Université de Montréal

Insights into the function of short interspersed degenerated retroposons in the protozoan parasite *Leishmania*

par
Martin Smith

Programme de bioinformatique
Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de *Magister Scientiæ*
en bio-informatique

Décembre, 2007

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulée :

Insights into the function of short interspersed degenerated retroposons in the protozoan parasite *Leishmania*

présenté par :

Martin Smith

a été évalué par un jury composé des personnes suivantes :

Dr Franz Lang, président-rapporteur

Dre Barbara Papadopoulou, directeur de recherche

Dre Gertraud Burger, codirecteur

Dr Mathieu Blanchette, codirecteur

Dr Stephen Michnick, membre du jury

RÉSUMÉ

Leishmania est un parasite membre des *Trypanosomatidae* qui cause la leishmaniose, une maladie à transmission vectorielle qui afflige des millions de personnes à travers le monde. Ces parasites comportent un grand nombre de processus cellulaires et moléculaires particuliers en raison de leur divergence ancienne dans l'évolution des eucaryotes. Dernièrement, une nouvelle famille de rétroposons courts intercalés et dégénérés (*SIDER*) a été identifiée dans le génome de *Leishmania major*. L'observation d'un grand nombre de SIDER dans les régions régulatrices et le fait que ces éléments peuvent déstabiliser l'ARNm démontrent l'assimilation apparente des éléments SIDER par le génome de *Leishmania*. Ces découvertes ont entraîné des analyses génomiques à grande échelle de la structure et de la fonction des SIDER. Dans la présente thèse, nous décrivons diverses approches comparatives dans le but de caractériser davantage ces répétitions intercalées. La première partie décrit une nouvelle méthode d'amélioration de la prévision des régions non traduites de l'ARNm chez *Leishmania*. La deuxième partie de la présente thèse détaille la création de profils statistiques élaborés pour l'optimisation de l'alignement de séquences et pour la représentation des sous-familles d'éléments SIDER. En combinant ces outils avec des recherches génomiques intraspécifiques et interspécifiques pour *L. major*, *L. infantum* et *L. braziliensis*, nous sommes en mesure de cibler des séquences fonctionnelles présumées pour des analyses *in-silico* additionnelles. Nous démontrons que deux classes de SIDER sont fragmentées et dispersées de façon disproportionnée à travers les génomes de trois espèces de *Leishmanias*. Les estimations antérieures de la distribution génomique des SIDER sont corrigées, tandis que des spéculations sont énoncées en ce qui a trait à la fonction évolutive de ces rétrotransposons assimilés.

Mots clés : ARN non-codant, génomique comparative, *Leishmania*, polyadénylation, rétrotransposons, *trans*-épissage.

ABSTRACT

Leishmania is the trypanosomatid parasite that causes leishmaniasis, a vector-borne disease that afflicts millions of people worldwide. These parasites bear many distinctive cellular and molecular processes ensued by their early divergence in the evolution of eukaryotes. Recently, a novel family of short interspersed degenerated retroposons (SIDER) has been identified in *Leishmania major*. The apparent assimilation of SIDER elements by the *Leishmania* genome has been substantiated by the observation that SIDERs are largely abundant in regulatory regions and by the fact that these elements can reduce mRNA stability. These findings have prompted an in depth genomic analysis of SIDER structure and function. In this thesis, we convey various comparative approaches with the aim of further characterizing these interspersed repeats. The first part describes a novel method for improving the prediction of mRNA untranslated regions in *Leishmania* species. In addition, the second portion of this thesis details the creation of refined statistical profiles for the optimal alignment and profiling of SIDER elements. Combining these tools with intra- and inter-genomic scans in *L. major*, *L. infantum*, and *L. braziliensis*, we are able to target putative functional sequences for further *in-silico* analyses. We show that two SIDER classes are fragmented and unevenly scattered throughout the genomes of three *Leishmania* species. Previous distribution estimates are rectified and possible evolutionary functions of these assimilated retroposons are discussed.

Keywords: comparative genomics, *Leishmania*, non-coding RNA, polyadenylation, retroposons, *trans*-splicing.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

CHAPTER I

LIST OF FIGURES

INTRODUCTION

CHAPTER I

CHAPTER II

LIST OF SYMBOLS AND ABBREVIATIONS

A	Adenine
aa	Amino Acids
bp	Base Pairs
BC	Before Christ
C	Cytosine
cDNA	Complementary DNA
CDS	Coding Sequence
DGC	Directional Gene Cluster
DNA	Deoxyribonucleic Acid
EM	Expectation Maximisation
EN	Endonuclease
EST	Expressed Sequence Tag
G	Guanosine
HMM	Hidden Markov Model
Indel	Insertion/Deletion
kDNA	Kinetoplast DNA
*Lb*SIDER	*Leishmania braziliensis* SIDER
*Li*SIDER	*Leishmania infantum* SIDER
LINE	Long Interspersed Element
*Lm*SIDER	*Leishmania major* SIDER
LTR	Long Terminal Repeats
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
ncRNA	Non-Coding RNA
nt	Nucleotide(s)

ORF	Open Reading Frame
PSSM	Position-Specific Scoring Matrix
RNA	Ribonucleic Acid
RNAi	RNA Interference
RNP	Ribonucleoprotein
RT	Reverse Transcriptase
SIDER	Short Interspersed DEgenerated Retroposon
SJ	Splice Junction
SL RNA	Splice-Leader RNA
snRNA	Small Nuclear RNA
snRNP	Small Nuclear RNP
SS	Splice Site
T	Thymine
TE	Transposable Element
tRNA	Transfer RNA
U	Uridine
UTR	Untranslated Region
Y	Pyrimidine
YY	Pyrimidine Dinucleotide

This work is dedicated to the millions
of people who are afflicted
by vector-borne illness and poverty.

ACKNOWLEDGMENTS

Firstly, this thesis would not have been written without the support and supervision of Dr Barbara Papadopoulou, my primary Master's supervisor. Although associated with another university, Dr Papadopoulou has always taken the time and gone through the various hassles required to ensure my proper academic and scientific development. Her expertise and research interests have propelled my passion for science at a higher level.

The supervision and guidance of Dr Gertraud Burger were most helpful throughout my Master's studies. I am grateful for her encouragement, support, and motivation as well as her constructive scientific input. Dr Burger took the time to assist me, regardless of her busy schedule.

I would also like to acknowledge the assistance of Dr Mathieu Blanchette who has provided thoughtful insight into my work from the first day we met. His scientific prowess was fundamental in developing my aptitudes and analytic capabilities. I consider myself fortunate to have experienced such awesome supervision throughout my Master's.

I would like to recognize the kind and compassionate assistance of Elaine Meunier in most administrative tasks related to my studies at the University of Montreal.

The dynamic scientific environment at the *Infectious Disease Research Center* at the *CHUL* in Quebec City has been an additional source of inspiration throughout my graduate studies. The many friends and colleagues I have met there had their share in my scientific development.

I express my regards to the open source community for Ubuntu Linux and many for useful free programs.

Many thanks to my parents who have provided continued support, without whom I would surely not have undertaken graduate studies.

INTRODUCTION

As DNA sequencing technology keeps progressing, the collection of sequenced genomes is rising at a staggering rate. Deciphering such copious amounts of genomic data undoubtedly requires sophisticated computational tools capable of high-throughput screening. The advent of bioinformatics is an outcome that now plays a crucial role in further understanding the concealed details and discrepancies of the genetic code. As a case in point, this work focuses on the application of particular computational tools to address specific biological problems entailed by the recent genomic sequencing of *Leishmania major*, *Leishmania infantum*, and *Leishmania braziliensis* [1].

1. THE PROTOZOAN PARASITE LEISHMANIA

The term protozoan originates from the Greek words *protos* (first) and *zoon* (living being). It designates a large group of single-celled eukaryotes that are members of *Protista*, one of the five kingdoms of life. Many of these organisms are among the most enigmatic of all known species, conspicuously at the level of cellular biology. A plethora of odd morphologies, specialized organelles, and purposeful structures substantiate their extensive evolution and environmental adaptation. The unusual genetics underlying such evolutionarily distant species make many protists ideal candidates for discovering novel mechanisms of genetic regulation.

Not all protozoa are free-living microorganisms, such as those that can be observed by examining pond water under a microscope. Indeed, the organisms spotlighted in this work form a strictly parasitic genus infamous for plaguing humans and other mammals in tropical and sub-tropical climates worldwide. The following subsections will present an overview of the epidemiology and molecular biology of the *Leishmania* genus; a member of the *Trypanosomatida* family and *Kinetoplastea* order.

1.1. Leishmaniasis, a vector-borne disease

Leishmania parasites cause a diverse spectrum of disease resulting from their pathenogenicity and from the host's immune response [2]. It is estimated that nearly 2 million children and adults develop symptomatic ailments, and the incidence of infection is considerable when considering asymptomatic infections [3]. Given that more than 90% of cases occur in third-world countries, leishmaniasis is considered a neglected disease as insignificant financial gain fails to encourage anti-leishmanial drug development [4]. Recent reports indicate that leishmaniasis is now an emerging zoonosis in the United States [5-7].

1.1.1. Historical considerations

Dum Dum, India – 1901. A Scottish professor from Glasgow named Sir William Leishman examines a pathological specimen from the spleen of a deceased British soldier. The latter was experiencing bouts of fever, anemia, muscular atrophy, and a swollen spleen; symptoms of what was then termed 'kala-azar', the Hindi word for 'black fever'. Upon inspection, Sir Leishman discovers ovoid bodies in the sample and describes the associated illness as 'Dum Dum fever'. His findings were published in 1903, almost simultaneously with the similar yet independent discoveries of Irish physician Charles Donovan (**Figure 1**). The parasitic species was dubbed *Leishmania donovani* and is now known as the causative agent of visceral leishmaniasis [8].

Figure 1. Sir William Leishman and Dr Charles Donovan
Adapted from [8].

Leishmaniasis presents a variety of pathologies, which range from spontaneously healing skin lesions, to horrific destruction of mucocutaneous membranes and deadly splenomegaly (**Figure 2**). Descriptions of lesions highly comparable to these symptoms are dispersed throughout historical records, the most ancient probably consisting of tablets from the library of the Assyrian King Ashurbanipal from the 7th century BC (which may in turn be derived from earlier texts dating back as early as 1500 to 2500 BC) [9].

Figure 2. Pathophysiology of leishmaniasis

Typical pathological manifestations of *Leishmania* spp infection: visceral (left), cutaneous (middle), and mucocutaneous (right). Adapted from [10].

1.1.2. Life cycle

Not before 1941 will the mode of transmission properly be identified as the bite of the hematophagous female *Phlebotomus* sand flies (male sand flies typically feed on plant nectar) [11]. These insects constitute the vector of the disease by transmitting the parasitic protozoa from one animal host to another, thus perpetuating its life cycle (**Figure 3**). Contaminated sand flies will inadvertently inject parasites into the tissue and/or bloodstream of the host when feeding. The extraneous protozoa are subsequently ingested by phagocytic leukocytes in which they will evade the host's immune response for an uncertain time period.

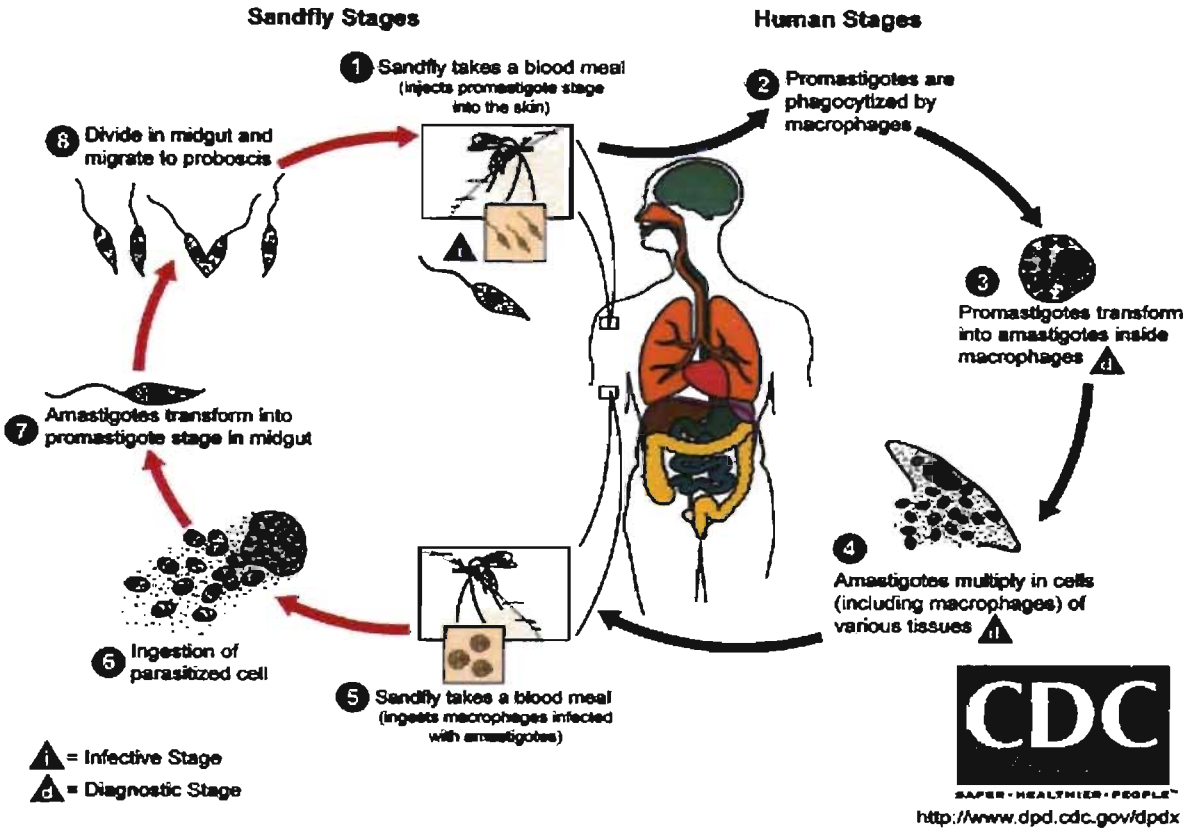

Figure 3. Life cycle of *Leishmania* sp.

Adapted from [10].

At this point, *Leishmania* cells experience acidic stress within the phagolysosome in addition to the temperature stress consequent to the change of carrier (ambient temperature to body temperature). The typically toxic environment inside the phagolysosome will prompt massive restructuring of the parasite's molecular makeup, ensuing in its differentiation into what is termed the amastigote stage. Remarkably, not only will the amastigote cells survive in such harsh environmental conditions, they will eventually multiply to a point where the enveloping cell will be breached, disseminating the pathogenic cells throughout the surrounding tissue.

When a roaming phlebotome ingests blood from an infected animal, any amastigotes present in the meal will respond to the new environment by differentiating back into the flagellated promastigote form and migrating to the sand fly's salivary glands. Hence, the cycle is complete.

1.1.3. Treatments and emerging drug resistance

There is hope for a vaccine against *Leishmania* based on the observed fact that successfully cured initial infections generally lead to protection against further infection [12]. However, current treatments are derived from chemotherapy and resistance to many drugs is approaching pandemic proportions [13]. The high-cost and toxicity of such drugs has propelled vaccine development for particular strains in humans and animals [14,15]. Hope for discovering novel drug and vaccine targets stems from the recently fully sequenced genomes of three *Leishmania* species.

1.2. On the edge of eukaryota

Given that an abundance of kinetoplastids are parasitic, including all trypanosomatids, these species are constantly subject to strong selection pressures in order to evade host resistance mechanisms. Evidently, parasitized host species must counter-adapt to the invader's exploitation by increasing the speed or efficiency of pathogen recognition and eradication. The entailing mutual antagonism encourages coevolution of both species; one of the many facets of exploiter-victim dynamics (reviewed in [16]). In spite of this, single-celled parasites have a slight edge in the ensuing arms race which arises from their relatively short generation time. Considering these facts, it comes as no surprise that trypanosomatids exhibit faster evolutionary rates than non-parasitic or asymbiotic species.

Leishmania and other members of the *kinetoplastea* order exhibit structural and biochemical peculiarities that alienate them from other eukaryotes (detailed in section 1.3). Phylogenetic analyses postulate that trypanosomatids and related species branch out near the base of the evolutionary tree of eukaryotes [17-19]. These declarations are derived from the consensus of multiple phylogenies, yet remain speculative in nature. The faster evolutionary rate of trypanosomatids may bias the construction of molecular phylogenies [20,21]. In fact, rooting the eukaryotic tree implies overcoming data sampling and methodological intricacies which may never truly elucidate the veritable branching point of ancestral eukaryotes [22,23]. Kinetoplastids are nonetheless generally considered to be an

ancient class of protozoa, an affirmation substantiated by organelles and molecular properties remnant of prokaryotic ancestors.

1.3. A peek at trypanosomatid biology

The most distinctive feature of trypanosomatids is the kinetoplast organelle; a dense mass of intermingled circular DNA contained within the single mitochondria of all *Kinetoplastidae* [24,25]. *Trypanosoma brucei* presents the only proven instance of genetic exchange in trypanosomatids. The species is also the only known eukaryote that inherits kinetoplast DNA (kDNA) from both parents via the poorly understood recombination of maxicircle kDNA [26]. Although the diploid nuclear genome of most trypanosomatids is reason for the existence of sexual reproduction [27], clonal procreation is seemingly the regular form of proliferation [28]. *Leishmania* reproduces by binary fission.

Trypanosomatids are model organisms for discovering and studying novel molecular, biochemical, and cellular mechanisms given their evolutionary standing. Many distinctive characteristics are shared between *Leishmania* and *Trypanosoma*, which differ mainly in the size of their average inter-CDS length at the genomic level [29]. The following subsections describe biological mechanisms of *Leishmania* and of most trypanosomatids that are pertinent to this work.

1.3.1. Directional gene clusters

The foremost genomic characteristic that alludes to *Leishmania*'s evolutionary seclusion is the tandem arrangement of genes (see chromosome maps at GeneDB website [30]), much like that of bacterial operons [31-33]. The *L. major* genome contains 133 clusters of tens to hundreds of coding sequences on the same strand of DNA, also known as DGCs, distributed throughout 36 chromosomes [34]. Resemblance to prokaryotes is striking when considering the virtual absence of introns. In fact, only four cases of genes containing introns have been reported to date in *Leishmania major* [34,35]. Albeit such resemblance, neighbouring genes generally display incongruent expression profiles in *Leishmania*.

1.3.2. Lack of transcriptional control

As a general conception, positive regulation of gene expression occurs mainly at the level of transcription initiation in higher eukaryotes (reviewed in [36]). *Leishmania* and other kinetoplastids are deviants of eukaryotic transcription with regards to this standard, as transcriptional control of protein-coding genes appears aberrant if not altogether absent [37,38]. *Leishmania* has homologues of all three RNA polymerase core sub-units, yet transcriptional activators, co-activators, basal transcription factors, and other polymerase components cannot be easily identified [34,39]. So far, no conserved promoters have been elucidated for trypanosomatid RNA polymerase II, which is accountable for synthesising mRNA [40,41]. Moreover, some experiments have demonstrated that transcription initiation is possible in the absence of promoters [38,42,43].

Transcription initiation, however, is not a random process in *Leishmania* protists. It has been shown that transcription initiation has a certain affinity for divergent strand switch regions; the 0.9- to 14-kb non-coding regions preceding opposite strand DGCs [32,44,45]. These locations display skewed sequence composition which may play a functional role in transcription initiation [46]. It has also been shown that convergent strand switch regions may be involved in transcriptional termination [44]. The RNA polymerase II complex transcribes DNA bidirectionally from these regions, generating long polycistronic RNA transcripts which can attain half a chromosome in length [31,32]. Individual mRNA transcripts are subsequently excised from the precursor RNA via *trans*-splicing (reviewed in section 1.3.3 and **Figure 4**). Although there is evidence for unspecific antisense transcription in *Leishmania* [38], nuclear run-on studies have demonstrated that transcriptional orientation is controlled by termination (i.e., RNA polymerase II aborts transcription promptly in anti-sense orientation) [47].

In light of these observations, it comes as no surprise that regulation of gene expression occurs largely at the post-transcriptional level. The discovery of many RNA-binding domains [34,48], regulated processing of cytoplasmic RNAs [49], and conserved regions in 3'UTRs [50-54] are hard evidence that corroborates this statement. Several studies have shown that sequences within 3'UTRs regulate differential expression of the

upstream gene mainly by modulating mRNA stability [53-58] or translational efficiency [50,59-62], although other mechanisms may exist (reviewed in [51,58]). It is noteworthy to mention that RNAi machinery has been identified and is apparently functional in *Leishmania braziliensis* [1,51]. Conversely, RNAi is believed to be absent in other *Leishmania* species where most antisense approaches to reverse genetics seem to work poorly [63].

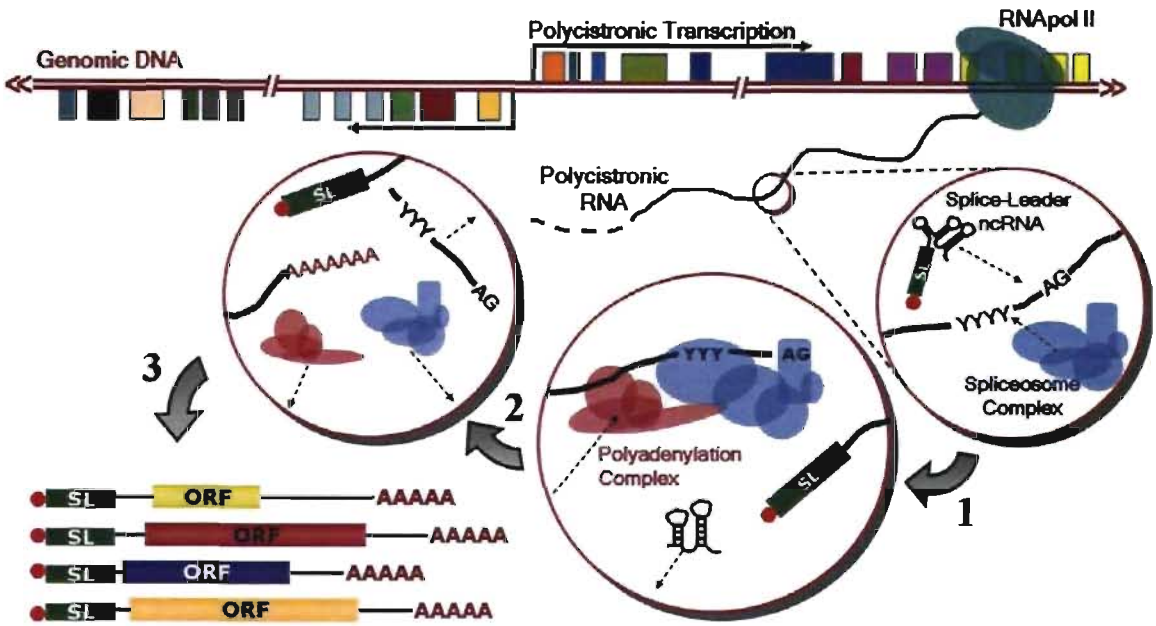

Figure 4. Overview of mRNA processing in *kinetoplastidae*

Following polycistronic transcription possibly mediated by strand-switch regions, (**1**) the intergenic region is targeted by the spliceosome complex, which binds to a poly-pyrimidine tract (YYY) and catalyses the ligation of a ncRNA downstream of an AG dinucleotide. (**2**) Once the ~39 nt splice leader is added, the hypothetical polyadenylation complex cleaves somewhere upstream and adds the poly(A) tail in a spliceosome-dependent manner. (**3**) The complexes free the trans-spliced mRNA, the intergenic spacer is released for possible degradation, and mRNA is exported to the cytoplasm.

1.3.3. *Trans*-splicing of mRNA transcripts

Trans-splicing consists of the joining of exons from two distinct RNA transcripts in order to yield a chimeric transcript. The process was first discovered in *Trypanosoma brucei*, the causative agent of African trypanosomiasis (a.k.a. sleeping sickness) [64]. When studying the mRNA transcripts of variant surface glycoproteins, the same 39 nt conserved

sequence was common to all 5' extremities of mRNA sequences. Subsequent discoveries showed that the conserved 5' sequence was present in all trypanosomatid mRNAs and that this leader sequence mapped to tandem clusters on a single chromosome [65,66]. Only when a Y-branch intermediate molecule was identified during pre-mRNA maturation experiments was *trans*-splicing properly understood [67,68]. Since this discovery, *trans*-splicing has been reported in many other species and is believed to carry out various biological functions, such as RNA processing irregularities, gene expression regulation, and generation of diversity (reviewed in [69]).

The 39 nt miniexon originates from a small, highly abundant ncRNA referred to as the SL RNA. These 135-147 nucleotides (nt) long structured RNAs are transcribed by RNA polymerase II and withhold a hyper-modified cap structure referred to as cap-4 since four nucleotides after the 7-methylguanosine are methylated [70]. SL RNAs in *Leishmania* are characterized by three stem-loop structures (**Figure 5**). The first contains the mini-exon which is bounded by the canonical GU splice donor site. This region may be implicated in SL-specific methyltransferase binding, as mutations in *Leptomonas seymouri* inhibited *trans*-splicing and proper cap formation [71]. The other two stem-loops are part of the intronic sequence and bear structural similarity to the Sm-binding sequence of U-rich snRNAs, although they lack sequence conservation [72,73]. The Sm-binding sequence of snRNAs is recognized by core proteins of the spliceosome complex, namely Sm proteins (reviewed in [73]). The catalytic process of *trans*-splicing is much similar to *cis*-splicing, in which two consecutive transesterification reactions are catalyzed by a large ribonucleoprotein complex called the spliceosome [74]. The main difference between *trans*- and *cis*-splicing resides in the formation of a Y-branch instead of a loop intermediate resulting from the fusion of the exogenous RNA transcript (**Figure 5**).

Several studies have demonstrated that polyadenylation is essentially linked to *trans*-splicing of the downstream gene [75-81]. When *trans*-splicing is inhibited, 3' end formation and polyadenylation are not observed [76]. Furthermore, a 'scanning' model has been proposed for 3' end formation and polyadenylation based on the discovery that the polyadenylation site moves in tandem with the *trans*-splicing position [75]. Together with the observation that polypyrimidine tracts are crucial for targeting both processes [77-82],

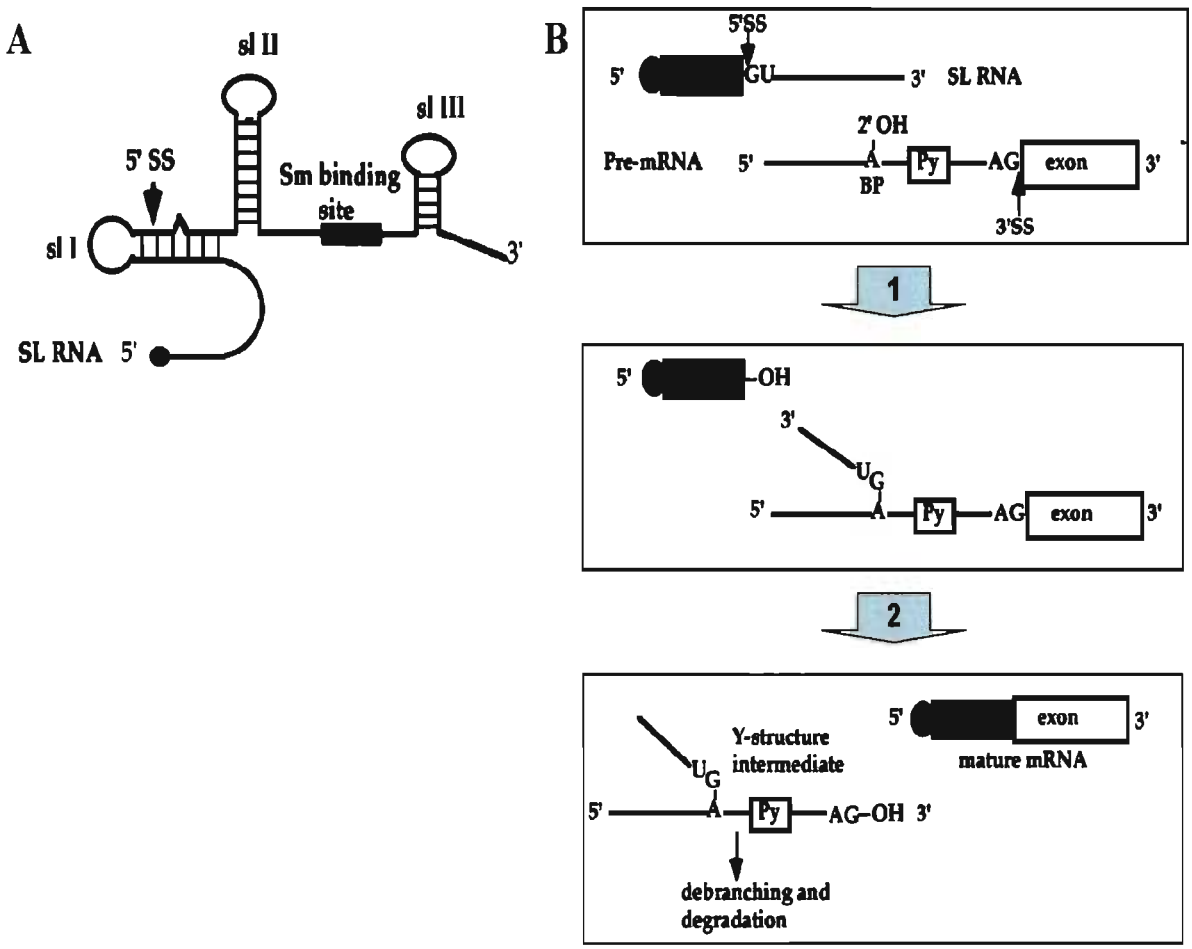

Figure 5. Structure of the SL RNA and *trans*-splicing in *Kinetoplastidae*

(**A**) Consensus secondary structure of the SL RNA in kinetoplastids. The approximate position of the GU 5' splice site (5' SS) is identified by the arrow in the first stem-loop (sl I). (**B**) Schematic, 2-step illustration of *trans*-splicing. The first step (**1**) shows branch-point formation by spliceosomal transesterification of the 5' SS with the hydroxyl group on the 2' carbon of the ribose backbone, commonly an adenosine. The second step (**2**) illustrates the transesterification reaction that swaps the 3' hydroxyl from the splice leader (SL) to the 3' end of the Y-structure intermediate, thus connecting the SL to the exon. The Y-structure intermediate will subsequently be cleaved before polyadenylation. Adapted from [73].

These findings suggest that the spliceosome complex interacts with the polyadenylation machinery in trypanosomatids. Although little is known about the molecular perpetrators of this association, it has been shown that the U1 snRNP is crucial to the coupling of *cis*-splicing and polyadenylation in higher eukaryotes [83], hence suggesting that a similar process may exist in *Leishmania*.

Contrasting with the AAUAAA cytoplasmic polyadenylation sequence in higher eukaryotes, trypanosomatids have no known consensus motif that drives polyadenylation downstream of coding regions [34]. Polyadenylation seems to occur within a specific range of the *trans*-splicing site [75], however there are reports that distant polypyrimidine tracts can drive polyadenylation further upstream [79,82].

2. REPETITIVE GENETIC SEQUENCES

DNA repeats (homologous DNA fragments that are present in multiple copies throughout the genome) have been subject to innumerable studies over the years. They were discovered 40 years ago via reassociation kinetics experiments and were immediately classified into two categories: 'highly' and 'middle' repetitive sequences [84]. These two groups bear analogy to the current classification of such elements as either tandemly repeated DNA or interspersed repeats. For sake of clarity,

2.1. Tandemly repeated DNA

Tandemly repeated sequences, commonly referred to as satellites, are relatively short and practically identical contiguous patterns of DNA ranging in size from 2. The term satellite originates from the behaviour of DNA in caesium chloride density gradients. The distinct nucleotide composition of short repeats produces a secondary or 'satellite' band when separated from genomic DNA. Satellite DNA is usually confined to specific chromosome locations propagating by replicational slippage and recombination [85]. Minisatellites and microsatellites are relatively small repeats of 1-5 nt and 5-25 nt respectively whereas macrosatellites are heavier repeats of 25 or more nucleotides. The classification of repeats gets taxonomically delicate when dealing with larger repeat units

which can also be replicated by gene conversion and transcriptional slippage [86], potentially duplicating regulatory regions and protein-coding segments as frequently observed in trypanosomatid genomes [1,29]. Tandemly repeated DNA is prone to slipped-strand mispairing, a phenomenon that occurs when complementary regions are mispaired during RNA replication. This process causes high variability in the total length of the tandem repeats among different individuals of the same species, making them supreme candidates for genetic profiling.

Repetitive DNA plays an important role in chromosome structure, namely by characterizing telomeric and centromeric regions. Telomeric regions consist of DNA tandem repeats located at chromosome extremities, which are essential for preventing loss of genetic information from incomplete DNA polymerase replication during the late S phase of mitosis. Interestingly, the reverse transcriptase (RT) domain of the telomerase enzyme encloses structural similarity to the RT of retroposons [87]. This implies that either the telomerase enzyme is a retroviral gene that has been domesticated by a eukaryotic ancestor, or that the retrovirus RT gene originated from a telomerase enzyme. Centromeres are important in cell division since they are responsible for proper chromosome pairing and kinetochore formation. The centromeres of higher eukaryotes are generally composed of large tracts of tandem repeats, although some exceptions in *S. cerevisiae* and parasitic protozoans indicate that centromeres are not necessarily satellites [88-90]. There is evidence that subtelomeric repeats may have centromeric properties in *Leishmania* [91,92].

2.2. Interspersed repeats

A second class of repetitive DNA encompasses a much broader and complex group of genetic elements known as interspersed repeats. In contrast to tandem repeats, the repeat units of interspersed repeats are separated by varying stretches of genomic sequence. Interspersed repeats are in most cases known as transposable elements (TEs), although not all such repeats are of mobile nature. TEs were first discovered by Barbara McClintock in 1950, three years before the molecular characterization of DNA by James D. Watson and Francis Crick [93,94]. Today, active TEs have been identified in almost every known species, with the possible exception of *Plasmodium falciparum* [87,95]. There are two great

categories of TEs: DNA transposons and retrotransposons. The latter is divided into two classes: long terminal repeat retrotransposons and non long tandem repeat (LTR) retrotransposons (or retroposons). All TEs can be either fully autonomous, relying on their own encoded machinery to transpose, or non-autonomous, relying on proteins hijacked from related autonomous TEs to carry out their transposition.

2.2.1. DNA transposons

DNA transposons are mobile genetic elements that can propagate themselves throughout the genome. In their simplest form, DNA transposons comprise three components: target site duplications at their extremities, coding sequence for the transposase enzyme, and inverted repeats between these components [96]. Once transcribed and translated into its protein form, the transposase binds to either a specific sequence or to unspecific sequences (depending on the type of transposase) and to the termini of the transposon DNA. It then makes a staggered cut at the target site producing 'sticky ends', cuts out the transposon and ligates it to the target site [97]. Genomic DNA polymerase and ligases complete the process by repairing the target sites. This cut and paste mechanism does not directly produce multiple transposon copies; duplication may occur during the S phase of the cell cycle when the original site has been duplicated but the target site has not.

There is a noticeable absence of DNA transposons from the sequenced nuclear genomes of five protozoan parasites (e.g., *Leishmania major, Trypanosoma brucei, Trypanosoma cruzi, Giardia Lamblia, and Plasmodium falciparum*) [90]. Although only five genomes were compared, it has been proposed that this DNA transposon deficiency may be a result of the tight control of cell membrane traffic in these species. Since mutations in transposon DNA directly affect their competence, frequent horizontal transfers to virgin genomes are important for transposons to maintain their function [98,99]. It is worth mentioning that there are additional varieties of DNA transposons in other eukaryotes which code for additional genes, such as DNA binding proteins. Some of these other DNA transposons include rolling-circle transposons named helitrons and complex, multi-gene encoding polintons (reviewed in [87]).

2.2.2. Retrotransposons

The majority of TEs in eukaryotic genomes transpose via an RNA intermediate. LTR retrotransposons are very similar to retroviral genomes in that they may encode three ORFs, including a reverse-transcriptase and/or integrase gene [100]. In general, upon transcription of the retrotransposon DNA from cellular polymerase II, the mRNA transcript is reverse-transcribed into cDNA that is subsequently inserted into the genome by an integrase [101]. These retrotransposons can form virus-like particles in which most of the transposition process takes place. No such retrotransposons have so far been identified in *Leishmania* species [1], although a class of LTR retroelements known as VIPER has been identified in some species of *Trypanosoma* [102,103].

2.2.3. Retroposons

Non-LTR TEs are commonly referred to as retroposons or long interspersed elements (LINE). They may contain one or two ORFs, generally RT and endonuclease (EN) genes and a RNA polymerase promoter region upstream [96]. The mode of replication of non-LTR retrotransposons is well studied and somewhat less intricate than that of LTR retrotransposons [104]. The EN protein cleaves a single strand of the target DNA that will be used as a primer for the direct reverse-transcription of the TE's mRNA into the chromosome. A second single-strand nick is performed by the EN nearby on the opposite strand in order for the cell's DNA polymerase II to synthesize the complementary strand, thus duplicating the genomic sequence between both nicks at each retroposon extremity. Target site duplication and the poly-A tail of the RNA intermediate, which is copied to the newly synthesised retroposon, are hallmark sequences of non-LTR retroelements.

Trypanosomatids abound with various retroposons. It is estimated that ~3% of the nuclear genomes of *Trypanosoma brucei* and *Trypanosoma cruzi* harbour TEs [105]. These active TEs generally display insertion site-specificity [106-108] and include SLACS/CZAR [108], *ingi*/RIME [109,110], and L1Tc/NARTc [111,112] retroposons. Two other potentially active retroposons have also been identified in *Leishmania braziliensis*: a SLACS/CZAR homolog and the telomere-associated transposable element [1]. *Leishmania major* and *Leishmania infantum* are believed to lack active retroposons, but harbour

remnants of active transposons [53,113]. One of which, the SIDER element, represents the most abundant TE presently characterized in trypanosomatids. The extinct SIDERs do not display apparent site-specificity for genomic integration, yet they are preponderantly distributed in the intergenic regions of DGCs [53]. **Figure 6** outlines the structural properties and similarities of various trypanosomatid retroposons.

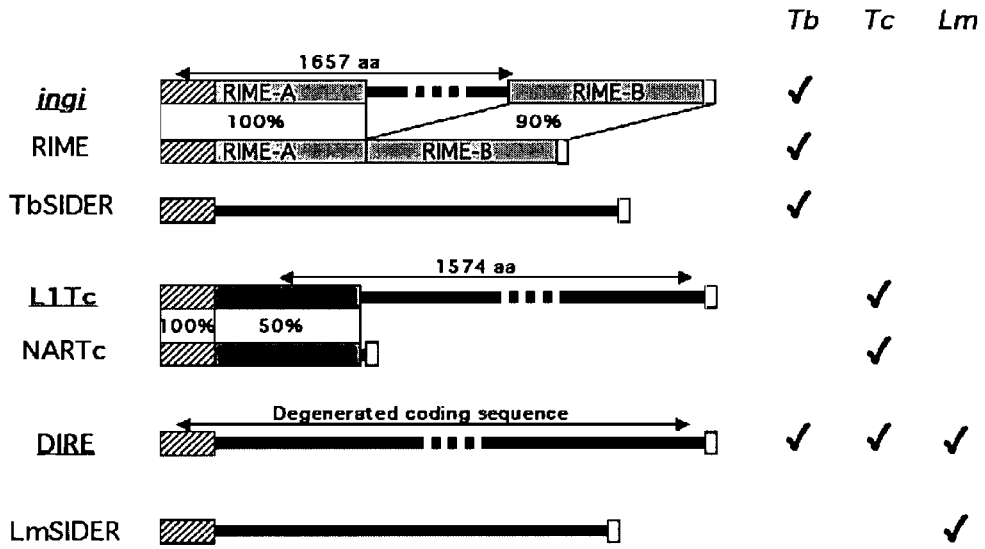

Figure 6. Trypanosomatid retroposons

Retroelement names underlined in bold represent autonomous elements, the others being non-autonomous truncated elements. Grey boxes represent conserved sequence identity; hatched boxes correspond to a 79 nt conserved 'signature' sequence whereas white boxes correspond to A-rich sequences remnant of a mRNA intermediate. Tb = *T. brucei*, Tc = *T. cruzi* and Lm = *L. major*. Modified from [53].

2.3. Evolutionary and regulatory considerations

The discovery of TEs and tandem repeats was somewhat ignored for many years due to the poor understanding of their features. They were initially called 'controlling elements' based upon their effect on phenotypic variation and were even speculated to modulate entire metabolic pathways by shuffling regulatory sequences [114,115]. However, little was known about their biological function prior to the era of DNA sequencing, and they were initially labelled as selfish DNA elements that survive by parasitizing genomes

[116,117]. It is now recognized that a significant portion of eukaryotic genomes are composed of TEs: In humans, 40% of the genome is composed of characterized mobile genetic elements, while coding sequences are estimated to encompass only 1,5% [118,119]. Similar proportions have been reported in maize [120] and at least 15% of the *Drosophiia* genome contains TEs [121]. It is estimated that ~3% of the trypanosome genome consists of TEs [105], notwithstanding that initial experiments presaged over 30% of the genome to contain tandem and/or interspersed repeats [122]. The degenerated SIDER elements of *Leishmania* cover ~3% of the genome alone, and their distribution strongly invokes functional assimilation [53]. The profusion of transposable elements in eukaryotes and increasing evidence that TEs supply novel regulatory and evolutionary material to genomes has altered the assumption that TEs are merely selfish genetic parasites [87,123].

The importance of TEs in forging eukaryotic genomes is highlighted by their contribution to gene regulation. In humans, there is strong evidence that thousands of fragments of mobile elements have undergone strong purifying selection in constrained non-exonic regions near genes involved in regulation of transcription and development [124]. It is evident that particular regions derived from interspersed repeats have been preferentially domesticated into *cis*-regulatory functions [124-126]. This evolutionary process seems to have also arisen in *Leishmania* species, where TEs are frequently located in 3'UTRs and regulate gene expression at the mRNA level [53]. Recent studies have also demonstrated that TEs conceal many transcription regulation signals in humans, such as transcription factor binding sites, that were not present in the original promoters of target genes [127]. Similar observations have been reported in *Trypanosoma cruzi* where the L1Tc (and the non-autonomous NARTc) retroposons drive transcription initiation via the mobile element's internal RNA polymerase II promoter [128]. Interestingly, the promoter region is contained within the 79 nt signature sequence of trypanosomatid retroposons (**Figure 6**), yet there are no similar reports in *Leishmania* as yet.

In addition to regulation of gene expression, repetitive DNA sequences are believed to perpetrate a variety of evolutionary phenomena, such as chromosome restructuring, somatic gene recombination, sexual determination, reproductive isolation, gene-silencing

mechanisms, and even host extinction through sex-ratio distortions (comprehensibly reviewed in [87,129]).

3. COMPUTATIONAL TOOLS FOR SEQUENCE ANALYSIS

The following sections cover some fundamental notions underlying the computational methods and data representations inherent to this work (chapter II in particular). Multiple sequence alignment (MSA) and hidden Markov models (HMMs) are of central importance to this work. The former is used to emphasize the similarities and variations of related genomic sequences (in this case, the *Leishmania sp.* SIDER elements), while the latter are employed for representing sequence data in MSAs and genomic search queries. In addition, this section presents a summary of various RNA structure prediction tools even though they are not implicitly incorporated into this thesis. Their description is included since RNA structure predictions is an ideal objective and a logical succeeding of the work presented in the following sections. These tools are presented in a straightforward manner given the generally arduous mathematical, statistical, and algorithmic concepts they embody. An overview of typical computational methods for prediction of mRNA boundaries in trypanosomatids is discussed in Chapter I.

3.1. Multiple sequence alignment

Arguably the most important tool for studying evolutionary relationships is the alignment of biological sequences. Certain genes can be conserved among divergent species, performing identical functions or acquiring novel characteristics in accordance with natural selection. The arrival of protein and nucleic acid sequencing enabled biologists to peer into the structure and composition of the molecules that carry out the fundamental processes of life. By aligning a group of related genes or sequences, the patterns of change can be analyzed to gain structural, functional, and evolutionary information on their nature.

On the whole, multiple sequence alignment algorithms attempt to optimise the amount of similar characters (usually nucleic or amino acid) in the same column of an alignment. Sequence conservation directly influences the efficiency of multiple alignment

algorithms. Highly similar sequences can be easily aligned without computational assistance, whereas sequences displaying profuse variation consume much more resources. Indeed, poorly conserved sequences require comparing massive amounts of combinations in order to provide an optimal alignment.

Since the biological function of RNA and proteins depends on three dimensions of structure, a perfect MSA should take such structural considerations into account for these molecules (discussed in section 3.3.2.). Also, the evolutionary relationship between sequences should be considered in such a program. Most MSA programs, however, only consider primary structure.

3.1.1. Dynamic programming

A global optimum can be obtained using multi-dimensional dynamic programming, however such algorithms are impractical for more than a few sequences as their complexity is exponential [130]. The *MSA* program implements such an algorithm which can be run on ~7 protein (or nucleic acid) sequences less than 300 aa in length in reasonable time [131].

3.1.2. Progressive alignment

Progressive alignment is probably the most popular MSA tool used by biologists, chiefly thanks to the CLUSTALW program [132]. These algorithms function by first pairing two sequences using pairwise alignment strategies. A third sequence, usually the most similar to the pair, is added to the fixed alignment represented as a nucleotide frequency matrix, otherwise known as a sequence profile [133]. This procedure continues until all sequences are aligned. Progressive alignment strategies are heuristic; they produce an approximation of the optimal alignment, significantly speeding up the process as a consequence. In spite of this, the main caveat of these algorithms is that sub-alignments are fixed and cannot be modified as additional information is incorporated into the alignment. For the abovementioned reasons, progressive strategies generate rather poor alignments when sequence conservation is poor [134].

3.1.3. Iterative refinement

To circumvent the limitations of progressive alignment strategies, iterative refinement strategies can additionally optimise sequence alignments. Once an initial alignment is obtained, typically from progressive methods, individual or groups of sequences are removed from the alignment and then re-aligned. This tactic is guaranteed to converge towards a local optimum because sequence space is finite [135]. A popular and accurate iterative refinement alignment program is *MAFFT* [136,137].

3.1.4. Probabilistic models

In essence, probabilistic models are a thoughtful manner of representing sequence data. A multiple sequence alignment can be represented as a bi-dimensional matrix of character frequencies which, in turn, can be transformed into a hidden Markov model profile (see section 3.2. below). The same can be done from unaligned sequences by applying a variety of algorithms developed specifically for HMMs [138]. The simulated annealing algorithm is worth mentioning since it provides faster execution times than the more rigorous Baum-Welch expectation maximisation (EM) algorithm for the global optimisation problem. Simulated annealing finds random 'nearby' solutions to an alignment with a probability that decreases proportionately to a temperature factor, consequently ensuring that the system moves towards a global optimum [139]. The alignment produced from the simulated annealing heuristic can subsequently be refined via the use of Baum-Welch EM [140].

3.2. Hidden Markov model profiles

Multiple sequence alignments, e.g. a gene family, contain a wealth of comparative information: position-specific conservation, exon/intron lengths, consensus structure, base composition bias, codon usage, etc. The heterogeneous nature of such data requires comprehensive modeling in order to perform successful computational biology analyses. This can be achieved with hidden Markov models: statistical representations that provide a conceptual toolkit for building complex models from an intuitive outline of a process [141].

HMMs consider all possible combinations of matches, mismatches and gaps in order to statistically characterize a multiple sequence alignment. Since they are fully probabilistic models, Bayesian probability theory can be used to manipulate their parameters and scores for a variety of analytical purposes. These properties justify why HMMs are integrated intoq several bioinformatics programs today.

HMMs are all based on a Markov process, in which the probability distribution of the current state is conditionally independent of the path of past states [142]. In a HMM, the state is not directly visible but the variables that influence the state are. In a MSA HMM profile, the different states are either 'match', 'insert' or 'deletion' (**Figure 7**). Each state has an emission probability for every possible outcome (i.e. a particular nucleotide residue). Also, each state has a transition probability to every following state in the model. The model allows for any combination of states (i.e. residues) to be generated and a particular score will be assigned for each combination (generally a bit-score).

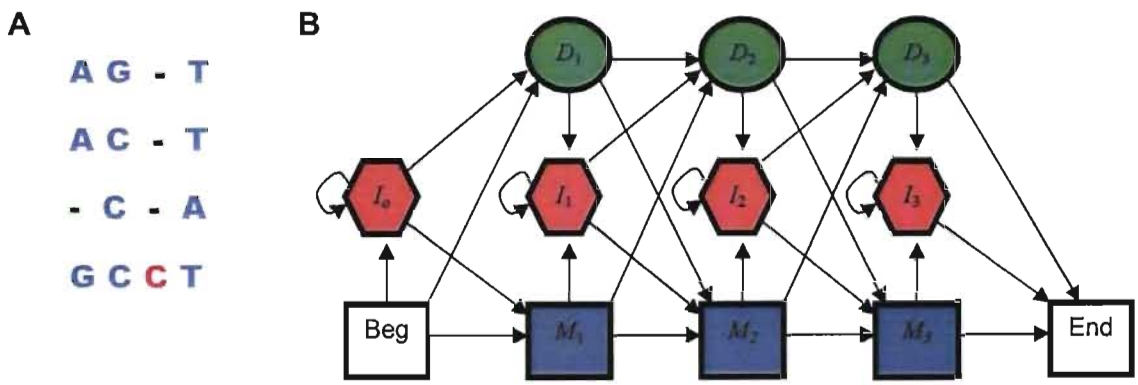

Figure 7. Relationship between sequence alignment and hidden Markov models

(**A**) A sample alignment of DNA nucleotides. (**B**) HMM profile model for MSA. Arrows indicate transition probabilities, and coloured boxes indicate states (M = match, I = insertion, D = deletion). Each state has its proper emission probability distributions, which determine the nature of the residue at a certain position. Interpreted from [143].

Besides their statistical features, HMMs are ideal computational models for biological sequences given their flexibility. Additional states, emission probabilities, and transition probabilities can be incorporated into a model, provided that the respective probabilities sum to one. However, since they are based on Markov chains, distant

correlations among non-consecutive residues are ignored. For example, HMMs are inappropriate for RNA secondary structure modelling since distant nucleotide pairings cannot be incorporated into the model [141].

3.3. RNA secondary structure prediction

In molecular biology, RNA has long been seen as a simple intermediate between the hereditary information encoded in DNA and protein molecules that carry out the structural, regulatory and catalytic functions in a cell. This perception has long been known as the central dogma of molecular biology [144]. However, the discovery of self-splicing RNA led to the recognition of the catalytic capabilities of RNA molecules [145], and so opened the door to innovative hypotheses on the origin of life [146]. There is now strong evidence that RNA is much more than a simple intermediate in the processing of genomic code, which has fuelled recent interest for non-coding RNA discovery (reviewed in [147,148]).

The flexible ribose backbone and single-stranded nature of RNA cause its bases to form hydrogen bonds with other neighbouring bases. Such base-pairing instigates tertiary structure formation founded on secondary structure scaffolds. In addition to the Watson-Crick and 'wobble' base-pairings (A=U, G≡C ; G≈U respectively), X-ray crystallography studies have identified many non-canonical structures in RNA molecules (reviewed in [149,150]). These distinctive structural motifs, tertiary interactions, and *trans*-acting molecules make precise 3D modeling unfeasible or at least extremely delicate.

Another interesting property of ncRNAs is that they are not subject to the same evolutionary constraints as protein-coding genes. There are lots of characterized ncRNA molecules that, in some instances, bear little or no sequence identity among each other, yet share matching structures (tRNAs for example [151]). Drastic changes in sequence are often tolerated provided that compensatory mutations maintain base-pairing complementarity; an occurrence termed covariation. The inherent mutational flexibility of ncRNA allows for faster evolutionary rates than proteins, which complicates multiple alignment and prediction analyses. For instance, standard sequence-based alignment strategies completely ignore structural information content, hence proving to be rather futile

for some ncRNA sequences [152]. By considering positional covariation and structural predictions, certain computational tools attempt to surmount these shortcomings.

3.3.1. Structural predictions from single sequences

Since its origins in 1980 [153], secondary structure prediction has evolved from maximising base-pairing in a given sequence via dynamic programming, to incorporation of energy rules and probabilistic calculations [154]. The *mFOLD* program predicts 2D structures by combining base-pair complementarity and experimentally acquired free-energy values for precise RNA structures [155]. Complementary regions are evaluated by a dynamic programming algorithm to predict the most thermodynamically stable conformation. For example, stacked bases which form a helix structure will contribute to lower the free-energy of a molecule, whereas a destabilizing loop will contribute to raise the free-energy score [156,157]. Although this approach can emit sub-optimal structures, it does not compute all the possible structures within a given energy range [154].

A different approach to predicting ncRNA structures is to consider the probability that each base-paired region will form based on principles of thermodynamics and statistical mechanics. Using the Boltzmann distribution to calculate the likelihood of all possible structures, the McCaskill algorithm can predict the most probable structure in addition to intermediate structures, base-pair opening and slippage, and temperature dependencies [158]. This methodology is implemented in the Vienna RNA software package [159,160] and can be used to create graphical representations of RNA structures and energy dot-plots. As exemplified in **Figure 8**, these popular programs do not always produce correct structures. This drawback is more important when the amount of queried sequence increases, as this lowers the probability that the correct structure is predicted. One way to surmount this problem is to consider a comparative approach in order to maximise the information content of the queried sequences.

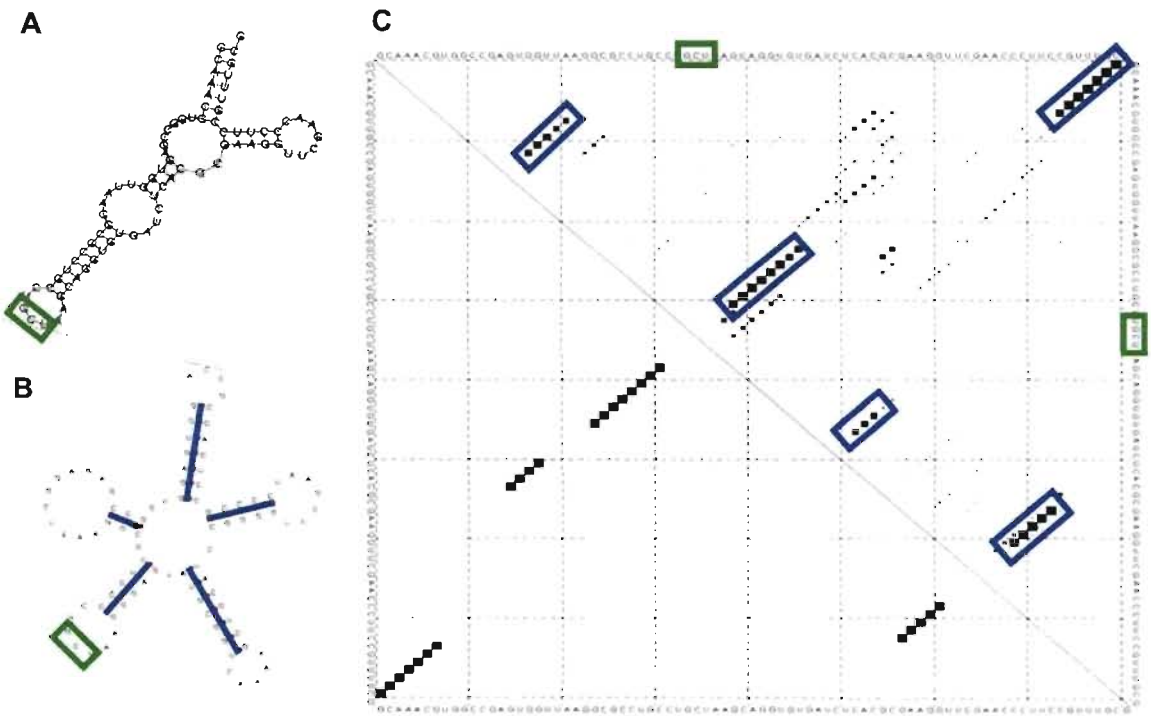

Figure 8. Graphical representation of RNA structural predictions

(**A**) Predicted secondary structure of a *Leishmania major* serine-tRNA using the RNAfold program from the Vienna RNA package [159]. The highlighted green box marks the GCU anticodon. (**B**) 'True' biological structure of an *E. coli* serine-tRNA obtained from the genomic tRNA database [151]. The 5-stem serine-tRNA structure is highly conserved throughout the tree of life [161, 162]. Stems are marked as blue lines. (**C**) Dot-plot illustrating most probable structure (bottom half) and all sub-optimal structure probabilities as predicted by the McCaskill algorithm in the RNAfold program (top half) [160]. The bottom half corresponds to the structure in A. The size of the dots represents the probability of a specific base-pairing between two positions (the query sequence borders the graph). In this example, the optimal structure is not the true structure, yet the valid structure can be distinguished among the sub-optimal pairing probabilities (highlighted blue boxes correspond to potential helices). The anticodon position is highlighted in green along the bordering sequence of the plot.

3.3.2. Conserved structure predictions

Another approach to predicting RNA structures is based on covariation analysis. Consider a group of related ncRNA sequences believed to share similar function and structure (i.e. a ncRNA gene family or *cis*-regulatory regions in a mRNA UTR). These sequences should therefore be subject to similar selection pressures. Covariation models inspect aligned ncRNA sequences for positions that vary together in order to form base-pairs in all, or most, of the sequences. This targeting can aid discrimination between structural (e.g. helix) and non-structural (e.g. single-stranded) positions in an alignment. There are several programs available for performing such analyses that use different strategies and data representations [163-174]. A relatively simple yet effective method for identifying a consensus secondary structure is the *RNAalifold* program [164,170], which integrates phylogenetic and thermodynamic information into a modified energy model. In a nutshell, it calculates a consensus sequence for a given alignment and folds the consensus using the unique energy model.

The majority of these tools emit predictions that strongly depend on the input alignments. Given that standard multiple alignment procedures focus on optimising sequence similarity, one can easily envision that ncRNA alignments may be distorted. This predicament is what motivates the use of structural alignment tools for identifying common ncRNA features. Recent implementations of the Sankoff algorithm for the simultaneous alignment of sequence and structure have proven to be quite successful at aligning poorly conserved RNA sequences [175-179]. *LocaRNA* [179] and *FOLDALIGNM* [178] are two very recent programs that align RNA sequences based on their structural properties by using the rich information content embedded in McCaskill probability matrices (c.f. **Figure 8**). There also exist structural alignment programs that exploit graph theory to align ncRNA sequences represented as trees [180,181].

4. OBJECTIVES

It is apparent that a large family of degenerated retroposons is widespread throughout the genome of *Leishmania* species, with a notable preponderance in 3'UTRs that are known to regulate gene expression [53]. These SIDER elements, subdivided into the SIDER1 and SIDER2 groups, have only been partially characterized in *Leishmania major* and seem to have an effect on either mRNA stability or translation [50,52,53].

The prime objective of this work is to gain additional knowledge on the features of these interspersed repeats in order to elucidate their role in the developmental regulation of gene expression. To do so, computational tools will be applied to investigate sequence conservation and genomic distribution in the three *Leishmania* nuclear genomes sequenced to date. This will require constructing reliable SIDER alignment profiles for intra- and inter-species genomic scanning.

Predicting structural motifs should facilitate the identification of underlying biological mechanisms. Given the abundance of SIDERs in the genome, discerning between potentially functional and non-functional sequences is fundamental for this purpose. To improve structural predictions, it is thus necessary to develop and/or refine tools that will improve the prediction of mRNA extremities with marked emphasis on the 3'UTR. This constitutes the second objective of this study.

CHAPTER I

1. ARTICLE PRESENTATION

A method for predicting *trans*-splicing and polyadenylation sites from genomic sequences was required as an initial phase for the proper characterisation of conserved elements in 3'UTRs. However, existing methods are not well suited for 3'UTR prediction in *Leishmania*. A scrutiny of public sequence libraries was performed in order to reveal insightful genomic characteristics that can be incorporated into refined scoring models. Section 2 details the findings and describes how they can contibute to improving the prediction of mRNA extremities in the context of rapid, large-scale genomic annotation. The study also introduces a cut down 5'UTR predictor with similar accuracy to a published method.

The manuscript was submitted for publication in the first week of November 2007 to the journal *BMC Bioinformatics*. Author contributions are exposed on page 42.

2. IMPROVING THE PREDICTION OF MRNA EXTREMITIES IN THE PARASITIC PROTOZOAN LEISHMANIA

(as submitted to BMC Bioinformatics, February 2008)

Martin Smith[1,2], Mathieu Blanchette[2], and Barbara Papadopoulou[1#]

1 – Research Centre in Infectious Diseases, CHUL Research Centre
2705 Laurier Blvd.
Quebec, QC
Canada G1V 4G2

2- McGill Center for Bioinformatics
3775 University St.
Montreal, QC H3A 2B4
Canada

#Corresponding author: Barbara Papadopoulou

Phone: (418)-654 2705

Fax: (418)-654 2715

Email: [redacted]

E-mail addresses:
MS [redacted]

MB [redacted]

Abstract

Background: *Leishmania* and other members of the *Trypanosomatidae* family diverged early on in eukaryotic evolution and consequently display unique cellular properties. Their apparent lack of transcriptional regulation is compensated by complex post-transcriptional control mechanisms, including the processing of polycistronic transcripts by means of coupled *trans*-splicing and polyadenylation. *Trans*-splicing signals are often U-rich polypyrimidine (poly(Y)) tracts, which precede AG splice acceptor sites. However, as opposed to higher eukaryotes there is no consensus polyadenylation signal in trypanosomatid mRNAs.

Results: We refine a previously reported method to target 5' splice junctions by incorporating the pyrimidine content of query sequences into a scoring function. We also investigate a novel approach for predicting polyadenylation (poly(A)) sites *in-silico*, by comparing query sequences to polyadenylated expressed sequence tags (ESTs) using position-specific scanning matrices (PSSMs). An additional analysis of the distribution of putative splice junction to poly(A) distances helped to increase prediction rates by limiting the scanning range. These methods were able to simplify splice junction prediction without loss of precision and to increase polyadenylation site prediction from 22% to 47% within 100 nucleotides.

Conclusion: We propose a simplified *trans*-splicing prediction tool and a novel poly(A) prediction tool based on comparative sequence analysis. We discuss the impact of certain regions surrounding the poly(A) sites on prediction rates and contemplate correlating biological mechanisms. This work aims to sharpen the identification of potentially functional untranslated regions (UTRs) in a large-scale, comparative genomics framework.

Background

Leishmania is a unicellular eukaryote that belongs to the *Trypanosomatidae* family; a strictly parasitic order of *kinetoplastida*. *Leishmania* is the causative agent of leishmaniases, vector-borne parasitic diseases with a large spectrum of clinical manifestations in humans ranging from self-resolving skin lesions to life-threatening visceral diseases [1]. Leishmaniasis is endemic in 88 countries mainly in tropical and subtropical regions with an estimated 12 million people presently infected worldwide and at least 350 million people being at risk of infection [2].

Trypanosomatid protozoan parasites have diverged early on in eukaryotic evolution [3]. Their evolutionary closeness to bacterial ancestors is delineated by intrinsic cellular characteristics such as tandem arranged genes [4], polycistronic transcription [5, 6], mitochondrial RNA editing [7], lack of transcriptional control [8], infrequent introns [9], and *trans*-splicing [10]. The latter consists of the 5' cleavage of polycistronic RNA precursors into individual mRNA transcripts by addition of an exogenous 39 to 41 base long capped RNA fragment, namely the splice leader (SL) or mini-exon, provided by a highly abundant SL-RNA [11], yet similar processes have also been discovered in nematodes and even in mammals [12, 13]. This process is somewhat similar to *cis*-splicing in other organisms, as RNA is cleaved at an AG dinucleotide downstream of a polypyrimidine stretch.

In addition to co-transcriptional *trans*-splicing, polyadenylation of the upstream transcript is also required in order to generate monocistronic mRNAs in these organisms. Trypanosomatid protozoa are believed to lack a conserved polyadenylation (poly(A)) signal, in contrast to other eukaryotes who generally require a cytoplasmic polyadenylation motif for successful polyadenylation [14]. Several studies support that polyadenylation is mechanistically coupled to *trans*-splicing and that it depends upon the presence of polypyrimidine tracts [15-19], thus leading to the belief that the spliceosome complex interacts with the polyadenylation machinery in trypanosomatids. It has also been reported that distant pyrimidine tracts may be responsible for polyadenylated positions further away

from the downstream 5' splice site in trypanosomes [17, 20]. These analyses also convey the non-specific nature of poly(A) site selection in trypanosomatids, as polyadenylation seems to occur in a given region rather than at a specific position.

The apparent heterogeneity of kinetoplastid mRNA polyadenylation and its dependence on successful *trans*-splicing make 3'-untranslated region (3'UTR) length predictions troublesome. Currently, there exists a 3'UTR prediction method for *Trypanosoma brucei* derived from the statistical analysis of mRNA transcript extremity lengths from expressed sequence tag (EST) data [20]. The prediction is essentially obtained by selecting the position located at an empirical distance (100 bases) upstream of the polypyrimidine tract closest to the open reading frame (ORF). The authors claim a 38% prediction rate within a 73-nucleotide window. These metrics are somewhat inappropriate for predictions in the *Leishmania* genus since the species flaunt larger intergenic (IR) sequences, higher average UTR lengths, and less stringent splice acceptors [4, 21].

In addition to the statistical analysis of transcript length distributions for 3'UTRs, 5'UTR prediction has been submitted to supplementary investigation [22-24]. Prediction algorithms that essentially focus on selecting the first AG dinucleotide after the longest polypyrimidine stretch can reportedly identify 62% of valid splice junctions in trypanosomes and 51% in *Leishmania* [20, 23]. For *Leishmania*, it has been shown that by fragmenting the non-coding sequence upstream of a start codon at every occurrence of AG, the AG following the longest fragment corresponds to a valid splice junction in 60% of the cases. When combining this method with a linear discriminant analysis of dinucleotide composition, the later method can obtain a prediction accuracy as high as 92% on selected high-scoring sequences [23].

Considering that regulation of gene expression in kinetoplastids occurs mostly at the post-transcriptional level, it has become apparent that UTRs bear essential regulatory tags [8, 25-29]. From the standpoint of computational motif discovery, it is imperative to discriminate between functional and non-functional sequences in order to successfully identify novel conserved regulatory regions. This premise is most important when dealing with non-coding RNA as it is exposed to less stringent evolutionary pressure than open

reading frames [30]. It can be expected that limiting sequence and structure motif searches to legitimate mRNA, UTRs will generate more informative results while reducing the inherent computational cost of search algorithms.

This paper aims to further improve the *in-silico* prediction of mRNA extremities in kinetoplastid organisms. We polish *trans*-splicing prediction in *Leishmania* by incorporating the pyrimidine content of intergenic regions into a previously developed scoring function, and propose a polyadenylation prediction method based on the global nucleotide composition observed in published expressed sequence tag (EST) data. The selection of different genomic regions surrounding the poly(A) site and their impact on prediction rates has validated the impact of adenosines and downstream polypyrimidines on trypanosomatid polyadenylation.

Results

Considering pyrimidine content increases splice-junction prediction accuracy

Previously, the best method to predict splice acceptor sites in trypanosomatids combined statistical analysis of dinucleotide composition with inter-AG fragment length assessment [23]. We simplified the procedure by discarding the statistical discrimination of inter-AG fragments based on dinucleotide composition, thus only considering the inter-AG fragment size for predictions. This approach was compared to two pyrimidine-bias scoring functions that rate inter-AG segments in proportion to their pyrimidine content in addition to their size (see Methods). Both functions are such that inter-AG fragments displaying lower than average pyrimidine content are proportionately penalised whereas those with higher than average content are rewarded.

Each scoring model's relative sensitivity with respect to a set of 214 known splice junctions is compared in **Table 1**. It appears that models that consider pyrimidine concentration can predict more valid splice junctions than those using the inter-AG length metric alone. The proportion of valid predictions is notably higher (+7%) when allowing a 25-nucleotide margin of error. This is not surprising as it is common for more than one AG dinucleotide to be in close range of each other near splice acceptor sites (data not shown).

The pyrimidine-bias scoring functions were compared to the full inter-AG and linear discriminant analysis using the same reported search space (400 nt upstream of the splice junction). Both methods offer similar predictions although the pyrimidine bias functions display slightly higher rates (+2%). The linear pyrimidine scoring function was chosen for subsequent analyses given its accuracy and simplicity.

Nucleotide composition shifts surrounding the genomic poly(A) site

Of the 12,052 *Leishmania* EST sequences in GenBank, 81% correspond to *L. infantum* and 19% to *L. major* cDNA. We filtered the data to collect sequences harbouring significant poly-A or poly-T stretches near their extremities in order to search for polyadenylation signals. Only 850 sequences (7% of initial data) satisfied our search constraints (see Methods) of which a mere 218 (1.8%) were successfully mapped to genomic intergenic regions of *L. infantum* (the accession numbers for the 218 ESTs can be viewed in **Additional File 1**). The *L. infantum* EST data contains several flagrantly erroneous and repeated sequences. Comparing the pair-wise identity of mapped ESTs revealed 4 pairs of highly similar sequences which, once aligned, proved to be the only example of alternatively polyadenylated sequence in our data (GenBank accession IDs: CV669949.1, CV670417.1, CV668181.1, CV665773.1, CV670284.1, CV668879.1, CV667130.1, CV661593.1).

The position-specific nucleotide frequencies of genomic regions aligned and centered at the mapped poly(A) position reveals prominent trends in global sequence composition (**Figure 1**). Adenosine residues are bountiful near the poly(A) site and an elevated concentration of pyrimidines is perceptible 300 to 600 bases downstream of it. Interestingly, thymine bases are almost twice as abundant around 50 bases upstream of the poly(A) site and their higher overall concentration is synonymous with that of pyrimidine dinucleotides. Not only are adenosine and pyrimidine nucleotides more abundant in polyadenylated regions, they also fluctuate more than that of randomly selected genomic sequences (**Table 2**). When comparing the standard deviations of residues near poly(A) sites, pyrimidine dinucleotides (YY) have a higher standard deviation than their individual

nucleotides alone. It is noteworthy to mention that the nucleotide frequencies tend to resemble that of the random control when extended further away from the poly(A) position.

Capturing such blatant genomic signals in addition to more discrete parameters, like progressive shifts in nucleotide and dinucleotide compositions, could be an effective means of identifying poly(A) sites in unresolved sequences. Such a comparative approach is appealing since conserved sequence motifs surrounding poly(A) sites in trypanosomatid species are not as common as in higher eukaryotes. Using motif detection programs such as MEME [31] did not yield conclusive results (data not shown). Indeed, the intergenic regions of *Leishmania* parasites are riddled with low-complexity regions (i.e., short consecutive repeats of 1-3 nucleotides) that can bias the scoring metrics of such programs. To surmount this shortcoming, we investigated over-represented motifs in the regions directly surrounding genomic poly(A) sites in *Leishmania* using the word enumeration program YMF [32] in combination with FindExplanator [33]. Hexamers that are over-represented in the regions directly flanking known genomic poly(A) sites were compared to those found in more distant regions (see **Additional File 2** for details). The highest-ranking motifs are present in only a fraction of all known poly(A) sites and appear to be randomly distributed within their vicinity (data not shown).

Poly(A) sites can be predicted using scanning matrices

We converted the genomic alignment into a position specific scoring matrix (PSSM) that can subsequently be used to scan non-coding sequences. The PSSM is aligned to every position within the intergenic sequence and emits a bit-score for each position (see Methods). The higher the score, the closer the current position in the intergenic sequence resembles the global composition of a polyadenylated region. We present the depicted prediction rates of a given PSSM as a measure of its sensitivity, or ability to detect valid poly(A) sites. Since the biological data is limited, sensitivity was determined using tenfold cross-validation (see Methods) which allows for unbiased testing, as the testing data is excluded from the training data [34]. The position displaying the highest PSSM score is retained as the poly(A) candidate. Seeing as only non-coding sequences are scanned, we omitted specificity testing on additional control sequences.

Given that the molecular mechanisms of kinetoplastid polyadenylation have yet to be completely demystified, we tested multiple PSSM lengths in order to elucidate which regions surrounding the cleavage site have an effect on polyadenylation. Matrix sizes were limited to regions where a meaningful base composition pattern was observed. The most precise predictions are obtained with small PSSMs encompassing the adenosine rich region directly surrounding the aligned poly(A) sites (**Figure 2A**). Using a prediction tolerance of ≤ 10 nucleotides, a PSSM of 25 bases upstream and 25 bases downstream of the poly(A) site (25A25) shows the highest sensitivity (21% average after 15 runs of 10-fold cross-validation), with a standard deviation of 1.1%. At lower resolutions, the same small PSSMs still display the best predictions, however longer matrices such as the 300 upstream and 600 downstream PSSM (300A600) offer similar sensitivities (**Figure 2C**). Overall, the surface plots show that the regions adjacent to the poly(A) site offer the highest close-range predictions when scanning entire intergenic regions, although larger matrices also display competitive detection rates provided that the margin of error is relaxed.

Limiting PSSM scanning range increases poly(A) site prediction rates

In order to maximize the sensitivity of poly(A) site targeting, we tested the impact of bounding the PSSM search space within a given confidence interval. To do so, the aforementioned refined splice-junction prediction method was applied to the intergenic sequences derived from the polyadenylated ESTs in order to obtain an approximation of the distances between both cleavage sites. The distribution of the putative intergenic spacers shows that 83% of the spacer sequences are shorter than 1500 bases (**Figure 3**), with a median value of 498.

Based on these observations, it is clear that distance is an important factor to incorporate into an mRNA extremity prediction algorithm. We tested the effect of predicting 3'UTR extremities using splice junction prediction combined to a fixed distance as the prime metric. The highest prediction accuracies using this approach are obtained by selecting the median value of spacer sequence sizes as a scanning limit (**Figure 4**). When allowing predictions to be within 100 bases of the valid poly(A) site, this tactic predicts

22% of valid splice sites. At this resolution, scanning the entire IR with PSSMs yields a 36% detection rate, more than double the distance-only value.

We subsequently scrutinized the prediction rates for all PSSMs using various scanning distance limitations, a handful of which are compared amongst themselves (**Figure 5**). The impact of limiting the scanning distance directly upstream of the putative splice-junction site produces a notable increase in sensitivity for most PSSMs. The overall highest sensitivities are obtained by limiting the scanning distance to within 1000 bases of the SJ. This is most notable for the longer matrices, some of which gained over 5% sensitivity within the 10-nucleotide range (**Figure 2B**), thus competing with the shorter matrices for the best prediction rate. At the 100-nucleotide range, limiting the scanning distance to within 1000 bases increased the sensitivity from 36% to almost 45% (**Figure 2D**). Curiously, matrices encoding the pyrimidine rich regions offer the highest sensitivities at this resolution whereas very small ones containing the A-rich region perform best within a 10 nucleotides error margin. When loosening the predictive resolution to within 250 bases, certain PSSMs (most notably 30A600 and 300A600) can identify slightly more than 60% of the mapped poly(A) sites (see **Additional File 3** for all sensitivities).

We tested the effect of combining the high-resolution accuracy of the 25A25 matrix with the low-resolution accuracy of a larger matrix on prediction sensitivity. Two algorithms were tested. The first involves an initial scan with the large matrix, where the highest scoring position and its surrounding sequence are then re-scanned with the smaller matrix. The highest score from this second scan is reported as the presumed poly(A) site. Similarly, the second algorithm combines the scores of both PSSMs but considers all large matrix positions instead of only the highest scoring one. This second algorithm (overviewed in **Figure 6B**) displays the best prediction rates when limiting the smaller matrix scanning to within 75 nt upstream and downstream of the larger matrix's position, with 2-4% higher sensitivity depending on the resolution (data not shown). Although similarly as effective as the 25A25 matrix within 10 nt, this approach displays a higher sensitivity when lowering the resolution to 100 nt (**Figure 5C**). Predictions are nonetheless higher than using individual matrices at any resolution. Including such an approach in a poly(A) prediction program is straightforward given its higher sensitivity.

In order to asses the selectivity of this approach, the highest scores obtained from annotated coding sequences (CDS) were compared to those of known splice-junction and poly(A) regions. The average highest score for poly(A) prediction in all 3789 *Leishmania infantum* CDS over 1500 nucleotides in length is 17.8 bits with a standard deviation of 7.8 bits. Using the same data, the average high score for SJ prediction is 704.1 units with a standard deviation of 401.1 units. The inherent properties of normal distribution statistics stipulate that over 95% of the high scores are within two standard deviations of the mean [35]. Thresholds corresponding to these values (e.g., 34 for poly(A) and 1506 for SJ prediction) were incorporated to the prediction algorithm, which then scanned all datasets. The resulting false-positive and true-positive detection rates are presented in **Table 3**. Integrating the scoring thresholds limits false-positive predictions to less than 5% while only slightly affecting specificity (predictions drop 5-6% for SJ and 1-2% for poly(A) predictions).

Discussion

The 5' splice junction prediction methods disclosed in this work were conceived to estimate *trans*-splicing sites for all input sequences using a simple and effective metric. Since pyrimidines play an important role in *trans*-splicing, including such a parameter into the inter-AG splice prediction model was forthright and can be warranted by the subsequent increase in sensitivity. Although rather effective, the inter-AG metric's principal hoodwink resides in its synthetic nature, as the underlying biological process is difficult to conceive. The assessment of polypyrimidine tract length was not considered in this work as it has been shown that the inter-AG metric is more powerful [23]. Even if our splice junction prediction results are encouraging, some uncertainty subsists when testing on unconfirmed sequences. This may potentially be a consequence of the parasitic nature of trypanosomatids, which coerces these protozoa to alternate between different life-stages depending on their insect and mammalian host. An additional level of complexity may be essential to improve *in-silico* predictions in view of the fact that *trans*-splicing of certain transcripts is developmentally regulated in trypanosomes [36, 37].

When compared to previously published *trans*-splicing prediction rates [23], the models we propose here appear to be just as effective at predicting known *trans*-splicing sites when tested on the same search space (**Table 1**). Their accuracy remains significant even when increasing the query sequence size (1.75x increase in search space at the cost of 0.9x accuracy). The augmented search space is in order to ensure that the full inter-AG fragments upstream of putative splice sites are considered. Overlapping into the downstream coding sequence is vindicated by erroneous genome annotations; it is not uncommon that the furthest in-frame ATG is selected as a start codon. Also, our scoring function rates all inter-AG fragments, unlike the previously proposed study that selects high-scoring fragments based upon their dinucleotide composition [23]. As shown in **Table 3**, a scoring threshold can be implemented to ensure that few false-positives are unsuitably identified as splice-junctions at the cost of slightly lower specificity. However, a threshold will necessarily neglect certain sequences, which may be objectionable when dealing with few or essential queries. Since our method is more dependent on correct annotations, it is conceivable that coupling it to linear discriminant analysis would generate even better predictions at the cost of higher complexity.

Predicting poly(A) sites with PSSM's have previously been shown to successfully predict poly(A) sites in humans [38]. Capturing the global nucleotide composition surrounding known poly(A) sites and utilizing it as a comparative predictor has also proven to be a successful prediction procedure in *Leishmania*. Albeit the public EST data appears to be of questionable quality, stringent screening has permitted to reveal specific polyadenylated sequence traits. Given the nature of the sequence data, smaller mRNA transcripts may be favoured and this should be considered when analyzing results. Nonetheless, PSSM scanning is more than 10 times more effective at identifying poly(A) sites than the distance-only approach when precision is fundamental (**Figures 2A** and **4**). This result can be interpreted as evidence that distance is not as powerful for targeting poly(A) sites in *Leishmania* than in trypanosomes.

For *Leishmania*, precision may not be essential when predicting 3'UTR extremities given that several mappings display heterogeneous poly(A) positions [15]. This observation motivates the use of an error margin, which is interpreted as lowering the resolution of

sensitivity testing in this work. Allowing correct predictions to be within a certain range of the mapped position emulates the identification of a polyadenylation region. We also tested a window scanning approach, where the cumulative bit-scores for a given range were averaged over the size of the window instead of considering each position independently. Such an approach yielded weaker overall predictions than the position-specific approach (data not shown), perhaps because the extent of polyadenylation regions varies among different transcripts.

The best 3'UTR predictions emanate from the grouping of distance limitation and scanning with dual PSSMs. Combining both metrics proved to be more effective than either one individually (**Figures 2, 4,** and **5**), a result that hints at the importance of each factor when predicting poly-A sites in *Leishmania*. For restraining PSSM scanning, we tested various distances instead of using a specific confidence interval since spacer sequences display somewhat of a bias towards longer fragments. Although the data is partially derived from estimations, such a shift in the distribution supports the notion that polyadenylation does not occur randomly in *Leishmania*. Poly(A) sites further away from the splice junction may be an effect of distant polypyrimidine tracts, a situation that has already been observed in trypanosomes [20]. One must also consider that the longer non-coding regions in *Leishmania* may contain non-annotated genes or provide alternative stage-specific polyadenylation sites, which could explain the longer spacer sequences. These are considerations that motivated the exclusion of intergenic sequences longer than 5000 nucleotides for sensitivity testing.

To our knowledge, no other method can predict poly(A) sites as effectively in *Leishmania spp.* as the one described in this work. Even enforcing a highly-selective threshold only faintly affects this method's specificity (**Table 3**). The rather unusual and non-specific nature of kinetoplastid polyadenylation is a line of reasoning to substantiate low computational prediction rates. Although over-represented A-rich hexamer motifs are found (**Additional File 2**), these are not however present in all the genomic poly(A) sites, which suggests that they may not play a central role in driving polyadenylation in *Leishmania*. In addition, the genomic alignment of polyadenylated EST mappings cannot be used to mark out a precise consensus sequence, as it is impossible to distinguish the

exact cleavage site among multiple consecutive adenosines on the unprocessed transcript. The heterogeneity of poly(A) sites in *Leishmania* mRNA transcripts is extra incentive for using PSSMs that embody a global trend in nucleotide composition. Furthermore, neglecting secondary structure and stage-specificity are additional factors that make it difficult to conceive obtaining higher prediction accuracies at this point.

Notwithstanding the possibility that no consensus motif drives polyadenylation in kinetoplastids, there is evidence for a biological model based on sequence context. The low sensitivity obtained from a poly(A) prediction algorithm based on spacing metrics alone is an evidence for a more dynamic biological model. Also, the correlation between certain regions of the genomic alignment and their respective prediction rates is most interesting, as best illustrated by the sensitivity surface plots (**Figure 2**). The data is presented in order to asses the innate characteristics that have an impact on poly(A) targeting.

Two main common sequence features appear to directly influence the prediction sensitivities. Firstly, the adenosine-rich region within close range to the mapped poly(A) site. Secondly, the pyrimidine-rich region 300 to 600 positions downstream. The latter, which represents the polypyrimidine tracts known to be crucial for *trans*-splicing, generates the best predictions when loosening the accuracy and bounding the search space. In turn, the A-rich region is responsible for the best predictions when precision is fundamental. Interestingly, the affluence of polypyrimidines (most notably thymines) in the -50 to -25 region (**Figure 1**) may play a role in 3'UTR cleavage since its exclusion from scanning matrices reduces the sensitivity at close range (**Figure 2**). The matrix encoding the sequence information of zero upstream bases and 25 downstream (0A25) is somewhat futile at predicting poly(A) sites, a rather surprising observation seeing as the adenosine concentration is comparable. Upon closer inspection, it is apparent that adenosine-rich regions are not a fundamental marker because many sequences do not contain profuse adenosine residues at their poly(A) site.

PSSMs can be regarded as a simplistic representation of the interaction between an enzymatic complex and a strand of nucleic acids. The highest scoring position corresponds to a region that is most similar to the consensus of all poly(A) sites, which relates to a high

affinity region for the polyadenylation complex. In this perspective and based on our results, it is enticing to contemplate a generic biological model where adenosine richness (possibly contrasted by a pyrimidine-rich upstream region) helps to direct the polyadenylation of specific positions downstream of polypyrimidine tracts in unprocessed mRNA transcripts. Deletion studies directed at these features followed by mapping the modified transcript's poly(A) site could shed additional light into the biological process. Moreover, *in-vitro* UV cross-linking could help identifying novel ribonucleoproteins (RNPs) that might interact with the *trans*-splicing/polyadenylation complexes.

The computational tools we describe in this work have been implemented in a small JAVA program named PRED-A-TERM (PREDicting poly(A) sites and TERMinal splice junctions) that can be downloaded from **Additional File 4**. It emits poly(A) and *trans*-splicing predictions from intergenic sequence input with partial coding sequence overlap and allows end-users the possibility to select various prediction parameters. The program is tuned for *L. infantum* but is suitable for other *Leishmania* species. Although trypanosomes have shorter average intergenic regions than *Leishmania*, both share similar *trans*-splicing machinery [39, 40]. Scanning *Trypanosoma* IRs will however, require additional sequence analysis and subsequent tuning of the model.

Conclusions

We present a simplified 5'UTR prediction function that can predict more than 65% of known *trans*-splicing sites within 25 nucleotides. This approach performs as good as previously published methods but it significantly reduces computational cost. We also present a novel 3'UTR prediction method for the trypanosomatid *Leishmania* that compares query sequences to known polyadenylated sequences using position specific scanning matrices. Such an approach is capable of predicting almost 50% of known poly(A) sites within 100 nucleotides, thus doubling the accuracy of the previous distance based approach. The final algorithm implemented in PRED-A-TERM is summarized in **Figure 6**.

By increasing the precision of large-scale transcriptome predictions in trypanosomatids, the prospective identification of novel regulatory non-coding RNA

structures is now within reach. The relatively recent fervour for investigating regulatory functions of non-coding RNA has propelled the emergence of multiple structural RNA detection algorithms [41, 42]. These modern computational methods combined with biological validation could facilitate the discovery of innovative targets for therapeutic treatments.

Methods

5' Splice junction prediction

After aligning the EST data to the genome, we extracted 500 nucleotides upstream of the coding sequence associated to the EST and the first 200 nucleotides downstream of the annotated start codon. *Trans*-splicing predictions are based upon the most recently published method [23]. Sequences are fragmented at every occurrence of "AG" and each fragment's size is calculated. In the simplest scoring scheme, the longest inter-AG fragment is retained and the sequence's final position is considered as a splice junction candidate. Linear and polynomial pyrimidine bias models calculate the relative pyrimidine concentration of inter-AG fragments and modify each fragment's score proportionately using the following functions:

$$\boldsymbol{L} = \lambda + 150 \bullet \lambda \bullet \delta$$

$$\boldsymbol{P} = \lambda + 150 \bullet \lambda \bullet \delta^3$$

where $\boldsymbol{L}$ and $\boldsymbol{P}$ are the linear and polynomial model scores respectively, λ is the inter-AG fragment length, and δ corresponds to the difference between the pyrimidine concentration of the inter-AG fragment and the average intergenic concentration (55%). In both cases, the last position of the highest scoring inter-AG fragment is retained as the putative splice junction. Optimal coefficients were determined by trial and error testing. Sensitivity testing was performed on the same 214 EST sequences from *Leishmania major* as reported in that article.

Data collection

Leishmania sequences for the poly(A) analysis were downloaded from GenBank's expressed sequence tag (EST) public database [43]. Data were filtered to retain sequences having at least 12 adenine (A) or thymine (T) residues at their 3' or 5' end, respectively. Poly-T sequences were subsequently reverse-complemented. *Leishmania infantum* sequences were aligned to the genome (version 3 downloaded from http://www.GeneDB.org) using BLAST with low-complexity filtering disabled [44]. Hits over 100 nucleotides long that displayed over 95% sequence identity were retained. We define an EST sequence as being polyadenylated if it satisfies the following criteria: (i) The last position of the best BLAST hit must immediately precede the poly(A) stretch. (ii) There should be no more than 9 "A" residues out of the next 12 genomic nucleotides following the BLAST hit. (iii) The last alignment match must not be a "N" in the genomic or EST sequence. The polyadenylation site is defined as the last non-"A" residue shared between the EST extremity and the genomic sequence. The full list of polyadenylated EST accession numbers can be viewed in the supplementary data (S1). All filtering steps were achieved using *ad-hoc* JAVA scripts.

Building poly(A) scanning matrices

The genomic sequences of the polyadenylated ESTs were aligned and anchored at the mapped poly(A) site, as previously defined. From this alignment, we calculated the specific nucleotide composition for each position relative to the poly(A) site. The resulting nucleotide frequencies were divided by their corresponding average genomic frequency (A=20.1%, T=20.2%, C=29.7%, G=30.0%) to create an odds matrix. The final position specific scoring matrix (PSSM) was obtained by log-transforming the odds matrix to generate bit scores for each matrix entry.

Poly(A) prediction using scanning matrices

The genomic intergenic regions (IRs) associated to the retained ESTs were extracted and extended 600 bases past the stop and start codons of the most recent *L. infantum* genome annotation (version 3). Only IRs inferior or equal to 5000 bases in length were retained. IRs of interest were scanned with PSSM sizes ranging from 1 to 300 upstream and 1 to 600 downstream of the poly(A) location. When scanning the entire IR, query sequences are scanned such that the position corresponding to the anchored poly(A) site in the PSSM is aligned to the first non-coding position downstream of the stop codon; at this point, the upstream matrix positions overlap the ORF. A cumulative bit-score is emitted for each given position and this step is repeated for every position of the intergenic sequence (the positions downstream of the matrix's poly(A) position may overlap the ORF when scanning the last positions). The positions with the highest scores are retained as putative polyadenylation sites. When predicting a polyadenylation region instead of a single position, the cumulative individual bit-scores are averaged over the length of the region scanned. The optimal prediction algorithm is summarized in **Figure 6**.

Ten-fold cross-validation sensitivity testing

The prediction accuracies presented in this work arise from cross-validation sensitivity testing, where the polyadenylated EST data are divided into 10 subsets. Nine of those are used as a training set (in this case, to build a PSSM) which are subsequently tested on the left-over subset. This step is repeated for all subsets and the results are averaged to obtain the mean sensitivity. The average and standard deviation of 15 runs of cross-validation were performed for PSSM scanning and 30 runs for distance-only predictions. All testing was performed using *ad-hoc* JAVA scripts.

Authors' contributions

MS conceived of the study, performed all computational analyses, and drafted the manuscript. MB and BP contributed to the conception and coordination of the study and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgments

BP is a member of a Canadian Institute of Health Research (CIHR) Group on Host-Pathogen Interactions and of a Fonds Québecois de la Recherche sur la Nature et les Technologies (FQRNT) Center for Host-Parasite Interactions. This work was supported by an operating grant from the CIHR (MOP-12182) awarded to BP.

References

1. Murray HW, Berman JD, Davies CR, Saravia NG: **Advances in leishmaniasis.** *Lancet* 2005, **366**(9496):1561-1577.
2. **World Health Organization** [http://www.who.int/leishmaniasis]
3. Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW: **The tree of eukaryotes.** *Trends Ecol Evol* 2005, **20**(12):670-676.
4. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C *et al*: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science* 2005, **309**(5733):404-409.
5. Imboden MA, Laird PW, Affolter M, Seebeck T: **Transcription of the intergenic regions of the tubulin gene cluster of Trypanosoma brucei: evidence for a polycistronic transcription unit in a eukaryote.** *Nucleic acids research* 1987, **15**(18):7357-7368.
6. Kooter JM, Borst P: **Alpha-amanitin-insensitive transcription of variant surface glycoprotein genes provides further evidence for discontinuous transcription in trypanosomes.** *Nucleic acids research* 1984, **12**(24):9457-9472.

7. van der Spek H, Arts GJ, Zwaal RR, van den Burg J, Sloof P, Benne R: **Conserved genes encode guide RNAs in mitochondria of Crithidia fasciculata**. *Embo J* 1991, **10**(5):1217-1224.

8. Clayton CE: **Life without transcriptional control? From fly to man and back again**. *Embo J* 2002, **21**(8):1881-1888.

9. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI *et al*: **A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA**. *Rna* 2000, **6**(2):163-169.

10. Sather S, Agabian N: **A 5' spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei**. *Proc Natl Acad Sci U S A* 1985, **82**(17):5695-5699.

11. Liang XH, Haritan A, Uliel S, Michaeli S: **trans and cis splicing in trypanosomatids: mechanism, factors, and regulation**. *Eukaryot Cell* 2003, **2**(5):830-840.

12. Krause M, Hirsh D: **A trans-spliced leader sequence on actin mRNA in C. elegans**. *Cell* 1987, **49**(6):753-761.

13. Vandenberghe AE, Meedel TH, Hastings KE: **mRNA 5'-leader trans-splicing in the chordates**. *Genes Dev* 2001, **15**(3):294-303.

14. Zhao J, Hyman L, Moore C: **Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis**. *Microbiol Mol Biol Rev* 1999, **63**(2):405-445.

15. LeBowitz JH, Smith HQ, Rusche L, Beverley SM: **Coupling of poly(A) site selection and trans-splicing in Leishmania**. *Genes Dev* 1993, **7**(6):996-1007.

16. Matthews KR, Tschudi C, Ullu E: **A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes**. *Genes Dev* 1994, **8**(4):491-501.

17. Hug M, Hotz HR, Hartmann C, Clayton C: **Hierarchies of RNA-processing signals in a trypanosome surface antigen mRNA precursor**. *Mol Cell Biol* 1994, **14**(11):7428-7435.

18. Schurch N, Hehl A, Vassella E, Braun R, Roditi I: **Accurate polyadenylation of procyclin mRNAs in Trypanosoma brucei is determined by pyrimidine-rich elements in the intergenic regions**. *Mol Cell Biol* 1994, **14**(6):3668-3675.

19. Vassella E, Braun R, Roditi I: **Control of polyadenylation and alternative splicing of transcripts from adjacent genes in a procyclin expression site: a dual role for polypyrimidine tracts in trypanosomes?** *Nucleic acids research* 1994, **22**(8):1359-1364.

20. Benz C, Nilsson D, Andersson B, Clayton C, Guilbride DL: **Messenger RNA processing sites in Trypanosoma brucei**. *Molecular and biochemical parasitology* 2005, **143**(2):125-134.

21. Clayton CE, Ha S, Rusche L, Hartmann C, Beverley SM: **Tests of heterologous promoters and intergenic regions in Leishmania major**. *Molecular and biochemical parasitology* 2000, **105**(1):163-167.

22. Requena JM, Quijada L, Soto M, Alonso C: **Conserved nucleotides surrounding the trans-splicing acceptor site and the translation initiation codon in Leishmania genes**. *Exp Parasitol* 2003, **103**(1-2):78-81.

23. Gopal S, Awadalla S, Gaasterland T, Cross GA: **A computational investigation of kinetoplastid trans-splicing**. *Genome Biol* 2005, **6**(11):R95.

24. Xu Y, Liu L, Michaeli S: **Functional analyses of positions across the 5' splice site of the trypanosomatid spliced leader RNA. Implications for base-pair interaction with U5 and U6 snRNAs**. *J Biol Chem* 2000, **275**(36):27883-27892.

25. Aly R, Argaman M, Halman S, Shapira M: **A regulatory role for the 5' and 3' untranslated regions in differential expression of hsp83 in Leishmania**. *Nucleic acids research* 1994, **22**(15):2922-2929.

26. Wu Y, El Fakhry Y, Sereno D, Tamar S, Papadopoulou B: **A new developmentally regulated gene family in Leishmania amastigotes encoding a homolog of amastin surface proteins**. *Molecular and biochemical parasitology* 2000, **110**(2):345-357.

27. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, Papadopoulou B: **A common mechanism of stage-regulated gene expression in Leishmania**

mediated by a conserved 3'-untranslated region element. *J Biol Chem* 2002, **277**(22):19511-19520.

28. Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, Papadopoulou B, Ghedin E: **Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania.** *PLoS pathogens* 2007, **3**(9):1291-1307.
29. Haile S, Papadopoulou, B.: **Developmental regulation of gene expression in trypanosomatid parasitic protozoa.** *Current Opinion in Microbiology* 2007:In press.
30. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2**(12):919-929.
31. Bailey TL, Williams N, Misleh C, Li WW: **MEME: discovering and analyzing DNA and protein sequence motifs.** *Nucleic acids research* 2006, **34**(Web Server issue):W369-373.
32. Sinha S, Tompa M: **YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nucleic acids research* 2003, **31**(13):3586-3588.
33. Blanchette M, Sinha S: **Separating real motifs from their artifacts.** *Bioinformatics* 2001, **17 Suppl 1**:S30-38.
34. Geisser S: **The Predictive Sample Reuse Method with Application.** *J Amer Stat Ass* 1975, **70**:320-328.
35. Ross SM: **Probability and Statistics for Engineers and Scientists, Introduction to.** . San Diego, Burlington, London: Elsevier Academic Press; 2004.
36. Benabdellah K, Gonzalez-Rey E, Gonzalez A: **Alternative trans-splicing of the Trypanosoma cruzi LYT1 gene transcript results in compartmental and functional switch for the encoded protein.** *Molecular microbiology* 2007, **65**(6):1559-1567.
37. Jager AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC: **mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes.** *Proc Natl Acad Sci U S A* 2007, **104**(7):2035-2042.

38. Legendre M, Gautheret D: **Sequence determinants in human polyadenylation site selection**. *BMC genomics* 2003, **4**(1):7.

39. Ullu E, Matthews KR, Tschudi C: **Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts**. *Mol Cell Biol* 1993, **13**(1):720-725.

40. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R *et al*: **The genome of the kinetoplastid parasite, Leishmania major**. *Science* 2005, **309**(5733):436-442.

41. Jossinet F, Ludwig TE, Westhof E: **RNA structure: bioinformatic analysis**. *Curr Opin Microbiol* 2007, **10**(3):279-285.

42. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E: **Bridging the gap in RNA structure prediction**. *Curr Opin Struct Biol* 2007, **17**(2):157-165.

43. **GenBank** [www.ncbi.nlm.nih.gov/Genbank]

44. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic acids research* 1997, **25**(17):3389-3402.

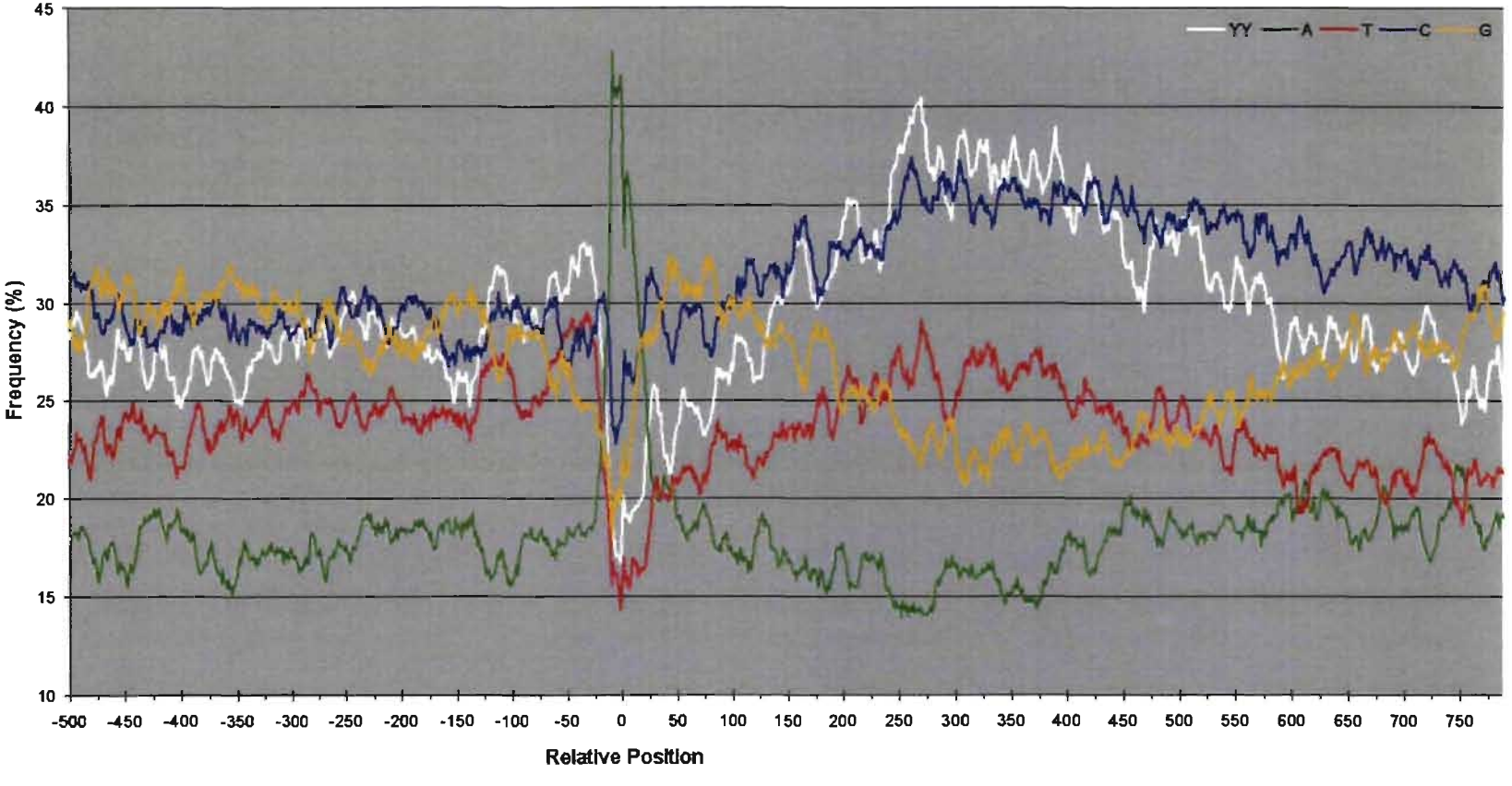

Figure 1. Nucleotide and pyrimidine dinucleotide frequencies surrounding the mapped polyadenylation site of 218 expressed sequence tags from *Leishmania infantum*.

Frequencies are averaged over an 11 nucleotide sliding window in order to smooth out the graph. Negative positions are 5' of the poly(A) site and positive positions are towards the downstream gene. Pyrimidine dinucleotides are considered to be any occurrence of consecutive C or T residues.

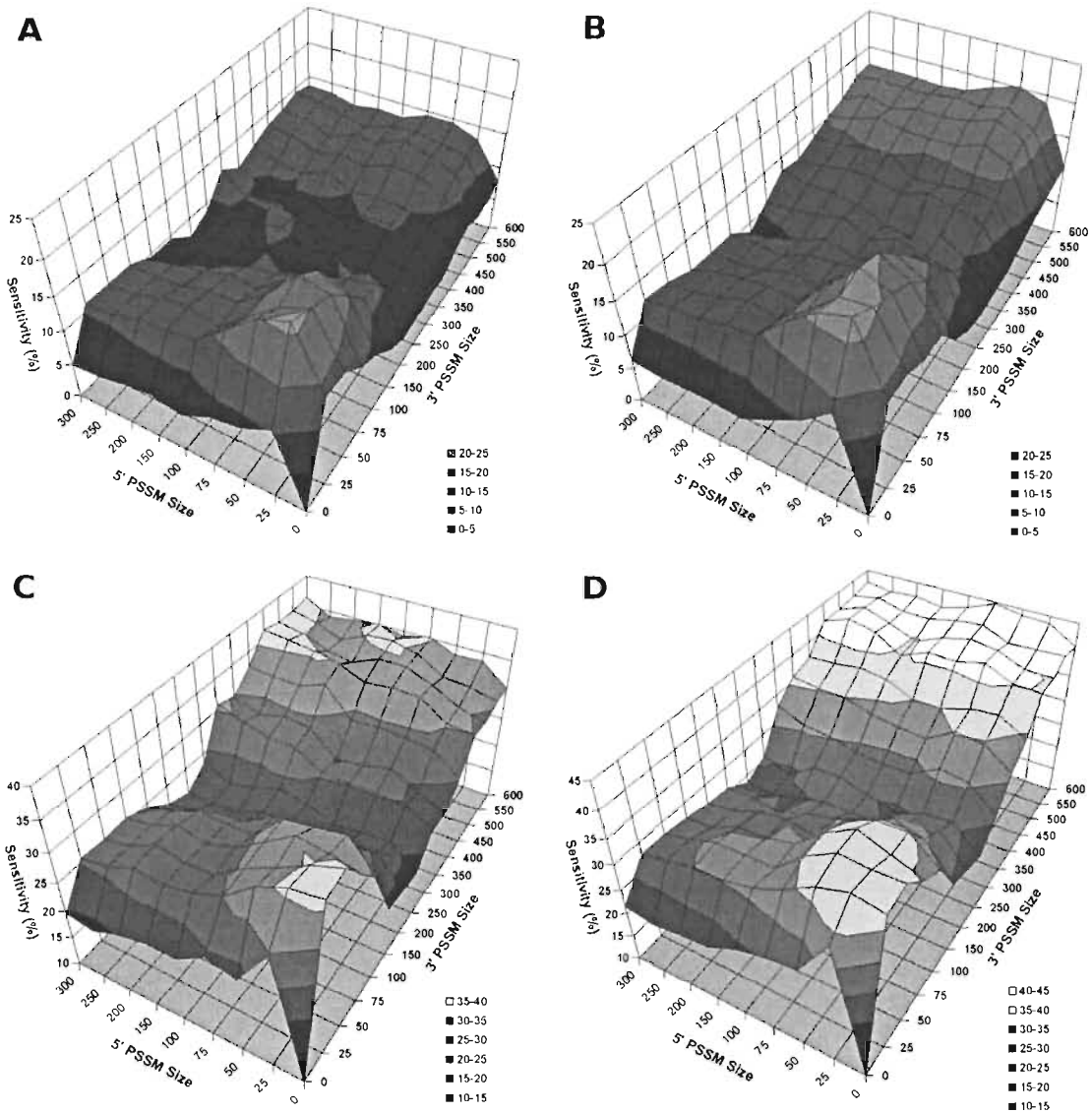

Figure 2. Surface plots of poly(A) prediction sensitivities as a function of various PSSMs (**A**) Predictions within 10 bases of sequenced poly(A) site when scanning the entire intergenic region (IR) and (**B**) when limiting scanning to 1000 bases upstream of the predicted splice junction. Predictions within 100 bases when scanning the whole IR (**C**) and within 1000 bases (**D**). Sensitivities are presented as the average of 15 runs of ten fold cross-validation for each PSSM. The 5' and 3' PSSM size axes correspond to the region upstream and downstream of the genomic alignment of mapped poly(A) sites, respectively. In order to amplify the resolution of regions directly surrounding the poly(A) sites, the scale for 5' and 3' matrix sizes inferior to 100 is decreased from 50 to 25.

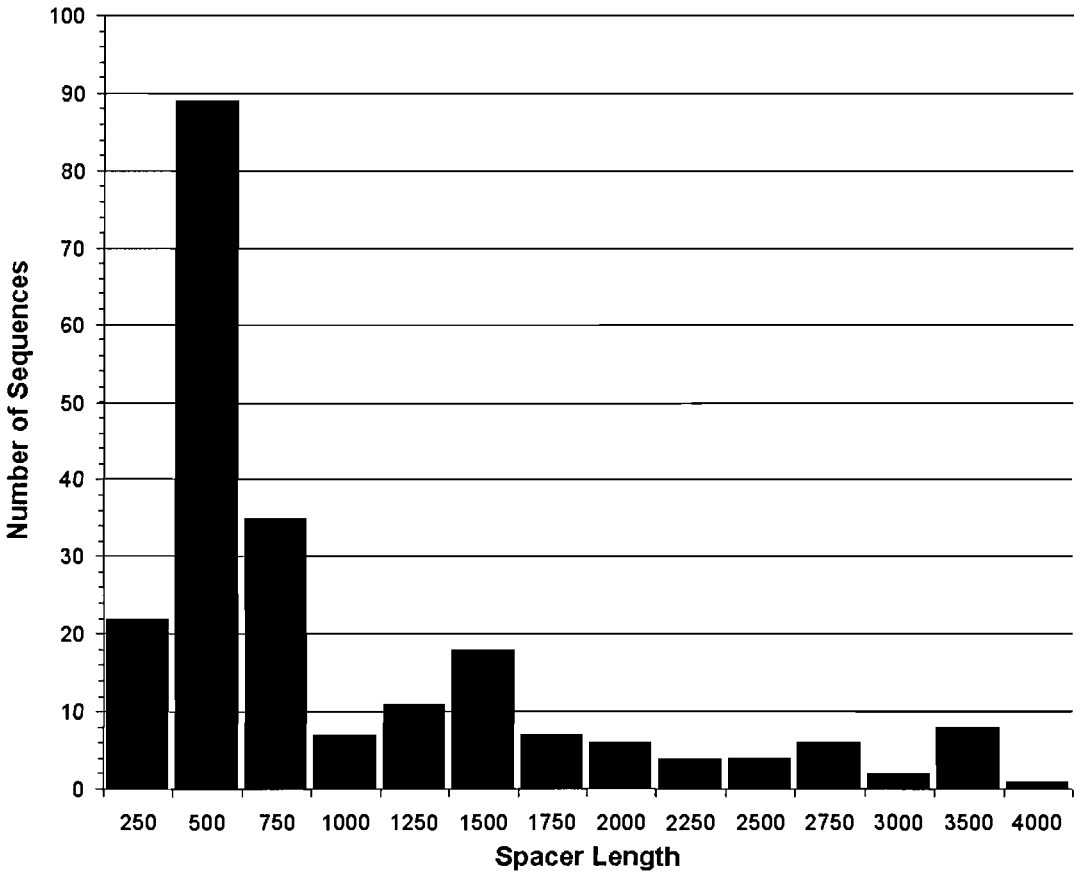

Figure 3. Distribution of spacer sequences

Distribution of the distance between the mapped polyadenylation site and the putative splice junction (spacer length) of 218 intergenic regions from *Leishmania infantum*. *Trans*-splicing positions were estimated using the linear pyrimidine bias function described in Methods.

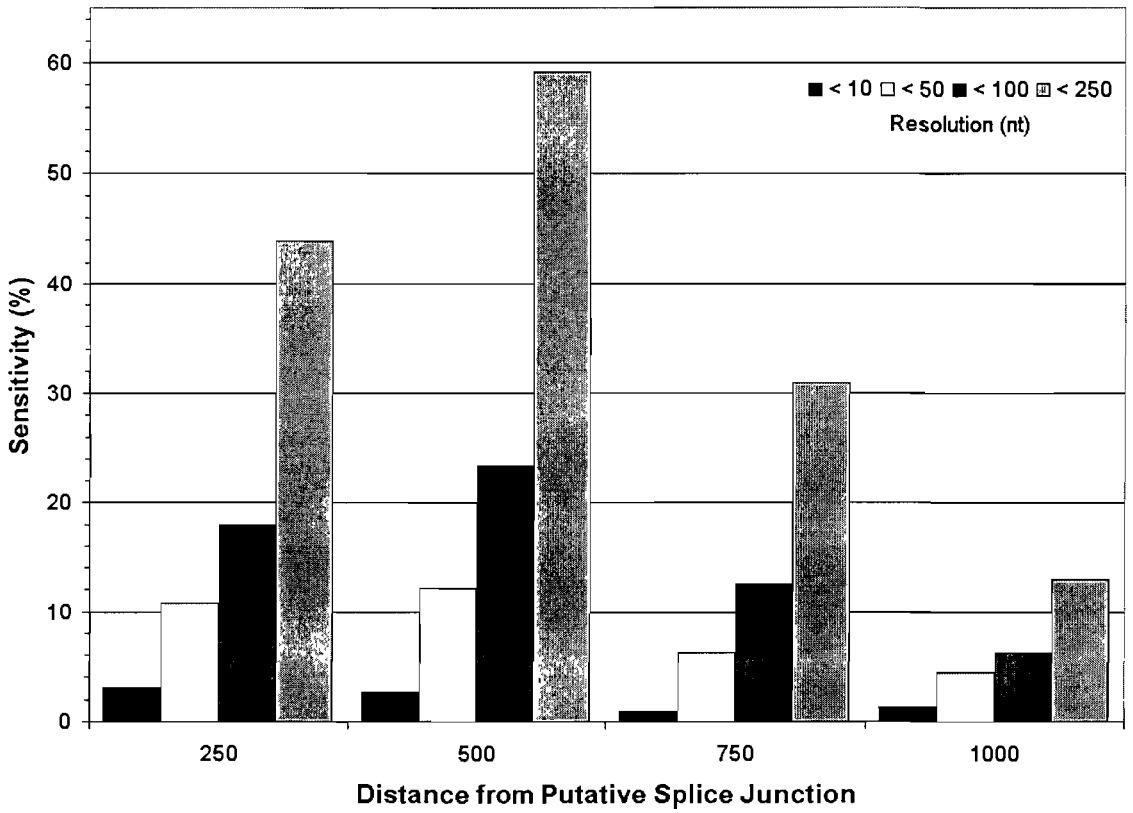

Figure 4. Prediction sensitivities using fixed distances

Sensitivity of poly(A) site predictions using fixed distances from the putative splice junction of 218 intergenic regions mapped from polyadenylated ESTs in *Leishmania infantum*. The resolution corresponds to the distance allowed between the true poly(A) site and the predicted poly(A) site. Standard deviations are denoted as the bars above each column.

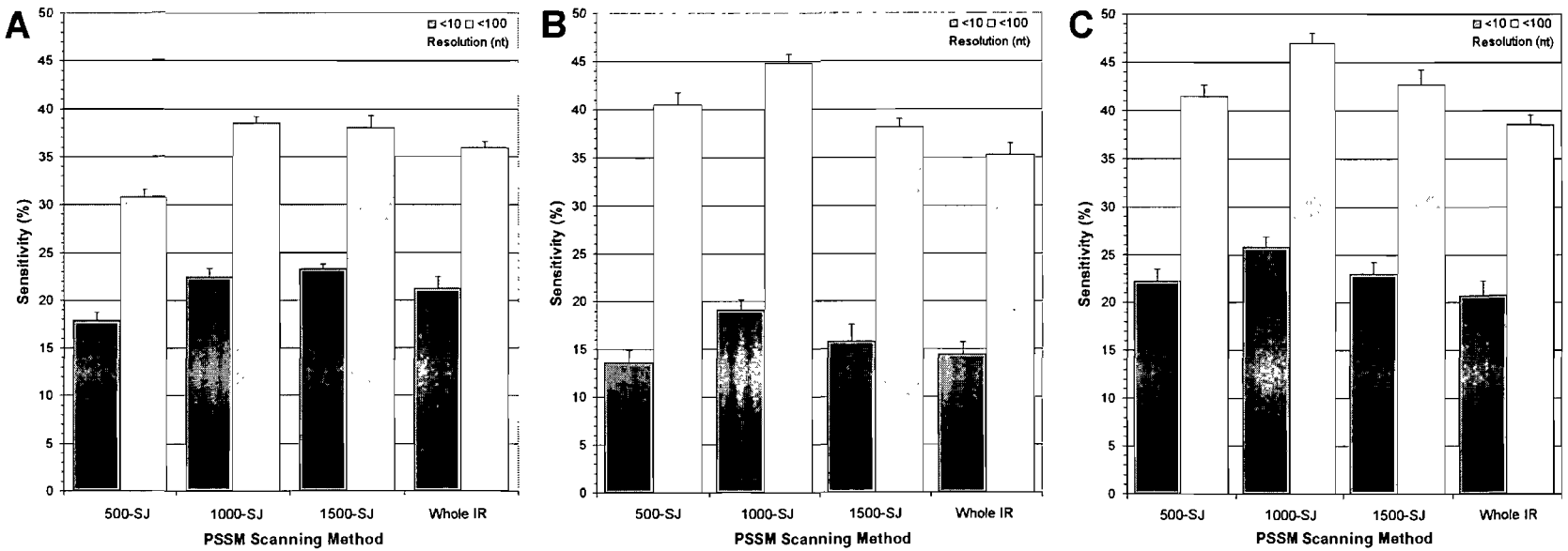

Figure 5. Comparison of poly(A) prediction sensitivities for chosen PSSMs using different scanning approaches

The mean sensitivities after 15 runs of tenfold cross-validation are presented for whole intergenic region scanning and for 3 limited scanning ranges (from the putative splice-junction to 500, 1000, and 1500 positions upstream. (**A**) Mean sensitivities using a PSSM size of 25A25 (25 bases upstream and downstream of the mapped poly(A) position). (**B**) Mean sensitivities using a PSSM size of 75A600. (**C**) Mean sensitivities using a combination of both PSSMs (see Results for details).

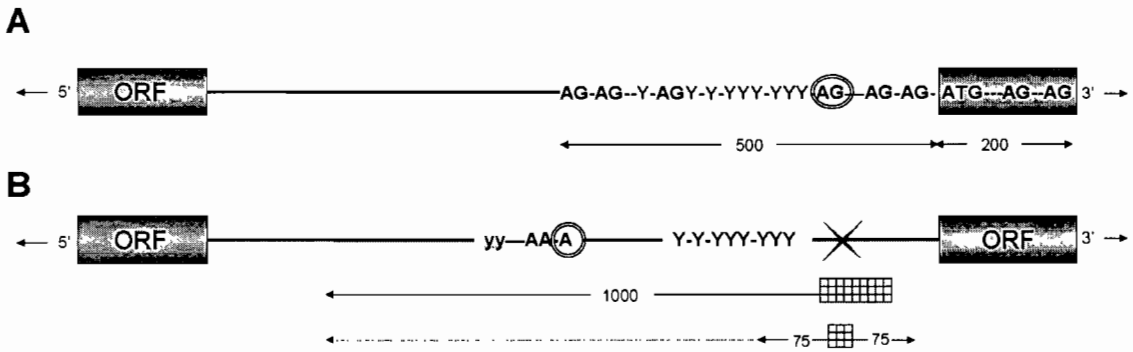

Figure 6. Summary of PRED-A-TERM program

(**A**) The putative trans-splicing site is first predicted by fragmenting 700 nucleotides at each occurrence of AG. The longest, pyrimidine rich inter-AG fragment is selected as the sequence upstream of the splicing site. (**B**) A large scanning matrix subsequently skims through 1000 nucleotides upstream of this position, identifying pyrimidine rich areas with adenosines upstream of them. For each large matrix position, a smaller matrix scans 75 positions in both directions from the larger matrix's hypothetical poly(A) site in order to pinpoint adenosine residues contrasted to pyrimidines. The position of the smaller matrix displaying the highest sum of bit-scores is retained as the putative poly(A) site.

Tables

Scoring Function	***Exact***	***< 25 nt***
Longest Inter-AG Length	49.5	58.9
Linear *Y*-bias	53.7	65.9
Polynomial *Y*-bias	53.3	64.0
Inter-AG Length + LDA*	58.6	72.1
Linear *Y*-bias**	58.9	74.3
Polynomial *Y*-bias**	60.7	74.3

* Interpreted from [23]

** Using same search space as Inter-AG + Linear Discriminant Analysis

Table 1. Splice junction prediction sensitivities of three different scoring models

Values are the ratio (%) of correct predictions among the 214 sequences in the test set, for exact predictions and predictions within 25 bases of the sequenced 5' splice junction. The upper half displays values obtained using a search space of 700 whereas the lower half displays predictions using the same query size as the linear discriminant analysis.

	Genomic Poly(A)					*Random Genomic*				
	A	**T**	**C**	**G**	**YY**	**A**	**T**	**C**	**G**	**YY**
Average	20.3	25.6	34.1	29.3	29.1	20.1	20.2	29.7	30.0	24.1
Median	20.8	25.7	34.0	28.6	29.6	20.0	20.4	29.7	30.0	24.2
Standard Deviation	3.8	2.7	2.6	2.8	4.0	1.0	1.0	1.0	1.0	1.1

Table 2. Global statistics of genomic sequences in *Leishmania infantum*

Values are derived from 218 sequences of 1601 nucleotides and represent the canonical DNA bases in addition to pyrimidine dinucleotides (YY). The genomic poly(A) sequences encompass 800 nucleotides surrounding the mapped poly(A) site. As a control, 218 genomic regions were selected at random via an *ad-hoc* JAVA script.

Additional files

Additional File 1 – Polyadenylted ESTs *(available in Appendix I)*

PDF document of all 218 polyadenylated EST accession IDs used to build scanning matrices in this work.

Additional File 2 – Over-represented hexamers

A Microsoft Word document containing the 10 highest scoring hexamers identified with YMF and FindExplanator programs. An alignment of 223 sequences was used to compare regions encompassing the [-125; +125] of genomic poly(A) sites to the [-800; -126] and [+126; +800] regions.

Additional File 3 – PSSM Scanning Results for Different Matrix Sizes and Scanning Distances

A Microsoft Excel spreadsheet containing all sensitivity results for single-matrix scanning approaches.

Additional File 4 – PRED-A-TERM Program

The prediction algorithm described in this manuscript has been implemented into a JAVA program which can be used to scan query sequences using any operating system. Once extracted, detailed usage instructions can be viewed in the README.txt file.

CHAPTER II

In a recent paper, we report the presence of a highly abundant family of extinct retroposons in *Leishmania major* [53]. The study centers on the characterization of the *Lm*SIDER2 subgroup for which an exhaustive manual alignment was performed. However, the *Lm*SIDER1 subgroup was somewhat ignored just the same as the incidence of SIDERs in other *Leishmania* species (albeit this was carried out for the related parasite *Trypanosoma brucei*). The following sections present various computational methodologies which further improve the representation of SIDER sequences in the context of full-scale comparative genomics.

1. SIDER PROFILING

The initial discovery of SIDERs was achieved by means of genome-wide local alignments and manual annotation [53]. The 79 nt signature sequence of trypanosomatid retroposons (c.f. Introduction, Figure 6) was used as a query for *BLAST* searches on the *L. major* genome [182]. The sequences downstream of relevant hits were aligned with *CLUSTALW* [132], which enabled the detection of retroposon-derived terminal poly(A) stretches. Significant sequences were then queried with *BLAST*, from which results were collected and re-submitted in an iterative manner. Any new, non-redundant hits were added to the resulting list of identified elements. Globally, the *BLAST* searches revealed two somewhat detached sets of results, thus giving way to a coarse labelling of the repeats into two subgroups: SIDER1 and SIDER2.

1.1. Determining an optimal alignment strategy

The initial dataset of degenerated interspersed repeats in *L. major* was quite difficult to align due to high sequence divergence and size polymorphism. Nonetheless, the *MUSCLE* program [183] was used to produce a workable MSA for SIDER2 elements that was refined by extensive manual editing from the author [53 – supplementary data]. However, this alignment contained 1 013 of the 1 073 identified *Lm*SIDER2s and no alignment was produced for *Lm*SIDER1 sequences given their

higher divergence. In the context of an investigation encompassing multiple genomes, an efficient and automated tool for aligning multiple sequences is fundamental. This program must be capable of processing hundreds of different sequences ranging from 300 to 1000 nucleotides relatively quickly. It should not fall short of producing reliable results when confronted with sequences displaying variable rates of conservation. Furthermore, it ought to be flexible enough to add new sequences into an existing alignment in order to quickly retrain a new alignment.

Several available programs were tested in order to determine the best candidate. Many MSA programs do not perform well with sequences displaying mean pair-wise identities around 50%, which is ironically the case among SIDERs. The most important condition required to properly characterize the primary structure of SIDERs is multiple alignment quality. Seeing as evaluating this outcome is very subjective, it was necessary to devise some means of comparing the results of various MSA tools. At the time of the experiment, a selection of 15 representative sequences from the robust manual alignment was available. This alignment was used as the 'gold standard' for benchmarking the quality of other alignments. De-gapped sequences were submitted to five programs, each based on different MSA algorithms, using their default parameters. The 40% consensus sequence (i.e. the sequence composed of the character that makes up 40% or more of each position) was calculated for each alignment with the *BIOEDIT* multiple alignment editor [184] and subsequently aligned with *CLUSTALW* [132]. The *MEGA3* program [185] was then used to calculate a neighbour-joining phylogenetic tree from the alignment of consensus sequences (**Figure 1**).

As foreseen, visual evaluation of the alignments obtained from each program proved to be tricky since *MAFFT* and *HMMER* produced alignments graphically similar to the manual reference. However, the tree in **Figure 1** clearly shows that *HMMER* produces a multiple alignment that is more closely related to the reference.

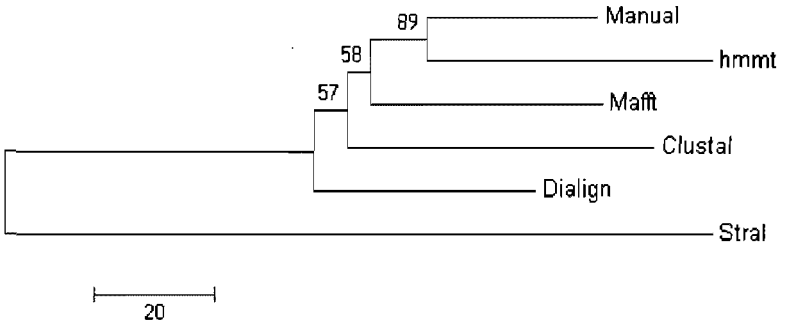

Figure 1. Neighbour-joining tree of the aligned consensus of six multiple sequence alignments

Five different algorithms are compared to an exhaustive manual alignment: *hmmt* – default simulated annealing and Viterbi HMM training implemented in *HMMER* [186]; *Mafft* – iterative refinement algorithm [136]; *ClustalW* – progressive alignment algorithm [132]; *Dialign* – segment pair graph-based algorithm [187]; *Stral* – structural alignment algorithm employing a heuristic scoring function [188]. Branch common to two methods represent their shared identity. Bootstrap values after 500 replicates are indicated above branching points.

1.2. SIDER alignments

Having elected a preferred alignment strategy, it was applied to all published SIDER sequences spanning between 400 and 700 nt (**Figure 2**). The sequences are encoded into a HMM via the *HMMER* program and submitted to two rounds of optimisation: (i) an initial alignment round using simulated annealing; (ii) an enhancement round using Baum-Welch EM. For the initial step, a ramp value of 0.992 is specified in order to ensure maximal iteration in simulated annealing or else Viterbi optimisation ensues, which would normally be desired. However the alignments produced with Viterbi optimisation were clearly erroneous due to either a software bug or to memory constraints from the large dataset (the program was executed in a 32-bit operating system).

The resulting alignment was submitted to phylogenetic analysis in order to determine the evolutionary relationship among SIDER sequences. Any sequence displaying over 95% sequence identity to another is removed in order to reduce base

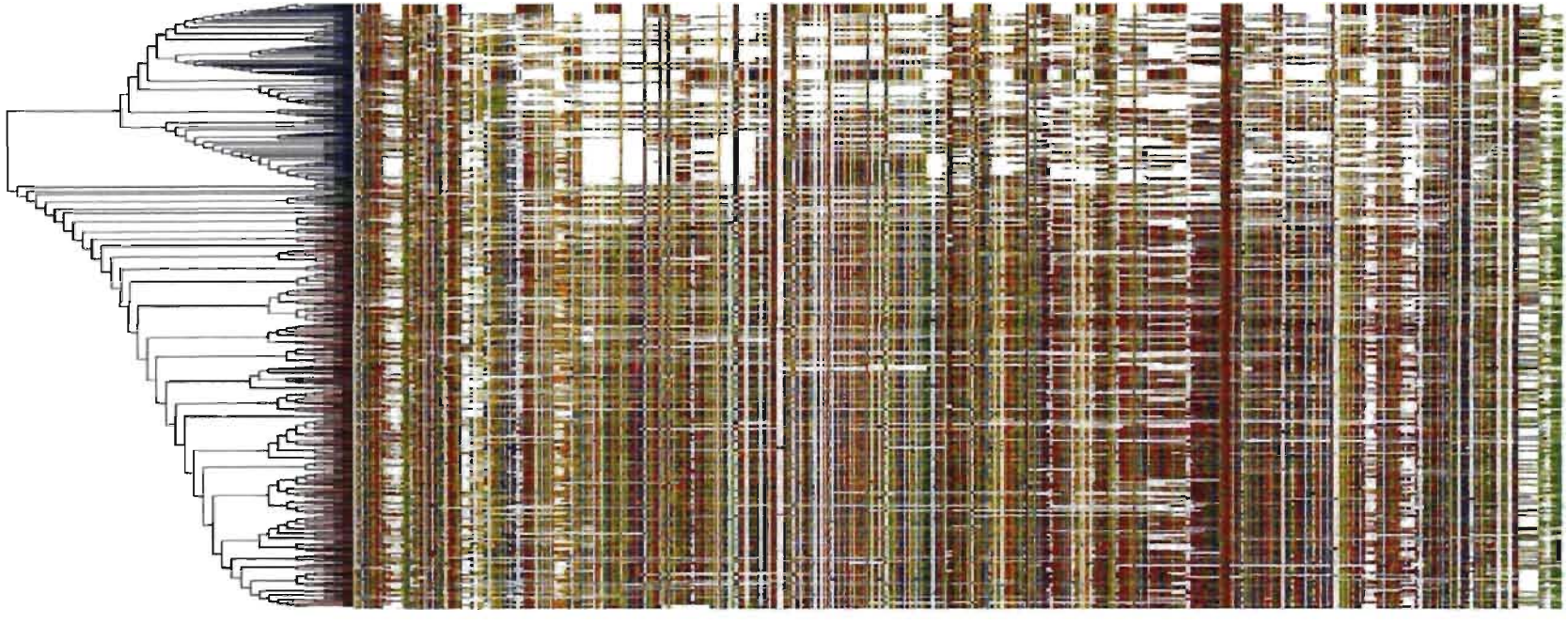

Figure 2. Multiple alignment and phylogenetic relationship of annotated *Lm*SIDERs

The 1351 annotated SIDERs longer than 400 nt and shorter than 700 for *Leishmania major* were aligned with version 1.8.5 of the *HMMER* software package [186]. Initial ligument was performed using the *hmmt* command and parameters *–k 10 –r 0.992*. Sequences displaying over 95% pairwise identity were removed from the alignment. The subsequent alignment was then enhanced with the *hmmt –b* command (Baum-Welch expectation maximization). Columns with abundant insertion/deletion characters (indels) were removed to ease viewing. To colour correspondence of alignment residues is the following: green = A, red = T, orange = C, blue = G. A minimum evolution phylogenetic tree was performed on the alignment using MEGA3 [185]. Sequences are sorted from the order in the tree. The alignment was graphically pasted and precisely scaled to the tree illustration. 92% of red tree branches are SIDER2s while 76% of blue braches are SIDER1s, black branches in the middle contain similar proportions of both.

composition bias. Using the minimum evolution algorithm from the *MEGA3* program, we constructed a phylogeny based on the number of differences contained within parsimonious informative columns of the input alignment [185]. A similar methodology was supplied for the paper revealing the discovery of SIDERs [53]. **Figure 2** illustrates the resulting tree scaled to the sorted alignment. From this perspective, it is quite obvious that SIDERs form two distinct groups when glancing at the global sequence composition of the 2 main clusters. It appears such grouping authenticates SIDER taxonomy as upon closer inspection, 92% of sequences in one cluster are annotated as SIDER2 whereas 76% of the other cluster is composed of sequences tagged as SIDER1. In spite of this, there is a small cluster of SIDERs that contains practically equal amounts of both subgroups. The original, un-normalized phylogenetic tree can be viewed in **Appendix II**.

1.3. Lm*SIDER profiles*

From the alignment disclosed in **Figure 2**, it appears that SIDER1 elements are generally shorter than SIDER2s and regularly lack certain portions in the alignment. They also share lower pairwise sequence identity than the SIDER2 group. It can now be postulated with fair confidence that SIDERs form two separate subgroups. But what truly defines each class of SIDER at the primary structure level? To answer this question appropriately, sequences from both main clusters in **Figure 2** were split into distinct datasets. Improperly labelled sequences were discarded to avoid uncertainty (e.g. a SIDER2 sequence in the SIDER1 cluster). Both subgroups were subsequently de-gapped and aligned using the same parameters as previously described. Any sequences displaying a pairwise identity over 90% to another sequence was discarded. All other alignments were produced using these specifications, unless mentioned otherwise. The initial SIDER2 alignment was governed by a HMM profile modeled on the published manual alignment, however, in order to take advantage of its meticulous content. The resulting HMM profile was submitted to Baum-Welch EM for sake of consistency.

An advantage of using the *HMMER* software package is that the HMM profiles intrinsic to the creation of multiple alignments can also be used to scan sequences.

Consequently, scanning the individual unaligned sequences with the profiles for both SIDER subgroups enabled the comparison of their relative scores (**Figure 3**). The formation of distinct point clusters validates the selective nature of the HMM profiles, since little high-scoring sequences seldom overlap. Based on this observation, it is possible to establish a cut-off for classifying SIDER sequences with regard to their relative score for both profiles. This is important because the SIDER1 profile assigns positive values to some SIDER2 sequences. Many sequences score poorly, which may be due to faulty annotation (e.g. the length of certain SIDERs may have been overestimated). When considering hits scoring over 100 bits, there is potential evidence for one improperly labelled SIDER2 and 29 falsely labelled SIDER1s in the training sets. The corresponding sequences were swapped into the proper subgroup and the profiles were realigned using the same methodology as described above.

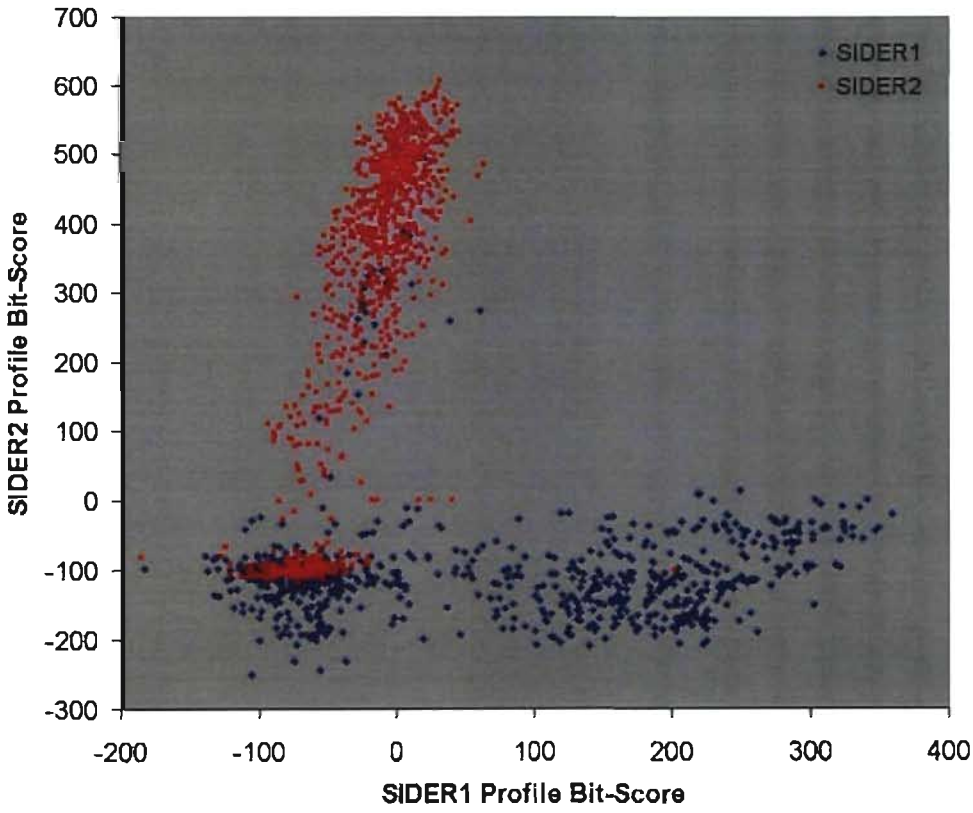

Figure 3. Selectivity scatter-plot of initial SIDER profiles

Unaligned input sequences were scanned with the initial HMM profile of 2 SIDER subgroups using the *hmms* global alignment command from *HMMER*. The bit-scores for each sequence are plotted in the bidimensional grid.

The final SIDER profiles expose additional information of the particularities of each subgroup. To illustrate these differences, the 40% consensus of both profiles was aligned using the global, pair-wise alignment tool in *BIOEDIT* (**Figure 4**). The SIDER1 consensus displays a notable gap when compared to the SIDER2 consensus, which corresponds to the second 79 nt retroposon signature previously reported for SIDER2s. It may not be evident when only considering **Figure 2,** but homologous positions between both subgroups can be isolated when comparing it with **Figure 4.** The region corresponding to the first 79 signature appears to be well conserved among both subgroups, just the same as the central portion of the alignments. The A-rich tail and a few residues preceding it are also well conserved among both profiles. Conversely, the second half of both SIDER profiles is quite divergent.

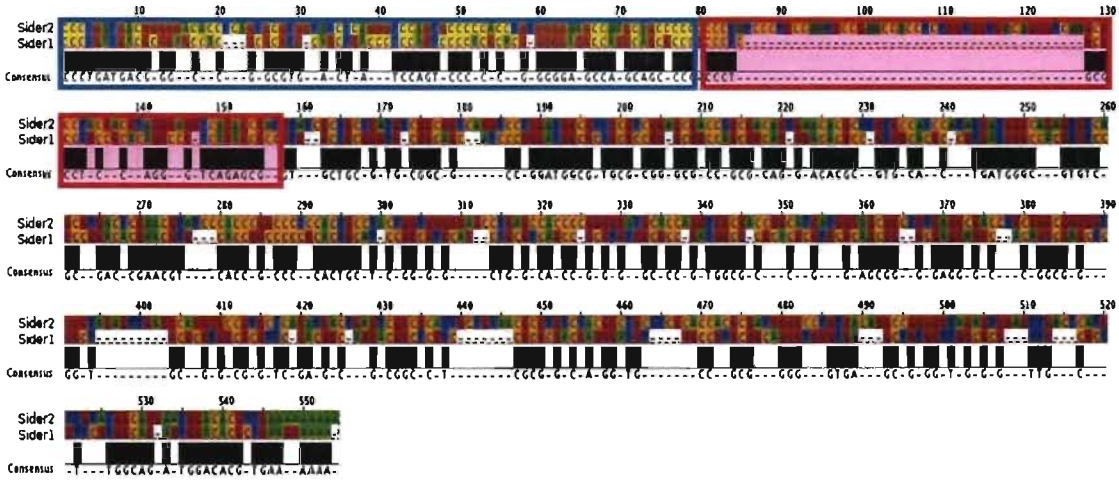

Figure 4. Pairwise alignment of distinct SIDER1 and SIDER2 subgroups

The aligned sequences each correspond to the 40% consensus of their respective subgroup. Black bars indicate an identical consensus position. The corresponding nucleotide appears under it. Shaded boxes highlight the dual 79 nt signature sequences described for *Lm*SIDER2 [53].

2. GENOMIC DISTRIBUTION

As mentioned above, HMM profiles can be used to scan sequences with the purpose of identifying regions that are homologous to the model. The main advantage of using HMM profiles resides in their rich information content. Instead of comparing one sequence to another, a group of sequences is compared to the target. As a consequence,

under-represented sequences can proportionately contribute to the search, hence increasing its sensitivity. Although the algorithms implemented in *HMMER* are much slower than search tools like *BLAST*, they are slightly less restrictive than the latter as *BLAST* requires that the query and target sequence share a small stretch of ungapped identity (usually 13 nt) [182]. These reasons justify the use of HMM profiles in the context of an iterative search strategy that aims to improve the identification and characterization of SIDERs in genomic data.

2.1. Building optimal search profiles

The SIDER profiles described in the previous section were used as a backbone to scan all three sequenced *Leishmania* genomes. We used the HMM fragment search command (*hmmfs*) included in the *HMMER* software package for all searches. This process returns a set of optimal, non-overlapping matches to a HMM in much the same way as the Smith-Waterman algorithm [186,189]. Since version 1.8.5 of *HMMER* does not provide expectation values (E-values) for search results, some means of testing the specificity of the search profiles was necessary. A 50 million base-pair synthetic genome was created randomly using nucleotide frequencies similar to that observed in the *L. major* genome (40% A/T; 60%GC). Both strands were scanned using the initial search profile for both SIDER subgroups. False positive statistics are displayed in **Table 1**. Results demonstrate that hits under 5 bits for the SIDER1 and the SIDER2 profile can be potential false positives. To test if false positives were caused by discrepancies in the HMM profile, the distribution of false positive hits in relation to the profile consensus were plotted (**Figure 5**). It appears that the initial portion of the profile detects more false positives than the rest of the profile. Indeed, almost all false-positives are short sequences less than 50 nt long (data not shown).

Table 1. False positive statistics for initial SIDER profiles

	Hits	**Average Bit-Score**	**Median Bit-Score**	**Standard Deviation**	**Maximum Value**
SIDER 1	67	1.25	1.10	1.01	5.05
SIDER 2	29	1.73	1.76	1.40	5.36

To accurately estimate the genomic distribution of SIDERs, search results from a first *hmmfs* round were used to create refined species-specific HMM profiles for both subgroups. The following genome versions were considered: *L. major* 5.2, *L. infantum* 3.0, *L. braziliensis* 2.0. Since *hmmfs* returns fragments of optimal matches, some hits are split into two or three results (large insertions in the target sequences are rejected by the algorithm). A custom JAVA script [190] was developed to concatenate such results.

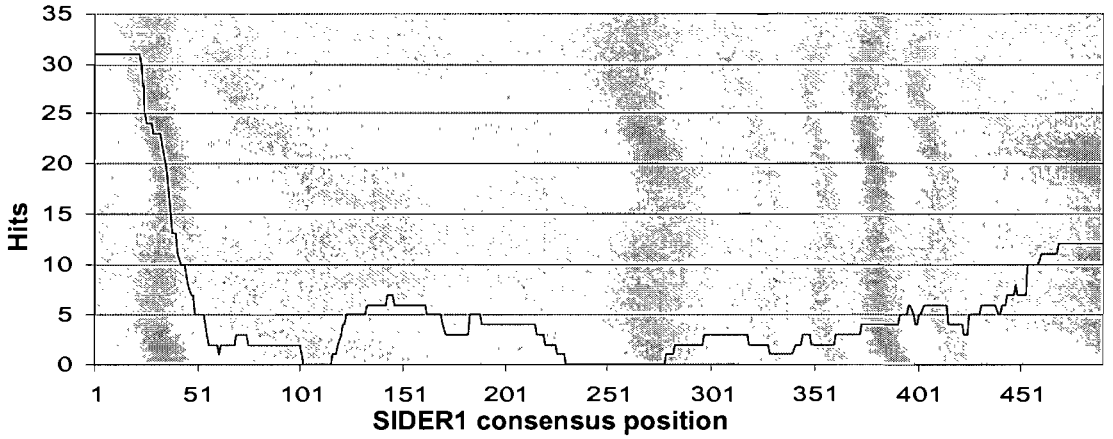

Figure 5. Position-specific profile susceptibility to false positives

The amount of false-positives is plotted relative to the consensus of the initial SIDER1 profile. SIDER2 profile hits are within the 1-50 consensus region of the SIDER2 profile (data not shown).

Three essential conditions were imposed for the incorporation of search hits into a species- and subgroup-specific refined HMM profile: (i) sequences must encompass 90% or more of the search profile's consensus; (ii) sequences must share less than 90% pairwise identity with any other sequence in the set of results, with the purpose of reducing compositional bias for an optimal iterative search profile; (iii) score over 50 bits. Empirical data suggests that subgroup discrimination based on the relative scores for each subgroup is reliable for overlapping hits over 50 (**Figure 3**). All hits were filtered and sorted using an *ad-hoc* JAVA script prior to being trained into the new refined HMM profiles. **Table 2** lists the amount of sequences included in all six refined profiles. The 40% consensus sequences of the refined profiles were aligned in order to evaluate the evolutionary relationship of SIDERs among the three *Leishmania* species (**Figure 6**).

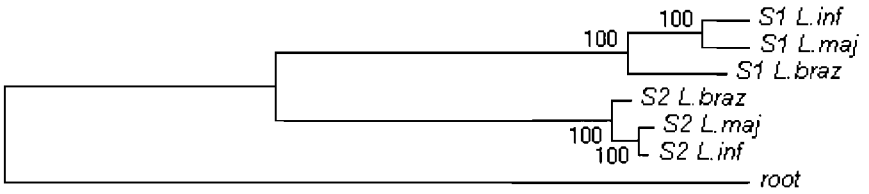

Figure 6. Inter-species SIDER similarity

Neighbour-joining phylogeny of the 40% consensus of the species-specific refined HMM profiles aligned with *CLUSTALW*. Bootstrap values for 500 replicates are indicated above branching points. A 550 nt random sequence was used to root the tree. S1 = SIDER1; S2 = SIDER2.

Table 2. Amount of full-length sequences in refined HMM profiles

	SIDER 1	SIDER 2
L. major	214	587
L. infantum	227	549
L. braziliensis	161	458

2.2. SIDER fragment distributions

The refined HMM profiles for both SIDER1 and SIDER2 subgroups were used to retrain genomic datasets for all three species. Results were concatenated and sorted as detailed above, although all hits were considered this round. A summary of results is presented in **Table 3**. As could be expected, more SIDER-related sequences are predicted with the refined HMM profiles than previously reported (785 SIDER1s, 1073 SIDER2s in *L. major* [53]).

Table 3. SIDER fragments in the genome of 3 *Leishmania* species

	SIDER1	SIDER2
Leishmania major	975	1278
Leishmania infantum	767	1284
Leishmania braziliensis	756	1467

Because a fragment search strategy was used, it is possible that certain profile portions are more abundant or conserved than others. The amount of hits per consensus position in the profile was studied to address this issue (**Figure 7**). It appears that SIDER2s are equally distributed throughout the genome in all three species. On the other hand, there appears to be a significant amount of small SIDER1 fragments bearing homology to the 400-450 region of the HMM profile in *L. major* and *L. infantum*, although not as abundant in the latter. This high copy number does not correlate with sequence conservation, as significantly fewer fragments remain when the threshold is raised. SIDER1 fragments in *L. braziliensis* appear to be uniformly distributed.

2.3. Genomic organization of SIDER fragments

It has been shown that *Lm*SIDER2s are preponderantly positioned in the intergenic regions of DGCs [53]. Most of these elements (i.e. 73%) have been postulated to lie within 3'UTRs. Such estimates are derived from mRNA extremity predictions based on an algorithm for the *Trypanosoma* genus [79]. By combining the more accurate algorithm described in chapter I with the search results from section 2.2, the proportion of SIDERs potentially enclosed in 3'UTRs can be assessed more assertively. **Figures 8**, **9** and **10** detail the occurrence of SIDER fragments throughout the three genomes, consequently assigning them to specific categories. On average, 58% of the SIDERs reported over the 5 bit threshold are predicted to be within the 3'UTR of the upstream coding sequence, regardless of their subgroup. This value varies less than 5% between both subgroups for all species.

SIDERs are present in both orientations of DNA. For sake of clarity, fragments in the same orientation as the CDS will be termed 'sense' (5'→3' in RNA), whereas 'antisense' fragments will designate those in the opposite orientation (complement 3'→5' in RNA). In *L. major* and *L. braziliensis*, there is roughly 10 times more sense SIDERs in predicted 3'UTRs than antisense SIDERs. This proportion reaches 5x in *L. infantum*. In contrast to this, SIDERs predicted to be in the IR downstream of the poly(A) site occur up to 3x more often as antisense, except for the *Lm*SIDER2 subgroup which presents equal proportions (**figure 8**). Interestingly, *L. infantum* intergenic SIDERs present the opposite ratios; 3x more sense fragments than antisense.

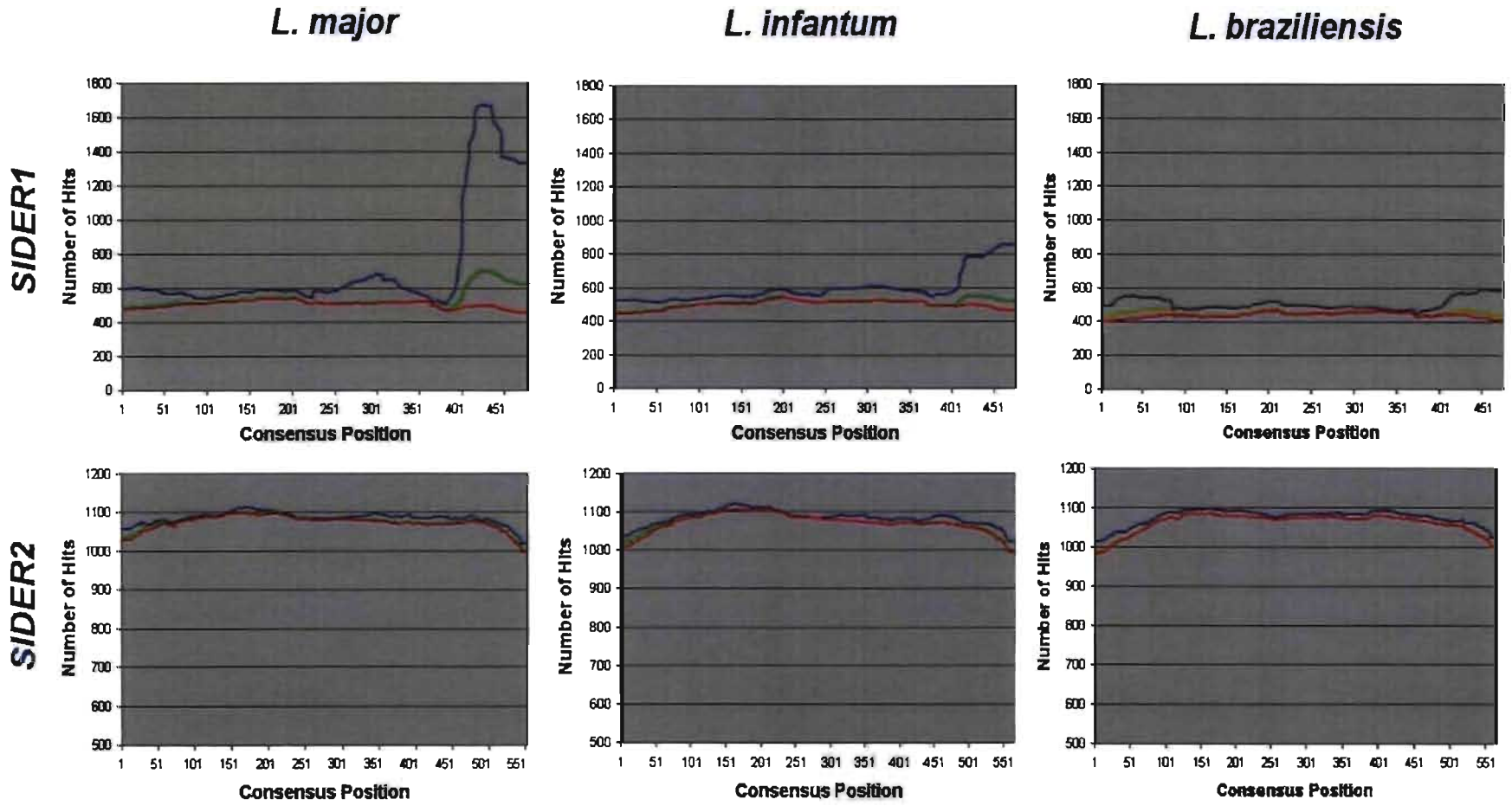

Figure 7. SIDER fragment distribution in the genomes of 3 *Leishmania* species

The amount of SIDER fragments are plotted relative to their position in the HMM profile consensus. The colours correspond to the amount of hits above a certain threshold: Blue = 0 bits; green = 5 bits; red = 10 bits.

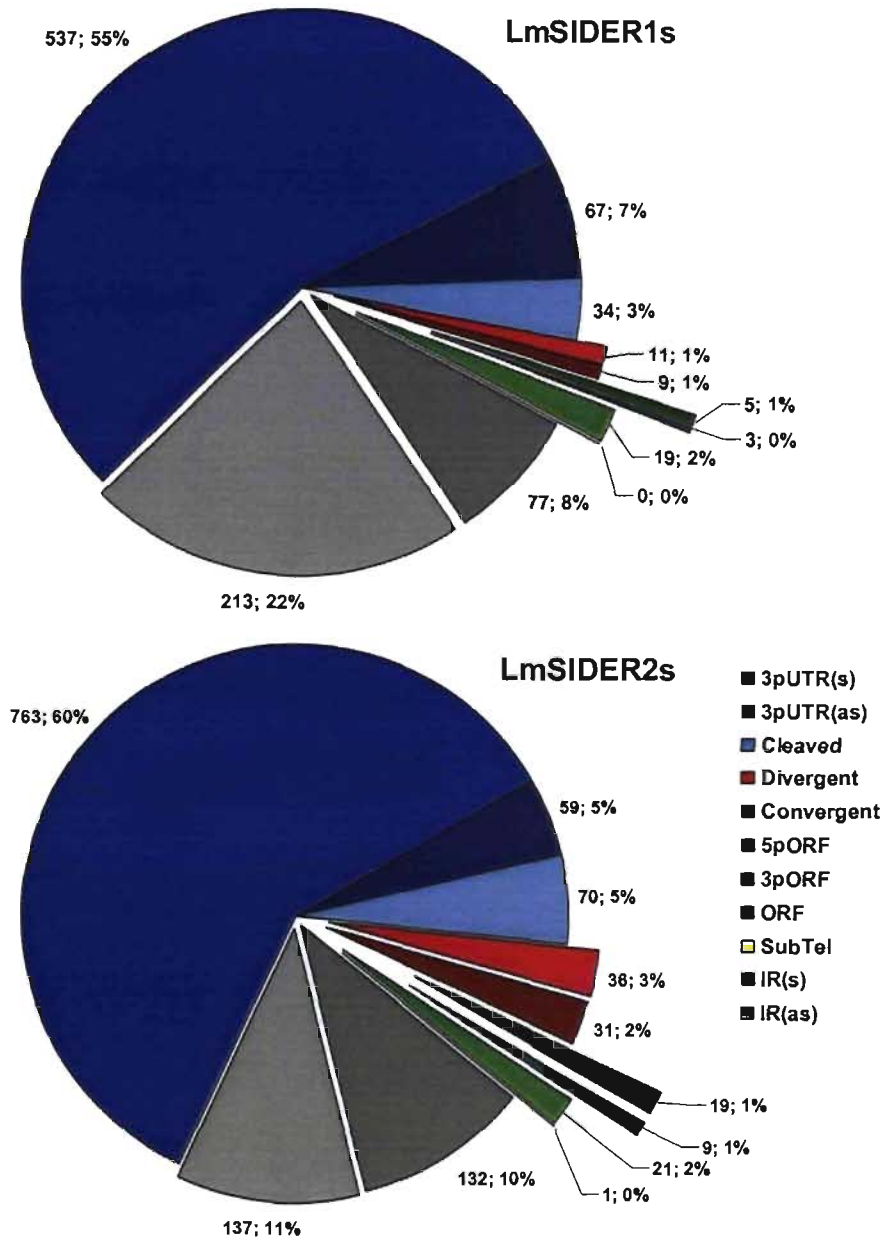

Figure 8. Genomic organization of *Lm*SIDERs

All SIDER fragments scoring over 5 bits are considered. UTR predictions made with the *PRED-A-TERM* program (c.f. Chapter I). Absolute and relative numerical values are provided next to each slice. Legend indicates virtual position of SIDER fragments in genome: *3pUTR(s)*, *3pUTR(as)* - in the predicted 3'UTR, either in same orientation as mRNA (*s*) or in opposite orientation (*as*); *Cleaved* – poly(A) site predicted in the SIDER (any orientation); *Divergent*, *Convergent* – divergent or convergent 'strand-switch' regions; *5pORF*, *3pORF*– overlapping the predicted ORF at 5' or 3' end respectively; *ORF* – enclosed in predicted ORF; *Subtel* – in the subtelomeric region; *IR(s)*, *IR(as)* – predicted outside of 3'UTR, in the intronic intergenic region.

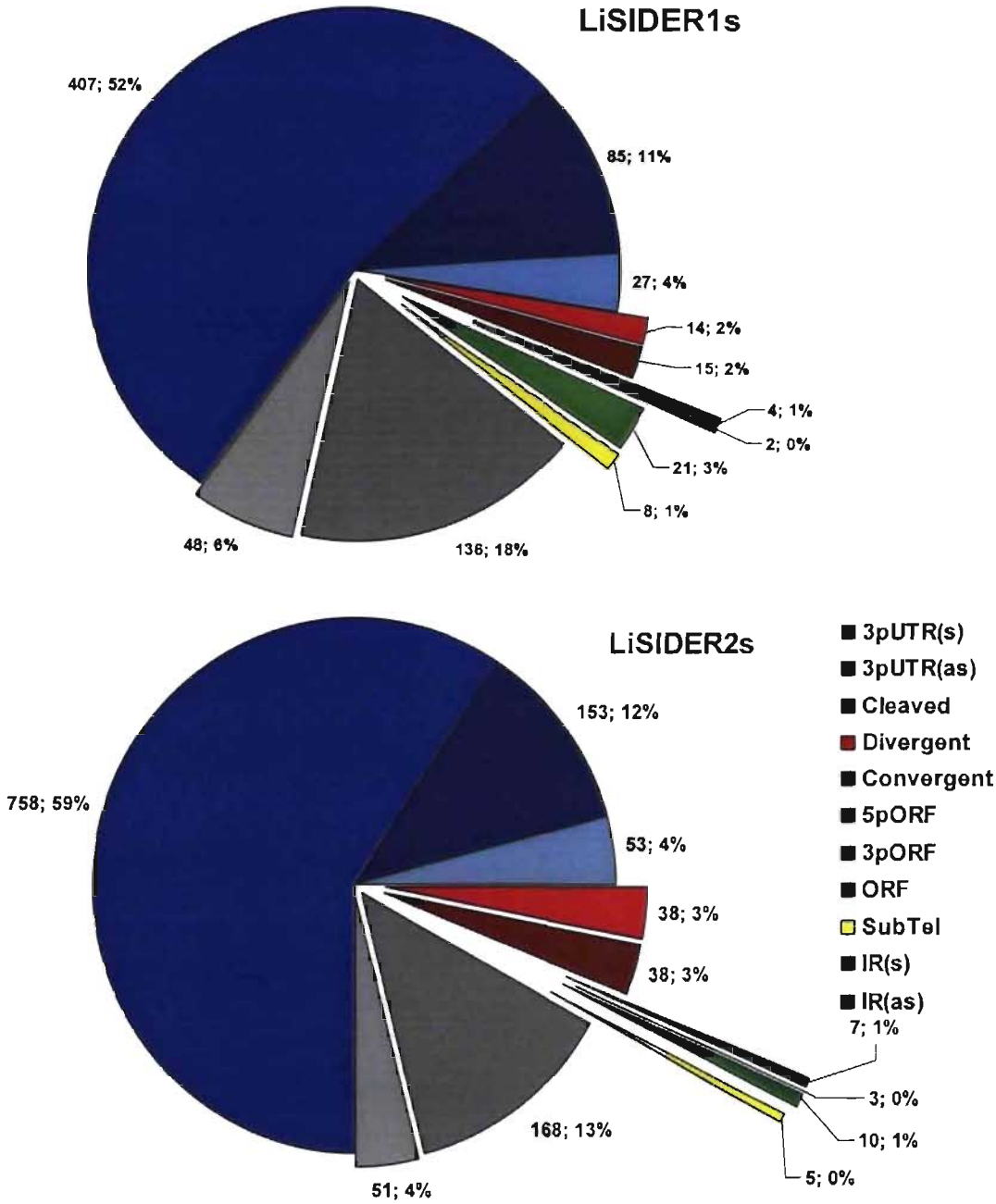

Figure 9. Genomic organization of *Li*SIDERs

All SIDER fragments scoring over 5 bits are considered. UTR predictions made with *PRED-A-TERM* program (c.f. chapter I). Absolute and relative numerical values are provided next to each slice. The legend symbols are the same as those in **figure 8**.

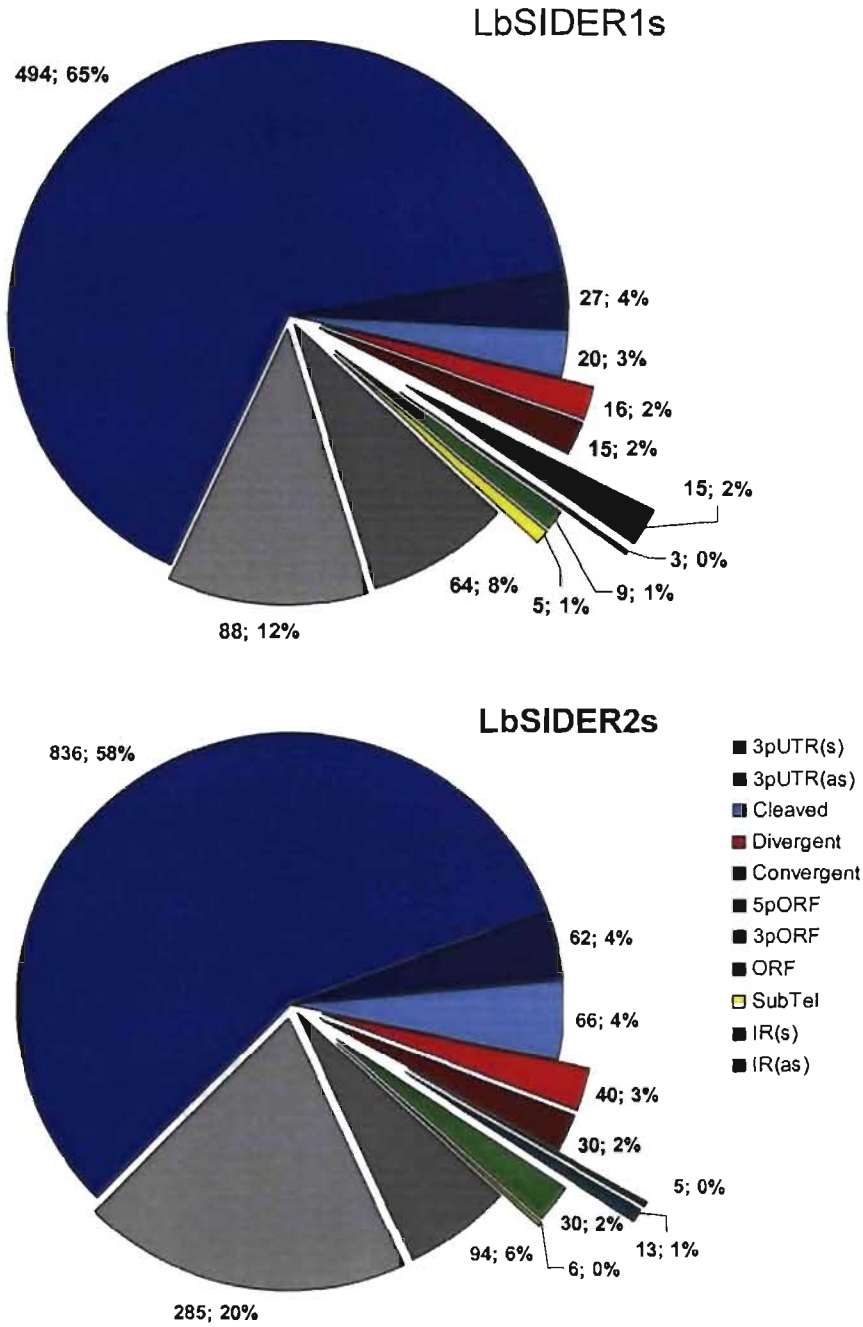

Figure 10. Genomic organization of *Lb*SIDERs

All SIDER fragments scoring over 5 bits are considered. UTR predictions made with *PRED-A-TERM* program (c.f. chapter I). Absolute and relative numerical values are provided next to each slice. The legend symbols are the same as those in **figure 8**.

Five to seven percent of SIDERs are within 'strand-switch' regions. *L. major* has approximately 133 such regions in its genome [34], 87 of which contain SIDERs (**Figure 8**). Values are higher for *L. infantum* (110) and *L. braziliensis* (141). These 'strand-switch' SIDERs are evenly distributed between convergent and divergent regions. Almost no SIDERs are present in subtelomeric regions; thus agreeing with previous reports [53]. There appears to be some occurrence of SIDERs in annotated coding sequences although the scores associated to these hits are weak. The upstream CDS are more often than not hypothetical proteins (data not shown).

3. DISCUSSION

The use of HMMs for optimizing multiple sequence alignments and conducting precise homology searches is very convenient for comparative analyses of biological sequences. Granted that HMM training is not the fastest method to align divergent sequences, it is seemingly the best approach to generate reliable SIDER alignments. The approach used to compare the effectiveness of the different alignment tools should not, however, be considered a rigorous benchmark of MSA quality. Ideally, entire alignments should be considered; not just the consensus. There do exist a handful of MSA benchmark tools (reviewed in [191]), but the intuitive nature of the presented analysis and its convincing outcome fulfil the problem at hand.

The alignment of all *Lm*SIDER sequences in **Figure 2** provides additional evidence that there are two subgroups of these assimilated retroposons. Grouping SIDER elements strictly on the provided phylogeny is a somewhat credulous insinuation, as the exceedingly divergent nature of SIDER sequences provides little sturdiness to the tree (**Appendix II**). In fact, this observation hindered bootstrapping validation and diminished overall confidence in the precision of the tree. Conversely, sorting sequences in the alignment synchronously with the tree topology allows for visual substantiation of the proposed SIDER classification based on two factors: (i) the proportion of SIDERs in both clusters; (ii) distinguishing patterns of sequence conservation. The first factor respects the initial subgroup annotation. The latter is validated by the abundance of indels in the top cluster. At this point, a rigorous

phylogenetic analysis was not deemed crucial to the analysis as the main purpose of this breakdown is for the construction of representative search profiles, a process that significantly reduces the amount of retained sequences (**Table 2**).

It was tempting to apply a purely statistical approach to SIDER subgroup modeling. For instance, this can be achieved by splitting sequences into two groups solely from phylogenetic data by the subsequent creation of optimal HMM profiles for both subgroups. Further isolation founded on the each sequence's relative score for both HMM profiles can yield discriminative profiles. Such an approach can be considered as an archetypal methodology for naïve sequences. On the other hand, since there is prior information on the nature of SIDER sequences in this experiment, the methodology described in the abovementioned sections attempts to combine established information with impartial statistical assessment of distinctive sequence traits. These considerations motivated the removal of incongruent sequences from both phylogeny-derived datasets.

Based on the consensus of the initial HMM profiles and on the alignment of all published *Lm*SIDERs sequences (**Figures 2** and **4**), *Lm*SIDER1s appear to lack the second 79 nt signature motif of *Lm*SIDER2. Representing the most conspicuous difference between subgroups, this deletion (or duplication) reveals a potential target for *in-vivo* functional studies. The 79 nt signature sequence is all the more appealing since it is known to harbour transcription initiation factors in *Trypanosoma* [128]. However, seeing that the consensus is obtained from initial search profiles, it could be that the refined profiles tell a different story. Fortunately, the multiple alignment of the refined profile consensus used to build the tree displayed in **Figure 6** exhibits very similar characteristics to **Figure 4** (data not shown). Another region that contributes to the differentiation of both subgroups is the divergent region between the middle and the end of the alignment. It is conceivable that this region may also confer different functions to both SIDER subgroups. **Figure 6** also suggests that SIDER1 and SIDER2 possibly diverged before speciation events; however this insinuation may be biased since *L. major* sequences were used for the initial search.

The overlapping hits observed when scanning the sequences in both training sets may appear problematic in that they cannot be associated to a particular profile

(**Figure 3**). The fact that almost all negative hits are overlapping hints to either the incorrect annotation of the associated SIDER elements or to the presence of large insertions or deletions. In fact, the scoring model used in the *hmms* program implements a global alignment algorithm, therefore any sequences containing superfluous nucleotide stretches will be penalized. The *hmmfs* command used to accomplish genomic scans uses a local search algorithm which unmistakably discriminated both subgroups within the training sequences (data not shown). All full-length sequences were nonetheless retained in the initial profiles in order to loosen the stringency of the initial searches. This may account for the presence of false positives in the control scan (**Table 1** and **Figure 5**). Nevertheless, these scores are not worrisome and their quantity may be over-represented as *Leishmania* genomes encompass ~32 million base-pairs. Since the refined search profiles consider only full-length results from the fragment search strategy, their specificity should definitely be stronger even though this was not tested (scanning 50 million base-pairs with *hmmfs* demands vast amounts of computation time). Doing so would provide custom threshold delineations for each species.

Considering the observation that SIDER1 sequences are largely characterized by the presence of only one 79 nt signature, the foremost portion of SIDER1 profiles can potentially recognize SIDER2 sequences. Indeed, some SIDER1s contain what can be considered as insertions after the first 79 nt signature, thus potentially aligning with the second signature motif of SIDER2s. In contrast, the more conserved nature of SIDER2s displays low incidence of deletions in the same region. Furthermore, since there are more SIDER2 sequences in the search profiles than SIDER1s, these regions are disproportionately represented. As a consequence, the SIDER1 profile picks-up many SIDER2 sequences in genomic scans. A custom *JAVA* script was created to compare any overlapping hits for both profiles; the score ratio of both hits is weighted and hits are classified as SIDER1 or 2 appropriately. Overlapping hits are almost always harshly unbalanced and scores are near null when this is not the case.

Genomic scanning using fine-tuned HMM profiles reveals that SIDER elements are more abundant than previously reported. Straightforward genomes queries using

single sequences and speedy heuristic search tools indubitably lack the depth of profile-based searches. Their use is most valuable for initial estimations of sequence homology and abundance, but their reliability weakens when precision is necessary. The fragment search tactic we employed identified ~20% more SIDER2-related sequences and ~30% more SIDER1-related sequences in the genome of *L. major* (**Figure 8**). A fragment search algorithm produces multiple hits when the genomic target contains large insertions, as this produces a higher score resulting from the low abundance of large insertions in the profile consensus. These regions may be neglected by global alignment strategies as they produce negative scores, whereas *BLAST* and related programs may overlook these regions altogether if the homologous regions do not share enough identity. A drawback of using the fragment search approach is that post-processing is required to concatenate these interrupted hits in the genomic data. Insertions were rarely larger than 500 nt, so a *JAVA* script was conceived to concatenate interrupted hits from the HMM consensus (e.g. two matches spanning positions 1-200 and 201-500 of the profile consensus separated by 150 nt are concatenate into one 650 nt hit and their scores are combined). By filtering all search results with this script, the fragment-induced bias was significantly reduced from the projected quantities.

An abundance of low-scoring SIDER1 termini has been identified in *L. major* and *L. infantum*. As mentioned above, refined HMM profiles should flaunt higher specificity than that of the initial search profile. It therefore seems unlikely that these matches are false positives: a supposition that is reinforced by the sheer profusion of the hits (almost 1700 copies). Considering the abundance of this region in contrast to its low conservation (i.e. low bit-score), we speculate that his portion of the SIDER1 consensus may harbour a conserved secondary structure. Granted profile specificity testing is required to ascertain this hypothesis, a conserved RNA structure may present weak primary structure conservation provided it forms a functional secondary or tertiary structure. For instance, structural components that interact with other molecules are submitted to different selection pressure than scaffolding components (e.g. the anticodon in tRNAs vs. any helix). This potential discovery is significant as it has been shown that some genes containing SIDER1 elements in their 3'UTR are developmentally regulated by conserved regions in 3'UTRs [50,52]. A similar

happening may take place *L. infantum* which displays ~850 low-scoring copies of the same region. The marginal amounts of such fragments in *L. braziliensis* can be explained by the exploitation of alternative regulatory mechanisms in this species. Indeed, RNA interference has been reported for this species which contains an apparent homolog of the Argonaut protein in its genome [34,51].

The most beneficial aspect of this thesis in regards to the functional characterization of SIDER elements is the improved accuracy of 3'UTR prediction. Using the PRED-A-TERM program described in Chapter I, the approximate proportion of SIDERs contained in 3'UTRs has been rectified to exclude potentially rubbish interspersed repeats. The initial estimate placed 73% of SIDER2s in 3'UTRs. This work advocates that the proportion is closer to 58% in spite of the higher incidence of SIDER fragments. The previous mRNA processing site predictor was developed from cDNA statistics for the *Trypanosoma* genus which is known to have much shorter intergenic regions and for whom polyadenylation occurs much closer to the splice junction. The resulting approximations for *Leishmania major* are skewed toward *Trypanosoma* parameters, thus justifying the higher value. Moreover, it is now clear that SIDERs of all genres are preferentially located in the 3'UTR for all three *Leishmania* species.

This work also reveals certain particularities concerning the orientation of SIDERs with regards to their genomic context. Most interestingly, the majority of reverse-complemented (or antisense) SIDERs are predicted to lie beyond the 3'UTR (**Figures 8, 9, 10**). In addition to the observation that 80-95% of SIDERs expected to reside in 3'UTRs are in the same orientation as the coding-sequence, it appears that correct orientation may be required for proper regulatory function (assuming that SIDER elements carry out such a role). Another possibility is that SIDERs might impact polyadenylation either by the interference of the terminal adenosine stretch or by some other unknown mechanism. Poly(A) predictions do not appear to be over-represented near SIDER extremities therefore we can safely reject the first hypothesis (data not shown). There are no known reports of SIDER-related mechanisms affecting polyadenylation or *trans*-splicing. Obviously, the *PRED-A-TERM* program does not emit perfect UTR predictions for reasons detailed in Chapter I. Assuming that the

presence of SIDERs in the IR does not affect polyadenylation and/or the prediction program, the UTR predictions made by *PRED-A-TERM* should not vary disproportionately to the true situation.

Slightly more SIDERs have been identified in strand-switch regions than previously reported. About 65% of strand-switch regions contain SIDER elements in *L. major*. This proportion is estimated from a previous version of the genome annotation, yet should not vary significantly. Such a high proportion correlates with the observation that the 79 nt signature can promote transcription initiation is *Trypanosoma* [128], as strand-switch regions are the general transcription initiation sites for trypanosomatids. It is also conceivable that, due to their repetitive nature, SIDERs may be accountable for homologous recombination events. Unfortunately, the total proportion of strand-switch regions containing SIDERs was not verified for the other two *Leishmania* species. It should nevertheless be similar seeing as total SIDER proportions vary slightly among all three species.

4. CONCLUDING REMARKS AND PERSPECTIVES

The short interspersed degenerated retroposons of *Leishmania* are abundant throughout all three sequenced genomes, yet present varying degrees of conservation and fragmentation. Such characteristics render *in-silico* analyses challenging and require accurate computational modeling to carry them out. The results set forth in this thesis convey the effectiveness of hidden Markov models towards surmounting this predicament.

Statistical optimization of multiple sequence alignments combined with phylogenetic support facilitated the characterization of two SIDER subgroups. Independent search profiles were created from these subgroups in order to scan all three sequenced *Leishmania* species in an iterative manner. This approach resulted in an increased estimation of the amount of SIDER fragments in *L. major* while providing reliable estimates for *L. infantum* and *L. braziliensis*.

Using a comparative approach, this work also presents a novel approach for predicting UTRs in *Leishmania* with more than double the accuracy of previous methods. This outcome is all the more commendable seeing as polyadenylation appears to be rather unspecific in trypanosomatids. Combined with the aforesaid search results, this tool allows for enhanced identification of SIDERs contained within mRNA transcripts. Given that the total amount of SIDER fragments is overwhelming for most structural motif detection programs, we can safely assert that, although not perfect, the *PRED-A-TERM* program helps target a subset of potentially functional SIDER elements.

Regrettably, no further investigation of conserved motifs is accounted for in this work due to time constraints. However, most preliminary analyses and pertinent tools for this purpose have been expounded. Evidently, perspective work should focus on further characterization of SIDER sequences predicted to be within regulatory regions at the level of primary and secondary structure, ideally in both an *in-silico* and *in-vivo* framework.

BIBLIOGRAPHY

1. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al.: **Comparative genomic analysis of three Leishmania species that cause diverse human disease**. *Nat Genet* 2007, **39**:839-847.
2. Weigle K, Saravia NG: **Natural history, clinical evolution, and the host-parasite interaction in New World cutaneous Leishmaniasis**. *Clin Dermatol* 1996, **14**:433-450.
3. Desjeux P: **Leishmaniasis: current situation and new perspectives**. *Comp Immunol Microbiol Infect Dis* 2004, **27**:305-318.
4. Guerin PJ, Olliaro P, Nosten F, Druilhe P, Laxminarayan R, Binka F, Kilama WL, Ford N, White NJ: **Malaria: current status of control, diagnosis, treatment, and a proposed agenda for research and development**. *Lancet Infect Dis* 2002, **2**:564-573.
5. Enserink M: **Infectious diseases. Has leishmaniasis become endemic in the U.S.?** *Science* 2000, **290**:1881-1883.
6. McHugh CP, Thies ML, Melby PC, Yantis LD, Jr., Raymond RW, Villegas MD, Kerr SF: **Short report: a disseminated infection of Leishmania mexicana in an eastern woodrat, Neotoma floridana, collected in Texas**. *Am J Trop Med Hyg* 2003, **69**:470-472.
7. Rosypal AC, Troy GC, Zajac AM, Duncan RB, Jr., Waki K, Chang KP, Lindsay DS: **Emergence of zoonotic canine leishmaniasis in the United States: isolation and immunohistochemical detection of Leishmania infantum from foxhounds from Virginia**. *J Eukaryot Microbiol* 2003, **50 Suppl**:691-693.
8. About Sir William Leishman on World Wide Web URL: http://leishman.cent.gla.ac.uk/william.htm
9. Cox FE: **History of human parasitology**. *Clin Microbiol Rev* 2002, **15**:595-612.
10. Center for Disease Control on World Wide Web URL: http://www.cdc.gov
11. Adler S, Ber, M. : **The transmission of *Leishmania tropica* by the bite of *Phlebotomus papatasi*.** . *Indian J. Med. Res.* 1941, **29**:803-809.
12. Ouellette M, Drummelsmith J, Papadopoulou B: **Leishmaniasis: drugs in the clinic, resistance and new developments**. *Drug Resist Updat* 2004, **7**:257-266.
13. Croft SL, Sundar S, Fairlamb AH: **Drug resistance in leishmaniasis**. *Clin Microbiol Rev* 2006, **19**:111-126.
14. Garg R, Dube A: **Animal models for vaccine studies for visceral leishmaniasis**. *Indian J Med Res* 2006, **123**:439-454.
15. Khamesipour A, Rafati S, Davoudi N, Maboudi F, Modabber F: **Leishmaniasis vaccine candidates for development: a global overview**. *Indian J Med Res* 2006, **123**:423-438.
16. Summers K, McKeon S, Sellars J, Keusenkothen M, Morris J, Gloeckner D, Pressley C, Price B, Snow H: **Parasitic exploitation as an engine of diversity**. *Biol Rev Camb Philos Soc* 2003, **78**:639-675.
17. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M, et al.: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba**. *Proc Natl Acad Sci U S A* 2002, **99**:1414-1419.

18. Baldauf SL: **The deep roots of eukaryotes**. *Science* 2003, **300**:1703-1706.
19. Burger G, Gray MW, Lang BF: **Mitochondrial genomes: anything goes**. *Trends Genet* 2003, **19**:709-716.
20. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading**. *Syst. Zool.* 1978, **27**:401-410.
21. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F: **Heterotachy and long-branch attraction in phylogenetics**. *BMC Evol Biol* 2005, **5**:50.
22. Gribaldo S, Philippe H: **Ancient phylogenetic relationships**. *Theor Popul Biol* 2002, **61**:391-408.
23. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life**. *Nat Rev Genet* 2005, **6**:361-375.
24. Lukes J, Guilbride DL, Votypka J, Zikova A, Benne R, Englund PT: **Kinetoplast DNA network: evolution of an improbable structure**. *Eukaryot Cell* 2002, **1**:495-502.
25. Shlomai J: **The structure and replication of kinetoplast DNA**. *Curr Mol Med* 2004, **4**:623-647.
26. Gibson W: **Sex and evolution in trypanosomes**. *Int J Parasitol* 2001, **31**:643-647.
27. Smith DF, Parsons, M.: *Molecular Biology of Parasitic Protozoa*: Oxford University Press; 1996.
28. Tibayrenc M, Ayala FJ: **Evolutionary genetics of Trypanosoma and Leishmania**. *Microbes Infect* 1999, 1:465-472.
29. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al.: **Comparative genomics of trypanosomatid parasitic protozoa**. *Science* 2005, **309**:404-409.
30. GeneDB : Leishmania major sequence database. on World Wide Web URL: http://www.genedb.org/genedb/leish
31. Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, et al.: **Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes**. *Proc Natl Acad Sci U S A* 1999, **96**:2902-2906.
32. Worthey EA, Martinez-Calvillo S, Schnaufer A, Aggarwal G, Cawthra J, Fazelinia G, Fong C, Fu G, Hassebrock M, Hixson G, et al.: **Leishmania major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene**. *Nucleic Acids Res* 2003, **31**:4201-4210.
33. Monnerat S, Martinez-Calvillo S, Worthey E, Myler PJ, Stuart KD, Fasel N: **Genomic organization and gene expression in a chromosomal region of Leishmania major**. *Mol Biochem Parasitol* 2004, **134**:233-243.
34. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, et al.: **The genome of the kinetoplastid parasite, Leishmania major**. *Science* 2005, **309**:436-442.
35. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI, et al.: **A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA**. *Rna* 2000, **6**:163-169.
36. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes**. *Annu Rev Genet* 2000, **34**:77-137.
37. Campbell DA, Thomas S, Sturm NR: **Transcription in kinetoplastid protozoa: why be normal?** *Microbes Infect* 2003, **5**:1231-1240.

38. Clayton CE: **Life without transcriptional control? From fly to man and back again.** *Embo J* 2002, **21**:1881-1888.
39. Palenchar JB, Bellofatto V: **Gene transcription in trypanosomes.** *Mol Biochem Parasitol* 2006, **146**:135-141.
40. Downey N, Donelson JE: **Search for promoters for the GARP and rRNA genes of Trypanosoma congolense.** *Mol Biochem Parasitol* 1999, **104**:25-38.
41. McAndrew M, Graham S, Hartmann C, Clayton C: **Testing promoter activity in the trypanosome genome: isolation of a metacyclic-type VSG promoter, and unexpected insights into RNA polymerase II transcription.** *Exp Parasitol* 1998, **90**:65-76.
42. Bellofatto V, Torres-Munoz JE, Cross GA: **Stable transformation of Leptomonas seymouri by circular extrachromosomal elements.** *Proc Natl Acad Sci U S A* 1991, **88**:6711-6715.
43. Curotto de Lafaille MA, Laban A, Wirth DF: **Gene expression in Leishmania: analysis of essential 5' DNA sequences.** *Proc Natl Acad Sci U S A* 1992, **89**:2703-2707.
44. Martinez-Calvillo S, Nguyen D, Stuart K, Myler PJ: **Transcription initiation and termination on Leishmania major chromosome 3.** *Eukaryot Cell* 2004, **3**:506-517.
45. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ: **Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region.** *Mol Cell* 2003, **11**:1291-1299.
46. Tosato V, Ciarloni L, Ivens AC, Rajandream MA, Barrell BG, Bruschi CV: **Secondary DNA structure analysis of the coding strand switch regions of five Leishmania major Friedlin chromosomes.** *Curr Genet* 2001, **40**:186-194.
47. Wong AK, Curotto de Lafaille MA, Wirth DF: **Identification of a cis-acting gene regulatory element from the lemdr1 locus of Leishmania enriettii.** *J Biol Chem* 1994, **269**:26497-26502.
48. De Gaudenzi J, Frasch AC, Clayton C: **RNA-binding domain proteins in Kinetoplastids: a comparative analysis.** *Eukaryot Cell* 2005, **4**:2106-2114.
49. Jager AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC: **mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes.** *Proc Natl Acad Sci U S A* 2007, **104**:2035-2042.
50. McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, Dube M, Papadopoulou B: **Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in Leishmania.** *J Biol Chem* 2005, **280**:35238-35246.
51. Haile S, Papadopoulou, B.: **Developmental regulation of gene expression in trypanosomatid parasitic protozoa.** *Current Opinion in Microbiology* 2007:In press.
52. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, Papadopoulou B: **A common mechanism of stage-regulated gene expression in Leishmania mediated by a conserved 3'-untranslated region element.** *J Biol Chem* 2002, **277**:19511-19520.
53. Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, Papadopoulou B, Ghedin E: **Members of a large retroposon family are**

determinants of post-transcriptional gene expression in Leishmania. *PLoS Pathog* 2007, **3**:1291-1307.

54. Holzer TR, Mishra KK, Lebowitz JH, Forney JD: **Coordinate regulation of a family of promastigote-enriched mRNAs by the 3'UTR PRE element in Leishmania mexicana.** *Mol Biochem Parasitol* 2008, **157**:54-64.
55. Charest H, Zhang WW, Matlashewski G: **The developmental expression of Leishmania donovani A2 amastigote-specific genes is post-transcriptionally mediated and involves elements located in the 3'-untranslated region.** *J Biol Chem* 1996, **271**:17081-17090.
56. Murray A, Fu C, Habibi G, McMaster WR: **Regions in the 3' untranslated region confer stage-specific expression to the Leishmania mexicana a600-4 gene.** *Mol Biochem Parasitol* 2007, **153**:125-132.
57. Aly R, Argaman M, Halman S, Shapira M: **A regulatory role for the 5' and 3' untranslated regions in differential expression of hsp83 in Leishmania.** *Nucleic Acids Res* 1994, **22**:2922-2929.
58. Clayton C, Shapira M: **Post-transcriptional regulation of gene expression in trypanosomes and leishmanias.** *Mol Biochem Parasitol* 2007, **156**:93-101.
59. Di Noia JM, D'Orso I, Sanchez DO, Frasch AC: **AU-rich elements in the 3'-untranslated region of a new mucin-type gene family of Trypanosoma cruzi confers mRNA instability and modulates translation efficiency.** *J Biol Chem* 2000, **275**:10218-10227.
60. Larreta R, Soto M, Quijada L, Folgueira C, Abanades DR, Alonso C, Requena JM: **The expression of HSP83 genes in Leishmania infantum is affected by temperature and by stage-differentiation and is regulated at the levels of mRNA stability and translation.** *BMC Mol Biol* 2004, **5**:3.
61. Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M: **Developmental regulation of heat shock protein 83 in Leishmania. 3' processing and mRNA stability control transcript abundance, and translation id directed by a determinant in the 3'-untranslated region.** *J Biol Chem* 2001, **276**:47922-47929.
62. Furger A, Schurch N, Kurath U, Roditi I: **Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of Trypanosoma brucei by modulating RNA stability and translation.** *Mol Cell Biol* 1997, **17**:4372-4380.
63. Beverley SM: **Protozomics: trypanosomatid parasite genetics comes of age.** *Nat Rev Genet* 2003, **4**:11-19.
64. Boothroyd JC, Cross GA: **Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end.** *Gene* 1982, **20**:281-289.
65. Parsons M, Nelson RG, Watkins KP, Agabian N: **Trypanosome mRNAs share a common 5' spliced leader sequence.** *Cell* 1984, **38**:309-316.
66. De Lange T, Liu AY, Van der Ploeg LH, Borst P, Tromp MC, Van Boom JH: **Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription?** *Cell* 1983, **34**:891-900.
67. Murphy WJ, Watkins KP, Agabian N: **Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing.** *Cell* 1986, **47**:517-525.

68. Sutton RE, Boothroyd JC: **Evidence for trans splicing in trypanosomes.** *Cell* 1986, **47**:527-535.
69. Mayer MG, Floeter-Winter LM: **Pre-mRNA trans-splicing: from kinetoplastids to mammals, an easy language for life diversity.** *Mem Inst Oswaldo Cruz* 2005, **100**:501-513.
70. Bangs JD, Crain PF, Hashizume T, McCloskey JA, Boothroyd JC: **Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides.** *J Biol Chem* 1992, **267**:9805-9815.
71. Lucke S, Xu GL, Palfi Z, Cross M, Bellofatto V, Bindereif A: **Spliced leader RNA of trypanosomes: in vivo mutational analysis reveals extensive and distinct requirements for trans splicing and cap4 formation.** *Embo J* 1996, **15**:4380-4391.
72. Bruzik JP, Van Doren K, Hirsh D, Steitz JA: **Trans splicing involves a novel form of small nuclear ribonucleoprotein particles.** *Nature* 1988, **335**:559-562.
73. Liang XH, Haritan A, Uliel S, Michaeli S: **trans and cis splicing in trypanosomatids: mechanism, factors, and regulation.** *Eukaryot Cell* 2003, **2**:830-840.
74. Burge CT, T., Sharp, P. A.: **Splicing of precursors to mRNAs by the spliceosome.** In *The RNA World.* Edited by: Cold Spring Harbor Press; 1999:525-559.
75. LeBowitz JH, Smith HQ, Rusche L, Beverley SM: **Coupling of poly(A) site selection and trans-splicing in Leishmania.** *Genes Dev* 1993, **7**:996-1007.
76. Ullu E, Matthews KR, Tschudi C: **Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts.** *Mol Cell Biol* 1993, **13**:720-725.
77. Matthews KR, Tschudi C, Ullu E: **A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes.** *Genes Dev* 1994, **8**:491-501.
78. Vassella E, Braun R, Roditi I: **Control of polyadenylation and alternative splicing of transcripts from adjacent genes in a procyclin expression site: a dual role for polypyrimidine tracts in trypanosomes?** *Nucleic Acids Res* 1994, **22**:1359-1364.
79. Benz C, Nilsson D, Andersson B, Clayton C, Guilbride DL: **Messenger RNA processing sites in Trypanosoma brucei.** *Mol Biochem Parasitol* 2005, **143**:125-134.
80. Clayton CE, Ha S, Rusche L, Hartmann C, Beverley SM: **Tests of heterologous promoters and intergenic regions in Leishmania major.** *Mol Biochem Parasitol* 2000, **105**:163-167.
81. Schurch N, Hehl A, Vassella E, Braun R, Roditi I: **Accurate polyadenylation of procyclin mRNAs in Trypanosoma brucei is determined by pyrimidine-rich elements in the intergenic regions.** *Mol Cell Biol* 1994, **14**:3668-3675.
82. Hug M, Hotz HR, Hartmann C, Clayton C: **Hierarchies of RNA-processing signals in a trypanosome surface antigen mRNA precursor.** *Mol Cell Biol* 1994, **14**:7428-7435.
83. Wassarman KM, Steitz JA: **Association with terminal exons in pre-mRNAs: a new role for the U1 snRNP?** *Genes Dev* 1993, **7**:647-659.

84. Britten RJ, Kohne DE: **Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms.** *Science* 1968, **161**:529-540.
85. Toth G, Gaspari Z, Jurka J: **Microsatellites in different eukaryotic genomes: survey and analysis.** *Genome Res* 2000, **10**:967-981.
86. Thomas EE: **Short, local duplications in eukaryotic genomes.** *Curr Opin Genet Dev* 2005, **15**:640-644.
87. Jurka J, Kapitonov VV, Kohany O, Jurka MV: **Repetitive sequences in complex genomes: structure and evolution.** *Annu Rev Genomics Hum Genet* 2007, **8**:241-259.
88. Sullivan BA, Blower MD, Karpen GH: **Determining centromere identity: cyclical stories and forking paths.** *Nat Rev Genet* 2001, **2**:584-596.
89. Sunkel CE, Coelho PA: **The elusive centromere: sequence divergence and functional conservation.** *Curr Opin Genet Dev* 1995, **5**:756-767.
90. Wickstead B, Ersfeld K, Gull K: **Repetitive elements in genomes of parasitic protozoa.** *Microbiol Mol Biol Rev* 2003, **67**:360-375, table of contents.
91. Dubessay P, Ravel C, Bastien P, Stuart K, Dedet JP, Blaineau C, Pages M: **Mitotic stability of a coding DNA sequence-free version of Leishmania major chromosome 1 generated by targeted chromosome fragmentation.** *Gene* 2002, **289**:151-159.
92. Tamar S, Papadopoulou B: **A telomere-mediated chromosome fragmentation approach to assess mitotic stability and ploidy alterations of Leishmania chromosomes.** *J Biol Chem* 2001, **276**:11662-11673.
93. Mc CB: **The origin and behavior of mutable loci in maize.** *Proc Natl Acad Sci U S A* 1950, **36**:344-355.
94. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 1953, **171**:737-738.
95. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al.: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419**:498-511.
96. Lewin B: **Retroviruses and Retrotransposons.** In *Genes VI.* Edited by: Oxford University Press; 1997.
97. Craig NL: **Unity in transposition reactions.** *Science* 1995, **270**:253-254.
98. Haring E, Hagemann S, Pinsker W: **Ancient and recent horizontal invasions of drosophilids by P elements.** *J Mol Evol* 2000, **51**:577-586.
99. Koga A, Shimada A, Shima A, Sakaizumi M, Tachida H, Hori H: **Evidence for recent invasion of the medaka fish genome by the Tol2 transposable element.** *Genetics* 2000, **155**:273-281.
100. Whitcomb JM, Hughes SH: **Retroviral reverse transcription and integration: progress and problems.** *Annu Rev Cell Biol* 1992, **8**:275-306.
101. Havecker ER, Gao X, Voytas DF: **The diversity of LTR retrotransposons.** *Genome Biol* 2004, **5**:225.
102. Vazquez M, Ben-Dov C, Lorenzi H, Moore T, Schijman A, Levin MJ: **The short interspersed repetitive element of Trypanosoma cruzi, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons.** *Proc Natl Acad Sci U S A* 2000, **97**:2128-2133.

103. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, et al.: **The genome of the African trypanosome Trypanosoma brucei**. *Science* 2005, **309**:416-422.
104. Luan DD, Korman MH, Jakubczak JL, Eickbush TH: **Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition**. *Cell* 1993, **72**:595-605.
105. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, et al.: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease**. *Science* 2005, **309**:409-415.
106. Bringaud F, Bartholomeu DC, Blandin G, Delcher A, Baltz T, El-Sayed NM, Ghedin E: **The Trypanosoma cruzi L1Tc and NARTc non-LTR retrotransposons show relative site specificity for insertion**. *Mol Biol Evol* 2006, **23**:411-420.
107. Bringaud F, Biteau N, Zuiderwijk E, Berriman M, El-Sayed NM, Ghedin E, Melville SE, Hall N, Baltz T: **The ingi and RIME non-LTR retrotransposons are not randomly distributed in the genome of Trypanosoma brucei**. *Mol Biol Evol* 2004, **21**:520-528.
108. Villanueva MS, Williams SP, Beard CB, Richards FF, Aksoy S: **A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of Trypanosoma cruzi**. *Mol Cell Biol* 1991, **11**:6139-6148.
109. Kimmel BE, ole-MoiYoi OK, Young JR: **Ingi, a 5.2-kb dispersed sequence element from Trypanosoma brucei that carries half of a smaller mobile element at either end and has homology with mammalian LINEs**. *Mol Cell Biol* 1987, **7**:1465-1475.
110. Murphy NB, Pays A, Tebabi P, Coquelet H, Guyaux M, Steinert M, Pays E: **Trypanosoma brucei repeated element with unusual structural and transcriptional properties**. *J Mol Biol* 1987, **195**:855-871.
111. Bringaud F, Garcia-Perez JL, Heras SR, Ghedin E, El-Sayed NM, Andersson B, Baltz T, Lopez MC: **Identification of non-autonomous non-LTR retrotransposons in the genome of Trypanosoma cruzi**. *Mol Biochem Parasitol* 2002, **124**:73-78.
112. Martin F, Maranon C, Olivares M, Alonso C, Lopez MC: **Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes**. *J Mol Biol* 1995, **247**:49-59.
113. Bringaud F, Ghedin E, Blandin G, Bartholomeu DC, Caler E, Levin MJ, Baltz T, El-Sayed NM: **Evolution of non-LTR retrotransposons in the trypanosomatid genomes: Leishmania major has lost the active elements**. *Mol Biochem Parasitol* 2006, **145**:158-170.
114. McClintock B: **Controlling elements and the gene**. *Cold Spring Harb Symp Quant Biol* 1956, **21**:197-216.
115. Britten RJ, Davidson EH: **Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty**. *Q Rev Biol* 1971, **46**:111-138.
116. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite**. *Nature* 1980, **284**:604-607.

117. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature* 1980, **284**:601-603.
118. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
119. Taft RJ, Pheasant M, Mattick JS: **The relationship between non-protein-coding DNA and eukaryotic complexity**. *Bioessays* 2007, **29**:288-299.
120. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al.: **Nested retrotransposons in the intergenic regions of the maize genome**. *Science* 1996, **274**:765-768.
121. Biemont C, Cizeron G: **Distribution of transposable elements in Drosophila species**. *Genetica* 1999, **105**:43-62.
122. Castro C, Craig SP, Castaneda M: **Genome organization and ploidy number in Trypanosoma cruzi**. *Mol Biochem Parasitol* 1981, **4**:273-282.
123. Jurka J: **Conserved eukaryotic transposable elements and the evolution of gene regulation**. *Cell Mol Life Sci* 2007.
124. Lowe CB, Bejerano G, Haussler D: **Thousands of human mobile element fragments undergo strong purifying selection near developmental genes**. *Proc Natl Acad Sci U S A* 2007, **104**:8005-8010.
125. Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashovi AS: **Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes**. *Genet Res* 2003, **82**:1-18.
126. Gentles AJ, Wakefield MJ, Kohany O, Gu W, Batzer MA, Pollock DD, Jurka J: **Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica**. *Genome Res* 2007, **17**:992-1004.
127. Thornburg BG, Gotea V, Makalowski W: **Transposable elements as a significant source of transcription regulating signals**. *Gene* 2006, **365**:104-110.
128. Heras SR, Lopez MC, Olivares M, Thomas MC: **The L1Tc non-LTR retrotransposon of Trypanosoma cruzi contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts**. *Nucleic Acids Res* 2007, **35**:2199-2214.
129. Hurst GD, Werren JH: **The role of selfish genetic elements in eukaryotic evolution**. *Nat Rev Genet* 2001, **2**:597-606.
130. Wang L, Jiang T: **On the complexity of multiple sequence alignment**. *J Comput Biol* 1994, **1**:337-348.
131. Lipman DJ, Altschul SF, Kececioglu JD: **A tool for multiple sequence alignment**. *Proc Natl Acad Sci U S A* 1989, **86**:4412-4415.
132. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.
133. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proc Natl Acad Sci U S A* 1987, **84**:4355-4358.
134. Nuin PA, Wang Z, Tillier ER: **The accuracy of several multiple sequence alignment programs for proteins**. *BMC Bioinformatics* 2006, **7**:471.

135. Richard Durbin SRE, Anders Krogh, Graeme Mitchison: *Biological sequence analysis* edn 9. Cambridge: Cambridge University Press; 1998.
136. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30**:3059-3066.
137. Katoh K, Kuma K, Miyata T, Toh H: **Improvement in the accuracy of multiple sequence alignment program MAFFT.** *Genome Inform* 2005, **16**:22-33.
138. Richard Durbin SRE, Anders Krogh, Graeme Mitchison: **Multiple alignment by profile HMM training**. In *Biological sequence analysis*, edn 9. Edited by: Cambridge University Press; 1998:149-159.
139. Eddy SR: **Multiple alignment using hidden Markov models.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:114-120.
140. Baum LE: **An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes.** *Inequalities* 1972, **3**:1-8.
141. Eddy SR: **What is a hidden Markov model?** *Nat Biotechnol* 2004, **22**:1315-1316.
142. Friedman A: *Stochastic differential equations and applications*, vol XIV. London, New-York, SanFrancisco: Academic Press; 1975.
143. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**:1501-1531.
144. Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227**:561-563.
145. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR: **Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena.** *Cell* 1982, **31**:147-157.
146. Gilbert W: **Origin of life: The RNA world.** *Nature* 1986, **319**.
147. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2**:919-929.
148. Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet* 2006, **15 Spec No 1**:R17-29.
149. Lee JC, Gutell RR: **Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs.** *J Mol Biol* 2004, **344**:1225-1249.
150. Leontis NB, Lescoute A, Westhof E: **The building blocks and motifs of RNA architecture.** *Curr Opin Struct Biol* 2006, **16**:279-287.
151. Genomic tRNA Database on World Wide Web URL: http://lowelab.ucsc.edu/GtRNAdb/
152. Durbin R, Eddy, S. R., Krogh, A., Mitchison, G.: **RNA structure analysis**. In *Biological sequence analysis*, edn 9. Edited by: Cambridge University Press; 1998:356.
153. Nussinov R, Jacobson AB: **Fast algorithm for predicting the secondary structure of single-stranded RNA.** *Proc Natl Acad Sci U S A* 1980, **77**:6309-6313.
154. Mount DW: **Prediction of RNA secondary structure**. In *Bioinformatics - Sequence and genome analysis*, edn 2nd edition. Edited by: Cold Spring Harbor; 2004:692.

155. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288**:911-940.
156. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.** *Proc Natl Acad Sci U S A* 2004, **101**:7287-7292.
157. Zuker M: **Prediction of RNA secondary structure by energy minimization.** *Methods Mol Biol* 1994, **25**:267-294.
158. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29**:1105-1119.
159. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31**:3429-3431.
160. Hofacker I, Fontana, W., Stadler, P. F., Bonhhoeffer, L. S., Tacker, M., Schuster, P.: **Fast folding and comparison of RNA secondary structure.** *Monatsh. Chem.* 1994, **125**:167-188.
161. Hartlein M, Cusack S: **Structure, function and evolution of seryl-tRNA synthetases: implications for the evolution of aminoacyl-tRNA synthetases and the genetic code.** *J Mol Evol* 1995, **40**:519-530.
162. Lenhard B, Orellana O, Ibba M, Weygand-Durasevic I: **tRNA recognition and evolution of determinants in seryl-tRNA synthesis.** *Nucleic Acids Res* 1999, **27**:721-729.
163. Mathews DH, Turner DH: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.** *J Mol Biol* 2002, **317**:191-203.
164. Hofacker IL: **RNA Consensus Structure Prediction With RNAalifold.** *Methods Mol Biol* 2007, **395**:527-544.
165. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**:2079-2088.
166. Siebert S, Backofen R: **Methods for Multiple Alignment and Consensus Structure Prediction of RNAs Implemented in MARNA.** *Methods Mol Biol* 2007, **395**:489-502.
167. Knudsen B, Hein J: **Pfold: RNA secondary structure prediction using stochastic context-free grammars.** *Nucleic Acids Res* 2003, **31**:3423-3428.
168. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci U S A* 2005, **102**:2454-2459.
169. Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2**:8.
170. Hofacker IL, Fekete M, Stadler PF: **Secondary structure prediction for aligned RNA sequences.** *J Mol Biol* 2002, **319**:1059-1066.
171. Wuchty S, Fontana W, Hofacker IL, Schuster P: **Complete suboptimal folding of RNA and the stability of secondary structures.** *Biopolymers* 1999, **49**:145-165.
172. Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF: **Automatic detection of conserved RNA structure elements in complete RNA virus genomes.** *Nucleic Acids Res* 1998, **26**:3825-3836.
173. Ding Y, Chan CY, Lawrence CE: **Sfold web server for statistical folding and rational design of nucleic acids.** *Nucleic Acids Res* 2004, **32**:W135-141.

174. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder--a covariance model based RNA motif finding algorithm**. *Bioinformatics* 2006, **22**:445-452.
175. Havgaard JH, Lyngso RB, Stormo GD, Gorodkin J: **Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%**. *Bioinformatics* 2005, **21**:1815-1824.
176. Sankoff D: **Simultaneous solution of the RNA folding, alignment and protosequence problems**. *SIAM J. Appl. Math.* 1985, **45**:810-825.
177. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J: **Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure**. *Genome Res* 2006, **16**:885-889.
178. Torarinsson E, Havgaard JH, Gorodkin J: **Multiple structural alignment and clustering of RNA sequences**. *Bioinformatics* 2007, **23**:926-932.
179. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R: **Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering**. *PLoS Comput Biol* 2007, **3**:e65.
180. Hochsmann M, Voss B, Giegerich R: **Pure multiple RNA secondary structure alignments: a progressive profile approach**. *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1**:53-62.
181. Liu J, Wang JT, Hu J, Tian B: **A method for aligning RNA secondary structures and its application to RNA motif detection**. *BMC Bioinformatics* 2005, **6**:89.
182. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.
183. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792-1797.
184. Hall T: **BioEdit**. Edited by; 2007:comprehensive sequence alignment editor.
185. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment**. *Brief Bioinform* 2004, **5**:150-163.
186. Eddy S, Birney, E.: **Hmmer - Biological sequence analysis using profile hidden Markov models**. Edited by; 2003.
187. Morgenstern B: **DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment**. *Bioinformatics* 1999, **15**:211-218.
188. Dalli D, Wilm A, Mainz I, Steger G: **STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time**. *Bioinformatics* 2006, **22**:1593-1599.
189. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**:195-197.
190. **JAVA programming language**. In *Sun Microsystems*
Edited by.
191. Edgar RC, Batzoglou S: **Multiple sequence alignment**. *Curr Opin Struct Biol* 2006, **16**:368-373.

APPENDIX I

GenBank accession IDs of polyadenylated ESTs for *Leishmania infantum*

CV670622	CV667147	CV669790	CV666576	CV664948	CV662751
CV670278	CV667104	CV669752	CV666568	CV664863	CV662543
CV670124	CV667102	CV669741	CV666518	CV664795	CV662518
CV670078	CV667085	CV669706	CV666500	CV664644	CV662516
CV670050	CV666990	CV669691	CV666454	CV664627	CV662488
CV670037	CV666905	CV669653	CV666429	CV664599	CV662290
CV670031	CV666863	CV669592	CV666427	CV664591	CV662231
CV670015	CV666798	CV669583	CV666389	CV664588	CV662152
CV670011	CV666794	CV669563	CV666323	CV664585	CV662127
CV669673	CV666789	CV669476	CV666258	CV664554	CV661979
CV669580	CV666769	CV669453	CV666121	CV664463	CV661923
CV669553	CV666760	CV669387	CV666101	CV664400	CV661857
CV665459	CV666735	CV669248	CV666093	CV664287	CV661842
CV669548	CV666724	CV669211	CV666087	CV664282	CV661832
CV669103	CV666719	CV669091	CV666084	CV664268	CV661828
CV668879	CV666609	CV668957	CV666073	CV664255	CV661695
CV668764	CV666605	CV668840	CV666071	CV664242	CV661552
CV668103	CV664280	CV668518	CV666056	CV664154	CV661459
CV667818	CV670675	CV668486	CV666038	CV664102	CV661168
CV667706	CV670639	CV668435	CV665948	CV664094	CV660981
CV667665	CV670623	CV668434	CV665911	CV664028	CV660962
CV667646	CV670467	CV668356	CV665909	CV664006	CV660927
CV667639	CV670463	CV668213	CV665901	CV663984	AJ276158
CV667625	CV670444	CV668181	CV665859	CV663977	
CV667619	CV670436	CV668178	CV665846	CV663968	
CV667005	CV670417	CV668130	CV665815	CV663966	
CV666662	CV670284	CV668096	CV665766	CV663941	
CV666468	CV670279	CV668014	CV665702	CV663871	
CV662176	CV670119	CV667999	CV665682	CV663850	
CV667598	CV670097	CV667777	CV665566	CV663778	
CV667575	CV669957	CV667565	CV665540	CV663715	
CV667542	CV669949	CV667515	CV665362	CV663605	
CV667358	CV669922	CV667398	CV665328	CV663591	
CV667347	CV669890	CV667385	CV665288	CV663579	
CV667338	CV669846	CV667347	CV665232	CV663564	
CV667325	CV669835	CV667232	CV665225	CV663464	
CV667210	CV669821	CV666826	CV665041	CV663132	
CV667167	CV669819	CV666690	CV665020	CV662998	
CV667149	CV669799	CV666586	CV664992	CV662943	

APPENDIX II

Minimum evolution phylogenetic tree corresponding to the topology of **Chapter II - Figure 2**. See text for details.