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.

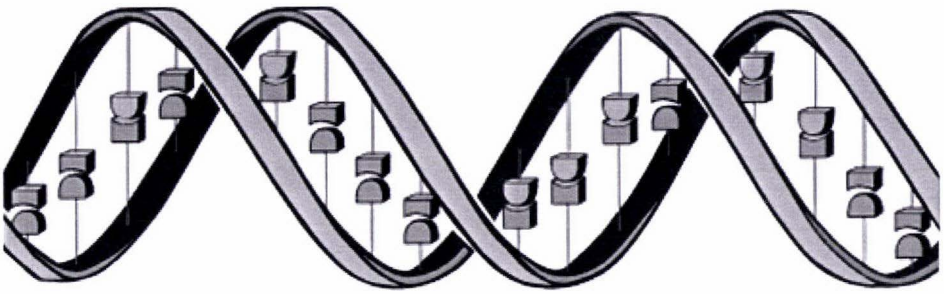
Thanks to the many people at the Allan Wilson Centre, especially Joy and Susan, for friendship, support and funding. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for the RT-PCR and sequencing of the *G. lamblia* ncRNA gene candidates over the course of some summers. A special thanks to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