the spliceosome evolved in the first eukaryote (Figure 1.5 page 6). Determining which snRNA and protein components may have been present in the first eukaryote would require searches of archaeal genomes (similar to searches in this study). Some spliceosomal proteins (including some U2snRNP-specific and Sm/Lsm proteins) have found in archaeal genomes (Anantharaman et al. 2002). The dataset of spliceosomal proteins compiled here to search basal eukaryotic genomes, could be applied to archaeal genomes.

5.6: Final Remarks

Computational genomics is where bioinformatics meets molecular biology. It is well known that the analysis of genomic data has not kept pace with the sequencing. As W. Martin et al. stated (2003): "In an ideal world, the analysis of genome sequences would have fully uncovered the history of life by now. But as it stands, genome sequencing has mostly uncovered that humans can efficiently sequence genomes". There is a treasure-house of information now scattered over the internet, but to make use of it we first need to find it, organise it, then analyse it. Bioinformatics will play as much a role in finding out how an organism works, as traditional biological sciences. As we struggle to comprehend biological systems in even the 'simplest' organisms, using high-end computing, massive sequence databases and reams of printed literature, it is nice to remember that our cells know what to do and have been doing so since our earliest eukaryotic ancestors.



References

Achsel, T., K. Ahrens, H. Brahms, S. Teigelkamp and R. Luhrmann (1998). "The human U5-220kD protein (hPrp8) forms a stable RNA-free complex with several U5-specific proteins, including an RNA unwindase, a homologue of ribosomal elongation factor EF-2, and a novel WD-40 protein." *Mol Cell Biol* **18**: 6756-66.

Adam, R. D. (2000). "The Giardia lamblia genome." *Int J Parasitol* **30**: 475-84.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, et al. (2000). "The genome sequence of Drosophila melanogaster." *Science* **287**: 2185-95.

Ajuh, P., J. Sleeman, J. Chusainow and A. I. Lamond (2001). "A direct interaction between the carboxyl-terminal region of CDC5L and the WD40 domain of PLRG1 is essential for pre-mRNA splicing." *J Biol Chem* **276**: 42370-81.

Allen, M., A. Friedler, O. Schon and M. Bycroft (2002). "The structure of an FF domain from human HYPA/FBP11." *J Mol Biol* **323**: 411-6.

Altman, S. (1989). "Ribonuclease P: an enzyme with a catalytic RNA subunit." *Adv Enzymol Relat Areas Mol Biol* **62**: 1-36.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." *J Mol Biol* **215**: 403-10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**: 3389-402.

Anantharaman, V., E. V. Koonin and L. Aravind (2002). "Comparative genomics and evolution of proteins involved in RNA metabolism." *Nucleic Acids Res* **30**: 1427-64.

Abrahamsen, M. S., T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu, C. A. Lancto, M. Deng, C. Liu, G. Widmer, S. Tzipori, et al. (2004). "Complete genome sequence of the apicomplexan, Cryptosporidium parvum." *Science* **304**: 441-5.

Andersson, S. G. and C. G. Kurland (1998). "Reductive evolution of resident genomes." *Trends Microbiol* **6**: 263-8.

Andrews, A. J., T. A. Hall and J. W. Brown (2001). "Characterization of RNase P holoenzymes from Methanococcus jannaschii and Methanothermobacter thermoautotrophicus." *Biol Chem* **382**: 1171-7.

Aravind, L., L. M. Iyer and V. Anantharaman (2003). "The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism." *Genome Biol* **4**: R64.

Archibald, J. M. and P. J. Keeling (2002). "Recycled plastids: a 'green movement' in eukaryotic evolution." *Trends Genet* **18**: 577-84.

Archibald, J. M., J. M. Logsdon, Jr. and W. F. Doolittle (2000). "Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes [In Process Citation]." *Mol Biol Evol* **17**: 1456-66.

Archibald, J. M., C. J. O'Kelly and W. F. Doolittle (2002). The Chaperonin Genes of Jakobid and Jakobid-Like Flagellates: Implications for Eukaryotic Evolution. *Mol Biol Evol* **19:** 422-431.

Arisue, N., L. B. Sanchez, L. M. Weiss, M. Muller and T. Hashimoto (2002). "Mitochondrial-type hsp70 genes of the amitochondriate protists, Giardia intestinalis, Entamoeba histolytica and two microsporidians." *Parasitol Int* **51**: 9-16.

Arndt, E., Kromer,W. and Hatakeyama,T. (1990). "Organization and nucleotide sequence of a gene cluster coding for eight ribosomal proteins in the archaebacterium Halobacterium marismortui." *J. . Biol. Chem.* **265**: 3034-3039.

Arnold, J. and N. Hilton (2003). "Genome sequencing: Revelations from a bread mould." *Nature* **422**: 821-2.

Auer, J., K. Lechner and A. Bock (1989). "Gene organization and structure of two transcriptional units from Methanococcus coding for ribosomal proteins and elongation factors." *Can J Microbiol* **35**: 200-4.

Badea, L. (2003). "Functional Discrimination of Gene Expression Patterns in Terms of the Gene Ontology." *Pacific Symposium on Biocomputing* **8**: 565-576.

Bahl, A., B. Brunk, R. L. Coppel, J. Crabtree, S. J. Diskin, M. J. Fraunholz, G. R. Grant, D. Gupta, R. L. Huestis, J. C. Kissinger, et al. (2002). "PlasmoDB: the Plasmodium genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished)." *Nucleic Acids Res* **30**: 87-90.

Bakatselou, C., D. Beste, A. O. Kadri, S. Somanath and C. G. Clark (2003). "Analysis of genes of mitochondrial origin in the genus Entamoeba." *J Eukaryot Microbiol* **50**: 210-4.

Baldauf, S. L. (2003). "The deep roots of eukaryotes." *Science* **300**: 1703-6.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, et al. (2002). "The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba." *Proc Natl Acad Sci U S A* **99**: 1414-9.

Bartels, C., H. Urlaub, R. Luhrmann and P. Fabrizio (2003). "Mutagenesis suggests several roles of Snu114p in pre-mRNA splicing." *J Biol Chem* **278**: 28324-34.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, et al. (2004). "The Pfam protein families database." *Nucleic Acids Res* **32 Database issue**: D138-41.

Bell, M. and A. Bindereif (1999). "Cloning and mutational analysis of the Leptomonas seymouri U5 snRNA gene: function of the Sm site in core RNP formation and nuclear localization." *Nucleic Acids Res* **27**: 3986-94.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp and D. L. Wheeler (2000). "GenBank." *Nucleic Acids Res* **28**: 15-8.

Bergeron, B. P. (2003). Bioinformatics computing. Upper Saddle River, New Jersey, Prentice Hall.

Berget, S. M. (1995). "Exon recognition in vertebrate splicing." *J Biol Chem* **270**: 2411-4.

Berman, H. M., T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig and J. Westbrook (2000). "The Protein Data Bank and the challenge of structural genomics." *Nat Struct Biol* **7 Suppl**: 957-9.

Berry, L. D. and K. L. Gould (1997). "Fission yeast dim1(+) encodes a functionally conserved polypeptide essential for mitosis." *J Cell Biol* **137**: 1337-54.

Birney, E. and R. Durbin (1997). "Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison." *Proc Int Conf Intell Syst Mol Biol* **5**: 56-64.

Bodenreider, O., J. A. Mitchell and A. T. McCray (2003). "Biomedical Ontologies." *Pacific Symposium on Biocomputing* **8**: 562-564.

Bonen, L. (1993). "Trans-splicing of pre-mRNA in plants, animals, and protists." *Faseb J* **7**: 40-6.

Boomershine, W. P., C. A. McElroy, H. Y. Tsai, R. C. Wilson, V. Gopalan and M. P. Foster (2003). "Structure of Mth11/Mth Rpp29, an essential protein subunit of archaeal and eukaryotic RNase P." *Proc Natl Acad Sci U S A* **100**: 15398-403.

Boue, S., I. Letunic and P. Bork (2003). "Alternative splicing and evolution." *Bioessays* **25**: 1031-4.

Breckenridge, D. G., Y. Watanabe, S. J. Greenwood, M. W. Gray and M. N. Schnare (1999). "U1 small nuclear RNA and spliceosomal introns in Euglena gracilis." *Proc Natl Acad Sci U S A* **96**: 852-6.

Brown, J. W. (1999). "The Ribonuclease P Database." *Nucleic Acids Res* **27**: 314.

Bruzik, J. P. and T. Maniatis (1992). "Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells." *Nature* **360**: 692-5.

Bult, C. J., J. A. Blake, J. E. Richardson, J. A. Kadin, J. T. Eppig, R. M. Baldarelli, K. Barsanti, M. Baya, J. S. Beal, W. J. Boddy, et al. (2004). "The Mouse Genome Database (MGD): integrating biology with the genome." *Nucleic Acids Res* **32** Database issue: D476-81.

Burge, C. B., R. A. Padgett and P. A. Sharp (1998). "Evolutionary fates and origins of U12-type introns." *Mol Cell* **2**: 773-85.

Burge, C. B., T. Tuschl and P. A. Sharp (1999). Splicing of precursors to mRNAs by the Spliceosomes. The RNA World : the nature of modern RNA suggests a prebiotic RNA. R. F. Gesteland, T. R. Cech and J. f. Atkins. Cold spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Canaday, J., L. H. Tessier, P. Imbault and F. Paulus (2001). "Analysis of Euglena gracilis alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation." *Mol Genet Genomics* **265**: 153-60.

Caudevilla, C., C. Codony, D. Serra, G. Plasencia, R. Roman, A. Graessmann, G. Asins, M. Bach-Elias and F. G. Hegardt (2001). "Localization of an exonic splicing enhancer responsible for mammalian natural trans-splicing." *Nucleic Acids Res* **29**: 3108-15.

Cavalier-Smith, T. (1989). "Molecular phylogeny. Archaebacteria and Archezoa." *Nature* **339**: 100-1.

Cavalier-Smith, T. and E. E. Chao (1996). "Molecular phylogeny of the free-living archezoan Trepomonas agilis and the nature of the first eukaryote." *J Mol Evol* **43**: 551-62.

Chamberlain, J. R., Y. Lee, W. S. Lane and D. R. Engelke (1998). "Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP." *Genes Dev* **12**: 1678-90.

Chan, S. P., D. I. Kao, W. Y. Tsai and S. C. Cheng (2003). "The Prp19p-associated complex in spliceosome activation." *Science* **302**: 279-82.

Chandrasekharan, U. M., S. Sanker, M. J. Glynias, S. S. Karnik and A. Husain (1996). "Angiotensin II-forming activity in a reconstructed ancestral chymase." *Science* **271**: 502-5.

Chang, B. S. and M. J. Donoghue (2000). "Recreating ancestral proteins." *Trends in Ecology and Evolution* **15**: 109-114.

Chawla, G., A. K. Sapra, U. Surana and U. Vijayraghavan (2003). "Dependence of pre-mRNA introns on PRP17, a non-essential splicing factor: implications for efficient progression through cell cycle transitions." *Nucleic Acids Res* **31**: 2333-43.

Chekanova, J. A. and D. A. Belostotsky (2003). "Evidence that poly(A) binding protein has an evolutionarily conserved function in facilitating mRNA biogenesis and export." *Rna* **9**: 1476-90.

Collins, L. J., V. Moulton and D. Penny (2000). "Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP." *J Mol Evol* **51**: 194-204.

Collins, L. J., A. M. Poole and D. Penny (2003). "Using ancestral sequences to uncover potential gene homologues." *Applied Bioinformatics* **2**: S85-S95.

Cornell, M., N. W. Paton, C. Hedeler, P. Kirby, D. Delneri, A. Hayes and S. G. Oliver (2003). "GIMS: an integrated data storage and analysis environment for genomic and functional data." *Yeast* **20**: 1291-306.

Courties, C., R. Perasso, M.-J. Chrétiennot-Dinet, M. Gouy, L. Guillou and M. Troussellier (1998). "Phylogenetic analysis and genome size of Ostreococcus tauri (Chlorophyta, Prasinophyceae)." *Journal of Phycology* **34**: 844-849.

Dacks, J. B., L. A. Davis, A. M. Sjogren, J. O. Andersson, A. J. Roger and W. F. Doolittle (2003). "Evidence for Golgi bodies in proposed 'Golgi-lacking' lineages." *Proc R Soc Lond B Biol Sci* **270 Suppl 2**: S168-71.

Dacks, J. B. and W. F. Doolittle (2001). "Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help." *Cell* **107**: 419-25.

Dacks, J. B., A. Marinets, W. Ford Doolittle, T. Cavalier-Smith and J. M. Logsdon, Jr. (2002). "Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang." *Mol Biol Evol* **19**: 830-40.

de la Cruz, J. and A. Vioque (2003). "A structural and functional study of plastid RNAs homologous to catalytic bacterial RNase P RNA." *Gene* **321**: 47-56.

Dehal, P., Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, et al. (2002). "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." *Science* **298**: 2157-67.

Dessen, P., M. Zagulski, R. Gromadka, H. Plattner, R. Kissmehl, E. Meyer, M. Betermier, J. E. Schultz, J. U. Linder, R. E. Pearlman, et al. (2001). "Paramecium genome survey: a pilot project." *Trends Genet* **17**: 306-8.

Dix, I., C. S. Russell, R. T. O'Keefe, A. J. Newman and J. D. Beggs (1998). "Protein-RNA interactions in the U5 snRNP of Saccharomyces cerevisiae." *Rna* **4**: 1675-86.

Djikeng, A., L. Ferreira, M. D'Angelo, P. Dolezal, T. Lamb, S. Murta, V. Triggs, S. Ulbert, A. Villarino, S. Renzi, et al. (2001). "Characterization of a candidate Trypanosoma brucei U1 small nuclear RNA gene." *Mol Biochem Parasitol* **113**: 109-15.

Dodgson, J. B. (2003). "Chicken genome sequence: a centennial gift to poultry genetics." *Cytogenet Genome Res* **102**: 291-6.

Donahue, W. F. and K. A. Jarrell (2002). "A BLAST from the past: ancient origin of human Sm proteins." *Mol Cell* **9**: 7-8.

Doolittle, W. F. (1998). "A paradigm gets shifty." *Nature* **392**: 15-6.

Dsouza, M., N. Larsen and R. Overbeek (1997). "Searching for patterns in genomic data." *Trends Genet* **13**: 497-8.

Durbin, R. (1998). Biological sequence analysis : probabalistic models of proteins and nucleic acids. Cambridge, UK New York, Cambridge University Press.

Ebel, C., C. Frantz, F. Paulus and P. Imbault (1999). "Trans-splicing and cis-splicing in the colourless Euglenoid, Entosiphon sulcatum." *Curr Genet* **35**: 542-50.

Eddy, S. R. (1998). "Profile hidden Markov models." *Bioinformatics* **14**: 755-63.

Eddy, S. R. (2001). "Non-coding RNA genes and the modern RNA world." *Nat Rev Genet* **2**: 919-29.

Eddy, S. R. (2002). "Computational genomics of noncoding RNA genes." *Cell* **109**: 137-40.

Eddy, S. R. and R. Durbin (1994). "RNA sequence analysis using covariance models." *Nucleic Acids Res* **22**: 2079-88.

Edvardsson, S., P. P. Gardner, A. M. Poole, M. D. Hendy, D. Penny and V. Moulton (2003). "A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction." *Bioinformatics* **19**: 865-873.

Eichinger, L. and A. A. Noegel (2003). "Crawling into a new era-the Dictyostelium genome project." *Embo J* **22**: 1941-6.

Elmendorf, H. G., S. M. Singer and T. E. Nash (2001). "The abundance of sterile transcripts in Giardia lamblia." *Nucleic Acids Res* **29**: 4674-83.

Embley, T. M. and R. P. Hirt (1998). "Early branching eukaryotes?" *Curr Opin Genet Dev* **8**: 624-9.

Embley, T. M., M. van der Giezen, D. S. Horner, P. L. Dyal, S. Bell and P. G. Foster (2003). "Hydrogenosomes, mitochondria and early eukaryotic evolution." *IUBMB Life* **55**: 387-95.

Eul, J., M. Graessmann and A. Graessmann (1995). "Experimental evidence for RNA trans-splicing in mammalian cells." *Embo J* **14**: 3226-35.

Fast, N. M. and W. F. Doolittle (1999). "Trichomonas vaginalis possesses a gene encoding the essential spliceosomal component, PRP8." *Mol Biochem Parasitol* **99**: 275-8.

Fetzer, C. P., D. J. Hogan and H. J. Lipps (2002). "A PIWI homolog is one of the proteins expressed exclusively during macronuclear development in the ciliate Stylonychia lemnae." *Nucleic Acids Res* **30**: 4380-6.

Fogel, G. B., V. W. Porto, D. G. Weekes, D. B. Fogel, R. H. Griffey, J. A. McNeil, E. Lesnik, D. J. Ecker and R. Sampath (2002). "Discovery of RNA structural elements using evolutionary computation." *Nucleic Acids Res* **30**: 5310-7.

Forch, P., L. Merendino, C. Martinez and J. Valcarcel (2003). "U2 small nuclear ribonucleoprotein particle (snRNP) auxiliary factor of 65 kDa, U2AF65, can promote U1 snRNP recruitment to 5' splice sites." *Biochem J* **372**: 235-40.

Forch, P., O. Puig, C. Martinez, B. Seraphin and J. Valcarcel (2002). "The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites." *Embo J* **21**: 6882-92.

Fortes, P., D. Bilbao-Cortes, M. Fornerod, G. Rigaut, W. Raymond, B. Seraphin and I. W. Mattaj (1999). "Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition." *Genes Dev* **13**: 2425-38.

Frank, D. N., C. Adamidi, M. A. Ehringer, C. Pitulle and N. R. Pace (2000). "Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA." *RNA* **6**: 1895-904.

Frank, D. N., H. Roiha and C. Guthrie (1994). "Architecture of the U5 small nuclear RNA." *Mol Cell Biol* **14**: 2180-90.

Frantz, C., C. Ebel, F. Paulus and P. Imbault (2000). "Characterization of trans-splicing in Euglenoids." *Curr Genet* **37**: 349-55.

Furuyama, S. and J. P. Bruzik (2002). "Multiple roles for SR proteins in trans splicing." *Mol Cell Biol* **22**: 5337-46.

Galagan, J. E., S. E. Calvo, K. A. Borkovich, E. U. Selker, N. D. Read, D. Jaffe, W. FitzHugh, L. J. Ma, S. Smirnov, S. Purcell, et al. (2003). "The genome sequence of the filamentous fungus Neurospora crassa." *Nature* **422**: 859-68.

Garcia-Blanco, M. A. (2003). "Messenger RNA reprogramming by spliceosome-mediated RNA trans-splicing." *J Clin Invest* **112**: 474-80.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, et al. (2002). "Genome sequence of the human malaria parasite Plasmodium falciparum." *Nature* **419**: 498-511.

Gautheret, D. and A. Lambert (2001). "Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles." *J Mol Biol* **313**: 1003-11.

Gesteland, R. F., T. Cech and J. F. Atkins (1999). The RNA world : the nature of modern RNA suggests a prebiotic RNA. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Gilson, P. R. and G. I. McFadden (1996). "The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns." *Proc Natl Acad Sci U S A* **93**: 7737-42.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, et al. (1996). "Life with 6000 genes." *Science* **274**: 546, 563-7.

Golas, M. M., B. Sander, C. L. Will, R. Luhrmann and H. Stark (2003). "Molecular architecture of the multiprotein splicing factor SF3b." *Science* **300**: 980-4.

Gonzalez-Santos, J. M., A. Wang, J. Jones, C. Ushida, J. Liu and J. Hu (2002). "Central region of the human splicing factor Hprp3p interacts with Hprp4p." *J Biol Chem* **277**: 23764-72.

Gopalan, V., T. W. Tan, B. T. Lee and S. Ranganathan (2004). "Xpro: database of eukaryotic protein-encoding genes." *Nucleic Acids Res* **32**: D59-63.

Gottschalk, A., G. Neubauer, J. Banroques, M. Mann, R. Luhrmann and P. Fabrizio (1999). "Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6.U5] tri-snRNP." *Embo J* **18**: 4535-48.

Gottschalk, A., J. Tang, O. Puig, J. Salgado, G. Neubauer, H. V. Colot, M. Mann, B. Seraphin, M. Rosbash, R. Luhrmann, et al. (1998). "A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins." *Rna* **4**: 374-93.

Graveley, B. R. (2004). "A protein interaction domain contacts RNA in the prespliceosome." *Mol Cell* **13**: 302-4.

Gribaldo, S. and H. Philippe (2002). "Ancient phylogenetic relationships." *Theor Popul Biol* **61**: 391-408.

Griffiths-Jones, S. (2004). "The MicroRNA Registry." *Nucleic Acids Res* **32** Database issue: D109-11.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna and S. R. Eddy (2003). "Rfam: an RNA family database." *Nucleic Acids Res* **31**: 439-41.

Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." *Nucl. Acids. Symp. Ser* **41**: 95-98.

Hall, T. A. and J. W. Brown (2002). "Archaeal RNase P has multiple protein subunits homologous to eukaryotic nuclear RNase P proteins." *Rna* **8**: 296-306.

Han, L. Y., C. Z. Cai, S. L. Lo, M. C. Chung and Y. Z. Chen (2004). "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach." *Rna* **10**: 355-68.

Harris, J. K., E. S. Haas, D. Williams, D. N. Frank and J. W. Brown (2001). "New insight into RNase P RNA structure from comparative analysis of the archaeal RNA." *Rna* **7**: 220-32.

Hastings, M. L. and A. R. Krainer (2001). "Functions of SR proteins in the U12-dependent AT-AC pre-mRNA splicing pathway." *Rna* **7**: 471-82.

Hendy, M. D. and D. Penny (1989). "A Framework for the Quantitative Study of Evolutionary Trees." *Systematic Zoology* **38**: 297-309.

Hennig, L. (1999). "WinGene/WinPep: user-friendly software for the analysis of amino acid sequences." *Biotechniques* **26**: 1170-2.

Henze, K. and W. Martin (2003). "Evolutionary biology: essence of mitochondria." *Nature* **426**: 127-8.

Herbert, A. (2004). "The four Rs of RNA-directed evolution." *Nat Genet* **36**: 19-25.

Hofacker, I. L. (2003). "Vienna RNA secondary structure server." *Nucleic Acids Res* **31**: 3429-31.

Horner, D. S. and T. M. Embley (2001). "Chaperonin 60 phylogeny provides further evidence for secondary loss of mitochondria among putative early-branching eukaryotes." *Mol Biol Evol* **18**: 1970-5.

Horowitz, D. S., E. J. Lee, S. A. Mabon and T. Misteli (2002). "A cyclophilin functions in pre-mRNA splicing." *Embo J* **21**: 470-80.

Huestis, R. and K. Fischer (2001). "Prediction of many new exons and introns in Plasmodium falciparum chromosome 2." *Mol Biochem Parasitol* **118**: 187-99.

Ismaili, N., D. Perez-Morga, P. Walsh, M. Cadogan, A. Pays, P. Tebabi and E. Pays (2000). "Characterization of a Trypanosoma brucei SR domain-containing protein bearing homology to cis-spliceosomal U1 70 kDa proteins." *Mol Biochem Parasitol* **106**: 109-20.

Jagadish, H. V. and F. Olken (2003). Data Management for the BioSciences: Report of the NSF/NLM workshop on Data Management for Molecular and Cell Biology. LBNL Report LBNL-52767.

Jain, R., M. C. Rivera and J. A. Lake (1999). "Horizontal gene transfer among genomes: the complexity hypothesis." *Proc Natl Acad Sci U S A* **96**: 3801-6.

Jarrous, N., P. S. Eder, C. Guerrier-Takada, C. Hoog and S. Altman (1998). "Autoantigenic properties of some protein subunits of catalytically active complexes of human ribonuclease P." *Rna* **4**: 407-17.

Jarrous, N., P. S. Eder, D. Wesolowski and S. Altman (1999). "Rpp14 and Rpp29, two protein subunits of human ribonuclease P." *Rna* **5**: 153-7.

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

Jiang, T. and S. Altman (2001). "Protein-protein interactions with subunits of human nuclear RNase P." *Proc Natl Acad Sci U S A* **98**: 920-925.

Jiang, T. and S. Altman (2002). "A protein subunit of human RNase P, Rpp14, and its interacting partner, OIP2, have 3'-->5' exoribonuclease activity." *Proc Natl Acad Sci U S A* **99**: 5295-300.

Jiang, T., C. Guerrier-Takada and S. Altman (2001). "Protein-RNA interactions in the subunits of human nuclear RNase P." *RNA* **7**: 937-41.

Jurica, M. S. and M. J. Moore (2003). "Pre-mRNA splicing: awash in a sea of proteins." *Mol Cell* **12**: 5-14.

Kabnick, K. S. and D. A. Peattie (1990). "In situ analyses reveal that the two nuclei of Giardia lamblia are equivalent." *J Cell Sci* **95 (Pt 3)**: 353-60.

Karaoz, U., T. M. Murali, S. Letovsky, Y. Zheng, C. Ding, C. R. Cantor and S. Kasif (2004). "Whole-genome annotation by using evidence integration in functional-linkage networks." *Proc Natl Acad Sci U S A* **101**: 2888-93.

Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, et al. (2001). "Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi." *Nature* **414**: 450-3.

Kaufer, N. F. and J. Potashkin (2000). "Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals." *Nucleic Acids Res* **28**: 3003-10.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, et al. (1998). "Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3 (supplement)." *DNA Res* **5**: 147-55.

Khan, A. U. and S. K. Lal (2003). "Ribozymes: a modern tool in medicine." *J Biomed Sci* **10**: 457-67.

King, R. D., P. H. Wise and A. Clare (2004). "Confirmation of data mining based predictions of protein function." *Bioinformatics*.

Kiss, T. and W. Filipowicz (1992). "Evidence against a mitochondrial location of the 7-2/MRP RNA in mammalian cells." *Cell* **70**: 11-6.

Klein, R. J. and S. R. Eddy (2003). "RSEARCH: Finding homologs of single structured RNA sequences." *BMC Bioinformatics* **4**: 44.

Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, et al. (1997). "The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus." *Nature* **390**: 364-70.

Kondrashov, F. A. and E. V. Koonin (2001). "Origin of alternative splicing by tandem exon duplication." *Hum Mol Genet* **10**: 2661-9.

Koonin, E. V., L. Aravind and A. S. Kondrashov (2000). "The impact of comparative genomics on our understanding of evolution." *Cell* **101**: 573-6.

Koonin, E. V., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, et al. (2004). "A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes." *Genome Biol* **5**: R7.

Koonin, E. V., Y. I. Wolf and L. Aravind (2001). "Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach." *Genome Res* **11**: 240-52.

Koonin, E. V., Y. I. Wolf and G. P. Karev (2002). "The structure of the protein universe and genome evolution." *Nature* **420**: 218-23.

Koshi, J. M. and R. A. Goldstein (1996). "Probabilistic reconstruction of ancestral protein sequences." *J Mol Evol* **42**: 313-20.

Kouzuma, Y., M. Mizoguchi, H. Takagi, H. Fukuhara, M. Tsukamoto, T. Numata and M. Kimura (2003). "Reconstitution of archaeal ribonuclease P from RNA and four protein components." *Biochem Biophys Res Commun* **306**: 666-73.

Kramer, A. (1996). "The structure and function of proteins involved in mammalian pre-mRNA splicing." *Annu Rev Biochem* **65**: 367-409.

Kuhn, A. N. and N. F. Kaufer (2003). "Pre-mRNA splicing in Schizosaccharomyces pombe: regulatory role of a kinase conserved from fission yeast to mammals." *Curr Genet* **42**: 241-51.

Kuhn, A. N., E. M. Reichl and D. A. Brow (2002). "Distinct domains of splicing factor Prp8 mediate different aspects of spliceosome activation." *Proc Natl Acad Sci U S A* **99**: 9145-9.

Labourier, E. and D. C. Rio (2001). "Purification of Drosophila snRNPs and characterization of two populations of functional U1 particles." *Rna* **7**: 457-70.

Labrador, M. and V. G. Corces (2003). "Extensive exon reshuffling over evolutionary time coupled to trans-splicing in Drosophila." *Genome Res* **13**: 2220-8.

Laferriere, A., D. Gautheret and R. Cedergren (1994). "An RNA pattern matching program with enhanced performance and portability." *Comput Appl Biosci* **10**: 211-2.

Lam, B. J. and K. J. Hertel (2002). "A general role for splicing enhancers in exon definition." *Rna* **8**: 1233-41.

Lambrix, P. and A. Edberg (2003). "Evaluation of Ontology Merging Tools in Bioinformatics." *Pacific Symposium on Biocomputing* **8**: 589-600.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* **409**: 860-921.

Lehner, B., G. Williams, R. D. Campbell and C. M. Sanderson (2002). "Antisense transcripts in the human genome." *Trends Genet* **18**: 63-5.

Li, L. and C. C. Wang (2004). "Capped mRNA with a single nucleotide leader is optimally translated in a primitive eukaryote, giardia lamblia." *J Biol Chem*.

Li, Y. and S. Altman (2003). "A specific endoribonuclease, RNase P, affects gene expression of polycistronic operon mRNAs." *Proc Natl Acad Sci U S A* **100**: 13213-8.

Liang, X. H., A. Haritan, S. Uliel and S. Michaeli (2003). "trans and cis splicing in trypanosomatids: mechanism, factors, and regulation." *Eukaryot Cell* **2**: 830-40.

Lin, X., Kaul,S., Rounsley,S.D., Shea,T.P., Benito,M.-I., Town,C.D.,, C. Y. Fujii, Mason,T.M., Bowman,C.L., Barnstead,M.E.,, T. V. Feldblyum, Buell,C.R., Ketchum,K.A., Lee,J.J., Ronning,C.M.,, H. Koo, Moffat,K.S., Cronin,L.A., Shen,M., VanAken,S.E., Umayam,L.,, L. J. Tallon, Gill,J.E., Adams,M.D., Carrera,A.J., Creasy,T.H.,, H. M. Goodman, Somerville,C.R., Copenhaver,G.P., Preuss,D.,, W. C. Nierman, White,O., Eisen,J.A., Salzberg,S.L., Fraser,C.M. and and J. C. Venter (1999). "Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana." *Nature* **402**: 761-768.

Liu, Q., X. H. Liang, S. Uliel, M. Belahcen, R. Unger and S. Michaeli (2004). "Identification and Functional Characterization of Lsm Proteins in Trypanosoma brucei." *J Biol Chem* **279**: 18210-9.

Lloyd, D., J. C. Harris, S. Maroulis, R. Wadley, J. R. Ralphs, A. C. Hann, M. P. Turner and M. R. Edwards (2002). "The "primitive" microaerophile Giardia intestinalis (syn. lamblia, duodenalis) has specialized membranes with electron transport and membrane-potential-generating functions." *Microbiology* **148**: 1349-54.

Lockhart, S. R. and B. C. Rymond (1994). "Commitment of yeast pre-mRNA to the splicing pathway requires a novel U1 small nuclear ribonucleoprotein polypeptide, Prp39p." *Mol Cell Biol* **14**: 3623-33.

Lorkovic, Z. J., D. A. Wieczorek Kirk, M. H. Lambermon and W. Filipowicz (2000). "Pre-mRNA splicing in higher plants." *Trends Plant Sci* **5**: 160-7.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." *Nucleic Acids Res* **25**: 955-64.

Lucke, S., T. Klockner, Z. Palfi, M. Boshart and A. Bindereif (1997). "Trans mRNA splicing in trypanosomes: cloning and analysis of a PRP8-homologous gene from Trypanosoma brucei provides evidence for a U5-analogous RNP." *Embo J* **16**: 4433-40.

Luo, H. R., G. A. Moreau, N. Levin and M. J. Moore (1999). "The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes." *Rna* **5**: 893-908.

Luo, M. L., Z. Zhou, K. Magni, C. Christoforides, J. Rappsilber, M. Mann and R. Reed (2001). "Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly." *Nature* **413**: 644-7.

Lybarger, S., K. Beickman, V. Brown, N. Dembla-Rajpal, K. Morey, R. Seipelt and B. C. Rymond (1999). "Elevated levels of a U4/U6.U5 snRNP-associated protein, Spp381p, rescue a mutant defective in spliceosome maturation." *Mol Cell Biol* **19**: 577-84.

Lygerou, Z., H. Pluk, W. J. van Venrooij and B. Seraphin (1996). "hPop1: an autoantigenic protein subunit shared by the human RNase P and RNase MRP ribonucleoproteins." *Embo J* **15**: 5936-48.

Lynch, M. and A. O. Richardson (2002). "The evolution of spliceosomal introns." *Curr Opin Genet Dev* **12**: 701-10.

Macke, T. J., D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case and R. Sampath (2001). "RNAMotif, an RNA secondary structure definition and search algorithm." *Nucleic Acids Res* **29**: 4724-35.

Mair, G., H. Shi, H. Li, A. Djikeng, H. O. Aviles, J. R. Bishop, F. H. Falcone, C. Gavrilescu, J. L. Montgomery, M. I. Santori, et al. (2000). "A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA." *Rna* **6**: 163-9.

Makarova, O. V., E. M. Makarov, S. Liu, H. P. Vornlocher and R. Luhrmann (2002). "Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6*U5 tri-snRNP formation and pre-mRNA splicing." *Embo J* **21**: 1148-57.

Makarova, O. V., E. M. Makarov and R. Luhrmann (2001). "The 65 and 110 kDa SR-related proteins of the U4/U6.U5 tri-snRNP are essential for the assembly of mature spliceosomes." *Embo J* **20**: 2553-63.

Malca, H., N. Shomron and G. Ast (2003). "The U1 snRNP base pairs with the 5' splice site within a penta-snRNP complex." *Mol Cell Biol* **23**: 3442-55.

Maniatis, T. and B. Tasic (2002). "Alternative pre-mRNA splicing and proteome expansion in metazoans." *Nature* **418**: 236-43.

Mann, B. J. (2002). "Entamoeba histolytica Genome Project: an update." *Trends Parasitol* **18**: 147-8.

Mann, H., Y. Ben-Asouli, A. Schein, S. Moussa and N. Jarrous (2003). "Eukaryotic RNase P: role of RNA and protein subunits of a primordial catalytic ribonucleoprotein in RNA-based catalysis." *Mol Cell* **12**: 925-35.

Manning-Cela, R., A. Gonzalez and J. Swindle (2002). "Alternative splicing of LYT1 transcripts in Trypanosoma cruzi." *Infect Immun* **70**: 4726-8.

Mansfield, S. G., R. H. Clark, M. Puttaraju, J. Kole, J. A. Cohn, L. G. Mitchell and M. A. Garcia-Blanco (2003). "5' exon replacement and repair by spliceosome-mediated RNA trans-splicing." *Rna* **9**: 1290-7.

Mathe, C., M. F. Sagot, T. Schiex and P. Rouze (2002). "Current methods of gene prediction, their strengths and weaknesses." *Nucleic Acids Res* **30**: 4103-17.

Matsuzaki, M., O. Misumi, I. T. Shin, S. Maruyama, M. Takahara, S. Y. Miyagishima, T. Mori, K. Nishida, F. Yagisawa, Y. Yoshida, et al. (2004). "Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D." *Nature* **428**: 653-7.

Mattick, J. S. (2001). "Non-coding RNAs: the architects of eukaryotic complexity." *EMBO Rep* **2**: 986-91.

McArthur, A. G., H. G. Morrison, J. E. J. Nixon, N. Q. E. Passamaneck, U. Kim, G. Hinkle, M. K. Crocker, M. E. Holder, R. Farr and C. I. Reich (2000). "The Giardia genome project database." *FEMS Microbiology Letters* **189**: 271-273.

McConnell, T. S., R. P. Lokken and J. A. Steitz (2003). "Assembly of the U1 snRNP involves interactions with the backbone of the terminal stem of U1 snRNA." *Rna* **9**: 193-201.

McConnell, T. S. and J. A. Steitz (2001). "Proximity of the invariant loop of U5 snRNA to the second intron residue during pre-mRNA splicing." *Embo J* **20**: 3577-86.

Meyer, C. and R. Giegerich (2002). "Matching and Significance Evaluation of Combined Sequence-Structure Motifs in RNA." *Z. Phys. Chem.* **216**: 193-216.

Miriami, E., R. Sperling, J. Sperling and U. Motro (2004). "Regulation of splicing: the importance of being translatable." *Rna* **10**: 1-4.

Miyokawa, T., T. Urayama, K. Shimooka and T. Itoh (1996). "Organization and nucleotide sequences of ten ribosomal protein genes from the region equivalent to the S10 operon in the archaebacterium, Halobacterium halobium." *Biochem Mol Biol Int* **39**: 1209-20.

Modrek, B. and C. J. Lee (2003). "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss." *Nat Genet* **34**: 177-80.

Mossel, E. (2003). "On the impossibility of reconstructing ancestral data and phylogenies." *Journal of Computational Biology* **10**: 669-676.

Mossel, E. and M. Steel (2004). "A phase transition for a random cluster model on phylogenetic trees." *Mathematical Biosciences* **187**: 189-203.

Murphy, M. W., B. L. Olson and P. G. Siliciano (2004). "The yeast splicing factor Prp40p contains functional leucine-rich nuclear export signals that are essential for splicing." *Genetics* **166**: 53-65.

Nagai, K., Y. Muto, D. A. Pomeranz Krummel, C. Kambach, T. Ignjatovic, S. Walke and A. Kuglstatter (2001). "Structure and assembly of the spliceosomal snRNPs. Novartis Medal Lecture." *Biochem Soc Trans* **29**: 15-26.

Nelissen, R. L., C. L. Will, W. J. van Venrooij and R. Luhrmann (1994). "The association of the U1-specific 70K and C proteins with U1 snRNPs is mediated in part by common U snRNP proteins." *Embo J* **13**: 4113-25.

Newman, A. J. (1997). "The role of U5 snRNP in pre-mRNA splicing." *Embo J* **16**: 5797-800.

Nickle, D. C., M. A. Jensen, G. S. Gottlieb, D. Shriner, G. H. Learn, A. G. Rodrigo and J. I. Mullins (2003). "Consensus and ancestral state HIV vaccines." *Science* **299**: 1515-8; author reply 1515-8.

Nilsen, T. W. (1995). "trans-splicing: an update." *Mol Biochem Parasitol* **73**: 1-6.

Nilsen, T. W. (2003). "The spliceosome: the most complex macromolecular machine in the cell?" *Bioessays* **25**: 1147-9.

Nixon, J. E., A. Wang, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus and J. Samuelson (2002). "A spliceosomal intron in Giardia lamblia." *Proc Natl Acad Sci U S A* **19**: 19.

Nott, A., H. Le Hir and M. J. Moore (2004). "Splicing enhances translation in mammalian cells: an additional function of the exon junction complex " *Genes Dev* **18**: 210-22.

Ohi, M. D. and K. L. Gould (2002). "Characterization of interactions among the Cef1p-Prp19p-associated splicing complex." *Rna* **8**: 798-815.

O'Keefe, R. T. (2002). "Mutations in U5 snRNA loop 1 influence the splicing of different genes in vivo." *Nucleic Acids Res* **30**: 5476-84.

Paine, C. T., M. L. Paine, W. Luo, C. T. Okamoto, S. P. Lyngstadaas and M. L. Snead (2000). "A tuftelin-interacting protein (TIP39) localizes to the apical secretory pole of mouse ameloblasts." *J Biol Chem* **275**: 22284-92.

Patel, A. A. and J. A. Steitz (2003). "Splicing double: insights from the second spliceosome." *Nat Rev Mol Cell Biol* **4**: 960-70.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." *Proc Natl Acad Sci US A* **85**: 2444-8.

Peng, R., B. T. Dye, I. Perez, D. C. Barnard, A. B. Thompson and J. G. Patton (2002). "PSF and p54nrb bind a conserved stem in U5 snRNA." *Rna* **8**: 1334-47.

Penny, D. and A. Poole (1999). "The nature of the last universal common ancestor." *Curr Opin Genet Dev* **9**: 672-7.

Pesole, G., S. Liuni and M. D'Souza (2000). "PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance." *Bioinformatics* **16**: 439-50.

Philippe, H. and A. Germot (2000). "Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution." *Mol Biol Evol* **17**: 830-4.

Philippe, H., A. Germot and D. Moreira (2000). "The new phylogeny of eukaryotes." *Curr Opin Genet Dev* **10**: 596-601.

Philippi, S. (2004). "Light-weight integration of molecular biological databases." *Bioinformatics* **20**: 51-7.

Pirrotta, V. (2002). "Trans-splicing in Drosophila." *Bioessays* **24**: 988-91.

Portal, D., J. M. Espinosa, G. S. Lobo, S. Kadener, C. A. Pereira, M. De La Mata, Z. Tang, R. J. Lin, A. R. Kornblihtt, F. E. Baralle, et al. (2003). "An early ancestor in the evolution of splicing: a Trypanosoma cruzi serine-arginine-rich protein (TcSR) is functional in cis-splicing." *Mol Biochem Parasitol* **127**: 37-46.

Pupko, T., I. Pe'er, R. Shamir and D. Graur (2000). "A fast algorithm for joint reconstruction of ancestral amino acid sequences." *Mol Biol Evol* **17**: 890-6.

Rappsilber, J., U. Ryder, A. I. Lamond and M. Mann (2002). "Large-scale proteomic analysis of the human spliceosome." *Genome Res* **12**: 1231-45.

Reed, R.(2003)."Coupling transcription, splicing and mRNA export." *Curr Opin Cell Biol* **15**: 326-31.

Richards, T., Hirt, R. P., Williams, B. A. P. and Embley, T. M. (2003) "Horizontal Gene Transfer and the Evolution of Parasitic Protozoa." *Protist* **154**: 17-32.

Rivas, E. and S. R. Eddy (2000). "Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs." *Bioinformatics* **16**: 583-605.

Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." *BMC Bioinformatics* **2**: 8.

Rivero, F. (2002). "mRNA processing in Dictyostelium: sequence requirements for termination and splicing." *Protist* **153**: 169-76.

Roger, A. J., S. G. Svard, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin and M. L. Sogin (1998). "A mitochondrial-like chaperonin 60 gene in Giardia lamblia: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria." *Proc Natl Acad Sci U S A* **95**: 229-34.

Romfo, C. M., C. J. Alvarez, W. J. van Heeckeren, C. J. Webb and J. A. Wise (2000). "Evidence for splice site pairing via intron definition in Schizosaccharomyces pombe." *Mol Cell Biol* **20**: 7955-70.

Roos, D. S. (2001). "Computational biology. Bioinformatics--trying to swim in a sea of data." *Science* **291**: 1260-1.

Scamborova, P., A. Wong and J. A. Steitz (2004). "An intronic enhancer regulates splicing of the twintron of Drosophila melanogaster prospero pre-mRNA by two different spliceosomes." *Mol Cell Biol* **24**: 1855-69.

Schneider, C., C. L. Will, O. V. Makarova, E. M. Makarov and R. Luhrmann (2002). "Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions." *Mol Cell Biol* **22**: 3219-29.

Schoof, H., P. Zaccaria, H. Gundlach, K. Lemcke, S. Rudd, G. Kolesov, R. Arnold, H. W. Mewes and K. F. Mayer (2002). "MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome." *Nucleic Acids Res* **30**: 91-3.

Sczyrba, A., J. Kruger, H. Mersch, S. Kurtz and R. Giegerich (2003). "RNA-related tools on the Bielefeld Bioinformatics Server." *Nucleic Acids Res* **31**: 3767-70.

Selenko, P., G. Gregorovic, R. Sprangers, G. Stier, Z. Rhani, A. Kramer and M. Sattler (2003). "Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP." *Mol Cell* **11**: 965-76.

Shav-Tal, Y. and D. Zipori (2002). "PSF and p54(nrb)/NonO--multi-functional nuclear proteins." *FEBS Lett* **531**: 109-14.

Shibata, K., Itoh,M., Aizawa,K., Nagaoka,S., Sasaki,N., Carninci,P.,, H. Konno, Akiyama,J., Nishi,K., Kitsunai,T., Tashiro,H., Itoh,M.,, N. Kikuchi, Ishii,Y., Nakamura,S., Hazama,M., Nishine,T.,, A. Harada, Yamamoto,R., Matsumoto,H., Sakaguchi,S., Ikegami,T.,, K. Kashiwagi, Fujiwake,S., Inoue,K., Togawa,Y., Izawa,M., Ohara,E.,, M. Watahiki, Yoneda,Y., Ishikawa,T., Ozawa,K., Tanaka,T., and S. Matsuura, Okazaki,Y., Muramatsu,M., Inoue,Y. and Hayashizaki,Y. (2001). "Functional annotation of a full-length mouse cDNA collection." *Nature* **409**: 685-690.

Sigrist, C. J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch and P. Bucher (2002). "PROSITE: a documented database using patterns and profiles as motif descriptors." *Brief Bioinform* **3**: 265-74.

Simpson, A. G., E. K. MacQuarrie and A. J. Roger (2002). "Eukaryotic evolution: early origin of canonical introns." *Nature* **419**: 270.

Simpson, A. G. and A. J. Roger (2002). "Eukaryotic evolution: getting to the root of the problem." *Curr Biol* **12**: R691-3.

Sjolander, K. (2004). "Phylogenomic inference of protein molecular function: advances and challenges." *Bioinformatics* **20**: 170-9.

Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, et al. (1997). "Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics." *J Bacteriol* **179**: 7135-55.

Sogin, M. L. (1991). "Early evolution and the origin of eukaryotes." *Curr Opin Genet Dev* **1**: 457-63.

Sogin, M. L. (1997). "Organelle origins: energy-producing symbionts in early eukaryotes?" *Curr Biol* **7**: R315-7.

Sorek, R., R. Shamir and G. Ast (2004). "How prevalent is functional alternative splicing in the human genome?" *Trends Genet* **20**: 68-71.

Spafford, J. D., A. N. Spencer and W. J. Gallin (1999). "Genomic organization of a voltage-gated Na+ channel in a hydrozoan jellyfish: insights into the evolution of voltage-gated Na+ channel genes." *Receptors Channels* **6**: 493-506.

Steel, M. and D. Penny (2000). "Parsimony, likelihood, and the role of models in molecular phylogenetics." *Molecular Biology and Evolution* **17**: 839-850.

Stevens, R. D., A. J. Robinson and C. A. Goble (2003). "myGrid: personalised bioinformatics on the information grid." *Bioinformatics* **19 Suppl 1**: i302-4.

Stevens, S. W., I. Barta, H. Y. Ge, R. E. Moore, M. K. Young, T. D. Lee and J. Abelson (2001). "Biochemical and genetic analyses of the U5, U6, and U4/U6 x U5 small nuclear ribonucleoproteins from Saccharomyces cerevisiae." *Rna* **7**: 1543-53.

Stevens, S. W., D. E. Ryan, H. Y. Ge, R. E. Moore, M. K. Young, T. D. Lee and J. Abelson (2002). "Composition and functional characterization of the yeast spliceosomal penta-snRNP." *Mol Cell* **9**: 31-44.

Stiller, J. W. and B. D. Hall (1999). "Long-branch attraction and the rDNA model of early eukaryotic evolution." *Mol Biol Evol* **16**: 1270-9.

Stoeck, T. and S. Epstein (2003). "Novel eukaryotic lineages inferred from small-subunit rRNA analyses of oxygen-depleted marine environments." *Appl Environ Microbiol* **69**: 2657-63.

Storz, G. (2002). "An expanding universe of noncoding RNAs." *Science* **296**: 1260-3.

Sullivan, P. M., P. Petrusz, C. Szpirer and D. R. Joseph (1991). "Alternative processing of androgen-binding protein RNA transcripts in fetal rat liver. Identification of a transcript formed by trans splicing." *J Biol Chem* **266**: 143-54.

Szkukalek, A., E. Myslinski, A. Mougin, R. Luhrmann and C. Branlant (1995). "Phylogenetic conservation of modified nucleotides in the terminal loop 1 of the spliceosomal U5 snRNA." *Biochimie* **77**: 16-21.

Takahashi, Y., T. Tani and Y. Ohshima (1996). "Spliceosomal introns in conserved sequences of U1 and U5 small nuclear RNA genes in yeast Rhodotorula hasegawae." *J Biochem (Tokyo)* **120**: 677-83.

Tamura, M., D. K. Hendrix, P. S. Klosterman, N. R. Schimmelman, S. E. Brenner and S. R. Holbrook (2004). "SCOR: Structural Classification of RNA, version 2.0." *Nucleic Acids Res* **32** Database issue: D182-4.

Tasic, B., C. E. Nabholz, K. K. Baldwin, Y. Kim, E. H. Rueckert, S. A. Ribich, P. Cramer, Q. Wu, R. Axel and T. Maniatis (2002). "Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing." *Mol Cell* **10**: 21-33.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." *Nucleic Acids Res* **29**: 22-8.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." *Nucleic Acids Res* **25**: 4876-82.

Tombes, R. M., M. O. Faison and J. M. Turbeville (2003). "Organization and evolution of multifunctional Ca(2+)/CaM-dependent protein kinase genes." *Gene* **322**: 17-31.

Tovar, J., G. Leon-Avila, L. B. Sanchez, R. Sutak, J. Tachezy, M. Van Der Giezen, M. Hernandez, M. Muller and J. M. Lucocq (2003). "Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation." *Nature* **426**: 172-6.

Tschudi, C. and E. Ullu (2002). "Unconventional rules of small nuclear RNA transcription and cap modification in trypanosomatids." *Gene Expr* **10**: 3-16.

Tsui, V., T. Macke and D. A. Case (2003). "A novel method for finding tRNA genes." *Rna* **9**: 507-17.

Tycowski, K. T. and J. A. Steitz (2001). "Non-coding snoRNA host genes in Drosophila: expression strategies for modification guide snoRNAs." *Eur J Cell Biol* **80**: 119-25.

Urlaub, H., K. Hartmuth, S. Kostka, G. Grelle and R. Luhrmann (2000). "A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snRNPs). Analysis of RNA-protein contacts in native U1 and U4/U6.U5 snRNPs." *J Biol Chem* **275**: 41458-68.

Valadkhan, S. and J. L. Manley (2000). "A tertiary interaction detected in a human U2-U6 snRNA complex assembled in vitro resembles a genetically proven interaction in yeast." *Rna* **6**: 206-19.

Valadkhan, S. and J. L. Manley (2001). "Splicing-related catalysis by protein-free snRNAs." *Nature* **413**: 701-7.

Valadkhan, S. and J. L. Manley (2003). "Characterization of the catalytic activity of U2 and U6 snRNAs." *Rna* **9**: 892-904.

van Eenennaam, H., D. Lugtenberg, J. H. Vogelzangs, W. J. van Venrooij and G. J. Pruijn (2001). "hPop5, a protein subunit of the human RNase MRP and RNase P endoribonucleases." *J Biol Chem* **276**: 31635-41.

van Eenennaam, H., G. J. Pruijn and W. J. van Venrooij (1999). "hPop4: a new protein subunit of the human RNase MRP and RNase P ribonucleoprotein complexes." *Nucleic Acids Res* **27**: 2465-72.

van Eenennaam, H., A. van Der Heijden, R. J. Janssen, W. J. van Venrooij and G. J. Pruijn (2001). "Basic Domains Target Protein Subunits of the RNase MRP Complex to the Nucleolus Independently of Complex Association." *Mol Biol Cell* **12**: 3680-9.

van Nues, R. W. and J. D. Beggs (2001). "Functional contacts with a range of splicing proteins suggest a central role for Brr2p in the dynamic control of the order of events in spliceosomes of Saccharomyces cerevisiae." *Genetics* **157**: 1451-67.

Vanacova, S., D. R. Liston, J. Tachezy and P. J. Johnson (2003). "Molecular biology of the amitochondriate parasites, Giardia intestinalis, Entamoeba histolytica and Trichomonas vaginalis." *Int J Parasitol* **33**: 235-55.

Vandenberghe, A. E., T. H. Meedel and K. E. Hastings (2001). "mRNA 5'-leader trans-splicing in the chordates." *Genes Dev* **15**: 294-303.

Vidal, V. P., L. Verdone, A. E. Mayes and J. D. Beggs (1999). "Characterization of U6 snRNA-protein interactions." *Rna* **5**: 1470-81.

Vivares, C. P., M. Gouy, F. Thomarat and G. Metenier (2002). "Functional and evolutionary analysis of a eukaryotic parasitic genome." *Curr Opin Microbiol* **5**: 499-505.

Wang, Q., K. Hobbs, B. Lynn and B. C. Rymond (2003). "The Clf1p splicing factor promotes spliceosome assembly through N-terminal tetratricopeptide repeat contacts." *J Biol Chem* **278**: 7875-83.

Wang, Q. and B. C. Rymond (2003). "Rds3p is required for stable U2 snRNP recruitment to the splicing apparatus." *Mol Cell Biol* **23**: 7339-49.

Watanabe, K. I. and T. Ohama (2001). "Regular spliceosomal introns are invasive in Chlamydomonas reinhardtii: 15 introns in the recently relocated mitochondrial cox2 and cox3 genes." *J Mol Evol* **53**: 333-9.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." *Mol Biol Evol* **18**: 691-9.

Wilihoeft, U., E. Campos-Gongora, S. Touzni, I. Bruchhaus and E. Tannich (2001). "Introns of Entamoeba histolytica and Entamoeba dispar." *Protist* **152**: 149-56.

Will, C. L., C. Schneider, A. M. MacMillan, N. F. Katopodis, G. Neubauer, M. Wilm, R. Luhrmann and C. C. Query (2001). "A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site." *Embo J* **20**: 4536-46.

Will, C. L., C. Schneider, R. Reed and R. Luhrmann (1999). "Identification of both shared and distinct proteins in the major and minor spliceosomes." *Science* **284**: 2003-5.

Williams, B. A. and P. J. Keeling (2003). "Cryptic organelles in parasitic protists and fungi." *Adv Parasitol* **54**: 9-68.

Wilson, R. K. (1999). "How the worm was won. The C. elegans genome sequencing project." *Trends Genet* **15**: 51-8.

Wishart, D. S., S. Fortin, D. R. Woloschuk, W. Wong, T. Rosborough, G. Van Domselaar, J. Schaeffer and D. Szafron (1997). "A platform-independent graphical user interface for SEQSEE and XALIGN." *Comput Appl Biosci* **13**: 561-2.

Wood, V., R. Gwilliam, M. A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, J. Hayles, S. Baker, et al. (2002). "The genome sequence of Schizosaccharomyces pombe." *Nature* **415**: 871-80.

Xiao, S., F. Houser-Scott and D. R. Engelke (2001). "Eukaryotic ribonuclease P: increased complexity to cope with the nuclear pre-tRNA pathway." *J Cell Physiol* **187**: 11-20.

Xiao, S., F. Scott, C. A. Fierke and D. R. Engelke (2002). "Eukaryotic Ribonuclease P: A Plurality of Ribonucleoprotein Enzymes." *Annu Rev Biochem* **71**: 165-89.

Xu, Y., H. Ben-Shlomo and S. Michaeli (1997). "The U5 RNA of trypanosomes deviates from the canonical U5 RNA: the Leptomonas collosoma U5 RNA and its coding gene." *Proc Natl Acad Sci U S A* **94**: 8473-8.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." *Comput Appl Biosci* **13**: 555-6.

Yang, Z., S. Kumar and M. Nei (1995). "A new method of inference of ancestral nucleotide and amino acid sequences." *Genetics* **141**: 1641-50.

Yazaki, J., K. Kojima, K. Suzuki, N. Kishimoto and S. Kikuchi (2004). "The Rice PIPELINE: a unification tool for plant functional genomics." *Nucleic Acids Res* **32**: D383-7.

Zhang, Y., T. Lindblom, A. Chang, M. Sudol, A. E. Sluder and E. A. Golemis (2000). "Evidence that dim1 associates with proteins involved in pre-mRNA splicing, and delineation of residues essential for dim1 interactions with hnRNP F and Npw38/PQBP-1." *Gene* **257**: 33-43.

Zhou, Z., L. J. Licklider, S. P. Gygi and R. Reed (2002). "Comprehensive proteomic analysis of the human spliceosome." *Nature* **419**: 182-5.

Zhu, W. and V. Brendel (2003). "Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome." *Nucleic Acids Res* **31**: 4561-72.

Ziehler, W. A., J. Morris, F. H. Scott, C. Millikin and D. R. Engelke (2001). "An essential protein-binding domain of nuclear RNase P RNA." *RNA* **7**: 565-75.

Internet Websites

Protein and Nucleotide databases:

NCBI Entrez	http://www.ncbi.nlm.nih.gov
Protein Data Bank	http://www.rcsb.org/pdb/
The RNaseP Database	http://www.mbio.ncsu.edu/RNaseP/main.html
Rfam: RNA families database	http://rfam.wustl.edu/
Pfam: Protein families database	http://www.sanger.ac.uk/Software/Pfam/
Prosite: Protein families and domains	http://kr.expasy.org/prosite/
KOG eukaryotic orthologous groups	ftp:ftp.ncbi.nih.gov/pub/COG/KOG
Ribosomal Database Project	http://rdp.cme.msu.edu/html

Internet-Based Analysis:

BLAST at NCBI	http://www.ncbi.nlm.nih.gov
BLink helpfiles	http://www.ncbi.nlm.nih.gov/sutils/static/blinkhelp.html

Software:

RNAmotif	ftp.scripps.edu/pub/macke/rnamotif-version.tag.gz
RSEARCH	http://www.genetics.wustl.edu/eddy/software/
ERPIN	http://tagc.univ-mrs.fr/pub/erpin
RNACad	http://www.cse.ucsc.edu/~mpbrown/rnacad/
WinPep Protein properties software	http://www.ipw.agrl.ethz.ch/~lhennig/winpep.html
PepTools	http://www.biotools.com
FastML	http://kimura.tau.ac.il
PamL	ftp://abacus.gene.ucl.ac.uk/pub/paml/
Paup*	http://paup.csit.fsu.edu/about.html

Miscellaneous:

Eukaryotic Tree (Figure 1.3)	http://hdes,biochem.dat.ca/Rogerlab/
The Gene Ontology Consortium ™ (GO)	http://www.geneontology.org/
Helix	http://helix.massey.ac.nz

Genomes:

Plasmodium falciparum	http://plamodb.org
Entamoeba histolytica	ftp.sanger.ac.uk/pub/pathogens/E_histolytica/
Dictyostelium discoideum: dictyBase	http://dictybase.org/
Giardia lamblia	http://jbpc.mbl.edu/Giardia-HTML/index2.html
Ciona intestinalis	http://genome.jgi-psf.org/ciona/
Drosophila melanogaster: Flybase	http://flybase.bio.indiana.edu/
Caenorhabditis elegans: WormBase	http://www.wormbase.org/

Appendix A.1: Abstract for ECCB' 2003

Presented at the European Conference for Computational Biology, Sept 2003, Paris France.

Searching for ncRNAs in protist genomes.

Lesley J. Collins[1*], Thomas J. Macke[2] and David Penny[1]

[1] Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 11222, Palmerston North, New Zealand.

[2] Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

Keywords: Noncoding RNA, RNAmotif, U5 snRNA, RNase P, *Giardia lamblia*.

Introduction

Non-coding RNAs make transcripts that function as RNA, rather than encoding proteins e.g. ribosomal-RNA (rRNA) and transfer-RNA (tRNA) [1]. They often form part of RNA-protein complexes (Ribonucleoproteins or RNPs) and play vital roles in essential cellular processes such as protein metabolism and splicing. Searching databases for homologs based on sequence similarity is only useful for the most slowly evolving or large ncRNAs like ribosomal RNAs, and becomes much less reliable for other snRNAs. If there is also large evolutionary distance between the species that is being searched and the species for which the gene is known, sequence similarity methods often fail to uncover any potential gene candidates for further analysis and confirmation as ncRNA gene homologs [1].

Noncoding RNAs usually fold into characteristic secondary structures and also can contain small sequence motifs. RNAmotif [2] is a program that uses this information to find ncRNA gene candidates with the design of an appropriate "descriptor" to model secondary structure and sequence motifs. However, in the past, descriptors that had to take inter-species structural variation into account could run into problems with overloading of the results file.

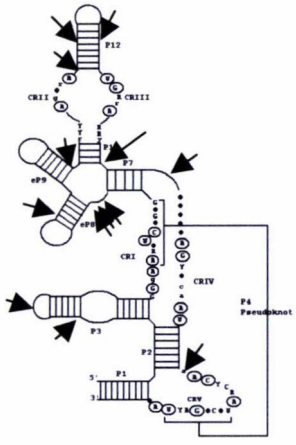
Here we show how the use of a user-defined scoring section, post-function commands and parallel implementation can help in reducing the problems associated with 'looser' descriptors. We describe the design and implementation of descriptors for two ncRNAs, the U5 snRNA and the eukaryotic RNase P RNA. These genes have conserved and variable sequence and structure areas which allowed the identification of gene candidates in some protist genomes such as *Giardia lamblia* [3] and *Encephalitozoon cuniculi* (a microsporidian).

Results and Discussion

Descriptors for the U5 snRNA were designed with a unique motif scoring section allowing important motif presence or absence to be seen at a glace in the results file. This is necessary because sometimes scoring regimes can add up individual motif scores in such a way that a sequence can be given a higher score even if an extremely important motif is missing. Spreadsheet sorting can also be used to detect sequences containing a certain motif. RNAmotif is a processor-intensive program. Even a short descriptor can run into problems when searching a large genome database and often the program will not run to completion in this situation. Parallel computing is one solution with large databases being split into smaller pieces, each piece run on a separate node, and then the results collated in a single result file. Getbest was also incorporated into the parallel implementation as a –post command, filtering the results from each worker node to give a more condensed results file. This reduced the space required for the results file and enabled realistic sequence analysis of the results

The U5 descriptors were tested by searching against the genomes of *E. cuniculi* and *Plasmodium falciparum* for which the U5 snRNA genes were already known. Application of this technique resulted in new U5 snRNA gene candidates from the genomes of *Ciona intestinalis* (sea squirt) and *Giardia lamblia*.

The RNase P RNA has a 'reasonably' conserved secondary structure with parts of this structure being highly conserved, even between all three kingdoms, and other parts are variable[4]. Figure 1 shows the consensus eukaryotic RNase P RNA secondary structure; circled nucleotides are conserved between all three kingdoms and arrows point to areas where helices may be inserted in some species. This secondary structure variability creates a challenge for designing a 'working' descriptor for the RNase P RNA. A descriptor was designed for part of the eukaryotic RNase P RNA secondary structure consisting of the P3-CRI-P7-P10 section of the whole secondary structure and was called the 'P7 descriptor'. It was found that descriptors designed for the full eukaryotic secondary structure were computationally-prohibitive, requiring weeks to search the simplest databases. Results from the RNAmotif scan were then analyzed for downstream conserved elements (e.g. CRV pseudoknot pairing and CRIV consensus sequence) to find a viable candidate sequence.



Testing of this descriptor against a database of known RNase P RNAs from all three kingdoms (Bacterial, Archaea and Eukaryotes) showed that it had specificity for the Eukaryotic RNase P RNA. Searches with the P7 descriptor against *E. cuniculi* genome recovered 14 sequence areas with only one having a viable CRV consensus sequence. When this sequence was BLASTed against GenBank, rat and mouse RNase P RNAs were returned with low scores. Genome searches against *G. lamblia* recovered eleven sequence areas with the top score, each having the Eukaryotic CRI consensus sequence. Of these only one had a viable downstream CRV area and upon further examination generally fitted the Eukaryotic RNase P RNA consensus secondary structure having the eukaryotic consensus sequences for the CR (I-V) regions in correct locations.

The RNAmotif genome searching procedure described in this study cannot guarantee finding a particular ncRNA in a particular genome. Other factors such as the quality of the genomic data and the phylogenetic distance between the species from which the known ncRNAs (the ones used to design the descriptor) and the genome being searched may also play a large role in the search outcome. This procedure, however, does offer a new way of searching for some sometimes hard-to-find ncRNA genes. Finding these genes in protist genomes may help in our understanding of the evolution of RNA metabolism from the earliest Eukaryotes.

Acknowledgements:

Many thanks to the administrators of the Helix parallel processing facility at Massey University, Albany, New Zealand for their help and advice. Many thanks to Mitchell L. Sogin, and Andrew G. McArthur and their teams at the *Giardia lamblia* Genome Project (funded by the NIAID/NIH under cooperative agreement AI 043273), Marine Biological Laboratory at Woods Hole for access to non-public data. Other genomes were downloaded from GenBank. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for performing preliminary experimental analysis on some of the ncRNA gene candidates. This work was supported by the NZ Marsden Fund.

References:

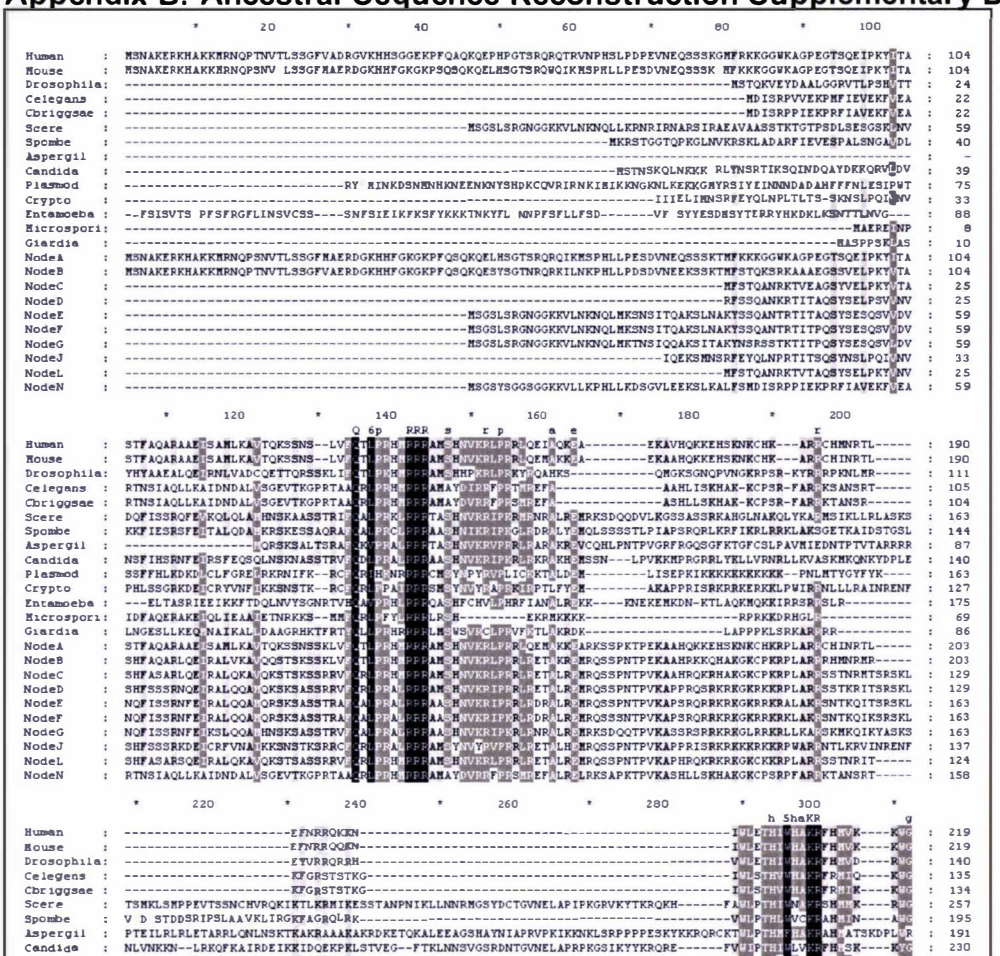
[1] Eddy, S. (1991) Non-coding RNA genes and the modern RNA world. Nature Reviews Genetics, **2**, 919-929.

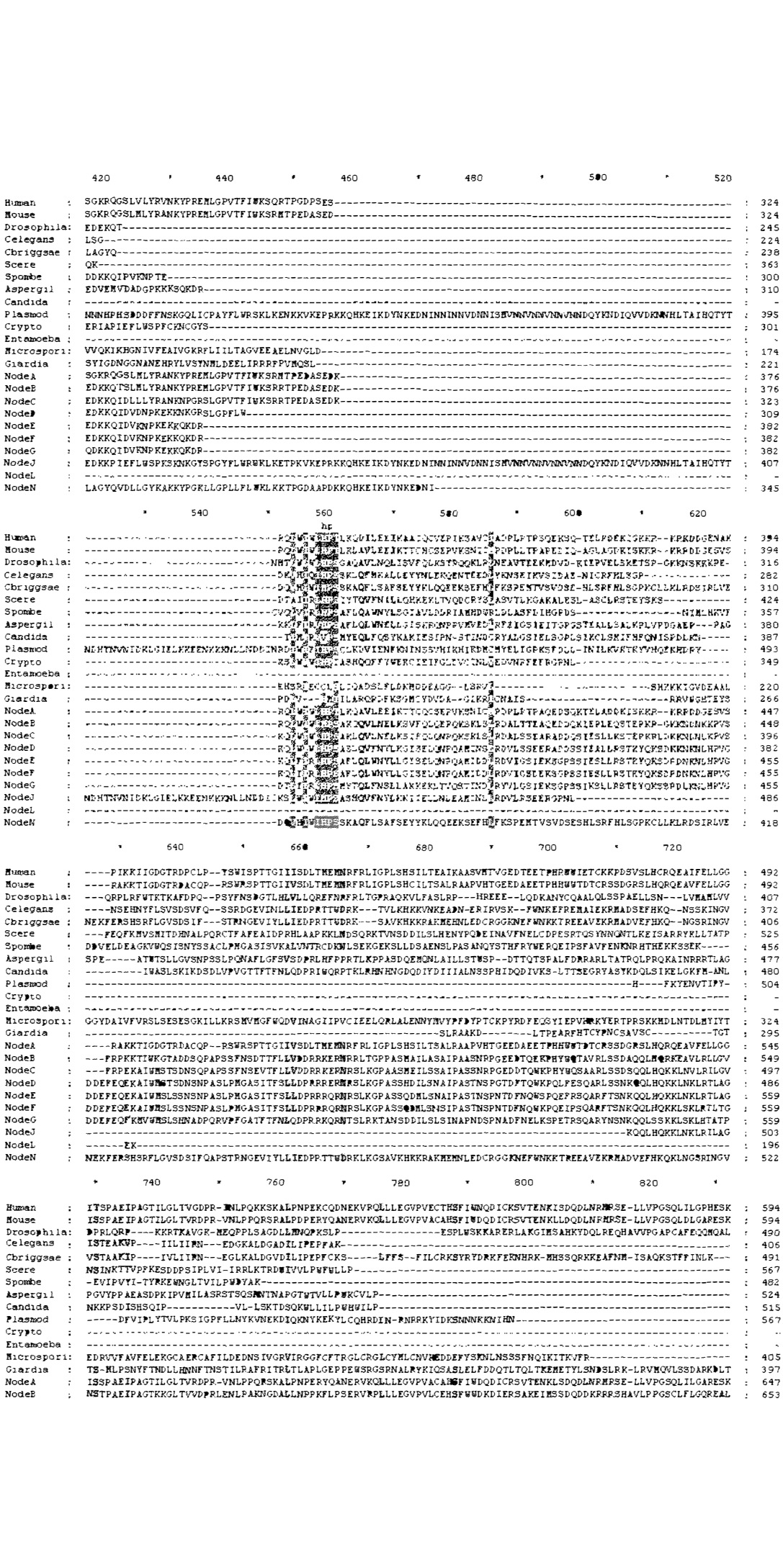
[2] Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D. A. and Sampath, R. (2001) RNAmotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res., **29**, 4724-4735.

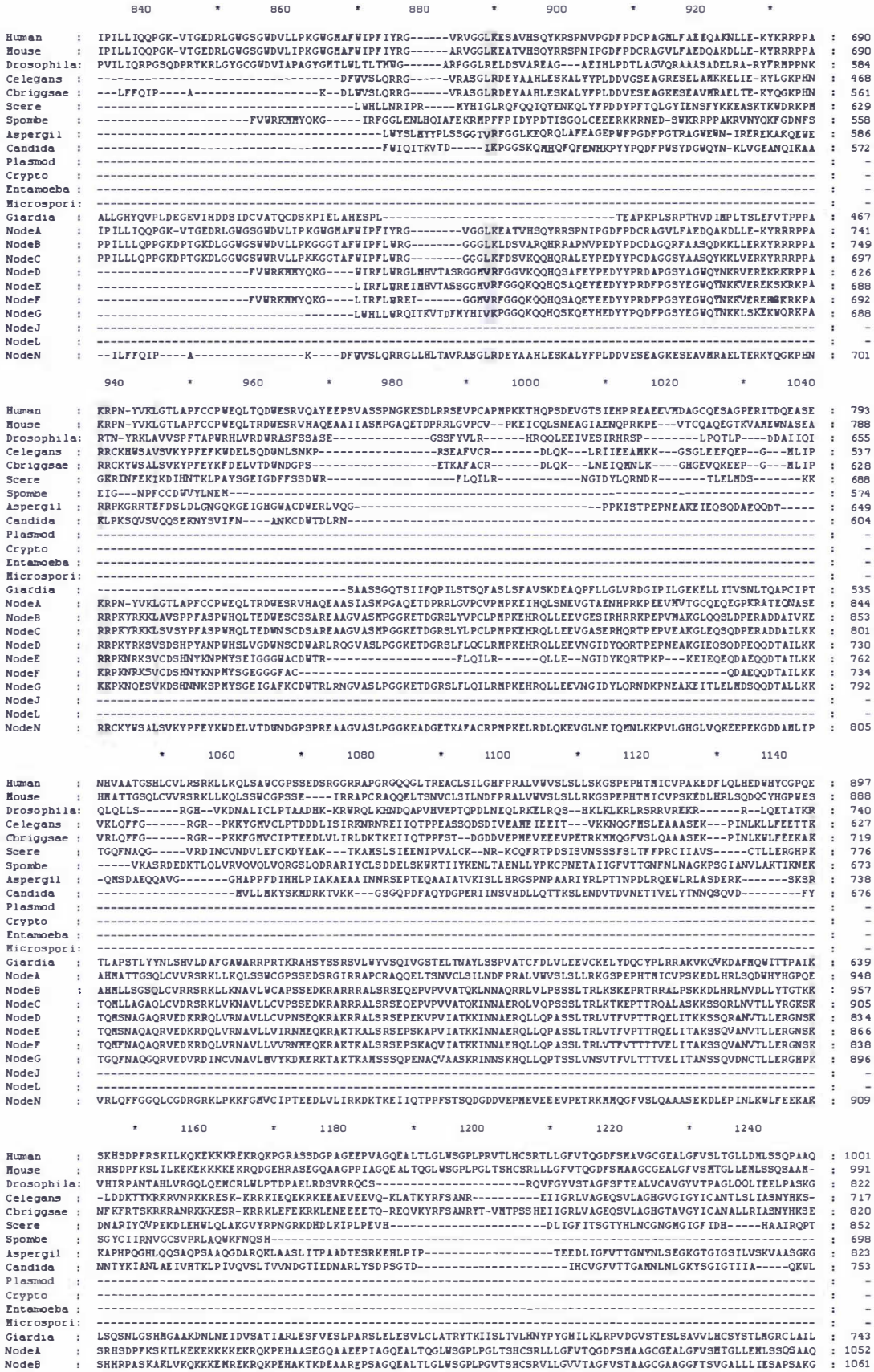
[3] McArthur, A.G., Morrison, H.G., Nixon, J.E.J., Passamaneck, N.Q.E., Kim and 10 others (2000) The *Giardia* genome project database. FEMS Microbiology Letters, **189**, 271-3.

[4] Frank, D.N., Adamidi, C., Ehringer, M.A., Pitulle, C. and Pace, N.R. (2000) Phylogenetic-comparative analysis of the eukaryal Ribonuclease P RNA. RNA. **6,** 1895-904.

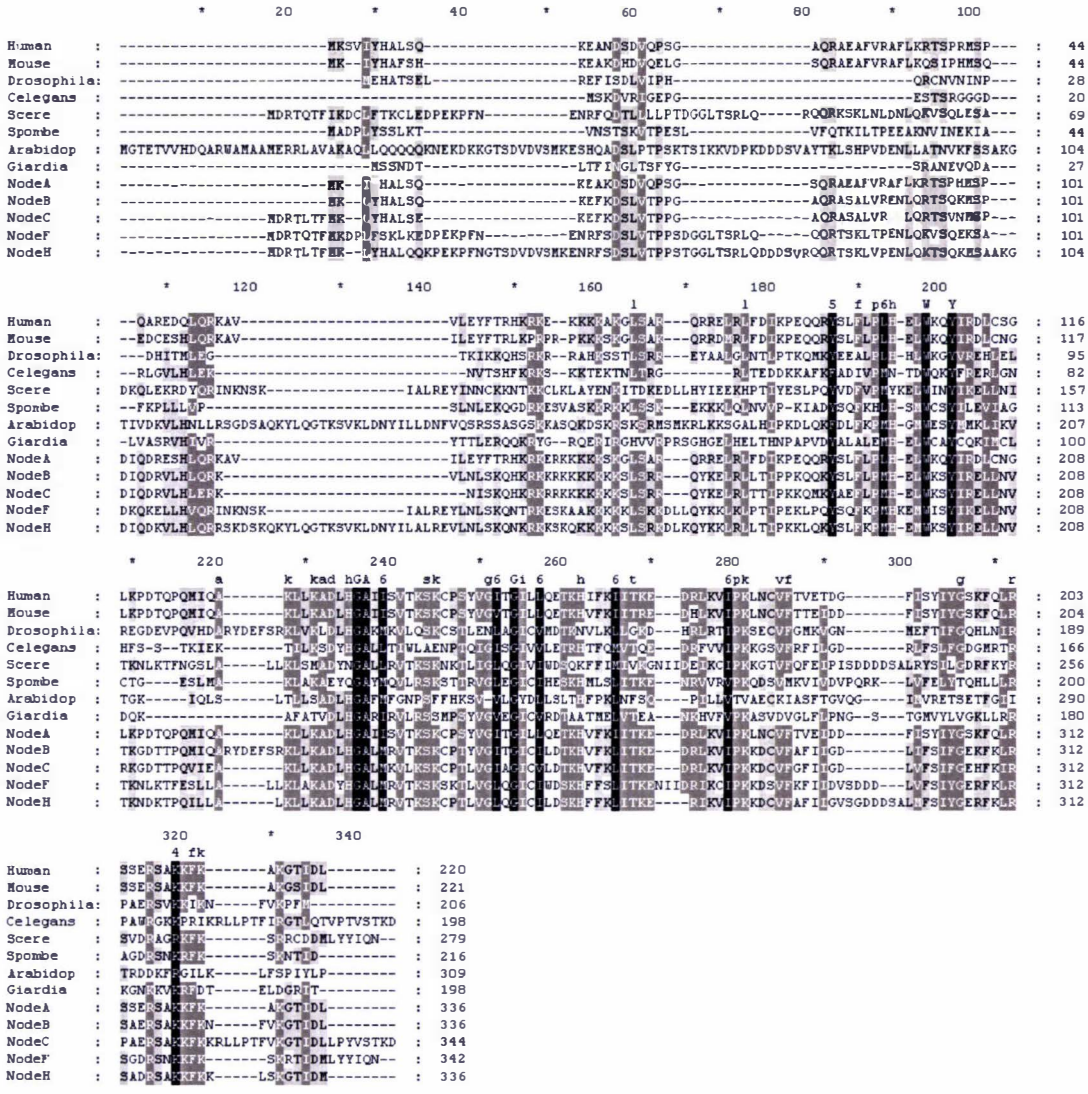
Appendix B: Ancestral Sequence Reconstruction Supplementary Data.



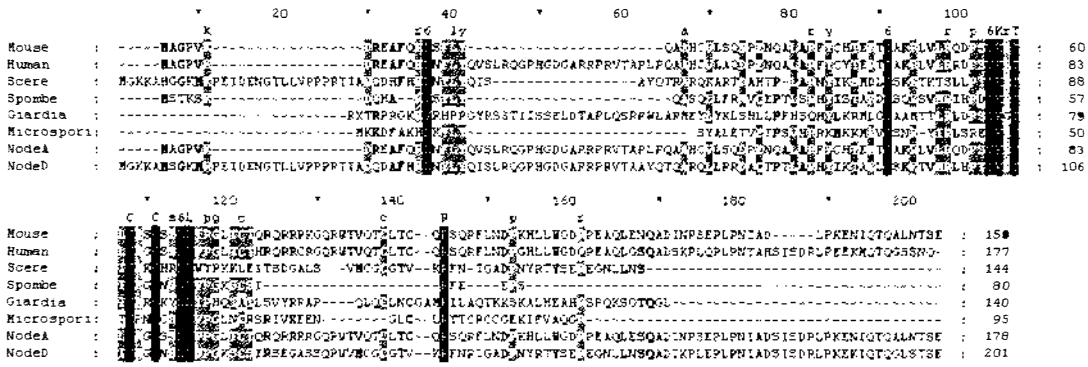




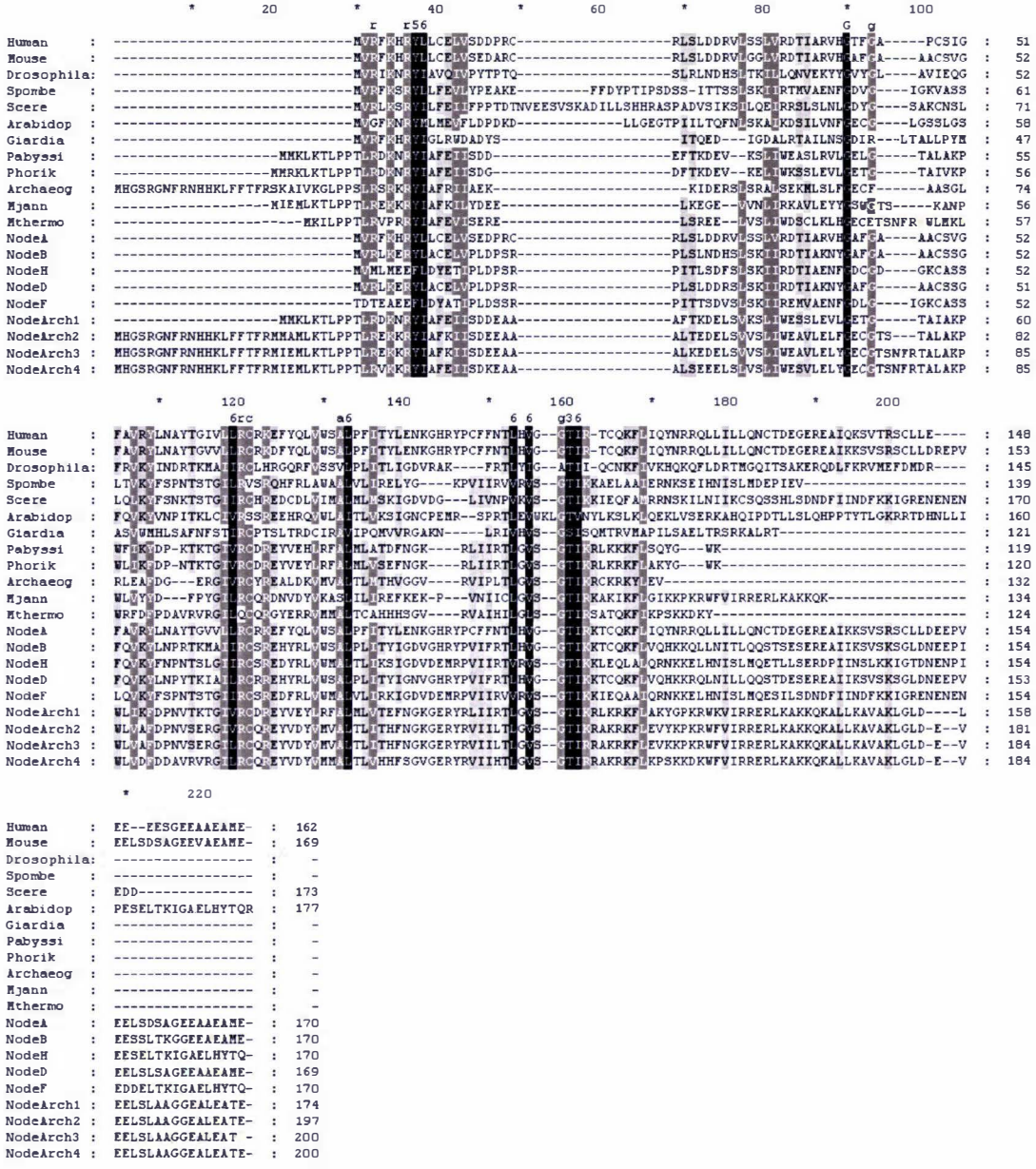
Supplementary Figure 1: Complete amino acid alignment of the *Giardia lamblia*, *Entamoeba histolytica* and *Encephilitozoon cuniculi* Pop1 candidates with eukaryotic Pop1 proteins and ancestral sequences. Node positions are shown in the main text as Figure 1. Key: Human – *Homo sapiens*; Mouse – *Mus musculus*; Drosophila – *Drosophila melanogaster*; C. elegans – *Caenorhabditis elegans*; C.briggsae – *Caenorhabditis briggsae*; Scere – *Saccharomyces cerevisiae*; S.pombe – *Schizosaccharomyces pombe*; Aspergil – *Aspergillus nidulans*; Candida – *Candida albicans*; Plasmod – *Plasmodium falciparum*; Crypto – *Cryptosporidium parvum*; Microspori – *Encephalitozoon cuniculi*; Giardia – *Giardia lamblia*. Darker shadings represent the increased amount of homology of the amino acid position in the alignment. Note that the alignment has only two areas of conservation, the "RRR-motif" and the "W-box", both of which are shown in greater detail in Figure 4.



Supplementary Figure 2: Complete amino acid sequence alignment of the *Giardia lamblia* Pop4 candidate with other eukaryotic Pop4 proteins and ancestral sequences. Key: Human – *Homo sapiens*; Mouse – *Mus musculus*; Drosophila – *Drosophila melanogaster*; C. elegans – *Caenorhabditis elegans*; Scere – *Saccharomyces cerevisiae*; S.pombe – *Schizosaccharomyces pombe*; Arabidop – *Arabidopsis thaliana*; Giardia – *Giardia lamblia*. Darker shadings represent the increased amount of homology of the amino acid position in the alignment.



Supplementary Figure 3: Complete amino acid sequence alignment of the *Giardia lamblia and Encephalitozoon cuniculi* Rpp21 candidates with other eukaryotic Rpp21 proteins and ancestral sequences. Key Human – *Homo sapiens*; Mouse – *Mus musculus*; Scere – *Saccharomyces cerevisiae*; S.pombe – *Schizosaccharomyces pombe*; Giardia – *G. lamblia*; Microspori – *Ecz. cuniculi*; Darker shadings represent the increased amount of homology of the amino acid position in the alignment.



Supplementary Figure 4: Complete amino acid sequence alignment of the *Giardia* Pop5 candidate with other eukaryotic and archaeal Pop5 proteins and ancestral sequences. Key: Human – *Homo sapiens*; Mouse – *Mus musculus*; Drosophila – *Drosophila melanogaster*; S.pombe – *Schizosaccharomyces pombe*; Scere – *Saccharomyces cerevisiae*; Arabidop – Arabidopsis thaliana; Giardia – *Giardia lamblia;* Pabyssi – *Pyrococcus abyssi*; Phorik – *Pyrococcus horikoshii*; Archaeog – *Archaeoglobus fulgidus*; Mjann – *Methanococcus jannaschii*; Mthermo – *Methanobacterium thermoautotrophicum*; NodeArch1 – Node between Pabyssi and Phorik; NodeArch2 – Node between NodeArch1 and Archaeog; NodeArch3 – Node between NodeArch2 and Mjann; NodeArch4 – Node between NodeArch3 and Mthermo. Darker shadings represent the increased amount of homology of the amino acid position in the alignment.

Known sequence or Node – Pop1	Target Genomes									
	S. pombe (PAML)		*S. pombe (FastML)*		Ecz. cuniculi		*Entamoeba*		*Giardia*	
	Score	*E-value*	*Score*	*E-value*	*Score*	*E-value*	*Score*	*E-value*	*Score*	*E-value*
Human	84	2e-16	59	4e-10	61	3e-10	-	-	35	0.13
Mouse	72	9e-13	70	3e-13	58	3e-09	-	-	37	0.045
D. melanogaster	81	1e-15	60	3e-10	50	9e-07	-	-	-	-
C. elegans	75	8e-14	46	3e-06	36	0.010	-	-	-	-
C. briggisae	64	3e-10	46	3e-06	38	0.003	-	-	-	-
A. nidulans	175	2e-41	89	4e-19	-	-	-	-	-	-
S. pombe	1362	0.0	171	7e-44	-	-	-	-	-	-
S. cerevisiae	162	6e-40	81	1e-16	33	0.099	-	-	-	-
C. albicans	168	5e-42	75	7e-15	34	0.051	-	-	-	-
C. parvum (partial)	62	3e-10	53	2e-08	34	0.022	-	-	39	0.002
P. falciparum (partial)	54	1e-07	-	-	46	5e-06	-	-	38	0.008
Node A	94	2e-19	70	3e-13	38	0.006	40	5e-04	43	8e-04
Node B	117	3e-26	74	1e-14	39	0.002	-	-	45	3e-04
Node C	132	7e-31	47	2e-06	41	7e-04	-	-	44	4e-04
Node D	140	4e-33	95	9e-21	46	1e-05	-	-	44	3e-04
Node E	177	2e-44	118	6e-28	41	7e-04	-	-	-	-
Node F	206	4e-53	120	2e-28	41	7e-04	-	-	-	-
Node G	123	4e-28	91	1e-19	37	0.010	-	-	41	0.008

Supplementary Table 1: Results from BLAST searches with known Pop1 sequences and ancestral sequences of the genome databases from *S. pombe*, *Entamoeba*, Ecz. cuniculi and *Giardia*. Node names are as indicated in Figure 1.
'-' Indicates no hits (E-values above 1.0 are reported as no hits).

Known sequence or Node – Pop4	Target Genomes							
	S. pombe		Ecz. cuniculi		*Entamoeba*		*Giardia*	
	Score	E-value	Score	E-value	Score	E-value	Score	E-value
Human	81	3e-16	-	-	-	-	48	6e-06
Mouse	74	3e-14	-	-	-	-	47	8e-06
D. melanogaster	-	-	-	-	-	-	41	8e-04
C. elegans	-	-	-	-	-	-	-	-
A. thaliana	32	0.25	-	-	-	-	-	-
S. pombe	371	e-104	-	-	-	-	46	2e-05
S. cerevisiae	49	1e-06	-	-	-	-	-	-
Node A	72	3e-13	-	-	-	-	51	5e-07
Node B	87	6e-18	-	-	-	-	-	-
Node C	75	4e-14	-	-	27	0.88	-	
Node H	94	9e-20	-	-	36	0.002	-	-
Node F	172	2e-43	-	-	31	0.058	57	9e-09

Supplementary Table 2: Results from BLAST searches with Pop4 known and ancestral sequences of the target databases from *S. pombe*, *Giardia, Entamoeba* and *Ecz. cuniculi.* Node names are as indicated on Figure 1. E-values above 1.0 are reported as: "–" (i.e. no hits). There were no sequences recovered with the archaeal Pop4 protein sequences or any ancestral sequence derived from them.

Known sequence or Node – Rpp21	Target Genomes							
	S. pombe		Ecz. cuniculi		*Entamoeba*		*Giardia*	
	Score	E-value	Score	E-value	Score	E-value	Score	E-valu
Human	54	3e-08	37	8e-04	-	-	36	0.006
Mouse	67	2e-12	39	1e-04	25	0.97	46	4e-06
S. pombe	163	6e-42	47	2e-07	22	5.4	50	1e-07
S. cerevisiae	46	5e-06	28	0.32	27	0.31	32	0.069
Node A	51	2e-07	39	3e-04	-	-	40	6e-04
Node D*	103	3e-23	48	7e-07	-	-	45	2e-05

Supplementary Table 3: Results from BLAST searches with Rpp21 known and ancestral sequences of the target databases from *S. pombe*, *Giardia, Entamoeba* and *Ecz. cuniculi.* Node names are as indicated on Figure 1.

E-values above 1.0 are reported as: "–" (i.e. no hits).

* This node is the "root" node for the tree of the above sequences.

HMM model	S. pombe		*Giardia*	
	Score	**E-value**	**Score**	**E-value**
Pop1 PAML sequences	162.82	4.3e-41	27.84	2.9e-03
Pop1 PAML and ancestral sequences	199.93	1.2e-53	27.84	3.3e-03
Pop1 PAML ancestral sequences only	163.88	5.5e-44	25.65	0.16
Pop1 FastML sequences	167.25	1.5e-35	18.73	0.014
Pop1 FastML and ancestral sequences	156.31	1e-34	21.03	0.039
Pop1 FastML ancestral sequences only	130.59	1e-30	17.07	0.11
Pop4 Eukaryotic sequences	97.16	5.2e-22	50.53	1.7e-07
Pop4 Eukaryotic and ancestral sequences	108.44	3.9e-30	49.40	1.7e-08
Pop4 Eukaryotic ancestral sequences only	70.53	1.1e-14	34.98	1.3e-04
Pop4 Eukaryotic and Archaeal sequences	70.56	1.5e-17	38.77	2e-05
Pop4 Eukaryotic, Archaeal and ancestral sequences	85.77	3.6e-19	49.78	9e-10
Pop4 Eukaryotic and Archaeal ancestral sequences only	50.63	4.7e-13	36.13	2e-06
Pop5 Eukaryotic sequences	111.88	5.2e-24	24.53	0.012
Pop5 Eukaryotic and ancestral sequences	151.60	5.1e-37	-	-
Pop5 Eukaryotic ancestral sequences only	145.07	8.3e-42	10.60	20
Pop5 Eukaryotic and Archaeal sequences	110.03	6.8e-28	-	-
Pop5 Eukaryotic, Archaeal and ancestral sequences	142.84	4.1e-37	-	-
Pop5 Eukaryotic and Archaeal ancestral sequences only	131.35	6.0e-35	-	-

Supplementary Table 4: HMM model results for Pop1 and Pop4 protein sequence alignments used in searches of the *S. pombe* and *Giardia* genome databases. ' – ' indicates that the candidate sequence was not found.

Appendix C.1: ncRNA Candidate Sequences:

Ciona intestinalis
CionaV1.fasta
http://genome.jgi-psf.org/ciona/

ncRNA	Scaffold	Region	+/-	Method
U1snRNA	388	48045-47973	-	BLAST
	73	3062-2966	-	
U2snRNA	241	132075-132129	+	BLAST
U5snRNA	112	58432-58331	-	RNAmotif
	71	40056-39955	-	
	1849	5070-5172	+	
	1028	15981-15885	-	
	108	18565-18656	+	
U4snRNA	42	351938-351852	-	BLAST
	300	45480-45566	+	
	396	28948-28862	-	
U6snRNA	245	112452-112352	-	BLAST
	51	274757-274857	+	
U12snRNA	99	278659-278642	-	BLAST
	19	293614-293586	-	
	244	32757-32779	+	
U6atac snRNA	563	22545-22589	+	BLAST
SL-RNA	1088	126-169	+	BLAST with
	94	275357-275400	+	sequence from paper.

Entamoeba histolytica
Database downloaded from: ftp.sanger.ac.uk/pub/pathogens/E_histolytica/
The Sanger Centre: Ent. histolytica BLAST server:
http://www.sanger.ac.uk/cgi-bin/blast/submitblast/e_histolytica

ncRNA	Contig	Region	+/-	Method
U2snRNA	Contig5621	1535-1491	-	BLAST
	Contig5402	6117-6161	+	
U5 snRNA	*Ent1359g08.p1k*	355-440	+	RNAmotif
RNaseP	*Ent1376f10.q1k*	375-103	-	RNAmotif

Giardia lamblia (AACB01000000)
http://www.ncbi.nlm.nih.gov

ncRNA	Contig	Region	+/-	Method
U5 snRNA	AACB01000156.1	17562-17453	-	RNAmotif
RNaseP	AACB01000012.1	65152-64918	-	RNAmotif

Dictyostelium discoideum
Dicty.reads
http://dictybase.org/

ncRNA	Contig	Region	+/-	Method
U5 snRNA	JC1b150d11.r1	211-340	+	RNAmotif

Encephalitozoon cuniculi
http://www.ncbi.nlm.nih.gov

ncRNA	Contig	Region	+/-	Method
RNaseP	Chr VIII	28735-27521	-	RNAmotif

Appendix C.2: Spliceosomal Protein Candidate Sequences

Note: In *Giardia lamblia* "contig" replaces AACB01000x.1" (e.g. AACB01000156.1 is stated as Contig156)
Group results are shown at the end of this appendix. Some results shown in Chapter 4 may superseed these tables.

Protein (Human)	(S.cere)	Giardia lamblia			Entamoeba histolytica			Plasmodium falciparum			Ecz. cuniculi		
		Rank	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*
U1 Proteins													
U1-70	SNP1	?	Group1	-	?	-	-	2	Chr 13-1	2575468-2575950	?	-	-
U1-A	MUD1	2	Contig17	C(15905-15342)	2	Contig4264	1711-1926	2	Chr 9	C(1367139-136837)	1	ChrIII	93747-94268
		2	Contig2	C(28719-28273)									
U1-C	YHC1	-	-	-	3	Contig5710	3807-3974	2	Chr 8	C(9200-709006)	1	ChrX	C(33886-34254)
	Prp39	1	Contig34	C(15910-15449)	?	-	-	2	Chr 4	C(5687-214671)	-	-	-
FBP11	Prp40	-	-	-	3	Contig3698	1484-1762	2	Chr 13-1	676016-677950	2	ChrIX	C(90361-91515)
	Snu56	-	-	-	-	-	-	-	-	-	-	-	-
	Snu65	-	-	-	-	-	-	-	-	-	-	-	-
	Nam8	4	Contig63	C(46366-46956)	2	Contig4121	C(2183-1659)	2	Chr 13-1	906995-907495	2	ChrVI	33038-33463
	Snu71	-	-	-	-	-	-	4	Chr 12	C(1216198-1215779)	?	-	-
	Usa1	-	-	-	-	-	-	-	-	-	-	-	-
U11-35		-	-	-	3	Contig4340	C(1224-1003)	2	Chr 8	714779-715000	2	Chr X	146374-146625
					3	Contig5730	C(4540-4274)						
U2 Proteins													
SAP155	HSH155	2 [1]	Contig11	C(18758-21781)	3	Contig3752	1483-1860	1	Chr3	C(378324-375808)	2	ChrXI	119099-121720
SAP145	CUS1	1	Contig21	C(25709-26638)	2	Contig5992	11608-12093	-	-	-	1	ChrXI	C(191499-192500)
							12170-12226						
SAP130	RSE1	-	-	-	3	Contig2941	80-367	1	Chr12	1452602-1455220	4	ChrXI	198727-199551
SAP114	Prp21	?	-	-	2	Contig5872	9559-10080	2	Chr6	C(154262-153732)	2[1]	ChrIX	101392-103239
SAP62	Prp11	1	Contig35	31467-32135	1	Contig5903	C(14555-14313)	1	Chr6	885554-886183	1	ChrIII	59553-60110
SAP61	Prp9	-	-	-	1	Contig4750	1712-3088	2	Chr9	1008158-1008529	2	ChrVII	138730-139842
SAP49	HSH49	1	Contig63	46676-47248	2	Contig4121	C(2444-1962)	2	Chr12	988664-98173	2	ChrVII	205566-206027
						Contig1888	20-319						
						Contig5903	C(14555-14313)						
U2 A'	LEA1	2	Group1		?	-	-	4	Chr9	C(1198038-1197610)	-	-	-
U2 B"	MSL1	2	Group1		2	Contig4264	1711-1926	3	Chr13-1	282275-282556	2	ChrII	30655-31395
p14	SNU17	2	Group1		?	-	-	2	Chr12	C(1011092-1010853	2	ChrII	C(62808-63089)
	Ist3	-	-	-	-	-	-	2	Chr13-1	2237549-2237815	?	-	-
								2	Chr12	988664-988972			
	Rds3	2	Contig4	26106-26372	3	Contig6003	5394-5615	2	Chr10	C(736511-736395)	2-B	ChrVII	111898-112134
					3	Contig5608	4210-4407						

Protein (Human)	(Scere)	Giardia lamblia			Entamoeba histolytica			Plasmodium falciparum			Ecz. cuniculi		
		Rank	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*
U5 proteins													
U5-220	Prp8	1	Contig99	C(455-4270)	2	Contig5950	C(9907-9017)	1	Chr4	303628-306867	1	ChrIV	C(95410-100635)
U5-200	Brr2	1	Contig103	26994-31778	1	Contig5603	C(3558-103)	1	Chr4	1036643-1042396	1	ChrVII	105267-107072
		1	Contig68	36454-39915							2	ChrVI	124051-126030
U5-116	Snu114	2	Contig123	C(30806-30123)	2-many	Contig6026	C(20229-18601)	1	Chr10	C(172723-176430)	?	-	-
						Contig5994	1973-3666						
U5-102	Prp6	-	-	-	1	Contig5535	4689-7274	2	Chr11	404067-405614	?	-	-
U5-100	Prp28	?	RNA	-	2 -S	Contig4846	C(4623-1528)	2	Chr5	C(764921-763539)	2	ChrVIII	121289-122485
			helicase			Contig4210	1498-3372				2	ChrII	82570-83703
U5-52	Snu40	-	-	-	2	Contig5323	C(2526-2359)	2	Chr10	1277885-1278175	-	-	-
U5-40		?	Group2	-	? -many	-	-	1	Chr8	C(338588-337680)	?	-	-
U5-15	DIB1	1	Contig42	56330-56761	1	Contig5824	C(12435-12115)	1	Chr12	1299589-1299843	2	ChrII	152286-152648
PSF/P54nrb		3	Contig2	154611-154829	3	Contig5711	3098-3898	3	Chr11	C(107611-108357)	3	Chr VII	178208-178891
U4-U6 proteins													
U4/U6-90	Prp3	-	-	-	2	Contig5888	C(6224-5058)	2	Chr13-1	C(386354-385443)	-	-	-
U4/U6-60	Prp4	?	Group2	-	2-many	Contig5892	5171-5995	3	Chr3	370027-370899	?-	-	-
						Contig5927	C(9941-9372)		Chr3	C(105687-105121)	many		
						Contig5804	C(5140-3875)						
						Contig5810	2967-3845						
U4/U6-61	Prp31	?	-	-	?	-	-	3	Chr8	520788-520651	2	ChrIX	C(484895-184086)
U4/U6-20	Cpr1	2	Contig1	C(10799-10320)	2	Contig5665	4934-5194	2	Chr11	C(587815-587330)	2	ChrX	C(243917-243516)
		2	Contig60	C(18080-17664)				2	Chr8	C(1044032-1043529)			
								2	Chr5	439293-439670			
U4/U6-15.5kD	Snu13	2	Contig26	C(32741-32385)	2	Contig5857	577-729	2	Chr11	942904-943059	4	ChrI	C(107490-107326)
RY-1		-	-	-	3	Contig3238	577-729	3	Chr8	520788-520651	-	-	-
	Spp41	-	-	-	-	-	-	-	-	-	-	-	-
	Snu23	-	-	-				4	Chr12	C(1814141-1813587)	-	-	-

Protein (Human) (Scere)		Giardia lamblia			Entamoeba histolytica			Plasmodium falciparum			Ecz. cuniculi		
		Rank	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*
U4/U6.U5 tri proteins													
SART-1	Snu66	-	-	-	3	Contig5246	C(2148-1894)	2	Chr3	C(988912-988220)	-	-	-
Tri-65	SAD1	3	Contig135	22619-21687	1	Contig5931	C(22493-21138)	2	Chr13-1	C(705289-704483	4	ChrIII	75437-76252
Tri- 27		-	-	-	-	-	-	2	Chr8	C(520806-520651)	-	-	-
	Aar2	-	-	-	-	-	-	-	-	-	-	-	-
Sm/ LSm core proteins													
Sm B/B'	SmB1	1	Contig8	C(42998-43387)	2	Contig4820	1434-1679	-	-	-	-	-	-
Sm D1	SMD1	1	Contig79	2464-2751	-	-	-	1	Chr11	C(998476-998158)	2	ChrI	C(69731-69438)
Sm D2	SMD2	1	Contig41	47464-47838	2	Contig5423	C(2390-2133)	1	Chr2	752962-753432	2	ChrX	C(218677-218432)
Sm D3	SMD3	2	Contig180	15992-16477	2	Contig3706	C(2054-1926)	1	Chr9	445779-446296	2	Chr III	C(102396-102691)
Sm E	SME1	2	Contig174	C(3433-3813)	-	-	-	-	-	-	1	Chr II	114779-114964
Sm F	SMX3	?	-	-	-	-	-	1	Chr11	C(1050730-1051057)	2	ChrIV	C(102881-102693)
Sm G	SMX2	-	-	-	-	-	-	1	Chr8	356025-356189	2	ChrV	C(120472-120332)
LSm2	LSM2	1	Contig103	C(18248-18571)	2	Contig4979	3589-3744	2	Chr5	838549-838716	2	ChrIX	215965-216246
LSm3	LSM3	-	-	-	2	Contig5767	C(1005-745)	-	-	-	2	ChrX	C(218713-218414)
LSm4	LSM4	2	Contig121	18571-18347	-	-	-	2	Chr11	C(289665-289913)	2	ChrIII	44793-45017
LSm6	LSM6	2	Contig80	C(27300-27536)	-	-	-	-	-	-	-	-	-
LSm7	LSM7	-	-	-	2	Contig5507	C(2348-2554)	4	Chr12	427705-427791	-	-	-
LSm1		-	-	-	-	-	-	-	-	-	2	Chr V	C(173905-173720)
LSm5		4	Contig128	19661-19876 B	2	Contig5768	C(2055-1936)	-	-	-	-	-	-
LSm8		-	-	-	-	-	-	1	Chr8	90616-90459	-	-	-

Protein (Human)	(Scere)	Giardia lamblia Rank	Contig	Area	Entamoeba histolytica Rank	Contig	Area	Plasmodium falciparum Rank	Contig	Area
Catalytic step II and late acting proteins										
Prp16		3	Contig41	C(11666-10902)	2	Contig2567	C(1450-458)	3	Chr 13-1	C(2477155-2474339)
		3	Contig41	C(13280-12747)				3	Chr 10	C(1227491-1225875)
		3	Contig92	C(34667-33465						
Prp22		3	Contig41	C(11567-11055)	2	Contig2567	C(1450-428	2	Chr 13-1	C(2476387-2474348)
		3	Contig92	C(34667-33465)				2	Chr 10	C(1227485-1225773)
Prp43		3	Contig41	C(13313-12747)	2-first	Contig2567	C(1450-446)	2	Chr13-1	C(2476387-2474348)
					2-end	Contig3548	C(1771-497)		Chr 10	C(1227485-1225773)
Slu7		-	-	-	2	Contig5855	C(10097-9750)	2	Chr 6	C(485987-484809)
Prp17		?	Group2	-	3	Contig4422	C(1550-933)	2	Chr 12	802804-804390
					3	Contig5435	1922-2539			
					3	Contig5662	6978-7427			
Prp18		4	Contig128	3325-3309	-	-	-	2	Chr 9	C(919570-919469)
Other DExD/H Proteins										
DDX16	Prp2	3	Contig41	C(11567-10893)	2	Contig2567	C(1450-428)	-	-	-
Abstrakt		2	Contig108	15562-16776	2	Contig5950	C(9922-8858)	1	Chr5	C(1154132-1155628)
		2	Contig15	C(41031-39916)						
p68	DBP2	3	Contig108	15583-16776	2	Contig5950	C(9919-8969)	?	Chr 5	360289-361530
		3	Contig15	C(79736-78129)						
Prp19-Proteins										
CDC5L	CEF1 (Cdc5)	3	Contig55	36747-37019	2	Contig5915	C(2972-2652)	2	Chr10	C(1348136-1346895)
Prp5		3	Contig108	115580-16713	2	Contig5950	C(9937-8951)	2	Chr5	360181-361587
		3	Contig15	C(41058-39880)						
fSAP33	ISY1 (Cwf12)	-	-	-	2	Contig1864	C(1002-436)	-	-	-
hCrn	CLF1	-	-	-	2	Contig5055	C(3297-2029)	1	Chr4	C(215573-214275)
PLRG1	Prp46	2	Contig146	C(17411-16488)	3	Contig5892	5156-5893	1	Chr3	C(105756-104662)
					3	Contig5804	C(4480-4001)			
					3	Contig6019	C(11671-11000)			
Prp19	Prp19	3	Contig76	C(22977-22411)	3	Contig5892	5321-5902	2	Chr3	369307-370902
					3	Contig2567	3285-3845			
EJC-associated Proteins										
Y14		-	-	-	4	Contig4121	C(2207-1962)	3	Chr13-1	2575681-2575899
Magoh		-	-	-	2	Contig6024	C(1742-1512)	1	Chr7	973948-974463
RNPS1		-	-	-	4	Contig5156	3652-3861	?	-	-
CBP80	Sto1	-	-	-	-	-	-	-	-	-
CBP20	Cbc2	-	-	-	?	-	-	2	Chr12	988661-988891

Protein (Human)	(Scere)	Giardia lamblia Rank	Contig	Area	Entamoeba histolytica Rank	Contig	Area	Plasmodium falciparum Rank	Contig	Area
SR proteins										
SRp75		-	-	-	-	-	-	3	Chr10	C(208649-208107)
SRp54	SFRS11	-	-	-	-	-	-	-	-	-
SRp55		3	Contig34	C(6372-6169)	3	Contig4121	C(1874-1692)	3	Chr5	C(725725-725246)
SRp40		-	-	-	-	-	-	-	-	-
ASF	SF2	4	Contig34	C(372-6169)	-	-	-	2	Chr5	C(942472-742251)
9G8		4	Contig34	C(372-6169)	-	-	-	2	Chr5	C(725728-725327)
SC35		-	-	-	4	Contig4121	C(2195-1968)	3	Chr12	681547-681768
SRp30c		3	Contig34	C(6372-6175)	3	Contig4121	C(1874-1707)	3	Chr10	C(942466-942269)
								3	Chr10	C(208658-208131)
hTra2		-	-	-	?	-	-	3	Chr13-1	907274-907495
								3	Chr13-1	2575678-2575908
SRp20		-	-	-	-	-	-	3	Chr13-1	2575693-2575920
SRm300		-	-	-	3	Contig5445	36-1148	3	Chr5	C(469825-469607)
SRm160		-	-	-	3	Contig5885	C(14368-13970)	3	Chr3	C(469825-469607)
Proteins associated with other cellular events										
UAP56	SUB2	2	Contig108	13688-14956	1	Contig5284	2521-3570	2	Chr	C(405384-404284)
TAT-SF1	CUS2	-	-	-	3	Contig3293	C(1155-868)	?	-	-
SKIP	Prp45	-	-	-	2	Contig5981	2498-3199	2	Chr 2	C(764764-763496)
THO2	Rlr1	-	-	-	3	Contig2172	C(422-21)	3	Chr12	C(2032733-2031924)
HPR1		-	-	-	-	-	-	-	-	-
hPrp4		2	Contig6	C(81160-80069)	3	Contig5840	2989-3981	?	-	-
kinase					3	Contig5926	9788-10549			
					3	Contig5145	C(4203-3652)			
TEX1		?	Group2	-	2	Contig4187	530-1117	2	Chr10	C(1111827-1111108)
					2	Contig5811	C(6107-5352)			
XAB2	SYF1	-	-	-	3	Contig5055	C(3210-2404)	3	Chr12	1499730-1498732
CA150		3	Contig6	88410-88943	3	Contig5733	8808-2238	3	Chr12	C(283935-283024)
						Contig5104	1373-2238			
CF1-68		-	-	-	-	-	-	-	-	-
CF1-25		-	-	-	2	Contig5868	C(2190-1612)	-	-	-
ASR2		-	-	-	3	Contig5919	C(3767-3339)	-	-	-
Aly	YRA1	2	Contig6	C(113452-112928)	4	Contig5094	C(1748-1290)	-	-	-
PABP	PAB1	2	Contig10	101001-101711	3	Contig4121	C(2444-1698)	2	Chr12	988709-989386
					3	Contig5548	2866-3771			
					3	Contig5002	C(5328-4297)			

Protein (Human)	(Scere)	Giardia lamblia Rank	Contig	Area	Entamoeba histolytica Rank	Contig	Area	Plasmodium falciparum Rank	Contig	Area
Other Splicing Proteins										
U2 F65	MUD2	2	Contig9	97171-97614	2	Contig5776	C(5560-4517)	-	-	-
U2AF35		-	-	-	1	Contig3998	1739-2167	2	Chr11	C(723353-722718)
SF1	BBP / Msl5	4	Contig12	C(91873-91559)	2	Contig3904	1420-2271	2	Chr6	1001806-1002711
fSAP118		1	Contig181	12740-14320	3	Contig5506	4140-5525	2	Chr6	143093-144709
								2	Chr9	448810-449526
TIP39	Ylr424w	-	-	-	-	-	-	-	-	-
SPF45		-	-	-	-	-	-	?	-	-
LUC7A		3	Contig110	30469-30837	-	-	-	-	-	-
MFAP1		-	-	-	2	Contig5776	C(3382-3047)	1	Chr13-1	973948-974463
IFN4	FAL1	2	Contig108	13811-14926	2	Contig5284	2503-3570	2	Chr4	1044928-1046079
		2	Contig15	C(41019-39850)						
RHA		3	Contig41	C(13343-12777)	2	Contig5368	C(6890-6330)	2	Chr13-1	C(2476420-2474381)
					3	Contig2567	C(1450-458)			
					3	Contig5430	3129-4166			
CCAP2	CWC15	-	-	-	2	Contig4623	C(1050-523)	2	Chr7	C(860018-859752)
SPF31		?	-	-	4	Contig5734	C(2684-2205)	2	Chr13-1	C(290519-289977)
					4	Contig5711	4034-4534			
RED		-	-	-	-	-	-	2	Chr13-1	C(280141-279965)
PUF60		3	Contig34	C(6375-6181)	3	Contig4121	C(2186-1690)	?	-	-
DGSI		-	-	-	-	-	-	3	Chr11	C(430444-429881)
fSAP15		4	Contig138	C(5966-5592)	-	-	-	2	Chr5	C(944568-944407)
fSAP29		4	Contig1	C(10580-10329)	4	Contig3453	C(1388-993)	2	Chr11	C(1520933-151545)
OTT		3	Group1	154611-154865	3	Contig4121	C(1928-1665)	-	-	-
IMP3		-	-	-	-	-	-	2	Chr6	268381-268776
fSAP94		-	-	-	-	-	-	2	Chr6	C(488129-487695)
fSAP59		3	Contig34	C(6378-6131)	3	Contig6040	C(10453-9929)	2	Chr13-1	906929-907501
GCFC		-	-	-	3	Contig5128	C(3547-2807)	-	-	-
fSAP57	PFS2	3	Contig76	C(23268-22414)	3	Contig5715	3496-4248	2	Chr13-1	C(440990-439431)
					3	Contig5804	C(4465-4103)			
fSAP164		?	-	-	3	Contig6018	C(5209-4424)	2	Chr13-1	2018781-2019893
fSAP11		-	-	-	2	Contig1864	C(1002-451)	-	-	-
fSAPa		3	Contig63	46712-46951	-	-	-	2	Chr14	C(96767-96168)
fSAP113		3	Contig49	C(58215-57265)	3	Contig5974	6446-7585	2	Chr7	C(1200475-1199300)
		3	Contig141	28647-29351				2	Chr10	C(240202-238580)
SPF38		3	Group2	-	3	Contig5804	C(4480-3998)	2	Chr8	C(338561-337809)
		3	Contig154	17516-18310	3	Contig5927	C(9968-9438)			
SPF27		-	-	-	-	-	-	2	Chr6	653020-653607
CrkRS	CTK1	3	Contig58	C(31897-31046)	3	Contig5630	C(2963-2076)	3	Chr4	C(795285-794230)
		3	Contig139	19418-20284	3	Contig5376	3367-4044			

Protein (Human) (Scere)	Giardia lamblia			Entamoeba histolytica			Plasmodium falciparum		
	Rank	*Contig*	*Area*	*Rank*	*Contig*	*Area*	*Rank*	*Contig*	*Area*
Other Splicing Proteins									
fSAP18	-	-	-	-	-	-	-	-	-
fSAP105	-	-	-	-	-	-	-	-	-
fSAP121	-	-	-	-	-	-	-	-	-
fSAP79	-	-	-	-	-	-	-	-	-
fSAP24	-	-	-	-	-	-	-	-	-
SPF30	-	-	-	-	-	-	-	-	-
SNP70	-	-	-	-	-	-	-	-	-
NAP	-	-	-	-	-	-	-	-	-
ZNF207	3	Contig72	C(20344-20123)	-	-	-	3	Chr10	381361-381645
fSAP71	-	-	-	-	-		3	Chr12	1936847-1937677
WTAP	-	-	-	-	-	-	3	Chr12	1264059-1264421
fSAP152	-	-	-	3	Contig5711	3782-4102	-	-	-
SHARP	3	Contig98	34205-34999	3	Contig5711	4337-5560	3	Chr12	C(284160-283015)
CIRP	4	Contig36	C(43284-43096)	4	Contig4121	C(2186-1965)	?	-	-
FBP3	-	-	-	-	-	-	3	Chr6	267421-267783
fSAP35	4	Group2	-	3	Contig5927	C(9845-9459)	3	Chr3	370486-370752
				3	Contig5437	2350-3333			

Group Results for Giardia lamblia Candidate Sequences.

Group1: Giardia lamblia Contig 2

Protein	Area within contig
U1_70	154611-154775
U2-A	154641-154847
U2_B	154641-154847
P14	154608-154829
OTT	154611-154865

Group 2: Giardia lamblia Contig 76

Protein	Area within contig
U5_40	23142-22411
Prp4	23268-22447
Prp17	23070-22417
fSap35	23367-22411
Tex1	23013-22396
fSap57	23268-22414
Spf38	23142-22489

Appendix C.3: RNaseP Protein Candidate Sequences:

Giardia lamblia (AACB01000000)

http://www.ncbi.nlm.nih.gov

Protein	Contig	Region	+/-	Method
Pop1	AACB01000017.1	65532-63193	-	ASR
Pop4	AACB01000007.1	78993-79586	+	BLAST
Pop5	AACB01000007.1	7351-7713	+	HMMer
Rpp21	AACB01000050.1	1091-1504	+	BLAST

Entamoeba histolytica

Database downloaded from: ftp.sanger.ac.uk/pub/pathogens/E_histolytica/

Protein	Contig	Region	+/-	Method
Pop1 (partial)	3539	210-698	+	ASR
Pop4	5866	4036-3758	-	ASR
	3512	2022-1708	-	

Encephalitozoon cuniculi

http://www.ncbi.nlm.nih.gov

Protein	Chromosome	Region	+/-	Method
Pop1	Chr VIII	28735-27521	-	BLAST
Rpp21	Chr X	233061-233273	+	BLAST

Appendix D: Perl Scripts developed for this thesis

Disclaimer: These Perl Scripts were written by myself during this project. They worked well on my computer but I make no claims that they will work on anyone elses.

D.1: BlastHits1.0.pl

```
# BlastHits1.0.pl
# Perl Program to parse BLAST output for analysis
#
# Author: Lesley Collins  -  Updated 15 January 2004
# Includes identity and positive percentages
#########################################################
#
# Input: BLAST standard output file
#
# Output: File formatted for input into Excel tables
#
#########################################################

print "BLAST2tab.pl\n============\n\n";
print "Input File: ";
$inputFilename = <STDIN>;
chomp $inputFilename;

open (INFILE, "$inputFilename") or die "Input file not opened";

print "Output file: ";
$outputFilename = <STDIN>;
chomp $outputFilename;
open (OUTFILE, ">$outputFilename") or die "Output file not opened";

print OUTFILE "BlastHits2.pl Program Output\n";
print OUTFILE "============================\n\n";
print OUTFILE "Blast Output file:\t$inputFilename\n\n";
#column headings
print OUTFILE "Query\tGene\tSpecies\tLength\tContig\tScore\tE-value\tFrame\t";
print OUTFILE "Identity\t\tPositives\t\t";
print OUTFILE "Query Region\t\tSubject Region\n\n";
#Set some original values
$dbfound = 0;
$inContig = 0;
$inAlign = 0;
$inQuery = 0;
$newQuery = 0;

#############################################################
#   Read each line of the input file
#   Check for the name and date of database
#   This is normally at the end of the blast output file
#   so dbfound indicates that the blast file is at the end.
#   QueryName is the start of each query.
#   Set inQuery to 1.
#############################################################

foreach my $line(<INFILE>)
{
  if ($line =~ /Posted date:\s*(.*)/)
  {
     $dbfound = 0;
     $postedDate = $1;
     chomp $postedDate;
  }
  elsif ($line =~ /Query=\s(.*)\s/)
  {
     $queryName = $1;
     chomp $queryName;
```

```
    if ($inQuery == 0){$newQuery = 1;}
    $inQuery = 1;
    $accession = "";
    $species = "";
    $contigName = "No Hits Found";
    if ($queryName =~ /(\w*)\s*(\w*)\s*(\w*\s\w*)/)
    {
     $accession = $1;
     chomp $accession;
     $gene = $2;
     $species = $3;
     $queryName =$accession;
     chomp $species;
     }
  }
  elsif ($line =~ /Database:\s(.*)/)
  {
     if ($dbfound == 0)
     {
        $database = $1;
        chomp $database;
        $dbfound = 1;
     }
  }

#################################################################
# New contig has been found,
# Print out any residual line of old contig
# Indicate that a new subject has been started
# Set inSubject to 1
# Can have multiple alignments under each contig
#################################################################

elsif ($line =~ />(.*)/)
  {
    if ($inQuery == 1)
    {
      $Qend = $Q2;
      $Send = $S2;
      $dbfound = 0;
    }
    if ($inContig == 1 & $dbfound ==1)
    {
       print OUTFILE "$queryName\t$gene\t$species\t$length\t";
       print OUTFILE
"$contigName\t$score\t$expect\t$frame\t$identityScore\t$identityPercent\t$positivesScore\t
 $positivesPercent\t";
      print OUTFILE "$Qstart\t$Qend\t$Sstart\t$Send\n";
      $inAlign = 0;
      $qstart = 0;
      $sstart = 0;
      $inContig = 0;
    }
   #reset values to avoid any floating values coming in with "no hits found".
    $Q1 = "";
    $Q2 = "";
    $S1 = "";
    $S2 = "";
    $Qstart = "";
    $qstart = "";
    $Sstart = "";
    $sstart = "";
    $score = "";
    $expect = "";
    $inAlign = 0;
    $Qend = "";
    $Send = "";
    $inqueue = 0;
    $frame = "";
    $identityScore = "";
    $identityPercent = "";
```

```
    $positivesScore = "";
    $positivesPercent = "";
    $contigName = $1;
    $inContig = 1;}

#############################################################################
#  For each new Alignment under a contig
#  Record.. score, expect, frame, Qstart, Qend, Sstart, Send
#
#
#
#############################################################################

 elsif ($line =~ /Score =\s(\d*.\d*)/)
  {
  #is already in an alignment - print out, reset values before proceeding
  if ($inAlign == 1)
        {
       $Qend = $Q2;
       $Send = $S2;
 print OUTFILE "$queryName\t$gene\t$species\t$length\t$contigName\t$score\t$expect \t";
         print OUTFILE
"$frame\t$identityScore\t$identityPercent\t$positivesScore\t$positivesPercent\t";
 print OUTFILE "$Qstart\t$Qend\t$Sstart\t$Send\n";
       $Q1 = "";
       $Q2 = "";
       $S1 = "";
       $S2 = "";
       $Qstart ="";
       $qstart = "";
       $Sstart = "";
       $sstart = "";
       $score = "";
       $expect = "";
       $frame = "";
        $identityScore = "";
        $identityPercent = "";
        $positivesScore = "";
        $positivesPercent = "";
       }
   $score = $1;
   if ($line =~ /Expect.* =\s(\d*.*)/)
   {
     $expect = $1;
     $inAlign = 1;
   }
 }
 elsif ($line =~ /Frame = (.\d)/)
  {$frame = $1;}

 elsif ($line =~ /Identities\s=\s(\d*.\d*)\s.(\d*%)/)
 {
   $identityScore = $1;
   $identityPercent = $2;
   if ($line =~  Positives\s=\s(\d*.\d*)\s.(\d*%)/)
   {
   $positivesScore = $1;
   $positivesPercent = $2;
   }
 }
 #start and stop of query and subject alignment

 elsif ($line =~ /Query:\s+(\d+)\s+\D+\s+(\d+)/)
 {
  $inQuery = 1;
  $Q1 = $1;
  $Q2 = $2;
  if ($qstart ==0)
  {
    $Qstart = $Q1;
    $qstart = 1;
```

```
  }
 }
 elsif ($line =~ /Sbjct:\s+(\d+)\s+\D+\s+(\d+)/)
 {
  $S1 = $1;
  $S2 = $2;
  if ($sstart == 0)
  {
   $Sstart = $S1;
   $sstart = 1;
  }
 }
 elsif ($line =~ /Number of letters/)
 {
   $Qend = $Q2;
   $Send = $S2;

  print OUTFILE "$queryName\t$gene\t$species\t$length\t$contigName\t$score\t$expect \t";
  print OUTFILE
"$frame\t$identityScore\t$identityPercent\t$positivesScore\t$positivesPercent\t";
  print OUTFILE "$Qstart\t$Qend\t$Sstart\t$Send\n";
    $Q1 = "";
    $Q2 = "";
    $S1 = "";
    $S2 = "";
    $Qstart ="";
    $qstart = "";
    $Sstart = "";
    $sstart = "";
    $score = "";
    $expect = "";
    $frame = "";
    $identityScore = "";
    $identityPercent = "";
    $positivesScore = "";
    $positivesPercent = "";
    $contigName = "No Hits found";
 }
elsif ($line =~ /(\d+)\sletters/)
  {
   $length = $1;
  }
}
print OUTFILE "Database:\t$database\n";
print OUTFILE "Date Posted:\t$postedDate\n";

*********************************************************************************
```

D.2: FindContig.pl

```
# FindContig.pl
# Perl program to extract 1 fasta formatted contig from a database txt file
#test program - working 26-02-03
#Author: Lesley Collins
print "\n\nFindcontig.pl\n";
print "**************\n\n";
print "Contig to be found:  ";
$contignumber = <STDIN>;
chomp $contignumber;
print "Database file:  ";
$infile = <STDIN>;
open (DATABASE, $infile) or die "File not opened";
open (OUTFILE, ">$ENV{USERPROFILE}\\Desktop\\outfile$contignumber.txt") or die "File not
opened";
$seq_line = 0;
$contig_found = 0;
foreach my $line (<DATABASE>) {
        if ($line =~ /^>/m) {
           if ($contig_found) {
           outputprint ();
           exit;
        }
        if ($seq_line) {
        $line =~ s.[\s0-9\/]//g;
        $seq .= $line;
        }
        if ($line =~ /^>(.*)$contignumber(:?\Z|\B)/mi) {
              $ident .=$line;
              $seq_line = 1;
              $contig_found =1;}
}
if (eof) {
        if (!$contig_found) {
           print "Contig not found in this file.\n";
            print "Press Enter to exit";
           $close = <STDIN>;
        }
outputprint ();}
exit;

sub outputprint {
        $seq = "\U$seq";
        print (OUTFILE "$ident");
        print (OUTFILE "$seq\n");
        #print "\n$ident";
        #print $seq;
        print "\n\nData has been printed to the file outfile$contignumber.txt on the
desktop\n\n";
        print "Press enter for program to end.\n";
        $close = <STDIN>;
        }

*****************************************************************************
```

D.3: RNAmotif_Count.pl

```
# Perl program to analyse RNAmotif output files.
# Program will output the name and number of output hits for each line
# within a certain number of nucleotides of the start position for that
# hit.

#open file to be examined.
print "Please enter file to be examined: ";
$inputFilename = <STDIN>;
chomp $inputFilename;
open (INPUTFILE, $inputFilename) or die "Input file not opened";

#open output file.
print "Please enter name for output file:";
$outputFilename = <STDIN>;
chomp $outputFilename;
open (OUTFILE, ">$outputFilename") or die "Output file not opened";

$position = 0;
$limit = 100;
$count = 0;
$name = "Initial Settings ";
print OUTFILE <"RNAmotif Output File Analysis\n">;
print OUTFILE <"==============================\n\n">;

print OUTFILE "Input file:  $inputFilename\n";
print OUTFILE "Sequence Position limit: $limit\n\n";
print OUTFILE sprintf "%-36s\t%-10s\t%-5s\n","Name","Position","Count";
print OUTFILE "================================================================\n\n";

foreach my $line (<INPUTFILE>) {
   if ($line =~ />(.*)/) {
      chomp $line;
      if ($line =~ /$name/) {
        }
      else {
      print STDOUT sprintf "%-40s\t%-5d\t%-5d\n", $name,$position,$count;
      print OUTFILE sprintf"%-40s\t%-5d\t%-5d\n", $name,$position,$count;
      $name = $1;
      $count = 0;
      $position = 0;
      }
      $count ++;
   }
   else {
   if ($line =~ /\s[\d]\s+(\d+)\s+/) {
    $positionNew = $1;
    $positionMin = $position-$limit;
    $positionMax = $position+$limit;
    if ($position != 0) {
      if ($positionNew > $positionMax || $position <$positionMin) {
    print STDOUT sprintf "%-40s\t%-5d\t%-5d\n", $name,$position,$count;
    print OUTFILE sprintf"%-40s\t%-5d\t%-5d\n", $name,$position,$count;
    $position = $positionNew;
    $count = 1;
    }
   }
   else {$position = $positionNew;}
   }
  }
}
print STDOUT sprintf "%-40s\t%-5d\t%-5d\n", $name,$position,$count;
print OUTFILE sprintf"%-40s\t%-5d\t%-5d\n", $name,$position,$count;


**************************************************************************
```

D.4: RNAmotif_Filter.pl

```
# Perl program to filter RNAmotif output files.
# Program takes an RNAmotif output file (either before or after getbest),
# then extracts all the lines containing scores above a user-defined
# limit.

#open file to be examined.
print "Please enter file to be examined: ";
$inputFilename = <STDIN>;
chomp $inputFilename;
open (INPUTFILE, $inputFilename) or die "Input file not opened";

#open output file.
print "Please enter name for output file:";
$outputFilename = <STDIN>;
chomp $outputFilename;
open (OUTFILE, ">$outputFilename") or die "Output file not opened";

print "Please enter maximum score value, (Default = 1.0): ";
$scoreMin = <STDIN> || 1.00;
chomp $scoreMin;
print "Minimum score = $scoreMin\n";
print OUTFILE <"RNAmotif Filter Program\n">;
print OUTFILE <"============================\n\n">;
print OUTFILE "Input file:  $inputFilename\n";
print OUTFILE "Minimum score = $scoreMin.\n";
print OUTFILE "=========================================\n\n";
$count = 0;

foreach my $line(<INPUTFILE>) {

   if ($line =~ /[>|#]/) {
     $count ++;
     }
   elsif ($line =~ /(^[\S*|\\]+)\s([\D|,]+)(\d+.\d+)/) {
     $name = $1;
     $motif = $2;
     $score = $3;
     print " Name = $name  Motif = $motif  Score = $score\n";
     $count++;
     if ($score >= $scoreMin ) {
        print OUTFILE "$line";
        }
     }
   elsif ($line = eof) {
      print "Count = $count.\n";
      print "eof reached - program finished\n";
      print "Please press enter to exit.\n";
      $exit = <STDIN>;
      exit;
    }
}
```

****

D.5: SplitDatabase.pl

```
#SplitDatabase.pl
# A program to take a fasta formatted database of contigs or chromosomes,and split
# it into smaller databases enabling speedier runs on parllel systems
# Author:  Lesley Collins
# Date: 6-03-03
# Some bug fixes by WTJW 27/3/2003.

print "SplitDatabase.pl\n";
print "***************\n\n";

$maxFiles = 40;
$outfileNumber = 1;
$contigNumber = 0;
#  User-defined input
print "Please enter fasta database to split:\n";
$infile = <STDIN>;
chomp $infile;
print "Please enter number of fasta files in each new file:\n";
$contigsEachFile = <STDIN>;
chomp $contigsEachFile;
print "Please enter name for output files:\n";
$outputName = <STDIN>;
chomp $outputName;
print "$outputName$outfileNumber\n";

open (DATABASE, $infile) or die "Input File not opened";
print "Infile = $infile\n";

open (OUTFILE, ">$outputName$outfileNumber.txt") or die "Output File not opened";
print "$outputName$outfileNumber.txt opened\n";

foreach my $line (<DATABASE>) {
        if ($line =~ /^>/m) {
                $contigNumber ++;
                #print "contigNumber =$contigNumber\n";
                if ($contigNumber > $contigsEachFile) {
                        close (OUTFILE);
                        print "$outputName$outfileNumber.txt closed\n";
                        $outfileNumber ++;
                        if ($outfileNumber > $maxFiles) {
                        print "Too many files would be produced! Stopping now.\n";
                        last;
                        }
                open (OUTFILE, ">$outputName$outfileNumber.txt") or die "File not opened";
                print "$outputName$outfileNumber.txt opened.\n";
                $contigNumber = 1;
                }
        }
        $seq = $line;
        print (OUTFILE "$seq");
}
close (OUTFILE);
print "EOF reached - $outputName$outfileNumber.txt closed\n";
print "Program finished - press enter to exit\n";
$exit = <STDIN>;
exit;

print "Program finished\n";
print "Press enter to exit \n";
$exit = <STDIN>;
exit;
```

****

D.6: GenPeptFile.pl

```
#   GenPeptFile.pl
# Perl program to parse a number of protein sequence files and extract
# the information required by the SpliceSite database.
#
# Author:  Lesley Collins - 09-July-2003
################################################################################
#
# Input file (Genpept file)
# Output file will be a tab delimited text file set up for direct input
# into the SpliceSite database protein table.
# The following fields are tabulated:
# Accession Number; Species; GI number; Protein definition; Protein length;
# Protein Type; Comments; Sequence.
# Note: The primary key in the SpliceSite protein table is the Accession Number.
#
################################################################################

#open file to be examined.
print "Please enter file to be examined: ";
$inputFilename = <STDIN>;
chomp $inputFilename;
open (INPUTFILE, $inputFilename) or die "Input file not opened";

#open output file.
print "Please enter name for output file:";
$outputFilename = <STDIN>;
chomp $outputFilename;
open (OUTFILE, ">$outputFilename") or die "Output file not opened";
print "Optional fields\n";
print "(Note: The same field entry will appear for all proteins in the input file.)\n";
print "Please enter protein name: ";
$name = <STDIN>;
chomp $name;
print "\nPlease enter protein type: ";
$type = <STDIN>;
chomp $type;
print "\nPlease enter comments: ";
$comment = <STDIN>;
chomp $comment;
$multiline = 0;
$count = 0;
$inseq = 0;
foreach my $line(<INPUTFILE>)
{
    if ($line =~ /\/\/\n/)
    {
      $inseq = 0;
      $seq =~ s/[\d | s | t]//g;
      print OUTFILE
"$accession\t$name\t$species\t$gi\t$definition\t$length\t$type\t$comment\t$seq\n";
     #print
"\n$accession\t$name\t$species\t$gi\t$definition\t$length\t$type\t$comment\t$seq\n";
      $count = $count+1;
      $seq = "";
      $definition = "";
      $accession = "";
      $gi = "";
      $species = "";
    }
    if ($inseq == 1)
    {$seq .= $line; }
   elsif ($multiline == 1)
   {
   if ($line =~ /ACCESSION/)
   {
      chomp $definition;
      $multiline = 0;
```

```
    }
   else
   {
    chomp $definition;
    $line =~ s/^\s*(.*?)\s*//g;
    $definition .= $line;
   }
 }
  elsif ($line =~ /LOCUS/)
  {
    $line =~ s/LOCUS[\s+]//g;
    $line =~ /(\w+)\b /;
    $accession = $1;
    $line =~ /(\d+)\s+[aa]/;
    $length = $1;
  }
  elsif ($line =~ /GI:(\d*)/)
      {$gi = $1; }
  elsif ($line =~ /ORGANISM/)
   {
   $line =~ s/ORGANISM[\s+]//g;
   $species = $line;
   $species =~ s/^\s*//g;
   chomp $species;
   }

  elsif ($line =~/DEFINITION[\s+](.*)/)
  {
    $multiline = 1;
    $definition = $1;
    $definition =~ s/^\s*(.*?)\s*//g;
  }

  elsif ($line =~/ORIGIN/)
  {$inseq = 1;}
}
print "\nNumber of genbank records processed: $count.\n";
print "Program Completed\n";
print "Output stored in file: $outputFilename.\n";
close(OUTFILE);
print "Hit enter to exit program\n";
$exit = <STDIN>;
exit;

*******************************************************************************
```

Appendix E Data management in the post-genomic era

"To err is human. To really foul things up requires a computer." – Anon

"Databases have recently become more specialized and better curated, but redundancies and database asynchrony is on the rise because of the distributed and collaborative type of research. Therefore, database-to-database comparisons are required for analysis and validation, which consume ever more compute cycles and storage, draining resources from more original research. Furthermore, in research, the databases do not just sit there, they get processed and transformed, calculations are run, and intermediate results are stored." - Briefing in Computational Biology: A white paper from Sun Global, Education and Research, Science and Engineering 2002.

During this project a number of data management issues emerged that are related to the information processes involved in studying ncRNA and associated proteins. Whatever the approach, gene prediction depends to a large extent on the current biological knowledge (Mathe et al. 2002). Searching for ncRNA and protein genes in genomic sequences involved the effective management of large amount of different types of data including protein sequences, whole genomes, literature, protein-motif information and analysed data results. It is evident that data management technology has not kept pace with data generation in biology and further research and development is needed to effectively use the large biological datasets that are now becoming available (Jagadish and Olken 2003). Although there are now a number of software packages available for managing large scale sequencing projects (such as ACeDB: http://www.acedb.org and Ensembl: Hubbard et al. 2002), there is at present little available for the management of small genomic analysis projects. LIMS (Laboratory information management systems) have been introduced to manage data for many applications including forensics and commercial laboratory testing, but these systems rely on standardised procedures and are often not applicable to strategic research laboratories. This chapter highlights the practical data management solutions used during this project to aid genomic searches for a moderate (10 – 50) number of genes.

Many molecular biology/biochemistry/genetics laboratories use the large international databases containing nucleotide (Genbank-(Benson et al. 2000)), protein (Entrez-http://www.ncbi.nlm.nih.gov, which includes SwissProt and PIR databases), protein domains and families (Prosite–http://tw.expasy.org/prosite/), transcription (UniGene-www.ncbi.nlm.nih.gov), literature (Entrez – www.ncbi.nlm.nih.gov) and 3-dimensional structure (PDB–(Berman et al. 2000), http://www.rcsb.org/pdb/) information. As well as public databases, laboratories may also have access to local research-specific information. Graphical interfaces have aided in the acquisition of data from remote databases e.g. Genbank was accessible by command line ftp access, but only after a graphical interface was included has the database become easily accessible. All

this information, often in a variety of conflicting formats, becomes extremely useful to a researcher, **only if** it can be put together in a coherent form. Figure E.1 shows linkages between the different data types in both preliminary information gathering and analysis of a gene. This is an area where data management strategies offer a great advantage.

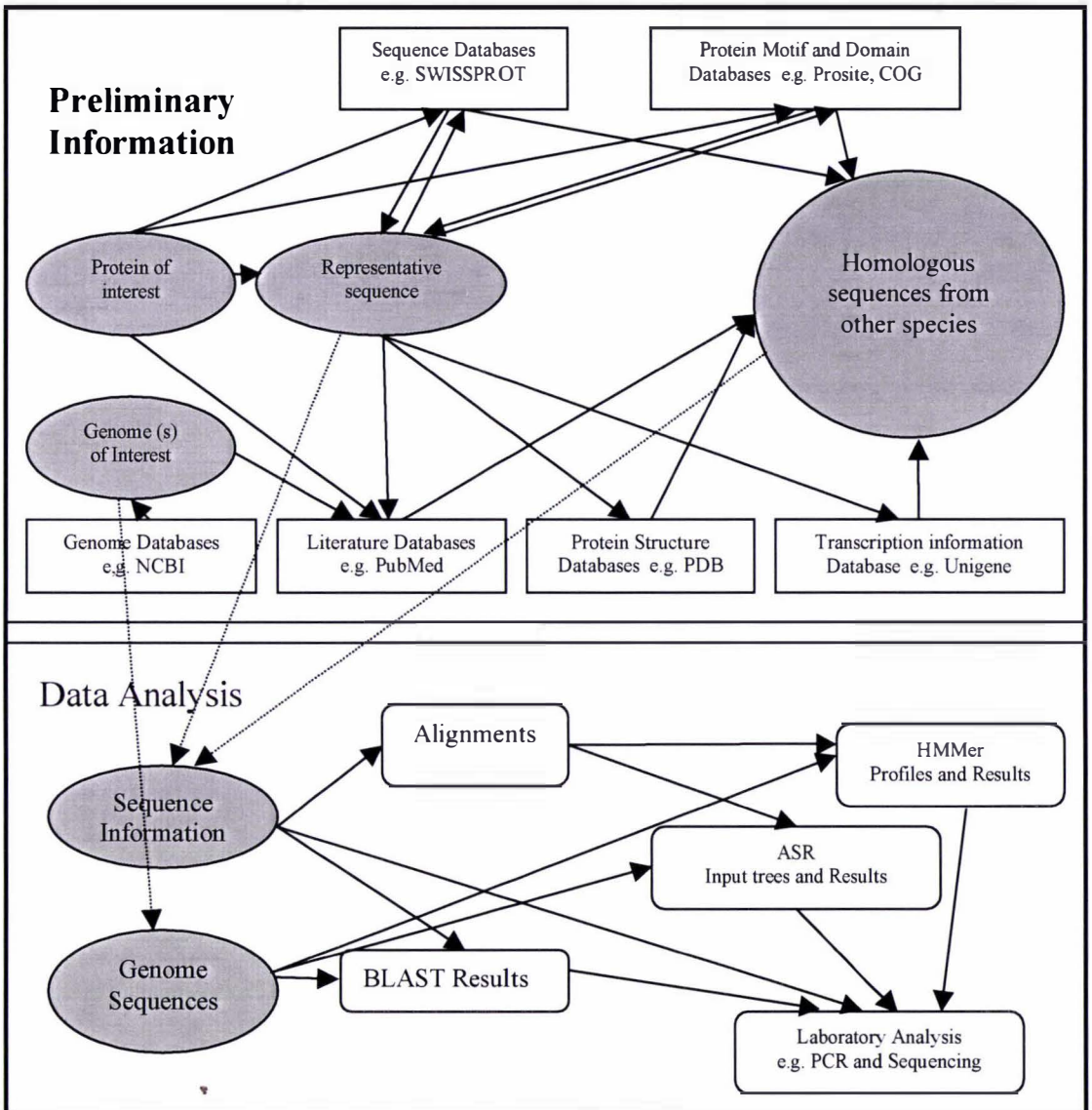


Figure E.1: "Oh, what a tangled web we weave..." (with apologies to Sir Walter Scott). Diagram showing how preliminary and analysis results data can be related and an indication on how the amount of data can quickly get out of hand. Ovals represent sequence information, including complete genomic sequences. Rectangles represent the databases from which a variety of information can be obtained and rounded rectangles represent analysis results that must also be managed for a genomic project. A relational database can organise and manage this information

A search for just one protein can involve a large amount of preliminary data. A protein may have extensive experimental data accumulated in the literature and often has multiple sequences lodged in the protein and nucleotide sequence databanks, each of which may be different in the amount and type of annotation. Thus a representative protein sequence must be selected for this species. The representative sequence is often the sequence mentioned in the literature or can be the same sequence obtained from a protein database containing more feature (and experimental evidence) annotation. This representative sequence can be then be used in a search of local and remote databases for the desired information. This is often not as straight-forward as it seems. The same gene can have many names, differing between species and often these "linked" names are not shown anywhere in the sequence's annotation (e.g. RNase MRP RNA is called 7-2 RNA in *Arabidopsis thaliana*). Another problem is that a protein sequence can be a member of a protein "family", a set of closely related sequences (but not necessarily the same sequence) meaning that it is difficult to select a homologue from a different species unless other information is taken into account (e.g. other small but indicative sequence motifs and/or experimental evidence, often from the literature). Sometimes a search of the annotation for a gene by name will not be productive, not because the name is different but because the sequence homologue may have not yet been identified in a particular species.

It was once thought that a well designed program would be able to access numerous designated databases and automatically return anything relevant to the gene query. There has been some move towards gene ontology methods to recover relevant information from the many different gene-information databases available. Gene ontology uses 'reasoning' and includes the use of natural language processing and logic programming (Bodenreider et al. 2003). A number of programs have been designed to this end (Badea 2003; Lambrix and Edberg 2003) however the biggest problem is not with the search software but in how the original data has been entered into the accessed databases. For example many genomes, although complete, are poorly annotated containing comments like "probable splicing protein" and "similarity to Arabidopsis protein" with no mention of what protein, that are not very helpful to the researcher and are not returned using a search based on a gene's name. Genome and individual gene annotation is often a mixture of data interpretation and experimental observation, sometimes with little indication as to which of the two has been used (Jagadish and Olken 2003). The goal of the Gene Ontology Consortium ™ (GO) (http://www.geneontology.org/) is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and

changing. As this system is being applied, the annotation of sequences and genome will become clearer and searches based on function, name and properties will become much more reliable than at present.

A useful preliminary search tool is the "BLink" application available for protein sequences lodged in the NCBI protein databases (references and website). BLink provides a graphical and statistical display of BLAST searches that have been done for every protein in the Entrez Proteins data domain and can reveal what protein sequences are similar to the query protein and the position and BLAST score of each hit. Homologous proteins can be found quickly, linked annotation examined for level of evidence (experimental, blast similarity or hypothetical) and relevant files downloaded. Thus, during a search for either a protein (or ncRNA) gene the amount of accumulated data can quite quickly get out of hand with the problem being magnified when a number of genes are involved in a project, such the components of the spliceosome. In this situation computerised data management becomes almost essential.

Simple forms of data management usually involve storing certain types of files together (e.g. one folder of alignments, one folder for genomes and one for results). When working with a very small number of genes, one folder may in fact be used for all the data associated with each gene. This is perhaps the most common type of data management seen in "simple" genomic searches and is used in many molecular biology laboratories. A folder containing all the data for one gene can have subfolders for alignments and results. However, these folders can become very large if there are a number of proteins involved and data can easily become misplaced and lost. A different approach is to use a custom designed database to import, sort, store and export the different types of data associated with small gene searching projects.

Perhaps the most useful aspect about storing genomic information in a database is that "data mining" (the efficient extraction of relevant data from a dataset) can be supported. Being able to recover information that has already been downloaded from an exterior database saves time and effort. Data can be stored in such a way that duplicate sequences can be avoided (i.e. with the accession number used as the primary key) and results that map to more than one query are also highlighted. Databases are also easily archived and enable easy backup of important genomic information.

Data integration and recovery is one of the core aspects of computational biology (Roos 2001), and there are now a number of database systems under development such as MyGrid (Stevens et al. 2003) and LIMBO (Philippi 2004) (Neither of these systems is readily

available as yet). GIMS (Cornell et al. 2003) is an integrated data storage and analysis system that has been applied to a number of genomes (presently *S. cerevisiae*, mouse and some bacterial genomes) and is accessible over the internet. Unfortunately each GIMS database system is custom built for each organism and the base system is not yet packaged or available for outside use.

Without any readily available genomic database system, an alternative was to develop a personal data management system to keep data 'explosion' under control. During this project, two databases were designed to contain information about the proteins and ncRNA that was being studied at the time. The first, P-MRPbase was a database containing information about RNaseP and MRP; two ribonucleoprotein complexes that contain ncRNA and share numerous protein components. The second was SpliceSite, a database holding information about proteins involved in the eukaryotic spliceosome (a "massive" ribonucleoprotein complex containing ncRNAs and a large number of proteins).

The key to the effective use of any database is in its design. Defining what information will be stored in the database and how it is related is essential, before even opening the database software (Bergeron 2003). If a database has been designed well then all the data can be managed from a single database file that includes tables in which to store data, queries to find and retrieve data, forms to view and update data in tables and reports for data output.

The databases in this project were created using the popular small database program (Microsoft Access), which is readily available and enabled small databases to be quickly developed and modified to handle the different types of data used. It is highly likely that there is other software that is able to handle gene-associated data in either a more user-friendly or platform-friendly way, or in a much more efficient way. However, Access is a good database system to start with when having to design small personal databases. Personal databases do not have to be large or contain graphical interfaces (although this can be readily achieved with Access and makes using the database easier) but can easily link different types of data from different input sources. The object in creating the P-MRPbase and SpliceSite databases was not to display data on the internet, or for sharing with multiple users (although it is possible to adapt these databases to do these things in the future), but for the personal management of a wide variety of information gathered during this project.

E.1: P-MRPbase

This database was created mainly to organise data files that had previously been stored in folders. It wasn't desirable at that stage to import all the information into the database but allow easy access to the data files using hyperlinks. A hyperlink needed only to be 'clicked' to open the appropriate file. Elementary data such as name, accession number, organism and sequence was stored as text as this data was frequently accessed for a number of analysis applications including construction of alignments and BLAST queries. Sequence, reference, alignment and results files were accessed using hyperlinks to their files.

Predicted physiochemical properties fields were included in the protein tables as these properties can be easily calculated from a protein sequence using such programs as PepTool (Wishart et al. 1997);'Lite' version available from http://www.biotools.com) and WinPep (Hennig 1999); available from http://www.ipw.agrl.ethz.ch/~lhennig/winpep.html). Properties such as length, isoelectric point (pI), relative ratios of singlets/pairs of residues and codon usage can give surprisingly strong clues about function of a sequence (King et al. 2004) and were extremely useful for comparison between known and candidate protein sequences.

RNaseP and RNaseMRP RNA information was also included in this database in separate tables from each other and their proteins. All protein data was entered into one table (accessed with the "Proteins Table" button) but entries referenced to any of the proteins could be accessed using the buttons for the individual proteins.

Screenshots from this database are shown in Figure E.2 showing (A) the entry screen (for easy access to the different RNase P/MRP proteins and RNA data), (B) an example of a protein form; and (C) BLAST results form showing data from one query. P-MRPbase presently contains 150 protein entries in the protein table, 116 entries in the RNaseP RNA table and 13 entries in the RNaseMRP RNA table. There are 45 alignment entries and 146 entries in the BLAST results table. 138 organisms are represented, 65 documents, 28 text files and 33 web pages are also linked to this database. Not all of the bacterial RNaseP RNA information has been included in the P-MRPbase as this information is available from the RNaseP database (Brown 1999) which also includes RNA secondary structural information. The RNaseP database however, contains only limited RNaseP protein information.

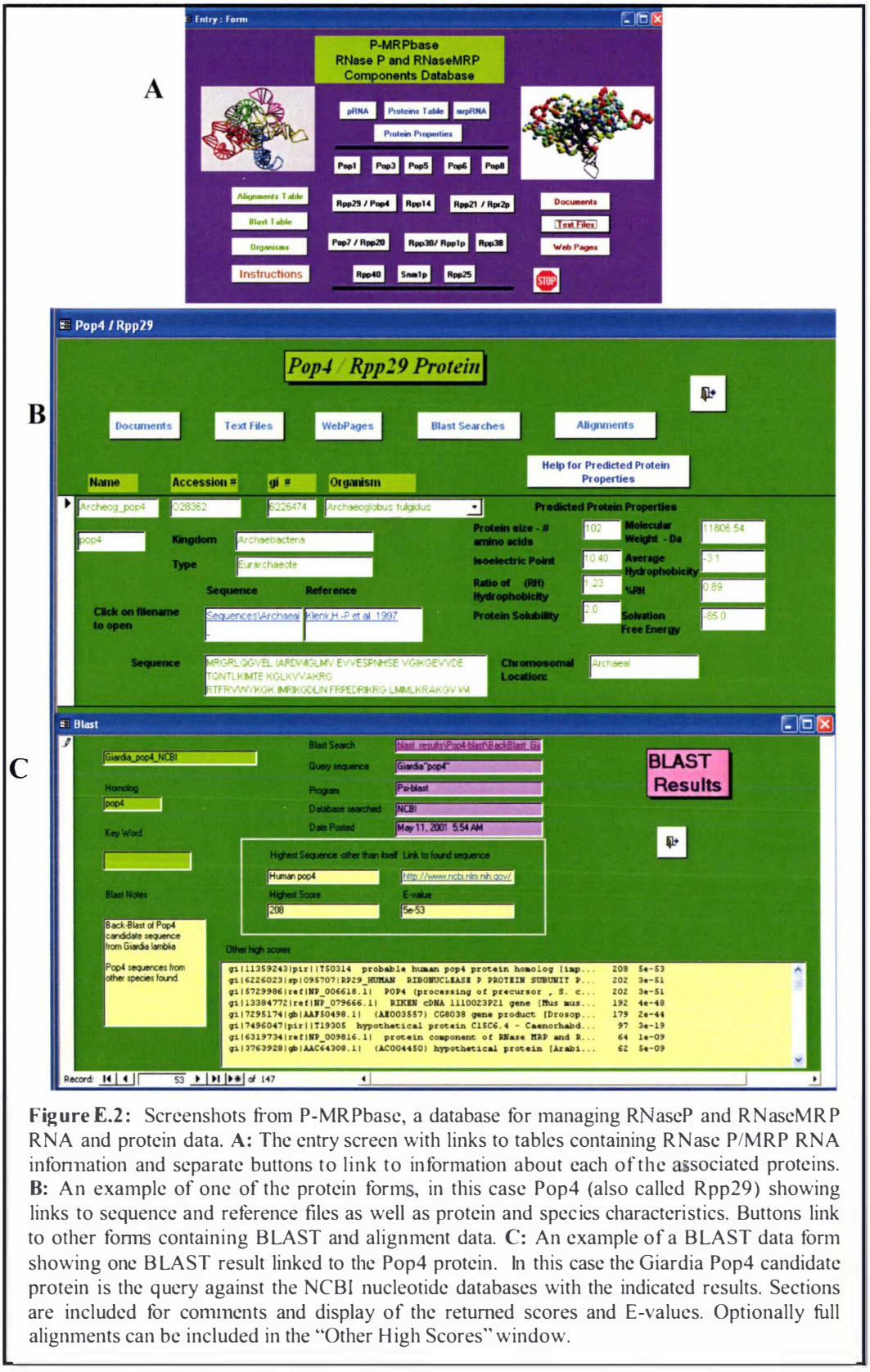


Figure E.2: Screenshots from P-MRPbase, a database for managing RNaseP and RNaseMRP RNA and protein data. **A:** The entry screen with links to tables containing RNase P/MRP RNA information and separate buttons to link to information about each of the associated proteins. **B:** An example of one of the protein forms, in this case Pop4 (also called Rpp29) showing links to sequence and reference files as well as protein and species characteristics. Buttons link to other forms containing BLAST and alignment data. **C:** An example of a BLAST data form showing one BLAST result linked to the Pop4 protein. In this case the Giardia Pop4 candidate protein is the query against the NCBI nucleotide databases with the indicated results. Sections are included for comments and display of the returned scores and E-values. Optionally full alignments can be included in the "Other High Scores" window.

E.2: SpliceSite

The SpliceSite database created to manage spliceosomal protein information was designed differently from P-MRPbase. This database was designed before any data was downloaded with the intention that data would be imported directly into database tables enabling efficient data mining and analysis. Database tables and data-relationships were created to ensure that the data from different types of files could be effectively related to each other. Protein files were downloaded in GenPept format (a standard format used for protein files in the NCBI databases) then converted to tabular format using the Perl script "genpeptfile.pl" (Appendix D) it was found that XML format (Bergeron 2003), also available at NCBI sites, was also useful for importing data into database tables. The advantages of using genpeptfile.pl was that not all the data needed to be imported into the 'protein' table and this program enabled the required data (including sequence, reference and literature data) to be reformatted into a tabular format that was easily imported into SpliceSite. Screenshots of the SpliceSite database are shown in Figure E.3.

Many of the proteins imported into this database belonged to protein families and sometimes a protein may be present in multiple BLink results. Importing the data into tables enabled these multiple results to be noted upon inclusion into the table by using the accession number of the sequence as the primary key. A comment could then be inserted with the query proteins that indicated close similarity to other proteins in the database. Proteins were also allocated to a 'group' applicable either to the proteins association with a snRNA (e.g. U1-specific protein) or to any common motifs (e.g. SR proteins). Not all the proteins known to be associated with the spliceosome were downloaded due to the shear numbers of proteins concerned (presently between 150 and 300 (Jurica and Moore 2003), but at last count 1594 individual sequences representing 153 proteins were stored in SpliceSite.

BLAST information from searches of local ("on site" databases as opposed to "global" international) databases (e.g. in this project *Giardia lamblia* and *Entamoeba histolytica* genomes) was formatted for direct import into the "BlastHit" table using the Perl program "BlastHits.pl" (Appendix D) and was cross-referenced to the individual proteins and to the protein group to which each protein belonged. Another table recorded information about the genome database, including date posted, open or restricted and any comments that could be directly related to the BLAST searches. There is also a table to record file storage (using hyperlinks) with the linked files stored in associated folders. A table pertaining to protein motifs and domains has been set up to enable cross referencing to Prosite entries (Sigrist et al. 2002); http://kr.expasy.org/prosite/) but this feature has not yet been activated.

By designing the SpliceSite database carefully before adding data, future relationships can be catered for without the potential loss of data that can occur if modifications are made to a database already in use.

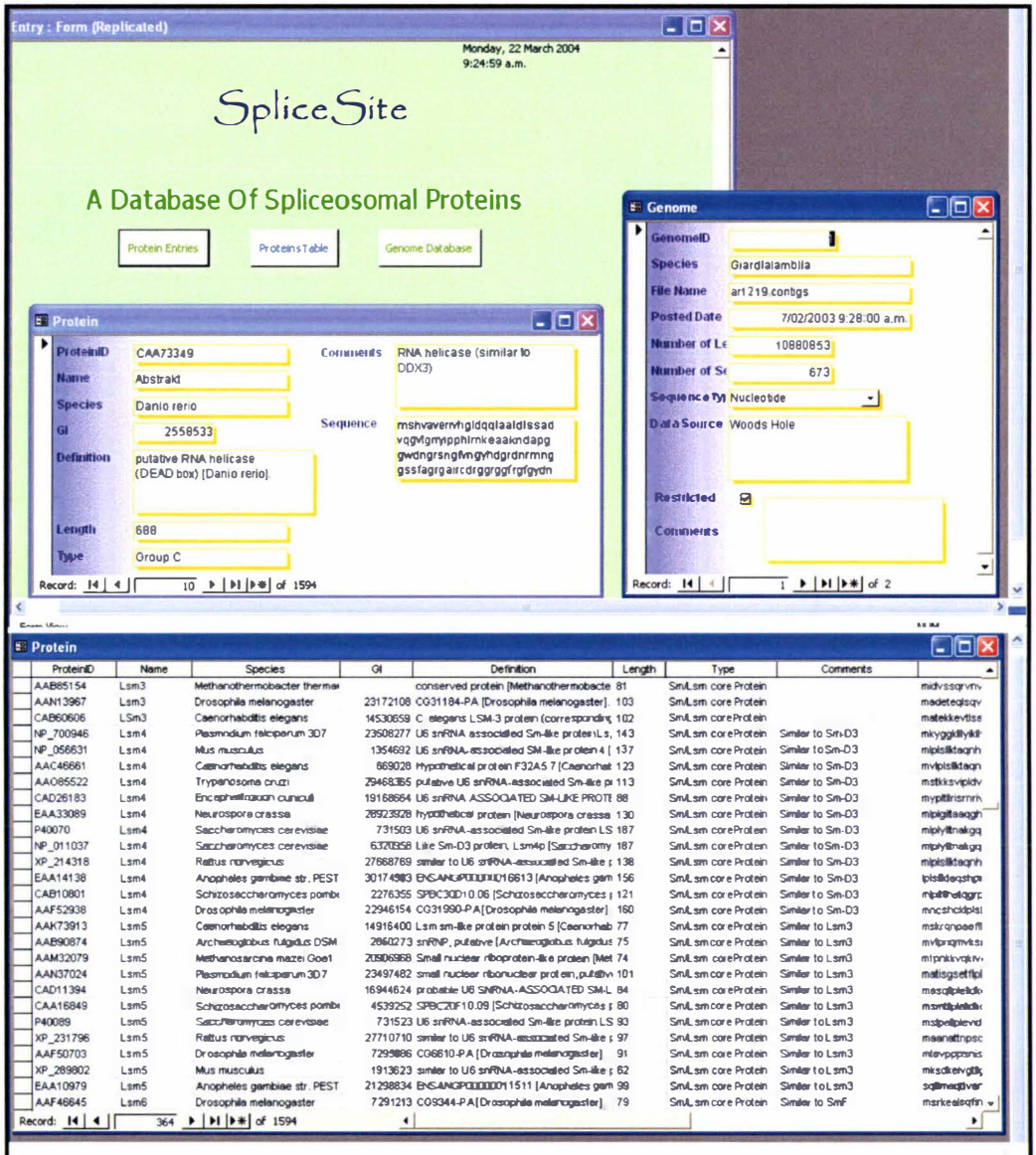


Figure E.3: Screenshots from the SpliceSite database created in this project to manage data from spliceosomal proteins. **A:** The entry screen with links to tables containing protein and genomic information. The protein information and genome database information forms are also shown.
B: The protein information table in spreadsheet format showing information from a number of spliceosomal proteins, Lsm3-6. Each protein is groups according to its 'Name' and 'Type' and included a comments field for any extra user-input information such as similarity to other proteins. Information from the protein table can be cut and pasted into text and formatted (e.g. Word) documents.

Neither P-MRPbase nor SpliceSite are complete in that they hold all the information about all the proteins involved in their separate complexes. There is still much data that could be added to these databases but even in their incomplete state, they enabled rapid recovery of gathered information, avoiding duplication of searches and loss of results. There are many improvements that could be made to both the P-MRPbase and SpliceSite databases, such as automatic data entry and automatic calculations of physiochemical properties. Features such as being able to search external databases (such as those at NCBI) for new information or integrated BLAST functions, are often built into large sequencing project databases and it might be possible to export some of these features to a small scale version. However, such features will require expert database management assistance and may not be required for personal management systems but would be extremely useful if the database was modified for multiple user or internet usage.

Not all information needs to be incorporated into a database. Some information can be stored in a spreadsheet which then could be included in a database at a later stage. One example is the spreadsheet containing the blast results with SpliceSite proteins against the *G. lamblia* genome. Information was extracted from the BLAST results using the Perl Script BlastHits1.0.pl (AppendixD). These spreadsheets were then imported into SpliceSite so that results could be directly related to the query protein information.

E.3: Future Directions

Even with expert design assistance, it is hard to determine, which features of a database will be useful in practice and which will not. Lessons learned from the design and use of the databases in this thesis will lead to the better design of future personal databases. An ideal project would be the design of a database "template" with features designed for the "average" gene hunter with perhaps separate features available for the storage of ncRNA data and protein sequence data. Added features could also include tables for the storage of laboratory experimental data (including gel documentation images, methods and PCR information), bibliographic searches and literature-gained information. This sort of database would best be designed with the aid of a practical laboratory willing to give this type of information storage system a trial.

Relational databases were used in this project for the main reason that this was the type of database with which I was most familiar. Other types of database structure such as

object-oriented or semi-structured (Bergeron 2003) may in fact be more efficient for this type of application but may require using different software for database construction.

Another useful application of this work would be to promote teaching of genomic-based small database design to those who would benefit from this skill; that is the researchers themselves. Commercial companies are available that can design specific databases; however this expensive option is often not available for academic researches. Courses are available for Microsoft Access and are impart the basics of database design. However available documentation (and course example data) is economy and small-business based, rather than science based and instructors are often unfamiliar with scientific, especially biological, terminology. The requirements of data representation and queries often present challenges which are often different than what is typically needed by business data processing (Jagadish and Olken 2003). As more researchers are taught how to manage the vast amounts of accessible data, this skill will become as common as checking one's e-mail is today.

Knowledge of Perl programming is a great advantage in the manipulation of data from one format to another, the major drawback to present data management systems. A number of Perl scripts (Perl programs are commonly called scripts) were written for this project transforming files such as GenPept and BLAST results files into formats that could be imported into the relevant database tables. The scripts that were written for this project are readily available for others to use (by request), however, accompanying documentation is not currently available, and it is a well known feature of Perl that it is a "write-only" language i.e. it is difficult to understand another Perl programmers scripts. Bioperl (open source software that is still under active development) is a collection of Perl modules that aid the development of Perl scripts for bioinformatics applications. These modules can be used in Perl scripts and offer fast solutions to programming problems associated with the manipulation of genomic information. Bioperl could be used to develop 'linking' programs with graphical interfaces that would for example, enable a file to be transformed into a database-friendly format by someone with little programming knowledge.

With the increasing focus on intellectual property, especially when there are possible commercial opportunities, it is more important than ever to have an effective system of data management. In the commercial world, lost information costs time and money as well as possible patent disqualification and yet there seems to be no "accepted" system of genomic data management with each laboratory/institution left to sort something in-house. Any genomic-based database system should be flexible enough for any institution to make changes to suit their research yet hold a framework that will allow effective data interaction.

However, trying to please everyone who requires genomic data management will create more chaos than is already present. Small, simple database design may be presently absent from the bioinformatics 'training' repertoire but provides a powerful data management tool that cannot now be ignored.

Figure E.4: A good reason for a data management system. This was the literature data alone used for the research carried out in Chapter **4**.
Guest appearance by Shannen Collins (Age 3.5).