From the Programme for Genomics and Bioinformatics
Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

Genomic Feature Identification in Trypanosomatid Parasites

Daniel Nilsson

Stockholm, 2006

Karolinska Institutet

Abstract

The trypanosomatid parasites cause death and suffering, among humans as well as livestock. Current drugs lack efficacy and cause severe side effects, and no vaccines are available. Increased knowledge of the biology of the parasites is vital for the development of new drugs. Research on these ancient eukaryotes has also already led to the discovery of mechanisms of broader relevance, such as RNA editing, *trans* splicing and antigenic variation. Post-transcriptional regulation is an important part of the regulatory networks of most higher organisms, including humans. In the kinetoplastids, only a very limited part of the control of gene expression is exerted at the transcriptional level. Genes are expressed as long polycistronic pre-mRNA, and individual messages are formed by *trans* splicing and polyadenylation. Even genes that are not coregulated can be on the same polycistronic pre-mRNA. The trypanosomatids can be regarded as models for post-transcriptional regulation, in relation to the more complex eukaryotes.

The progress of the human and other genome projects shows the opportunity provided by a complete genomic sequence to increase the efficiency of traditional molecular biology. Use of computer-aided and fully automated genome sequence analysis tools allows novel feature discovery as well as the direction of hypothesis driven experiments.

We have sequenced the genome of *Trypanosoma cruzi* as part of a three-centre collaboration, and provided an extensive annotation that identifies biologically interesting features. To this end we have used available informatics tools where possible, and developed some new programs. Focus was on integrating current molecular biology knowledge in large scale analyses, and arriving at experimentally testable hypotheses.

This thesis is based on five papers (I-V). Paper I describes a program for gene-finding and annotation that we constructed for the annotation of the genome, described in paper III. Here we collaborated with experts in several areas to investigate the gene content of *T. cruzi*. In paper II we present global base skew features in the genome. In paper IV we describe a model of *trans* splicing in *Trypanosoma brucei*, and the application of it at the genome level. In paper V, we apply the *trans* splice model to predict message boundaries in *Trypanosoma cruzi*, and based on these predictions, we find that upstream open reading frames are common. We hypothesise that these generally repress translation.

Keywords: *Trypanosoma cruzi*, genome sequence, bioinformatics, gene finding, *Trypanosoma brucei*, *Leishmania major*, strand asymmetry, *trans* splicing, polyadenylation, post-transcriptional control, uAUGs.

ISBN 91-7140-789-8

Printed by
Larserics Digital Print AB

Elin – *light and love*

Publications included in this thesis

The thesis is based on the following papers, referred to by the Roman numerals I-V.

I. **Daniel Nilsson**, Björn Andersson.
A graphical tool for parasite genome annotation.
Computer Methods and Programs in Biomedicine, 73:55–60, 2004

II. **Daniel Nilsson**, Björn Andersson.
Strand asymmetry patterns in trypanosomatid parasites.
Experimental Parasitology, 109:143–149, 2005.

III. Najib M. El-Sayed, Peter J. Myler, Daniella C. Bartholomeu, **Daniel Nilsson**, Gautam Aggarwal, Anh-Nhi Tran, Elodie Ghedin, Elizabeth A. Worthey, Arthur L. Delcher, Gaëlle Blandin, Scott J. Westenberger, Elisabet Caler, Gustavo C. Cerqueira, Carole Branche, Brian Haas, Atashi Anupama, Erik Arner, Lena Åslund, Philip Attipoe, Esteban Bontempi, Frédéric Bringaud, Peter Burton, Eithon Cadag, David A. Campbell, Mark Carrington, Jonathan Crabtree, Hamid Darban, Jose Franco da Silveira, Pieter de Jong, Kimberly Edwards, Paul T. Englund, Gholam Fazelina, Tamara Feldblyum, Marcela Ferella, Alberto Carlos Frasch, Keith Gull, David Horn, Lihua Hou, Yiting Huang, Ellen Kindlund, Michele Klingbeil, Sindy Kluge, Hean Koo, Daniela Lacerda, Mariano J. Levin, Hernan Lorenzi, Tin Louie, Carlos Renato Machado, Richard McCulloch, Alan McKenna, Yumi Mizuno, Jeremy C. Mottram, Siri Nelson, Stephen Ochaya, Kazutoyo Osoegawa, Grace Pai, Marilyn Parsons,Martin Pentony, Ulf Pettersson, Mihai Pop, Jose Luis Ramirez, Joel Rinta, Laura Robertson, Steven L. Salzberg, Daniel O. Sanchez, Amber Seyler, Reuben Sharma, Jyoti Shetty, Anjana J. Simpson, Ellen Sisky, Martti T. Tammi, Rick Tarleton, Santuza Teixeira, Susan Van Aken, Christy Vogt, Pauline N. Ward, Bill Wickstead, Jennifer Wortman, Owen White, Claire M. Fraser, Kenneth D. Stuart, Björn Andersson.
The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease.
Science, 309(5733):409–415, 2005 Jul 15.

IV. Corinna Benz, **Daniel Nilsson**, Björn Andersson, Christine Clayton, D. Lys Guilbride.
Messenger RNA processing sites in *Trypanosoma brucei*.
Molecular and Biochemical Parasitology, 143(2):125–134, October 2005.

V. **Daniel Nilsson**, Anh-Nhi Tran, Marcela Ferella, Jenny Eklund, Fang Wang, Mariana Potenza, Björn Andersson.
Kinetoplastid parasite *trans* splice site predictions reveal translational control by uAUGs.
Manuscript.

Other publications

i. Johan Elf, **Daniel Nilsson**, Tanel Tenson, Måns Ehrenberg.
Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage.
Science, 300(5626):1718–1722, 2003 Jul 15.

ii. Erland L. Ljunggren, **Daniel Nilsson**, Jens G. Mattsson.
Expressed sequence tag analysis of Sarcoptes scabiei.
Parasitology, 127(Pt 2):139–145, Aug 2003.

iii. Najib M. El-Sayed, Peter J. Myler, Gaëlle Blandin, Matthew Berriman, Jonathan Crabtree, Gautam Aggarwal, Elisabet Caler, Hubert Renauld, Elizabeth A. Worthey, Christiane Hertz-Fowler, Elodie Ghedin, Christopher Peacock, Daniella C. Bartholomeu, Brian J. Haas, Anh-Nhi Tran, Jennifer R. Wortman, U. Cecilia M. Alsmark, Samuel Angiuoli, Atashi Anupama, Jonathan Badger, Frédéric Bringaud, Eithon Cadag, Jane M. Carlton, Gustavo C. Cerqueira, Todd Creasy, Arthur L. Delcher, Appolinaire Djikeng, T. Martin Embley, Christopher Hauser, Alasdair C. Ivens, Sarah K. Kummerfeld, Jose B. Pereira-Leal, **Daniel Nilsson**, Jeremy Peterson, Steven L. Salzberg, Joshua Shallom, Joana C. Silva, Jaideep Sundaram, Scott Westenberger, Owen White, Sara E. Melville, John E. Donelson, Björn Andersson, Kenneth D. Stuart, Neil Hall.
Comparative genomics of trypanosomatid parasitic protozoa.
Science, 309(5733):404–409, 2005 Jul 15.

Contents

* Previously unpublished material.

Abbreviations

ARE AU-rich element

BAC bacterial artificial chromosome

CAI codon adaption index

DTU discrete typing unit

EST expressed sequence tag

FN false negative

FP false positive

GFP green fluorescent protein

GPI glycosylphosphatidylinositol

GRE G-rich element

GSS genomic survey sequence

mRNA messenger ribonucleic acid

MSP major surface protease

ORF open reading frame

SL spliced leader

snRNA small nuclear ribonucleic acid

snoRNA small nucleolar ribonucleic acid

sp. one unspecified species

spp. several unspecified species

TN true negative

TP true positive

tRNA transfer ribonucleic acid

TS *trans*-sialidase

uAUG upstream start codon

uORF upstream open reading frame

UTR untranslated region

VSG variable surface glycoprotein

Part I

Introduction

"Imagine a world where parasites control the minds of their hosts, sending them to their destruction.

Imagine a world where parasites are masters of chemical warfare and camouflage, able to cloak themselves with their hosts' own molecules.

Imagine a world where parasites steer the course of evolution, where the majority of species are parasites.

Welcome to earth."

– Carl Zimmer, promoting his book *Parasite Rex*

Chapter 1

The trypanosomatid parasites

The trypanosomatid parasites cause death and morbidity in man, as well as livestock. Some 26 million people are infected with trypanosomatids, which directly cause 113 000 deaths each year (table 1.1), but also much disability as well as economic difficulty due to loss of livestock[1].

An organism that lives off the resources collected by another organism, without giving any benefit back, is called a parasite. The name trypanosoma derives from the greek trypanon, meaning borer, and soma, meaning body, and is suggestive of the rotational swimming movements of the parasites. The trypanosomatids are classified as protozoan euglenozoan kinetoplastids with a single flagellum. As unicellular eukaryotes, they are complex organisms compared to many other pathogens. They differentiate to adopt to the contrasting environments of their many hosts. This obligatory progression of cellular states is termed life cycle (figure 1.1). The parasite morphology changes considerably over the life cycle. The different stages are characterised by the relative position of nucleus and flagellum. Extracellular stages are comparatively large, 25 μm, and flagellated. The intracellular stage, amastigote, is smaller, about 5 μm, with a rudimentary flagellum. The trypanosomatid parasite species are

	Infected	Deaths yr^{-1}	New infections yr^{-1}
Leishmania spp.	>12 000 000	51 000	2 000 000
Trypanosoma brucei	3-500 000	48 000	3-500 000
Trypanosoma cruzi	13 000 000	14 000	200 000

Table 1.1: WHO estimates on the number of people currently infected with the trypanosomatid parasites, number of deaths caused annually and new infections caused annually. The *Leishmania* figures are uncertain since disease declaration is required in only 32 of the 88 endemic countries.

numerous and can be found in a broad range of hosts. This thesis is focused in particular on *Trypanosoma cruzi*, and to a lesser extent also on *Trypanosoma brucei* and *Leishmania major*.

The kinetoplastids are extraordinary organisms, with many features that are unique or at least uncommon among eukaryota. The single kinetoplastiod mitochondrion differs from other eukaryotic mitochondria. The defining characteristic is a suborganellar structure, the kinetoplast. It contains a braided network of circular DNA which forms the organellar genome, kDNA (reviewed in [2]). mRNAs in the organelle are extensively edited with the help of *trans*-acting guide RNAs that carry the information to change, add or remove certain U bases. Another most particular organisational feature is the compartmentalisation of glycolysis to glycosomes [3, 4].

A spliced leader RNA is *trans* spliced to each mRNA, as a memeans to get unicistronic mature messages from the multicistronic transcripts produced from the unidirectional transcription clusters found in the genomes [5, 6, 7]. Very few RNA polymerase promoters are known (reviewed in [8]), and no traditional polII promoters, not even at the start of each gene cluster, although transcription initiation is more common there than at any other investigated positions of the genome[9, 10]. The genomes have uniquely modified bases, DNA-J, which could be involved in epigenetic inheritance and gene silencing (reviewed in [11]).

The coat proteins of *Trypanosoma brucei* are not only exchanged as the parasite leaves the tsetse fly gut to enter the human bloodstream. Periodically, a subpopulation of the parasites within the host switch to a different surface antigen, so that when the adaptive immune response finally recognises epitopes from the last major parasite coat, the subpopulation can grow unaffected (reviewed in [12]). Many surface molecules in the trypanosomatids, including these variable surface antigens, are anchored to the cell membrane via addition of glycosylphosphatidylinositol (GPI) [13]. These discoveries have benefited basic research in a more general way, since other eukaryotes also use GPI anchoring, RNA editing and *trans* splicing.

Though similar in appearance, an estimated 100 million years of evolution has passed since the last common ancestor between Leishmania and Trypanosoma [14]. Their molecular peculiarities are as fascinating as their common features. The following sections describe some unique features, and summarise the life cycles, of each parasite species complex.

1.1 *Trypanosoma cruzi*

Trypanosoma cruzi, the etiological agent of Chagas disease, is prevalent in Latin America. The symptoms of this American trypanosomiasis are varied. After an acute phase of a few weeks, which sometimes shows symptoms and can even be lethal, the infection can remain relatively quiescent for as long as 10-20 years. Untreated *T cruzi* infection is often lifelong [15]. 25-30% of the chronically infected die by heart failure or failure of the digestive tract[1].

The parasite proliferates in the gut of triatomine bugs as an epimastigote,

migrates to the hindgut and progresses to a metacyclic trypomastigote (figure 1.1). When the insect bites a vertebrate, and ingests a bloodmeal, parasites can enter the wound via insect faeces. The *Trypanosoma cruzi* trypomastigotes do not switch coat proteins, but rather express a multitude of diverse surface molecules; notably *trans*-sialidases [16], mucins [17] and proteases [18]. The *trans*-sialidases transfer host sialyl moieties to the mucins. Mucin diversity may help avoid an immune response, and also mediate attachment to a broad range of cells[19]. The trypomastigotes enter cells by being endocytosed and escaping the phagocytic vacuole. The parasite is initially so successful in avoiding a host cell response that it has been called "the stealth parasite" [20]. The trypomastigotes differentiate into small amastigotes. Now the cell cycle is once again released, and the amastigotes can proliferate. The amastigotes again express different proteins, such as the surface glycoprotein amastin[21]. The amastigotes differentiate into trypomastigotes, via an epimastigote like intermediate stage [22], and burst out of the host cell. The trypomastigotes in the bloodstream are infective and can spread via the blood, e g via new insect bites. Upon differentiation to epimastigotes, the surface proteins are replaced by a different set [23].

Trypanosoma cruzi is a heterogenous species [24]. Several useful classification models have been proposed, based on different diagnostics and on data from a varying number of isolates. Current nomenclature differentiates between two major groups, *T. cruzi* I and *T. cruzi* II, but recognises the existence of additional isolates that do not fit these [25]. *T. cruzi* I and *T. cruzi* II can be further subdivided based on nucleotide sequence diversity into six different discrete typing units, DTU I and DTU IIa-IIe [26]. Two major ecological circulations are recognised. The silvatic group largely lives in wild animals, while *T. cruzi* strains from the other group circulate in domestic animals and human. The two can be locally connected, e.g. via animals that belong to both circulations, such as mice, rats and bats. Infectious species can be found also in the silvatic circulation.

Mixed infections and heterogenous natural isolates complicate the picture further. While propagation is predominantly clonal, hybrids can form *in vitro* [27] under selective pressure, and ample genetic evidence of hybrid formation *in vivo* exist [28]. The large genetic variation of *T. cruzi* will be discussed further in section 2.3.1.

1.2 *Trypanosoma brucei*

Trypanosoma brucei causes sleeping sickness in man and Nagana in livestock. It is spread by the tsetse fly on the African continent. Sleeping sickness is severely debilitating and fatal if untreated. The human form is caused by two subspecies, *T. b ghambiense* and *T. b rhodiense*. The genome project and much molecular biology research is focused on the livestock pathogen *T. b brucei*, believed to be similar to *T. b rhodiense*.

The lifecycle is complex (figure 1.1). When trypomastigotes arrive in the

midgut of the tsetse fly, they differentiate into proliferative procyclics. When they leave the midgut, they change into epimastigotes, which multiply in the salivary glands of the fly. The epimastigotes transform into infective metacyclic trypomastigotes. These can be injected into the mammal when the fly bites. The metacyclics then differentiate into the bloodstream forms. The long, slender bloodstream trypomastigotes can divide, but upon sensing high population densities they enter an arrested short, stumpy stage that is tsetse infective[29]. *Trypanosoma brucei* has no intracellular stage, and since life and proliferation occurs in the bloodstream of the vertebrate host, the parasite is directly exposed to the host immune system. The surface of the bloodstream form is coated with variable surface glycoproteins. While these are immunogenic, a parasite infection can still persist since occasionally a subpopulation will express a different VSG. The VSG switching is achieved by moving quiescent VSG variants to bloodstream form active expression sites. A large number of VSG genes exist in the genome, the inactive ones mostly in the form of pseudogenes [30]. The insect gut is quite different from the human bloodstream, and *T. brucei* responds by shedding of the VSG coat and expression of procyclins, a group of GPI anchored acidic glycoproteins, upon differentiation to the procyclic stage (see also chapter 6).

T. brucei genetic exchange differs from that of *T. cruzi*. It is also rare, but appears to involve meiotic division rather than hybridisation with aneuploidy [31].

1.3 *Leishmania* spp.

At least 20 different species of *Leishmania* are known to be infective in human, but even more are being recognised as a threat to immunosuppressed individuals [1]. The different diseases caused are collectively called leishmaniases. These include visceral leishmaniasis, mucocutaneous leishmaniasis and oriental sores. An absolute link between parasite strain and the type of disease caused has not been established. Differences between infected individuals, in particular in the immune system, affect the outcome.

Leishmania spp. promastigotes live and divide in the sand fly, differentiate to cell-cycle arrested metacyclics that when taken up by vertebrates can invade phagocytic macrophages. In the phagolysosome, the amastigotes proliferate, lyse the macrophage and reinfect others. Amastigotes can be taken up in insect blood meals, and once again transform into procyclics to complete the life cycle (see figure 1.1).

In differentiation to the metacyclic form, *Leishmania* promastigotes express larger amounts of GP63, a GPI anchored surface protease. The structure of the other major surface glycoprotein, lipophosphoglycan is also altered. This releases the parasites from the insect gut epithelium and, as the GP63 protease, is important for host cell attachment and survival in the phagolysosome (reviewed in [32]).

1.4 Chemotherapeutics and disease control

The success in the last years in reducing the disease impact owes much to vector control programs. Traps, nets, insecticides, urbanisation, better housing and blood donor screens limit the transmission, although the gap between rich and poor areas in endemic countries is large[1]. The trypanosomatids continue to cause diseases of the poor. New, effective and ideally cheap chemotherapeutics are much needed.

Against *Trypanosoma cruzi* two drugs, nifurtimox and benznidazole, are currently recommended. They are effective in the early stages of disease, immediately after infection. An efficacy of 80 % is reported [33], although the side effects can be severe. However, there are no drugs for the later stages of infection.

Against *Trypanosoma brucei*, two drugs are available for the early stage, pentamidine and suramin. One new drug against African trypanosomiasis has been introduced, eflornithine, effective against T. brucei gambiense. It has less severe side effects than the older melarsoprol, and is also effective in the late stage. The death rate from melarsoprol administration alone is about 5 %[34]. Eflornithine was initially too expensive for broad use. Since this drug also removes unwanted facial hair, eflornithine production has been scaled up [35], and is also made available for the African trypanosomiasis endemic countries via a donation program [1].

The standard treatment against visceral leishmaniasis, pentavalent antimonials, is loosing effectiveness due to increasing drug resistance. A liposomal formulation of amphotericine B is effective but expensive [33]. Recently, hope for treatment of visceral leishmaniases has been associated with miltefosine, a new orally administered drug - a failed antitumor agent - with a high cure rate and relatively mild side effects [36].

To develop a vaccine for african trypanosomiases seems difficult, when the mechanisms of antigenic coat switching are considered. Significant progress has been made on vaccine development in Leishmaniasis, but none are as yet in production. Generally, vaccine development for trypanosomatids must be considered difficult, as not even inocculation with live parasite is sufficient to induce protective immunity [37].

A handful of new potential therapeutic drugs are in clinical trials [33], but there is a tremendous need for new active substances. Increased knowledge of parasite biology should accelerate the development. Although one should note that previous attempts at rational drug design in the trypanosomatids have failed [33], new informatics attempts[38] where structure homology models and molecular dynamics are used to arrive at target models for rational inhibitor selection are interesting.

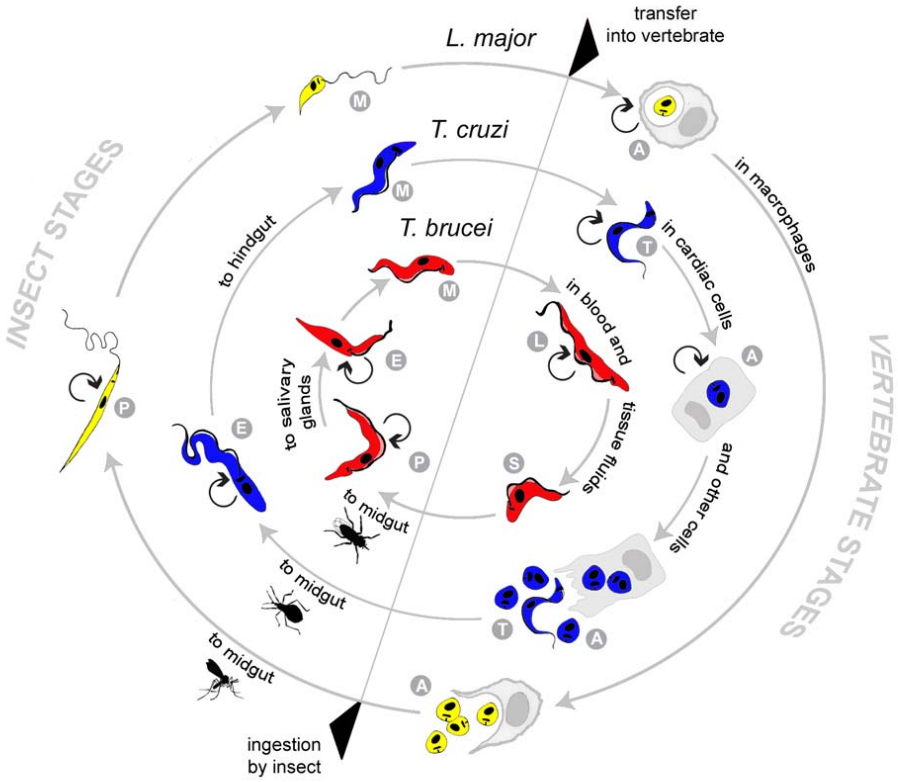

Figure 1.1: The life cycles of the TriTryps are complex and involve many large morphological changes, as well as changes of surface components and other adaptions to the differences between insect, bloodstream and intracellular environments. (A) Amastigote, (T) Trypomastigote (L) Long, slender (S) Short, stumpy (M) Metacyclic, (E) Epimastigote. Black, semi-circular arrows indicate proliferating stages. Reproduced from [iii, figure S1] with permission.

Chapter 2

Genome sequencing

2.1 Genomics

Genome sequencing can be defined as the complete determination of the hereditary nucleic acid material of an organism.

The genomics era began in earnest with the completion of the genome sequence of the bacteriophage MS2 [39]. The first complete DNA based genome ΦX170 [40] was possible after additional method development. These gene-dense phages of 3569 nt and 5375 nt, with overlapping genes and regulatory regions, confirmed contemporary knowledge about transcription and translation, and raised new questions. New techniques for sequencing were invented [41, 42], that made the sequencing of larger genomes feasible. Modified versions of the Sanger method for DNA sequencing with chain terminators [42] are todays method of choice. It's primary limitation of sequencing single stranded DNA only was solved [43] by use of an intermediate step of cloning into bacteriophages, and later by cycle sequencing with thermostable polymerases.

Briefly, a DNA polymerase elongates the template for sequencing from a primer site in a reaction mix with small amounts of dideoxytriphosphate nucleoside analogues together with the four ordinary nucleosides that inhibit polymerisation. Four separate reactions are run, each with a nucleotide specific reaction and a labeled primer, or in modern protocols, one reaction with a different terminator fluorophore for each inhibiting analogue. The ladder of sequences that is produced after several polymerisation cycles is resolved according to size in a gel, and the sequence of bases can be read.

The human mitochondrial genome was completed [44], and almost as a side product the genome of bacteriophage λ [45], a genome that has served as a model for gene regulation.

The Human Genome Project [46, 47], one of the great scientific endeavours of our time, drove technology forward in earnest. The project plan was ambitious and well ahead of its time when formalised in 1990. But, automation and innovative solutions made sequencing faster, cheaper and more accurate than hoped

for. In consequence, bacterial genomes were sequenced in 1995, *Haemophilus influenzae* [48] and *Mycoplasma genitalium* [49]. The first eukaryote sequenced was the unicellular *S. cerevisiae* [50], a genome with 6000 genes in 12 Mbp.

2.1.1 Shotgun sequencing

Although other strategies exist, the methods of shotgun sequencing (figure 2.1) are employed in almost all sequencing projects. Multiple copies of the target DNA are sheared to pieces, filtered to a relatively well defined size and subcloned into a phage or plasmid vector. The insert sequences are sequenced using the Sanger method with primers for so called universal primer sites just outside the insert in the subcloning vector. By sequencing enough randomly selected subclones, each base in the original clone has a good probability of being sequenced. Computer programs for sequence assembly are used to tile the sequence pieces together based on the near identical overlaps.

This procedure lends itself to automation, and scales well to a very large format due to its parallel nature once libraries have been constructed. The shotgun procedure is employed in two main strategies, clone-by-clone - aka hierarchical shotgun - or whole genome shotgun (reviewed in e.g. [51]). In clone-by-clone shotgun, a physical map is first constructed, and clones are chosen using this map to cover the entire genome with limited redundancy. The chosen clones are subsequently sequenced as previously described. A map can also be generated underway. In this case, start points in the form of seed clones are chosen on different chromosomes. New clones, overlapping these seed clones, are picked, sequenced and used as new seed clones in an iterative fashion. Sequencing and mapping can in this way progress in parallel.

In whole genome shotgun, the mapping and cloning step is left out. The entire genome is sheared and subcloned. To aid sequence assembly and reduce the need for laborious finishing, distance constraints via paired forward and reverse clone end reads can be added [52]. Different size inserts can be used to simplify assembly of repeat regions of different sizes.

Hybrid approaches can also be effective, such as whole-chromosome shotgun, used for four chromosomes in the *Trypanosoma brucei* genome project [30], where most chromosomes were sequenced in a clone-by-clone fashion.

2.1.2 Sequence assembly

The shotgun assembly problem is computationally difficult. The basic strategy assumes that all bases are correct, and each subsequence in the project above a minimum overlap length is unique. By determination of the overlaps between reads, the problem is reduced to a travelling salesman like shortest path finding exercise in a graph that represents this connectivity. The assumptions of correct bases must be relaxed, and are replaced by error probabilities gained from basecalling. Here, the electropherograms are scrutinised by a program that estimates the quality of each base on local properties of the elution curve [53, 54]. The base calling error probabilities are also used to help estimate the total error

in the finished sequence. The overlap graph becomes probabilistic, with each overlap conditional on sequence error probability, and methods to find the most probable solution can be used. To find completely unique overlaps would require a huge minimum overlap length. In practice, minimum overlaps of less than 100 bp are used. Repeats of this length, and even more than a read length, are not uncommon. Here, length constraints from inserts of different size can be used as additional constraints to prune the graph. Repeats with some polymorphisms between repeat copies can often be resolved by statistical methods [55]. The problem grows more complex if the repeat units are nearly identical. Polymorphism between homologous copies of the repeat from different chromosomes also deepens the complexity. Repeats are a common source of gaps between contigs. Sometimes these can be tentatively bridged by the clone end pairs, sequence-mapped gap. Chains of such ordered contigs are often called assembly scaffolds. Accurate sequence-mapped gaps can often be closed by simply sequencing the remainder of the clone that bridges it (see 2.1.1).

2.2 Pathogen Genomics

The drive to know the human genome is almost self evident. Even if the benefits to medicine, now manifest mainly in new genetic tests, may not initially have been as large as anticipated, curiosity as to our basic building blocks, and as to what makes us who we are and what makes individuals different, is a fundamental drive. To understand model organisms at a very detailed level is naturally important. Results from experiments can be tied to their genomic causes. The sequencing of pathogen genomes follows almost as a predictive method, with the model organisms as training examples. Once the genome is known, further experimentation and characterisation is accelerated. To find and clone a gene of interest is made very easy, and various forms of large and small scale functional genomic experiments become possible. Features can be inferred from similarity to those known from other organisms. Also, discovery and analyses of general phenomena are greatly simplified. What was once a deductive procedure with need for much experimental effort, or a serendipitous occurrence observed by a prepared mind, is opened up to exploration and feature discovery.

Pathogen sequencing has had a given place on the genomics agenda. Consequently, virus sequencing was early, notably the 5224 bp SV40 [56]. Sequencing and genome analysis in pathogenic microbes has had a tremendous impact [57].

The mass of completely known genomes is increasing rapidly, with 361 published non-viral genomes, and some 1600 ongoing[1]. Complete sequences are available for some 1200 virus types, some of which have been sequenced many times over.

The pioneering work in protozoan parasite genomics was done in the apicomplexans. The human malaria parasite *Plasmodium falciparum* [58] and the rodent malaria parasite *Plasmodium yoelii* [59], *Cryptosporidium hominis* [60] and *Cryptosporidium parvum* [61] and lately two bovine parasites *Theileria*

[1] http://www.genomesonline.org, April 1 2006 update

spp.[62, 63] have been sequenced at high coverage. Other recently completed protozoan genomes include *Entamoeba histolytica* [64] and *Cryptococcus neoformans*[65]. Several other projects are underway or near completion (reviewed in e g [66]), and much interesting data is being produced by means of low coverage sequencing of species closely related to the already sequenced ones as well as survey sequencing and expression studies.

2.3 The *Trypanosoma cruzi* Genome Initiative

The basis for the *Trypanosoma cruzi* Genome Initiative was established at the 1994 WHO/TDR Parasite Genome Network Planning Meeting at Fiocruz, Rio de Janeiro. The reference strain, previously known as F11F5, was chosen by a workgroup of 15 researchers and named CL Brener, after Professor Zigman Brener, who isolated it [67]. The first Genome Initiative progress, financed in part by seed money from WHO/TDR, involved several projects, many carried out in the endemic countries [68]. Macro-biological parameters, cytogenetics, EST sequencing, genomic survey sequencing and a pilot chromosome sequencing endeavour preceded the eventual Genome Project.

CL Brener was shown to grow and differentiate in culture, as well as to be infective, clinically relevant, susceptible to drugs and cytogenetically stable [69, 70].

2.3.1 Cytogenetics

The nuclear chromosomes of *Trypanosoma cruzi* condense poorly. This makes karyotype determination difficult. Pulse field gel electrophoresis techniques [71] coupled with hybridisation of gene probes or densiometric analysis has been used extensively [72, 73, 74, 75, 76, 77] in many strains. The variance in chromosome number and total genome size estimates is large between strains and between laboratories, and the complexity of the analyses leads authors to claim unspecified uncertainty in the results. The genome of CL Brener appears to be 87 Mb in 64 chromosomes [78, 75], $>$ 100 Mb in $>$ 40 chromosomes [74] and 78 Mb in 55 chromosomes [77]. Other estimation methods arrive at a DNA content of approximately 110 Mb [79]. Aside from unspecified method error, the kDNA content is possibly not included in the genome content for some of the karyotype determination studies.

Results indicate that CL Brener is largely diploid, with possible partial aneuploidy. Interestingly, the homologous chromosome pairs show large size variation [78, 74]. The homologs are heterogenous also at the sequence level, showing considerable allelic variation and correlation to comparative data indicates that CL Brener is a hybrid [28].

The genome of *T. cruzi* is highly repetitive[75], as are many other eukaryotic genomes. The long tandem arrays of housekeeping genes are a hallmark of the *T. cruzi* genome, and were encountered during the pilot sequencing[5]. The surface antigen molecules are repeated in large numbers [19], and have nearly as

many pseudogene copies at that. Dispersed repeats, in the form of non-tandem arrayed gene families and transposable elements, are common. Oligonucleotide repeats and other low-complexity sequences are also found[5, 80].

2.3.2 Transcriptome and genomic survey sequencing

My first experience with he Genome Initiative was with the final stages of the *T. cruzi* EST project, carried out as collaboration between several laboratories, with about half of the sequences produced at Uppsala University [81, 82, 83]. 4.3 Mb random shotgun fragments were also sequenced in genome survey sequencing [84].

Genome sequencing started with the pilot sequencing of chromosome 3 [5]. Clones from a cosmid library [79] that hybridised to the chromosome 3 homologs ($\sim$ 0.67 Mb and $\sim$ 1.1 Mb) were selected for shotgun sequencing. While a core, gene-dense strand switch part of the chromosome could be closed by sequenced cosmid overlaps, the project remains unfinished due to the large repetitive regions encountered.

2.3.3 The *Trypanosoma cruzi* CL Brener Genome Project

Large scale sequencing commenced in 2001 in a NIH/NIAID financed consortium of three groups; El-Sayed's at The Institute for Genomic Research, Stuart and Myler's at Seattle Biomedical Research Institute and Andersson's at Uppsala University, which moved to Karolinska Institutet in 2001. The initial approach of BAC-by-BAC shotgun, map-as-you-go proved difficult. All centers failed to find clear extensions for their seed BACs. Instead, sequencing was switched to whole genome shotgun. Since the genome is highly repetitive, clone end pairs from libraries with different insert sizes, but a high proportion of large inserts, were sequenced. Low coverage sequence from Esmeraldo, another *T. cruzi* strain, was obtained to help resolve the issues arisen from haplotype allelic variance. Further background and results can be found in chapter 7.5 and [III].

2.4 The TriTryp initiative

Early on in the *Trypanosoma cruzi* genome project, it was clear that a collaboration with the other two trypanosomatid genome projects could be mutually beneficial. The gene order was found to be well conserved between the three TriTryps [85]; later, large blocks of conserved syntenous genes could indeed be identified [86, 87]. Comparisons between the three related parasites was deemed an important field of research. Also, three of the four centers were directly involved in more than one of the sequencing projects, albeit in different constellations. This simplified collaboration.

The structural part of the comparative genomics effort [iii] focused on alignments of *L. major* and *T. brucei* chromosomes. By alignment of the *T. cruzi*

contigs to the *L. major* and *T. brucei* ones, some of the *T. cruzi* genome structure could also be used in the comparison.

For instance, some larger scaffolds in *T. cruzi* were found to have telomeric sequences and conspicuous subtelomeric gene arrays at synteny break points that coincided with chromosome end telomeric sequences in *L. major*, but not in *T. brucei*. This was one of several lines of evidence in suggesting that the *T. brucei* karyotype has formed by fusion of ancestral chromosomes after the split between the *Leishmania* and *Trypanosoma* lineages.

A core proteome of the three organisms comprising 6158 proteins was established by clustering of gene orthologs. The parasite specific commonalities are of particular interest for drug development, as these could possibly be multi-organism targets for a single compound.

The comparative annotation was also a major added benefit to the individual genomes. False negative genes from one genome that were present in the other two would leave an obvious hole in the synteny chain for the annotators to curate. This was especially useful to the larger shotgun parts of the projects: the *T. cruzi* genome, the groups of chromosomes not sufficiently separated by PFGE, "blobs", of *L. major* and the last large chromosomes of *T. brucei*. These had not been manually annotated at the steady clone-by-clone pace. For *T. cruzi*, functional annotation was tentatively transfered from *T. brucei* where homology was sufficient. These tentative functional assignments were manually curated, but at higher pace than would otherwise have been possible.

An integral part of the project was to continuously disseminate the gained information and integrate it with other knowledge [88]. This was accomplished by release of sequence data daily to the web and ftp sites of the genome centers, by inclusion of the TriTryp data on genedb [89], and upon publication by submission to GenBank [90]. The genedb resource is still actively maintained, and for *Trypanosoma cruzi*, the parallel TcruziDB also integrates proteomic data [88, 91].

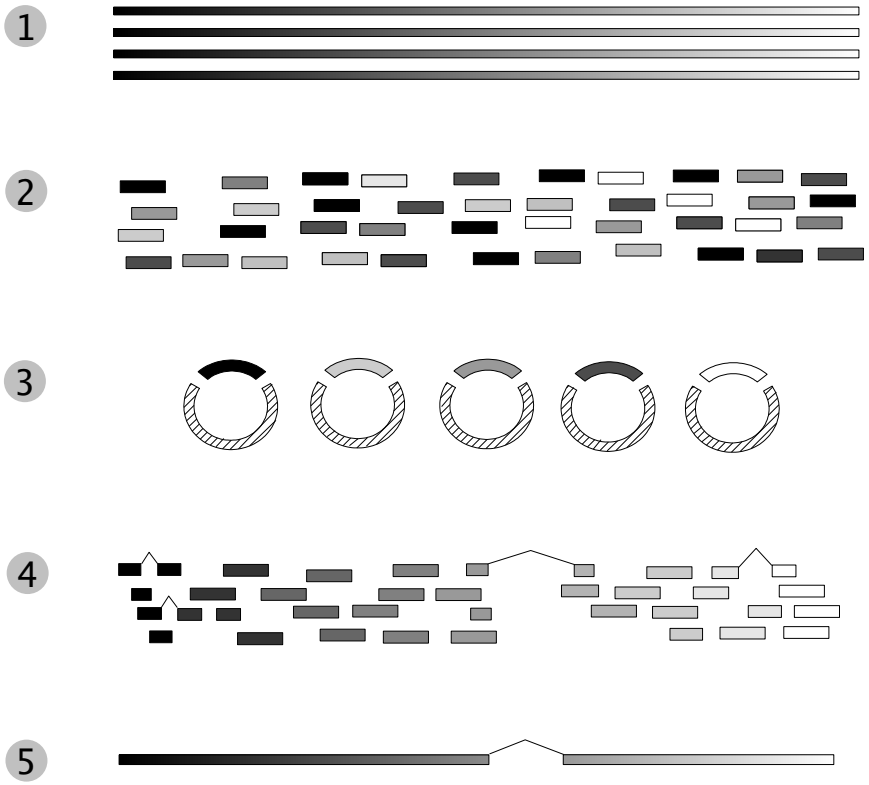

Figure 2.1: Shotgun sequencing and assembly. Multiple copies (1) of the target DNA is sheared into fragments of a desired size (2) and cloned into sequencing vector (3). Typically, the sequencing equipment cannot sequence the entire insert, rather short sequences from both ends of the inserts are sequenced. The resulting "reads" are processed and assembled *in silico* (4) into contigous sequences, contigs. Contigs can be ordered by virtue of sequence mapped gaps into scaffolds (5).

"Computers are useless. They can only give you answers."

– Pablo Picasso

Chapter 3

Bioinformatics

The fundamental theories of computation were established just prior to the modern revolution in molecular biological methodology. The development was to a large extent driven by the need to dechipher wartime crypto, the need to quickly and accurately solve differential equations and simulate rocket trajectories. Molecular biology initially benefitted from the possibility to store and manipulate the large amount of data produced in protein sequencing and later DNA sequencing. Catalogs of sequences, as pioneered by Dayhoff[92], could be stored digitally. The sequence databases and tools to handle them were the first components of bioinformatics, and remain important. The Internet has since considerably simplified access to biological information through various databases. The field of cybernetics[1] also later led to the means to search the sequence data in an evolutionarily meaningful fashion by means of homology based searches (see section 4.1). Many other experimental methods that generate large bodies of data have been established, and bioinformatics methods are used to handle and make sense of this data. The development of computer hardware capable of handling the biological information has been fortunately timely - or perhaps it was rather that the new machines enabled new experiments, such as in the case of shotgun sequencing and sequence assembly (see section 2.1).

In general, the analysis of data can be made in an explorative fashion with the use of statistical and probabilistic methods to find commonalites, connections and exceedingly rare events. This can in turn lead to the formulation of hypotheses, and the generation of information. Data analysis can also be driven directly by a hypothesis. Simple or elaborate models can be expressed mathematically, implemented and tested *in silico*. This allows us to test the bounds of our understanding of what caused the observed data, much as if these conjectures had been tested by real *in vivo* experiments. Predictions, qualitative or quantitative, that are not contradicted by the data at hand, can so be made. For example, a predictive method founded in our understanding of peptide bio-

[1] Control theory, originally inspired by neurobiology, and pioneered by Norbert Wiener.

chemistry is the prediction of protein features from primary sequence, spanning from relatively simple ones such as local hydrophobicity, to complex, such as complete three-dimensional structures and substrate specificities. An early application of qualitative predictive bioinformatical methods was that of finding genes in DNA sequence data (see chapter 4).

Bioinformatics can also pertain to the use of computers to perform these tasks, rather than the use of pen and paper or other aids. A distinction between bioinformatics and computational biology is often made[2]. If so, bioinformatics is typically said to encompass data storage and algorithms for feature identification and pattern recognition, as well as the large body of evolutionary models while computational biology deals with simulations, often via differential equation models of biological systems [93]. A distinction between qualitative data, which belongs to the field of bioinformatics, and quantitative data, which belongs to the fields of computational biology or systems biology.

It is difficult to imagine the field of genomics without the tools of bioinformatics. Base-calling, vector trimming, sequence storage, sequence assembly, gene prediction, similarity searches, functional predictions, multiple alignments and phylogenetic trees and many other bioinformatical approaches are intimately connected with the success of genomics.

[2]Such as the NIH Working Definintion of Bioinformatics and Computational Biology, July 17, 2000, http://www.bisti.nih.gov/CompuBioDef.pdf.

```
<TU>
  <FEAT_NAME>199.t00333</FEAT_NAME>
  <CHROMO_LINK>39.t00012</CHROMO_LINK>
  <DATE>Jan  9 2001  3:56PM</DATE>
  <GENE_INFO>
    <LOCUS>28H13.55</LOCUS>
    <PUB_LOCUS>Tb927.2.3410</PUB_LOCUS>
    <COM_NAME CURATED="1">hypothetical protein, unlikely</COM_NAME>
    <COMMENT>Giuliani, Richard L., D.D.S. Oral Surgery Dentists Only
             Drs. Breen &amp;amp; Giuliani, D.D.S.,
             P.A. 5530 Wisconsin Ave, Ste 640 Chevy Chase,
             MD 20815 (301) 652-7372</COMMENT>
    <IS_PSEUDOGENE>0</IS_PSEUDOGENE>
    <FUNCT_ANNOT_EVIDENCE TYPE="CURATED">
    </FUNCT_ANNOT_EVIDENCE>
    <DATE>May 14 2002  9:52AM</DATE>
  </GENE_INFO>
  <TRANSCRIPT_SEQUENCE>
    ATGTCGGGGCTTATTGACCGCGGGGCAAAGTGCAACAAAAACACAGTCTCAATGTCTTGT
    GCGTTGCCTCCTTCAACGCAAGTTGTGTATGGACATTCGAGAAAGAATTTAAAAAATAAA
    AAAAAGTATGAGTGCGGGGGAGCAATTTTAATCTGCACCTCTCTTCGTGAGATTGTAATT
    TGTCATGGTTCCTTTCTTTCTTTCTTTGTTGGAAATTATTGGAGAACTGCACTGCGGTAA
  </TRANSCRIPT_SEQUENCE>
</TU>
```

– Anonymous annotator, *Trypanosoma brucei chromosome II*, 2002

"[Gene annotations] provide not only a mechanism for researchers to focus searches on genes that interest them but also a framework upon which 'big picture' analyses can be built. Good genome annotation reflects the collective knowledge of many scientists but distributed over the entire genome. By providing tools and database infrastructure, this diffuse knowledge can be harnessed; the data can be interrogated and new hypotheses built."

– Matt Berriman, *Parasitology* 128 p. S23, 2004

Chapter 4

Identifying and annotating genes

To find genes in a long string of nucleotides by computational methods is still an open problem. Partial solutions explicitly or implicitly exploit evolutionary conservation, constraints from the function of the gene product and biochemical- and biological constraints that arise from the gene expression machinery. The problem of assignment of a biological function to the identified genes is related, but more complicated.

The terms gene identification and gene finding are used interchangeably and refer to a number of tasks, such as determination of the protein coding nucleotides in a genome or of the mature RNA coding ones and may also include the identification of *cis*-acting regulatory elements (reviewed in [94, 95, 96, 97]). The delineation of the exon-intron boundaries of genes, sometimes called structural annotation, adds complexity to the gene identification. Only very few *cis*-spliced genes have been found in the trypanosomatids [98, 99]. In the TriTryp genome projects, the automated gene finding did not consider introns.

The core components of gene finding methods can be divided into two broad categories, database lookup and *ab initio* methods. While this is also true for RNA gene finding, we shall treat this separately.

4.1 Database lookup methods

Database lookup based on sequence homology at the protein or even DNA level is useful for gene identification, in particular when a large number of genes are already known in the species or in a closely related species. For the TriTryps, ∼35% of the called genes could have been identified via database lookup [66].

The most common database lookup methods are based on pairwise alignments of the query sequence against many candidates. The problem of pairwise alignment is solved by a dynamic programming algorithm which will find the optimal alignment, given a set of match/mismatch/gap scores. The op-

timality principle employed is attributed to R. Bellman. A global alignment approach was suggested by Needleman-Wunsch [100], and a locally optimal version by Smith-Waterman [101]. A match/mismatch score is given to each pair of residues in an alignment, that reflect the odds of finding such a pair in unrelated sequences, and gaps are treated as another possible symbol in the pair. This appears most simplistic, when all the complexity involved in the evolution of proteins is considered. Nevertheless, the approach is often successful [102]. The match and mismatch scores can be represented by a matrix. The PAM family of score matrices are based on estimates of evolutionary distance between sequenced protein. The original PAM family matrices are somewhat dated, based on only few proteins, and computed iteratively from these using a simple model of evolution. The BLOSUM[103] family is derived from statistics on alignments of structurally similar proteins.

Less sensitive, but much faster methods for database lookup via pairwise alignment have been developed. These use computationally cheap exact word matching of short words against a database, and follow up matches over certain threshold criteria with a more computationally costly pairwise alignment. The most widely used must be FASTA [104] and BLAST[105, 106]. The usefulness of the database lookup methods hinges on the availability of estimates of error probabilities. The most commonly used are Karlin-Altschul statistics [107, 108], developed for the BLAST tools, that allows estimation of the probability of seeing a random match to the database from the query sequence given an alignment score and but a few parameters derived from the database.

Sensitivity/specificity tradeoffs are introduced in the choice of method, matrices and gap penalties. Parameters more suited to finding also distantly related sequences will often result in many false positives, while more selective approaches in general lower sensitivity. The optimal parameter set for detection varies from protein to protein given a certain database [102].

Database lookup methods are implemented in several gene finding tools, typically using careful rulesets to weed spurious matches from the actual genes. Other hybrid approaches use results from similarity searches to delineate exonic structure[109]. Similar methods are very useful for delineation of the exon structure by cDNA sequences of mature messages where *cis*-splicing is an issue.

Other database reliant approaches include motif searching, by means of motif specific score matrices [106], profile HMMs[110] or regular expressions [111]. This in particular, but also the other database methods, constitutes a pillar of functional annotation which will be introduced in section 4.5.

4.2 *Ab initio* methods

Ab initio gene finding, also called *de novo* or extrinsic gene finding, at least partially addresses the major weakness with the database lookup methods: our ability to identify novel genes, that are not previously part of the sequence databases. They are based around pattern recognition techniques, and the assumption that gene sequence differs from non-gene sequence in a consistent

manner. This assumption appears valid, given constraints from the transcription and translation machineries and constraints from protein function. If the cellular machinery can tell the difference, a computational model could be able to do the same.

The models are typically built around a number of sensors sensing statistical properties of local sequence, that are integrated into a decision method.

Sensors include simple features of the translational machinery, such as open reading frames, bounded by start and stop codons, periodicities of base three, that arise from the codon grouping of the nucleotides and synonymous codon usage bias, and biochemical DNA strand properties such as bendability. Slightly longer periodicities, hexanucleotides or longer subword frequencies, arising from protein function, as well as biochemical properties of conceptually translated peptides can also be exploited to separate coding regions from non-coding.

Transcription control signals, such as RNA promoter elements, transcription factor binding sites etc, form inputs to some gene finding models. Also, transcription related maturation signals, such as splice sites and polyadenylation sites have been used. Regulatory signals that pertain tot ranscript stability can influence the base composition of translated and untranslated gene regions, and be sensed.

Signals, signal strengths and measures of coding potential based on nucleotide composition are integrated and a decision is made on the classification of an ORF by an algorithm with certain parameters. Upon integration, contextual information and typical ordering of the sensed elements - the gene syntax, can be considered. Linear discriminants, Neural networks, Markov models and hidden Markov models are some common algorithms used by gene finding programs. The model parameters, in sensors and decision boundary, are adjusted using a training material, and tested to estimate the performance.

Codonusage [112] decides on which reading frame to use based on a codon usage table for an organism. Synonymous codon usage in a genome will vary between genes in an organism. Highly expressed genes must be well adapted to the tRNA isoacceptor pools, and use codons for which charged tRNAs are readily available. Calibration by organism average codon usage matrices [113] is common.

Testcode [114] classifies a given sequence as protein coding, uncertain or not protein coding, in a certain window. It employs a content measure

$$\frac{\max(N_{XinCP1}, N_{XinCP2}, N_{XinCP3})}{1 + \min(N_{XinCP1}, N_{XinCP2}, N_{XinCP3})}$$

drawing from the observation that any particular nucleotide in a coding sequence is much more likely to recur at a distance of $2+3n$, where n is an integer, than at distance $3n$ or $3n+1$. This pattern is rare in non-coding sequence. Probabilities of the sequence being coding are estimated for some intervals of the content measure. Likewise, coding probabilities for some intervals of single nucleotide frequency for each base are estimated. For a sequence window, these eight probabilities are weighted by one parameter each and added to form a measure

of coding potential, the Testcode measure. The method could be retrained to suit a particular organism, but in practice the original parameters are most often used. This is motivated by good performance on an early test on some 250 kb of sequence from different organisms.

GeneScan [115] uses a Fourier-analysis measure, and identifies portions of a sequence with high signal to noise ratio for a three-periodicity.

Glimmer [116], as several other earlier programs, uses the frequencies of different oligomers. Long oligomers are powerful in discriminating coding from non-coding sequence, but occur seldom in a training set. By careful weighting of different oligomers, treated as Markov chains, good use is made of the available samples. The program also considers limited gapping [117], by releasing the requirement that the nucleotides be sequential. A tree-like "Interpolated Context Model" is used, and a mutual information measure is applied to select what nucleotide positions within a context to use. Glimmer has shown very good performance in prokaryotes. A splicing enabled extension to Glimmer has been used with some success in the *Plasmodium falciparum* genome [118].

Combining predictions from different gene finders can give more accurate results than even the best component genefinder [119]. A requirement of agreement of all predictions will result in a conservative annotation, and weed out false positives, while including any prediction made by any program will give high sensitivity. The parameters of the combiner method can again be estimated from training data for a specific genome.

AutoMagi, a gene finder used for gene finding in *L. major* and *T. cruzi*, combines CodonUsage, Testcode, Glimmer and Genescan predictions by automation of the approach in [120] and the addition of an additional step of GC-skew based strand estimation[1].

4.3 RNA genes

RNA molecules fill important roles in the cell, even if messenger RNA is not considered. In the trypanosomatids, RNA components are used in at least the machineries for replication, transcription, splicing, RNA modification, message regulation, translation, and membrane translocation [99].

Separate methods are available for finding RNA genes. The database lookup methods can be used at the nucleotide level. Since RNA structure is largely defined by basepairing, methods that consider the pairwise covariaton between bases are useful. Stochastic context free grammars are well suited to finding RNA genes. Several methods for detecting tRNA are available, e.g. tRNAscan [121] which was used in the TriTryp analyses. A probabilistic method for detection of snoRNA [122] was also used. SnoRNAs are involved in several aspects of rRNA maturation, some of them acting as guides for sequence specific methylation.

Attempts to find non-protein coding genes based on RNA structure properties have been made, but the results indicate that RNA structure is not much

[1] http://apps.sbri.org/netmagi

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Table 4.1: A true positive is a correctly predicted actual feature, in this case a gene. A false negative is a missed actual gene. A false positive is a prediction of a gene that does not exist. A true negative is neither predicted nor real.

different from that which would be expected from the sequence composition bias [123].

4.4 Measuring performance

Most sensors and integration methods require calibration, or training, from example data, such as a set of experimentally verified genes. Most mentioned methods have organism specific parameters. For instance, the average synonymous codon usage bias varies between different organisms. A subset of verified genes can be excluded from training and used to estimate model performance in a testing or validation procedure (see table 4.1). The results are often reported as sensitivity

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

and specificity,

$$\mathrm{Sp} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

A number of combined performance measures weighting these have been devised. We will use one such, accuracy,

$$\mathrm{Ac} = \frac{\mathrm{TP} - \mathrm{FP}}{\mathrm{TP} + \mathrm{FN}},$$

that measures the number of correct predictions relative to the number of actual positives and is 1 if all cases are correctly predicted, with no missed cases or erroneous calls.

These measures can be applied at different resolutions - at the nucleotide, exon or gene level. The gene level is predominantly used in the present investigation, with a few measures taken at the nucleotide level for the prediction of splice sites and polyadenylation sites.

For an organism where little of the protein complement is known, accuracy estimation can be difficult. The gene finding methods would ideally be trained and tested on a few contiguous regions from the genome, where it is completely clear which nucleotides are coding and non-coding, which constitute regulatory signals and so forth. In practice it is especially hard to rule that a particular sequence is not expressed under any circumstances. It can still be difficult to

find such regions even for the most studied model organism genomes. Another interesting objective for these automated methods is to perform as well as an expert annotator. Then the test set can be a "gold standard" that reflects the opinion of expert annotators, and the method is judged according to how well it can automatically perform the gene calling task instead of the annotator.

Much of the development in gene finders of late has been in resolving intron-exon structure, to make the boundaries more exact and to incorporate related genomes to enhance the predictions, to construct state of the art methods such as TwinScan [124]. In the intron scarce trypanosomatid genomes, older programs as well as approaches originally intended for prokaryotic gene finding are useful.

4.5 Functional annotation

Once structural gene models have been established, functional annotation can start. This is usually taken to encompass the entire process where biological functions are inferred from the sequence information.

The evidence connected to a sequence can be plentiful and diverse. Much biological knowledge is not well represented in the many databases and tools available. It is reasonable to assume that educated guesswork by experts on certain gene families, pathways or organisms, well supported by what information can be extracted by automatic means, generally constitutes the best possible method without additional experiments. The approach is sometimes referred to as semi automatic annotation.

A common decision task set before the annotator is to inspect a large set of sequence based predictions, including database and motif matches, and the literature regarding these, predicted features of the translated protein, from hydrophobicity, pH and isoelectric point to protein cellular localisation, based on signal peptides, nuclear import and export signals, membrane spanning regions, secondary and tertiary structure, and regulatory signals. The annotator would judge the quality of database matches and motifs based on the alignments, taking not only scores into account - which an automaton could easily do - but also structure-function relationships and a broad biological perspective. Likewise the other sources of information would be weighed together based on the various likelihoods associated with different predictors, ideally to form a comprehensive picture of gene function. The function can subsequently be expressed in a controlled vocabulary to simplify downstream analyses. An interesting option is Gene Ontology terms, into which some parasite specific terms have been introduced [125].

Once all genes have been so analyzed, an expert annotator can proceed to establish if all expected components of pathways, molecular machine complexes are present. Often glaring holes in metabolic pathways can be filled by careful search for a particular function. If not, new hypotheses of alternative or substitute pathways can be formulated.

If closely related genomes are available, a transitive annotation can be made, where close homologs automatically receive function assignments from the re-

lated genome.

The current paradigm has it that sequence and structural similarity imply or is equivalent to functional similarity. However, it is also known that small differences in sequence or structure can cause a difference in function. The change of even a single catalytic amino acid in an active site, or an amino acid that determines substrate specificity can have dramatic effects [126]. The Janus peptide showed that proteins in the 40 % identity range, which will often receive good homology scores, can have completely different structures [127]. Near perfect sequence conservation is thus certainly not a guarantee for conservation of function. Also, a protein can have different functions in different situations and at different times. Conversely, function can be similar regardless of sequence conservation, as in convergent evolution. Functional annotation should not be taken as experimental evidence, except in the cases where genes have been previously studied, but rather as conjectures and hypotheses to be tested. Nevertheless, the statistically measured performance of many of the prediction methods used is already impressive, and the additional expert knowledge should increase the accuracy further.

4.5.1 Annotation platforms

For semi automatic annotation, an interactive graphical presentation of the different lines of evidence is natural. Modern alternatives suitable for protozoan genomes include Artemis [128], which was extensively used by WTSI for the *L. major* and *T. brucei* annotation. A GUI [I], which was used for some parts of the *T. cruzi* annotation at KI, is another example. These are sequence centric tools, suitable in particular for additional structural annotation where gene models are not already well defined.

In contrast, gene centric tools rely to a greater extent on the an automated pipeline or initial semiautomatic structural annotation using other tools. Examples include the two companion packages Manatee and Sybil, that were originally developed at TIGR. Manatee[2] is designed for single genomes and Sybil[3] enables comparative annotation against multiple Manatee-compatible genome databases. They were built employing the Coati[4] architecture with three abstraction layers: a bottom, database layer, a top, graphical presentation, layer and a middle interface layer connecting the two by abstraction of the database layer.

Manatee was used for functional annotation at TIGR, and via Sybil in the TriTryp comparative annotation. ACT[129], a comparative extension to Artemis, was also used by the WTSI.

[2] http://manatee.sourceforge.net/

[3] http://sybil.sourceforge.net/

[4] Named after the coatis met in Foz do Iguassu, Brasil, at the 2001 TriTryp meeting, where the initial structure was agreed upon.

"The universe is asymmetric."

– Louis Pasteur

Chapter 5

Strand asymmetry

5.1 Chargaffs rules

Chargaff postulated four rules [130] based on observation of the properties of cellular DNA. Two of these are called the 'parity rules'. The first parity rule states that in DNA, the total base content of A is equal to that of T, and that of C to that of G. The basis of this was not understood[131], but became evident two years later with Watson and Crick's model of the DNA structure [132]. Chargaff also observed that pyrimidines and purines tend to cluster in the sequence and often occur as oligonucletides - the 'cluster rule', and that GC content tends to be constant within a species, but vary between species - the 'GC rule'.

The second parity rule states that if the two DNA strands are separated, the A = T, C = G rule is still approximately correct within each of the strands, i e the intra-strand compositions share the pairing of the inter-strand ones. Genome sequences give us a good, albeit approximate, confirmation of this. The genome sequence - one strand - of *Eschericia coli* K12 [133] has 1142228 A and 1140970 T, but 1176923 G and 1179554 C. In *Leishmania major* Friedlin [99] chromosome 36, there were 781455 G and 777798 C but 565517 A and 557413 T. In *Trypanosoma brucei* chromosome 1 [30], there were 290021 A and 293719 T but 247081 G and 233651 C.

It is possible to derive the second parity rule from the first. With no bias between the DNA strands, the number of unique single nucleotide substitution rates are halved due to the Watson-Crick base pairing between strands [134]. Under equilibrium assumptions, this leads to inter-strand A = T, C = G [135].

5.2 Skew

While the second rule holds true for most genomes and chromosomes, there are local variations. These can be measured by counting the GC skew $(G - C)/(G + C)$ and AT skew $(A - T)/(A + T)$ in sliding windows over the genome

sequence [136]. For virii and bacteria the sign of these measures often coincide well with directions of replication and transcription. This becomes evident when cumulative GC skew curves [137] are used. These have optima at the origin and terminus of replication for many bacteria.

The genetic code leaves ample room for variation in the third codon position. Here, selection acting at the protein level will have little impact. The GC or AT skew is often prominent in the third codon position and in intergenic regions.

The replication fork is asymmetric with regard to strand. This can lead to different substitution rates on the leading and lagging strand, as is the case in *E. coli*[138]. Also other cellular machineries involved in the DNA metabolism differentiate between strands. In transcription, the non-coding strand - the template - is more protected from mutation than the coding strand [139]. Also, it also undergoes transcription coupled repair [140]. The substitution rates of the two strands can thus be different, which could lead to skewed nucleotide distributions [141].

In bacteria, there is a tendency for transcription and replication to progress in the same direction [142]. Colliding polymerases are problematic, and would at least slow the growth of bacteria. Functions for resolving collisions exist in bacteriophages.

Eukaryotes typically have less conspicuous skew curves, with many local maxima and minima, possibly due to multiple origins of replication [143, 144]. The striking pattern found in *Leishmania major* [145], with maxima and minima of the cumulative skew curves coinciding with strand switches, was unexpected.

"All organisms adapt to changes in their environment by adjustments in gene expression, and in all organisms, from Escherichia coli to man, the most important control point is at transcription initiation. All, that is, except those belonging to one very small family of early-branching eukaryotes, which seems to have completely lost the ability to regulate transcription by RNA polymerase II."

– Christine Clayton, *The EMBO Journal 21, p1881* 2001

Chapter 6

Gene expression

A genome, without the machinery to interpret and express it, is a dead molecule. The genome contains the information required to build more of the expression machinery, and signals that determine under what conditions and in what quantities the genes are expressed, but a system for expression must be present to interpret this.

As most other eukaryotes, the trypanosomatids have three principal DNA dependent RNA polymerases, but the three polymerases of trypanosomes have some unusual features (reviewed in [11, 146]). PolI transcribes some surface antigen genes in *T. brucei* [147] in addition to the expected rRNA, and polII the SL RNA [148, 149] genes as well as protein coding pre-mRNA. PolIII transcribes all U-rich snRNA in addition to the tRNAs.

Promoters are known for polI transcribed rRNA-genes, as well as the *T. brucei* bloodstream stage VSGs and insect stage procyclins. The regulation of the VSG genes is not fully understood. But, a single polI containing nuclear body [150] is located at one expression site per cell. This indicates a choice of transcription site, which is consistent with the apparently epigenetic regulation [151].

The polII transcribed SL RNA genes also have promoters, although these are not typical for polII. While it appears clear that polII transcription of the bulk of the pre-mRNA originates in strand switch regions, as shown in *L. major* [9, 10], no known promoter or otherwise overrepresented conserved sequence elements have been detected. Some larger areas of unusual sequence composition are present on the strand switches. In *T. cruzi*, GC-islands in certain strand switch regions have been shown to be centromeres [152]. *L. major* has AT-rich strand switches instead[153]. The coding strand appears about 10 fold more transcribed than the non-coding. The latter may however still be important, at least as anti-sense RNA, if not protein coding [154]. Transcriptional terminators are known, and can contribute in determining the strandedness of transcription.

Many housekeeping genes in the trypanosomatids, and in *T. cruzi* in particular, are repeated. This might constitute a primitive mechanism of control of the transcript concentration [5]. Naturally, this could be evolutionarily useful,

since pressure to keep the individual copy intact is reduced and variation between the copies can be allowed. While a copy number mechanism only offers crude differential control, the concentration of many housekeeping proteins does not need to vary much. This is consistent with data from a recent *T. cruzi* proteomics study [155]. This study detected 1500-2000 proteins from each stage. 30% of the proteome was found to be expressed in all tested lifecycle stages. Detection counts are related to concentration, although precise measurement is precluded, and houskeeping genes were among the most abundant in all stages. This speaks in favour of a limited variation in expression for at least this gene complement.

6.1 Transcript maturation

The trypanosomatid transcript maturation process is unusual. Genes are expressed as polycistronic messenger pre-mRNAs. Essentially only monocistronic transcripts are known to be translated, except in cases where internal ribosomal entry sites exist on the transcript to allow ribosomal entry at a downstream start codon (e.g. [156]). This mechanism has so far not been found in trypanosomes. Ribosomes are also known to re-initiate on certain transcripts even with multiple upstream open reading frames [157].

The 5' end of each mature transcript is provided by a small capped[158] spliced leader (SL) transcript[159]. The SL miniexon is important for export from the nucleus, at least during SL RNA biogenesis [160], as well as for translation [161]. The SL RNA's length is different in different species, around 95 nt to 135 nt. The exonic part of this is between 35 and 39 nt, and is added to the pre-mRNA by a *trans* splicing reaction[162, 163, 164, 165] (reviewed in [7], and depicted in figure 6.1). This is a two step process in which the SL RNA 3' end is attached to the branch site just upstream of the splice acceptor of one of the genes in a pre-mRNA, forming a Y-shaped intermediate [166], and the 5' splice donor attaches to the start of the pre-mRNA exon, at the 3' splice acceptor. The Y-shaped intronic sequences are degraded.

The recognition of the 3' *trans* splice acceptor appears similar to that of *cis*-splice acceptors in other eukaryotes. A pyrimidine stretch is required for splicing [167]. A pyrimidine stretch of more than 7 nt can be found between the branch site and the splice acceptor AG dinucleotide of efficient splice acceptors [168]. The exact minimum requirements are not known. Alternative *trans* splicing, where mature transcripts of different lengths are produced, is known. Examples include *T. brucei* procyclins [169] and the *T. cruzi LYT1* gene[170].

The process of *trans* splicing is directly connected to that of polyadenylation of the message 3' end [171, 172]. While the location of polyadenylation depends on an upstream splice-signal [173, 174], the exact sites can not yet be predicted. Polyadenylation is also directed by sites with a strong resemblance to the splice acceptors, but that are not major splice acceptors of any known downstream genes [174]. There is no clear consensus sequence at the polyadenylation site, as in e.g. yeast. Mapping studies sometimes identify multiple sites ([171] and

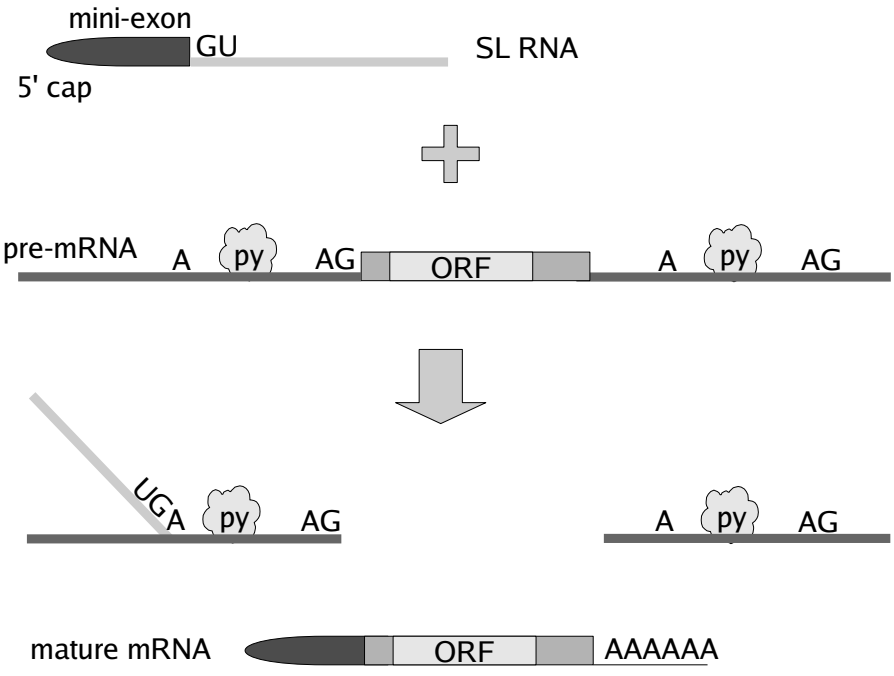

Figure 6.1: Trypanosomatid transcript maturation occurs via *trans*-splicing of a mini-exon sequence to the pre-mRNA and polyadenylation. Both processes are directed by signals including a branch site adenosine, a pyrimidine stretch and an AG dinucleotide.

references therein) for the same gene. Different distances are also found for different genes, even across studies where the same major site is observed for many transcript clones.

6.2 Post-transcriptional control

The trypanosomatid parasites achieve orchestration of cell cycle progression and differentiation into morphologically and biochemically distinct states. Yet, for most genes investigated, control of gene expression is not predominantly transcriptional (reviewed in [8]).

The procyclins in *T. brucei* give a well studied example. Expression of procyclin molecules on the surface while in the bloodstream is potentially lethal to the parasite, while expression of at least some of them is important for survival in the insect stage [175]. The regulation of the procyclins is tight [176] - no procyclin has been detected in bloodstream forms. Although the procyclin

genes have a transcriptional promoter, the down-regulation of transcription in bloodstream forms is only about 5-10 fold [177, 178]. Further levels of post-transcriptional regulation function in addition to the promoter to give the full dynamic range of expression. A procyclic stage stabilising *cis*-element [179, 180] on the 3'-UTR gives a 10-fold factor. The transcripts are rapidly degraded in the bloodstream form. Additional stabilisation occurs in the procyclics. Changes in translational efficiency, via known elements in the 3'-UTR, and possibly also other translational blocks, result in even lower translation levels in the bloodstream form. The post-transcriptional contribution to the regulation is in the end greater than the transcriptional control.

Transcript maturation, in particular the *trans* splicing and polyadenylation step, can affect the abundance of transcripts[167]. The *trans* splicing signals direct maturation with greatly varying efficiency. The maturation step could also be differentially regulated, as with the transcripts from the *T. cruzi* LYT1 gene[170]. One of the alternative transcripts showed a drastically reduced abundance in epimastigotes. Although the experiments did not ask for expression explicitly, this still indicates an important role for transcript maturation in post-transcriptional control.

Transcript stability can be modified by *cis*-acting elements elements. A *T. cruzi* amastin 3'-UTR element binds a *trans*-acting protein, which gives a 7 times longer halflife[181]. Such mechanisms are also used to achieve differential regulation. Transcripts of the amastin gene were 50 fold more abundant in amastigotes as compared to epimastigotes and trypomastigotes[21], while they were transcribed at an equal rate. AU-rich and G-rich elements on the 3'-UTRs have been shown to affect the stability of mRNAs in *T. cruzi* also differently in different stages [182]. The TcUBP1 and TcUBP2 proteins bind to the 3'-UTR motifs and form a stabilising complex [183, 184].

DNA microarrays have been used to examine the message levels of a larger range of genes in different cellular states, mainly in the life cycle stages, but so far met limited success. Only very limited control of transcript levels has been found, i.e. the levels of most genes tested show only small changes between life cycle stages. This is perhaps consistent with what can be expected with little transcriptional control [185]. Still, while transcription levels may be relatively similar for many genes, different transcripts have different halflives, as is even the case for some identical transcripts under different circumstances. Similar studies in *L. major* showed somewhat larger transcript level variations [186]. Recent *Trypanosoma cruzi* proteomics data [155] indicate more dynamics in gene expression at the protein level than do similar DNA microarray experiments [187]. Both studies are limited in extent. The DNA array was further limited by the incomplete state of the genome and limited gene annotation available in 2003, and the proteomics approach gives only limited quantitative information. A careful analysis of both transcription and translation at a large scale would be of interest. Likely, mRNA expression and stability will have a larger impact for some genes, and controlled translation and protein degradation for others. There are also indications that motifs - even the same motifs - in the 3'-UTRs affect translation in addition to transcription [179, 180]. Developmental regulation

of protein expression can also be conveyed by translational control mechanisms (e.g. [188]).

The codon usage of genes affects the translation of proteins due to differences in tRNA isoacceptor concentrations (see e.g. [i] and references therein). An early study on codon usage identified a narrow spectrum of within genome variation in *L. major*, but *T. brucei* and *T. cruzi* appeared more diverse [189] (see also 7.4.3).

Upstream open reading frames, AUGs on the exonic but normally untranslated 5'-UTRs of transcripts, generally down regulate translation of a transcript (reviewed in [157]). When the ribosome encounters an uAUG, one of several things can happen, depending on the state of the ribosome and on the sequence context of the AUG (figure 6.2). The ribosome can begin translation, translate the uORF and unload the transcript. The transcript can also under some circumstances be reloaded downstream of the uAUG. The uAUG can be ignored - leaky-scanned. uAUGs can also induce mRNA degradation. There are examples of transcripts with multiple uORFs that still express protein, albeit at a much lower level than with the uAUGs removed, e.g. the fungal *GCN4* with four uORFs, which is conditionally regulated on starvation [190].

uAUGs are common in eukaryotes. 15-52% of the 5'-UTRs had uAUGs, as estimated in 10 eukaryotic species[191]. uAUGs are also conserved between species in mammalia[192]. Upstream AUGs have been reported to down regulate translation of a gene in *Trypanosoma cruzi* [193] and on a reporter gene in *Leishmania major* [168].

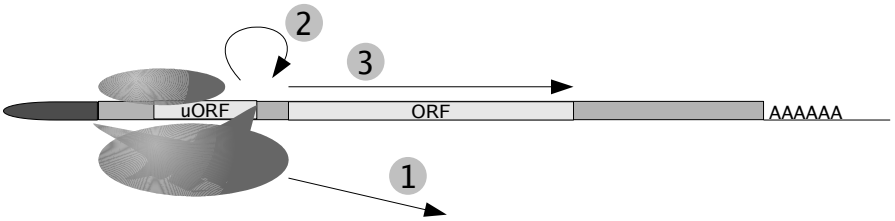

Figure 6.2: When the ribosome encounters an upstream AUG, one of several things can happen. The uORF can be translated, and the transcript unloaded (1). Ribosome stalling and transcript degradation are possible. The ribosome can translate the uORF, reload the transcript (2) and resume scanning to later translate the downstream gene (3). The ribosome can also leaky-scan past the uAUG without initiation and translate the gene (3).

"If we knew what it was we were doing,
it would not be called research, would it?"

– Albert Einstein

Chapter 7

Present investigation

7.1 Aims

- To participate in the sequencing of the genome of *Trypanosoma cruzi*.
- To provide a functional annotation of the gene content.
- To identify further features of the genome, especially with regard to post-transcriptional control.
- To develop algorithms and construct computer tools to aid the above as needed.

7.2 A graphical tool for parasite genome annotation (I)

We have constructed a tool for semi-automatic annotation and genome sequence visualization [I]. Visualization of predicted genes from several different gene finding programs, database homology search results, together with DNA and protein sequence properties provides a human annotator with ample decision support for calling genes. By combining the results from several different gene finding programs as well as database matches, a good basis for further annotation can be achieved. Based on organism specific knowledge we have also developed additional heuristics to facilitate the annotation process without loosing manual control. The program has been in routine use, and is provided free of charge for non-profit use.

7.3 Strand asymmetry patterns in the kinetoplastids (II)

We have investigated the base skews of three kinetoplastid parasites [II]. We report overall skew patterns similar to those found in bacteria, with optima coinciding with strand switches of unidirectional gene clusters. This skew is not directly caused by a skew inherent in the codon usage, since it can also be observed in intergenic regions, but can rather be hypothesized to have affected the codon usage. In particular, the derivative of the cumulative GC-skew of *T. cruzi* and *T. brucei* has the opposite sign to that of *L. major*. By analogy with bacteria, it is tempting to suggest that *T. cruzi* and *T. brucei* have origins of replication at the cumulative skew optima. The same analogy would place termini of replication at the corresponding strand switches in *L. major*. But, the three organisms share a strong conservation of synteny. Such a difference would require a major divergence in the replication machinery. Furthermore, it seems unlikely that regulatory signals are causing the difference seen between *T. cruzi*, *T. brucei* and *L. major*. We could not rule out simple, strand asymmetric repeats as a cause of the observed difference. Nevertheless, it seems likely that a difference in the transcription coupled repair or replication machinery is causing the difference in skew.

7.4 Gene finding in *Trypanosoma cruzi* (III)

7.4.1 Automating A GUI*

The tools available for semi-automatic annotation in A GUI were automated to allow for whole genome predictions in a limited time frame. Glimmer, Testcode, splicemodel, high GC-content and long ORF based strandedness predictions were combined using many different combiner settings.

Since a gene to be expressed by the polII machinery in the trypanosomatids must be spliced and polyadenylated, we used a version of the splicemodel [IV] trained on *T. cruzi* [V] as a gene prediction tool. It is interesting to note that a successful approach to score the coding potential of *T. brucei* genes essentially detects the presence of splicing signals [194].

The strand filtering predictions used a non-iterative version of the strandedness heuristic[I]. The GC-content score was not used. The procedure is described in figure 7.1. The combinations are made either serially or by majority vote. Serial combination is akin to "and" type combination, although the order of the operations has a slight impact in the cases of strandfinder and splicemodel. This is a conservative combination of the included gene finders, compared to predictor majority vote.

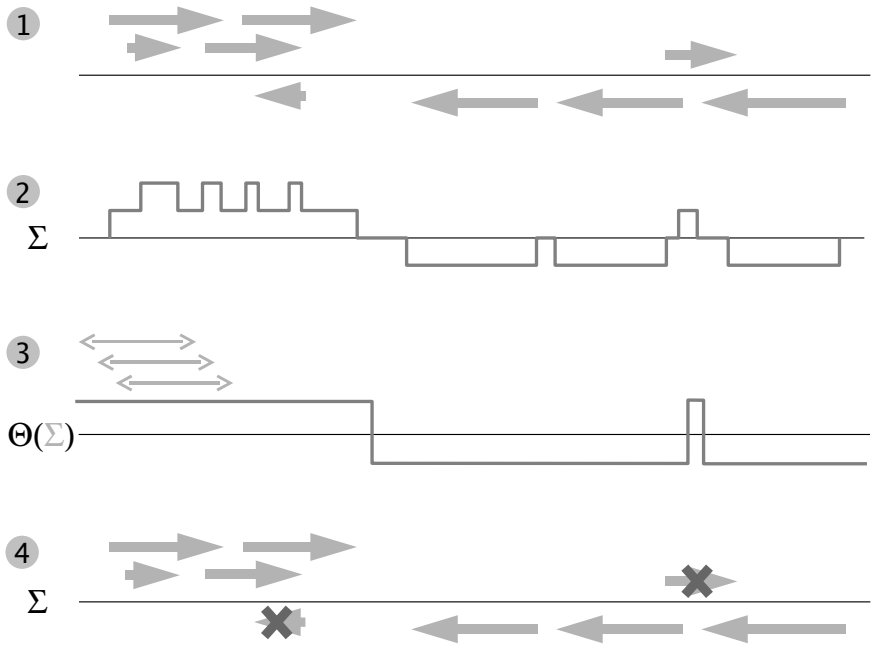

Figure 7.1: A re-implementation of the strand finder heuristic. ORFs on the Watson strand are assigned a score of 1 while Crick strand ORFs receive -1 (1). The scores are summed for each sequence position (2). The score is smoothed using a sliding window step function (3). ORFs are given a new score as the sum of the smoothed scores in each constituent sequence position (4). If the sign of this new ORF score disagrees with the sign of the ORF strand, the ORF is discarded.

7.4.2 Have we found all genes?*

Before the gene prediction on the whole genome started, around 1500 previously studied genes were already present in GenBank. A subset of these (~ 100) had been used as a training set for Glimmer, and even more for the estimation of codon usage tables. Thus they do not formally constitute a validation set, in the sense that extrapolations to the performance for other genes could be made. It is, however, certainly informative to know how many of the already established genes the gene finding methods missed. Many of the GenBank entries were obtained from resequencing of essentially the same gene. If these were used in the test set, uneven weight would be given to performance on a few genes. We thus clustered the 1507 genes with NCBI blastclust, at default settings, to produce 770 non-redundant representatives. 180 of these were shorter than the minimum 300 nt set for the shortest gene. Some of those were fragments and could potentially still be useful, and were initially retained in the analysis.

	Found genes						Red
		relative		FN relative		Pred.	vs
	#	Scaff	ORFs	Scaff	ORFs	num.	ORFs
T. cruzi annotation set	615	1.00					
ORFs > 299 nt	603	0.98	1.00	0.02		93070	
Glimmer	582	0.95	0.97	0.05	0.03	34152	0.63
Glimmer · Strandfinder	572	0.93	0.95	0.07	0.05	28674	0.69
Glimmer · Splicemodel	563	0.92	0.93	0.08	0.07	27920	0.70
Glimmer · Strf · Sm	563	0.92	0.93	0.08	0.07	23572	0.75
Orffinder > 299 nt	606	0.99	1.00	0.01	0.00	93432	
Orffinder · Strf	594	0.97	0.98	0.03	0.02	49071	0.47
Orffinder · Sm	591	0.96	0.98	0.04	0.02	63134	0.32
Orffinder · Sm · Strf	582	0.95	0.96	0.05	0.04	34147	0.63
Orffinder · Strf· Sm	577	0.94	0.95	0.06	0.05	33579	0.64
Testcode (w no opinion)	589	0.96	0.97	0.04	0.03	60794	0.35
Testcode (confident)	494	0.80	0.82	0.20	0.18	36063	0.61
G+O·(Strf+Sm)(≥ 2)	583	0.95	0.96	0.05	0.04	38045	0.59
G+O·(Strf+Sm)($= 3$)	557	0.91	0.92	0.09	0.08	21779	0.77
G+O·(Strf+Sm+T)(≥ 3)	573	0.93	0.95	0.07	0.05	26823	0.71
G+Strf+Sm+T(≥ 3)	581	0.94	0.96	0.06	0.04	28503	0.70

Table 7.1: · denotes sequential combination, + denotes majority vote according to the number in braces. The Found column has the number of genes identified out of 770 possible. G Glimmer, O Orffinder, Sm Splicemodel, Strf Strandfinder, T Testcode.

The protein sequences were searched against conceptual translations in all six reading frames of the target data, the genomic scaffolds longer than 5 kb. GenBank entries not found in the assembly - often old entries sequenced with less reliable techniques - are not useful for gene prediction comparisons[1]. 615 genes were found by BLAST search in the annotation set. These formed the baseline for further comparison of the prediction methods. Estimates of method FN and Sn were obtained by comparison to these genes. The reduction in number of predictions from the more sophisticated methods compared to a naïve orffinder were used as a tentative estimate of Sp. These approximative numbers agreed well with the average Sn and Sp derived from comparing a few automated predictions to a manual annotation gold standard on three reference sequences[2] and were used during the development phase to evaluate performance (shown in columns "Found rel Scaff" $\sim$ Sn and "Red vs ORFs" $\sim$ Sp in table 7.1).

As a final test, five contigs from the genome assembly were annotated man-

[1]On a related issue, initial worries that the assembly was incomplete were mitigated by comparisons between assembly and the clone-by-clone approach, as well as inspection of the missing genbank entries.

[2]The sequences were TIGR BAC 42O19, a SBRI cosmid from Chr 8 and the UU/KI Chr3 strandswitch[5], and the unpublished annotations were made by P. Myler, SBRI.

	Genes	FP	FN	Ac
Manual	168	0	0	100%
KI	200	47	16	69%
AM (2 of 3)	168	29	29	65%
AM (2of 4 with skew)	231	70	8	66%
AM (3 of 4 with skew)	143	11	35	68%

Table 7.2: Comparison of AutoMAGI at different settings with the automated A GUI approach against a manual gold standard annotation on five contigs.

ually by the SBRI team. For this analysis, the KI prediction from Glimmer, Strandfinder, Splicemodel and Testcode at qualified majority (at least three out of the four in agreement) was chosen since it showed good trade off in increased specificity in return for a small decrease in sensitivity as compared to the most sensitive methods, although it did not have the best overall accuracy.

168 genes were found in total. The KI approach was tied with AutoMAGI for overall accuracy (table 7.2), but at the same accuracy, AutoMAGI had a better Sp whereas the KI approach had better Sn. The previous was ultimately considered preferable, since in the following comparative annotation we could catch many genes missing from *T. cruzi* based on their homology to *T. brucei* and *L. major*, whereas the work with weeding out false positives would have been considerable. The annotation may possibly have lacked some *T. cruzi* specific genes due to this decision.

Observing the rather high FP rate required for AutoMAGI to reach a better Sn than the tested approach, it could be considered unfortunate that we did not evaluate the predictions with our gene finder at settings for higher specificity. But, since the AutoMAGI predictions were satisfactory and a round of comparative and manual annotation lay ahead, this was acceptable.

An additional source of information used in the AutoMAGI predictions, aside from CodonUsage and GeneScan, was GC-skew optimum based strandedness predictions. This should have made the assignment of strandswitches more exact than in the strandfinder approach. The analysis in [II] was performed on annotations made without use of any GC-skew directed strand choice.

It is interesting to note that there is a considerable number of ORFs with predicted splice-sites on the strand assigned as the non-coding (going from Orffinder splicemodel to Orff·Sm·Strf in table 7.1). This is consistent with results from later careful studies of gene expression in *L. major*, where two stable processed transcripts were detected from the non-coding strand, in addition to 10 from the coding strand [154].

7.4.3 Codon usage groups*

Codon usage bias has traditionally been estimated by selecting a set of ribosomal and other highly expressed proteins and counting the codon usage for these. A recent algorithm [195] attempts to find such a dominating codon usage

group *de novo* without consideration of annotation. The algorithm discovers a "dominant" group of genes that gives a high CAI when used for computing CAI weights by an iterative procedure. There is no guarantee that the most dominant bias is a translational one. The correlation to GC content, GC3 bias, GC-skew or other possible sources of biased codon usage must be investigated in order to be certain that a group of genes with actual translational bias has been found. It is possible to check what genes conform to this codon usage, and test whether they are highly expressed.

This algorithm was reimplemented, and found a dominant codon usage groups for each of the TriTryp genomes. The average codon usage was similar between *T. brucei* and *T. cruzi*. Codon usage in the dominant groups was also similar, and close to the average codon usage in *L. major* (figure 7.2).

The new codon adaption index(CAI) values showed low correlation to GC-skew ($R< 0.01$). A pre-genome sequence study on codon usage found a strong positive correlation between G+C content in the third codon position (GC3) and gene expression in the TriTryps[189]. As expected we found a correlation to GC3 ($R=0.91$) in *T. cruzi*. A comparatively weak correlation to GC content was also found ($R=0.33$), but the GC3 and GC unsurprisingly confounded ($R=0.41$). The results in *T. brucei* and *L. major* were similar. However, in *L. major*, the GC3 and GC content is not correlated; the variation in GC content is small although the GC3 is highly variable and strongly correlated to CAI. Consequently, no correlation between CAI and GC was found in *L. major*.

The functional annotation of the of the high CAI scoring genes indicate that these are highly expressed genes, such as histones, ubiquitin, ribosomal proteins, elongation factors, heat shock proteins, calmodulin, tubulin, paraflagellar rod protein and some glycolytic enzymes. A translational bias towards this group seems possible. The dominant set also includes genes for tryparedoxin peroxidase as well as putative genes for dynamin, ubiquitin hydrolase, IgE-dependent histamine-releasing factor, acetyltransferase and sets of genes for putative calcium binding and flagellar calcium binding proteins. While not expected, it is reasonable to assume a stress response gene such as tryparedoxin peroxidase to be efficiently translated. The high scoring set also includes DGF-1. DGF-1s are encoded by relatively large genes - often over 10 kbp. It is tempting to speculate that the use of abundant isoacceptors is a necessity to ensure timely and error free translation of a gene much larger than the average. Five hypothetical proteins were present in the dominant set. These were all located in close proximity on one contig[3].

The results for *L. major* and *T. brucei* are similar. The dominant set for *T. brucei* contained most of the expected genes found for *T. cruzi*, but lacked calmodulin and calcium binding proteins. In additon genes for putative universal minicircle sequence binding protein (UMSBP) and membrane protein KM-11 were in the set. It seems reasonable to assume that these are also highly expressed in some circumstances. A small group of six conserved hypothetical

[3]Tc00.1047053508153 with gene numbers 40, 70, 80, 90 and 110

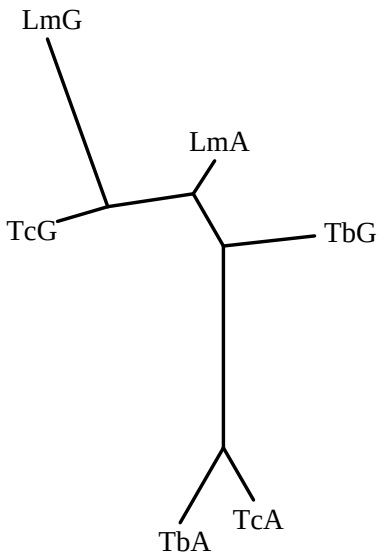

Figure 7.2: Relations between the codon usage tables of Lm *L. major*, Tb *T. brucei*, Tc *T. cruzi*, A all genes, G dominant codon usage group. The average codon usage for all genes is similar in *T. brucei* and *T. cruzi*, but the codon usages of the dominant groups are similar to the average codon usage of *L. major*. Drawn using EMBOSS cusp and codcmp[196], PHYLIP Neighbour and Drawtree[197].

genes were also in this set [4].

The *L. major* dominant set of genes is also similar to that of *T. cruzi*, but includes UMSBP and the activated protein kinase c receptor (LACK) genes. Eight hypothetical protein genes have a codon usage that indicates that they are highly expressed under certain conditions [5].

7.5 The genome sequence of *Trypanosoma cruzi* (III)

Together with our close collaborators at TIGR and SBRI, we presented the gene content and insights into the genome structure of *T. cruzi* CL Brener [III]. CL Brener was concluded to be a hybrid, showing considerable haplotype differ-

[4] Tb10.26.0240, Tb11.01.5740, Tb11.02.2540, Tb11.03.0660, Tb927.3.2560 and Tb09.211.1240

[5] LmjF24.0410, LmjF32.2420, LmjF34.3140, LmjF35.1980, LmjF36.3820, LmjF36.4820, LmjF32.2520 and LmjF06.0710

ence. The differences are clustered in blocks, with intervening regions of higher conservation. The genome is also highly repetitive. These factors made an accurate structural assembly very difficult. Current scaffolds are mostly short, but show good correlation to molecular karyotype markers. The structure of the telomeric regions, home to some of the surface antigen molecules, was further analyzed. 49 contigs with telomere repeats were found present in the genome sequence. The large repetitive surface molecule families, of keen interest for the understanding of host interaction and immune responses, were investigated. A novel large family, named mucin associated surface protein, was discovered and subsequently confirmed to be expressed [155]. The repetitive character of the genome was investigated, also in relation to *L. major* and *T. brucei*. Signalling pathways were analyzed, in particular the phosphatases and kinases, showing some promise as potential drug targets.

My personal involvement consisted of setting up and running the computer infrastructure in Uppsala and at the Karolinska, in-house database development and interfaces between these and the sequencing crew as well as towards the main *T. cruzi* sequence repository at TIGR. I have also taken part in the continuing discussions on the analysis of genome content and structure. Aside from the work on gene-finding, annotation software, message boundary predictions and uORF predictions, which is presented elsewhere in this thesis, I have been directly involved in the annotation of the *T. cruzi* surface molecules and telomeric regions, the TriTryp kinases and RNA binding proteins. While the genome analysis of *T. cruzi* revealed many more interesting features, the remainder of this section will expand upon these areas. I also took part in the analyses of the *T. cruzi* repeat content, in particular in the development of software for the estimation of putative collapsed tandem repeats [III, table S4], and contributed to the comparative structural and functional annotation of the TriTryp genomes[iii].

7.5.1 Surface molecules

The *T. cruzi* surface is covered with extensively glycosylated GPI anchored proteins of different kinds. These have previously been partially analyzed by cloning and sequencing of individual members, as well as non-stringent hybridisations with probes to estimate copy numbers. A large heterogeneity in the surface molecules was expected, and also found.

Trans-sialidases

T. cruzi cannot synthesise sialic acid; rather parasite shed surface *trans*-sialidases transfer sialyl moieties to surface mucins. These molecules appear to be important for parasite survival (see e.g. [198]). Initially, we worked from the hypothesis of six distinct *trans*-sialidase subgroups [16] (table 7.3), with 1a, 1b and 1c groups belonging to the *trans*-sialidase group, and group 2-4 to the *trans*-sialidase like group. Using BLAST searches with representative sequences from these subgroups as well as shorter peptide sequence signatures, such as SAPA

TS group	Function or notable members
1a	enzymatically active; SAPA repeats and Tyr active site residue
1b	enzymatically inactive; SAPA repeats and His active site residue
1c	enzymatically active; SAPA repeats and Tyr active site residue
2	GP85-GP90
3	FL160 (CEA/CRP)
4	Tc13 with EPKSA-repeat

Table 7.3: *Trans*-sialidase subgroups as initially defined for searches.

repeats, the sialidase SXDXGXTW, the N-terminal FRIP motif and the subterminal motif VTVxNVfLYNR, a matrix of similarity to the subgroups was established. The diversity was larger than expected, with many putative proteins that defied previous category boundaries. Also, many genes had good overall homology, but lacked what was previously thought to be key signature motifs.

Due to the within-group and between-group diversity, the six-group schema was abandoned. We classified the TS into one active TS category, with high similarity to the active form and the key tyrosine residue conserved [199], and one large TS-like group [III, table S13], in accordance with the previous major group division [16].

The active group was smaller than anticipated from previous estimates by about an order of magnitude. Collapse of repeat copies may explain a large part of this. The read coverage is deep on the active molecules, but without actually resolving the repeats, an accurate count is difficult.

Much variation was present in the TS-like group (as can also be seen from the TribeMCL clusters at different similarity cutoffs [III, table S5]), and it will be interesting to see further functional studies on subgroups of these. Several immune-response related effects have been shown for different catalytically inactive TS molecules (see e.g. [16] and references therein).

Mucins

BLAST searches using some 40 partial mucin sequences from different categories allowed the annotation of the mucins according to established subgroups of TcMUC[19] and TcSMUG[23] (see [III, table S14]).

The insect vector expressed Tc SMUG showed low variation and few assembled copies, although the copy number is almost certainly underestimated due to collapses of nearly identical repeat copies [III, table S4]. In contrast, a large variation was seen among the *TcMUC*. This is interesting, as these are expressed in the mammalian host and exposed to the immune system. No homologs were found in the other TriTryps, but eight of the PSA-2 GPI anchored surface proteins in *L. major* were found to be structurally similar to the *T. cruzi* mucins, and possess repeats similar to the T_7KP_2 otherwise found in TcMUC group I.

GP63

A massive expansion of putative GP63 metalloprotease genes, better known as *MSP* genes, was found in *T. cruzi* (see e.g. [III, table S5] and [99], table S17). GP63 is primarily known as a *Leishmania* spp. surface dominant protein and facilitator of complement lysis defense [18], but is also known to be expressed in both *T. cruzi* and *T. brucei*. A reason for the expansion in *T. cruzi* is not known, but due to the importance of GP63 in *Leishmania* sp. virulence, further investigation is warranted. The expansion follows the general tendency of expanded surface molecule gene families in *T. cruzi*.

7.5.2 Protein kinases

The TriTryp kinases [III, table S12] were identified using hmmer[200] searches for PFAM[201] kinase domain (PF00069), BLAST searches against the Sugen-Salk kinase database[6], multiple alignments of catalytic domains and manual expert inspection to verify classifications (see also [202]). The putative kinases were further classified into the major eukaryotic protein kinase groups[203, 204].

While the kinome is large for a protozoan, the protein tyrosine kinase groups (TK and TKL) are missing [III, table 4]. These are involved in many aspects of signaling and differentiation in eukaryotes, with an emphasis on relaying intercellular signals. The CAMK and AGC groups were underrepresented. These groups contain members sensing Ca^{2+} and secondary messengers. Only a few TriTryp PKs have predicted transmembrane regions. Taken together, it appears that the TriTryp kinases are not primarily involved in transmembrane signalling. The TriTryps showed relatively high numbers of CMGC, STE and NEK kinases. The CMGC group contains e.g. CDKs and MAP kinases, important in regulation of cell cycle progression. STE in turn contains activators of MAP kinases. The NEK group is less well studied, but NEK PKs function in cell cycle and cytoskeleton[202]. Perhaps most surprising, auxiliary domains normally present to convey signalling specificity were scarce. In all, the relatively large kinomes appear suited to enable the trypanosomatid parasites to respond to environmental changes via orchestration of the cell cycle and differentiation.

There were about 20 atypical PKs in each genome, not found in human, which is somewhat promising from a medical point of view (see e.g. [205]).

7.5.3 RNA recognition motif proteins

The RRM motif proteins are interesting in transcriptional and post-transcriptional regulation. Some are involved in the general RNA metabolism, and some bind specific elements in the 3'-UTRs of mRNAs and alter message stability and the differential regulation of expression[206].

We searched the genomes for the known TcUBP/RBP-proteins and some RRM motifs using BLAST, as well as using hmmer with the PFAM RRM motif (PF00076). Regular expression type patterns were used to confirm the presence

[6] http://kinase.com

of RNP signatures. Results are found in [99] and the analysis was subsequently expanded in [207]. In the context of RNA binding proteins, it is also interesting to note the expansion of a group of zinc-finger proteins as compared to yeast [99].

A large number of putative RNA binding proteins without assigned function opens for speculation, and is indeed consistent with post-transcriptional mechanisms acting at the levels of transcript stability and translation control.

7.5.4 The telomeric regions

We identified telomeric contigs and scaffolds by searching for inexact matches to the telomeric hexamer repeat. As the telomeric and subtelomeric regions are highly - and inherently - repetitive, assembly problems were expected. 49 contigs were found, and all but one were in scaffolds [III, table S6]. The telomeric regions were rich in putative retrotransposon hot spot protein, *trans*-sialidase, and and DGF-1 genes. Kinases, N-acetyltransferase ARD1 subunit, glycosyltransferase and DEAD box helicase genes were also present in many of the telomeric contigs. Many of these were annotated as pseudogenes. The 189 bp junction [208] and part of the "GP85 5'-UTR"-element is conserved among all contigs (a total of approximately 400 nt). Most also contain a further conserved region ending with a first RHS protein. This is consistent with other sequenced telomeric regions[209].

7.6 Messenger RNA processing sites in *Trypanosoma brucei* (IV)

We developed a simple computer model of kinetoplastid splicing and polyadenylation and applied this to the delineation of transcript boundaries [IV]. Expressed Sequence Tag sequences from *T. brucei* were aligned (mapped) to the genomic sequence to obtain a set of known splice sites and polyadenylation sites. These were used to calibrate and test the model in a leave-many-out cross-validation study. We found that the model was sufficient to exactly predict the correct splice sites in 2/3 of the cases, but that the exact polyadenylation site location is currently poorly understood (only 4% were exactly predicted, whereas 90% of the predictions were within a short distance from the polyadenylation site). We categorised the features of a typical splice site. We also made initial attempts to localise regulatory motifs in the mapped and predicted 3'-UTRs. A naïve linear motif finding approach did not uncover motifs that could explain developmentally regulated mRNA expression results sufficiently well. The regulation may depend on tertiary structure rather than linear motifs.

7.7 Kinetoplastid parasite *trans* splice site predictions reveal translational control by uAUGs (V)

The model of kinetoplastid splicing was validated and applied to the genome of *T. cruzi* [V]. Using both predictions and mapped cDNA, we found that a large proportion of the genes have upstream open reading frames. Previously, only one gene with a uORF has been studied in *T. cruzi*, but, just as for the many such genes documented in higher eukaryotes, its expression was down regulated by the uORF. Using the same model of splicing, a similar proportion of uORFs was predicted for *T. brucei* and *L. major*. A set of orthologous genes were predicted to have uORFs in all three species. We have proceeded to test predicted uORFs experimentally, and expect them to repress translation in GFP reporter constructs in *T. cruzi* epimastigotes. These experiments are still ongoing.

7.7.1 Other elements on predicted UTRs*

Predictions of transcript boundaries are directly useful for identification of regulatory elements. The uORFs provide an example of this. We also identified two subsets of transcripts with AU-rich elements (ARE) and G-rich elements (GRE) in the 3'-UTRs. Since the 3'-UTRs could not be predicted to the exact nucleotide, and a model for how the polyA tail is directed to one particular out of several alternative downstream splice signals, a conservative prediction with the shortest possible 3'-UTRs was used.

A set of ARE and GRE motifs were located in the predicted 3'-UTRs of the *T. cruzi* genome. To this end, a simple and again relatively conservative approach was used. ATTTATTTATTTATTTATTTA, ATTTATTTATTTATTTA, WATTTATTTATTTAW and CGGGGCGGGG sequences, with at most one mismatch each were located by fuzznuc[196], a IUPAC code enabled program for inexact nucleotide matching. In order to minimise the overlap in motifs, the longest motifs were first masked from the UTR sequences, and subsequent searches with the next shorter motifs were done on the masked sequences. A search in the predicted 3'-UTRs of all *T. cruzi* contigs $\geq 10kbp$ uncovered AREs for 973 genes, with 210 3'-UTRs that matched the longest motif form. GREs were found in 687 genes. Hypothetical genes, mainly conserved ones, were the most common; AREs were found in 605, and GREs in 345.

ARE and GRE motifs have previously been shown to regulate transcript stability in a set of mucin genes[210]. AREs promote stage specific rapid transcript degradation in trypomastigotes, whereas GREs specifically stabilise transcripts in epimastigotes.

While this approach revealed only 2 AREs in putative mucins (TcMUCII), GREs were found in the 3'-UTRs of 51 putative mucins, 27 *trans*-sialidases and 24 MASPs. GREs were also found in 11 GP63 genes.

A set of 49 kinases have AREs, and another 25 GREs. AREs were found in

17 chaperone genes, 11 with the longest form of the motif, in 14 phosphatase genes and in 11 RNA-binding factor genes.

We postulate that these transcripts are under differential control of mRNA stability. Experimental work to test this could prove interesting. While incomplete, a picture emerges where the stability of certain surface molecule transcripts may be under direct control, as well as sets of proteins potentially regulating many others, such as kinases, phosphatases, chaperones and RNA-binding proteis.

As both 3'-UTR boundary predictions and this motif finding approach are rather conservative, more AU/G rich elements must be present on expressed transcripts. This is supported by the small number of AREs found in mucin genes, and that no transcrips with both ARE and GRE were predicted, while they have previously been described [183]. These previously studied *TcMUC* transcripts had the GREs upstream of the AREs on the 3'-UTRs, which is consistent with overly conservative 3'-UTR predictions.

Given the predicted alternative sites, one can also speculate in a mechanism of alternative polyadenylation site choice, where the shorter form contains only the GRE, and the longer variant contains also ARE, which in turn would affect the stage specific stability of the transcripts.

"Science is what you know.
Philosophy is what you don't know."

– Bertrand Russell

Chapter 8

Concluding remarks

In this chapter, I will venture some thoughts of a more conjectural nature.

8.1 Base skew in the TriTryps

Given the striking difference between base skews in *L. major* vís-a-vís *T. cruzi* and *T. brucei*, I would have expected to find a large difference between *L. major* and *T. brucei*/*T. cruzi* in one of the strand asymmetric processes. From the replication, transcription and repair machineries [99] of the TriTryps, this is not yet evident. Perhaps instead small differences between the systems lead to the skews over time. Still, such a discrepancy could be experimentally tractable. It also seems that the story will turn more complex, before a solution is found. The GC3 of *L. major*[1], is also in conflict with current evolutionary theories.

8.2 The *Trypanosoma cruzi* genome

The *Trypanosoma cruzi* genome project yielded more structural information than could be expected with the current methodologies. The whole genome shotgun revealed the gene content, as expected, and core parts of chromosomes could be assembled. The libraries showed good coverage of the chromosomes, as exemplified by the large number of telomers assembled. The assembly was also tested against previously sequenced BACs. With the extreme repeat and polymorphism problems, a complete structural assembly was not to be expected. Low coverage sequencing of a second strain in addition to CL Brener allowed separation of regions where the haplotypes differed, and identification of regions where they merged. Alignment of the fragmented contigs to the structurally nearly complete *L. major* and *T. brucei* genomes allowed analyses at a structural level, although partially conjectural.

[1]Submitted work by Necsulea and Lobry, electronic preview at http://biomserv.univ-lyon1.fr/~necsulea/repro

While the most pressing objectives of the genome project have been fulfilled, finishing of the repeated and polymorphic genome of CL Brener could still yield valuable information on the evolution of *T. cruzi*. The variability and repetiveness is certainly not only a nuisance to shotgun assembly: it constitutes an inherent and important biological characteristic of *T. cruzi*. Ignoring it could be a mistake.

It would be interesting to see if an optical map could be used to order and bridge at least some of the repeats/polymorphic breaks. Optical maps were used for the *L. major* [99] and *T. brucei*[30] genomes. A new assembly program suited for a polymorphic genome, allowing bubbles of differences, or splitting of identical regions between homologs, would be useful. It would likely have to use information from both comparative analyses, related sequences, as well as current sequence mapped gaps, karyotype data and other physical information.

The functional annotation of the genome is by necessity incomplete. While dissatisfying at some level, even more detailed informatics analyses and more experimental work will over time contribute to filling our gaps in knowledge. The genome provides a lab notebook to connect results to. Only with the involvement of experts from various fields and their help did the annotation reach the current quality.

Publication of in particular the *T. cruzi* gene content was possibly delayed due to the decision to incorporate comparative data. On the other hand, this led to a level of structural and functional information that would otherwise not have been attainable at the time. The sequencing data were continuously made available to the community, albeit under restrictions against genome scale publications, which must have mitigated most problems with a delayed publication.

8.3 The usefulness of parasite genome sequences

Perhaps few in the post-genome era would raise objections to the cost-effectiveness of genome sequences. This was however not the general sentiment in the early days of the TriTryp project. The availability of the first few parasite genomes has had an impact on research.

The availability of a few high coverage parasite genomes opens up the possibility for low coverage comparative sequencing. This enables near full genome analyses at a lower cost. Such projects are well underway for each of the TriTryps - notably *L. infantum*, *L. braziliensis*, *L. chagasi*, one of the human infective *T. brucei* subspecies, *T. b. gambiense*, as well as the livestock parasites *T. congolense*, *T. vivax*, the *T. cruzi* related apathogenic *T. rangeli* and the extracellular fish parasite *T. carassii*.

The genome sequence enables large scale expression studies. One example is the University of Georgia proteome project [155], verifying the differential expression of some 2800 proteins from four life cycle stages (metacyclic trypomastigotes were separately recognised).

The availability of the genome sequence accelerates research in *T. cruzi* in several other ways as well. Researchers that work with the basic biology of the

trypanosomatids and some of those that strive to identify new drug and vaccine candidates report significantly more efficient work due to the availability of the genome data[2].

Studies of gene function, in turn, aid rational drug design. While effective drugs have mainly been discovered serendipitously, the many rational approaches do produce some interesting inhibitors for different biochemical pathways of the parasites. With further derivatization and formulation, potent drugs may appear. As drug resistance evolves, making old chemotherpeutics obsolete in parts of the world, understanding the genetic basis of these mechanisms must be worthwhile.

Metagenomics - to study by large scale sequencing all organisms in an environmental samples, all virii in a diseased person, or for that matter all protozoans in a complex infection, such as hyperendemic malaria - has become feasible, at least for virii and bacteria. This allows the elucidation of population structure, variation and interaction [57]. To sequence environmental isolates of parasites and samples from individuals with complex infections would undoubtedly further our knowledge - and deepen our questions - about the trypanosomatids.

8.4 Trypanosomatid regulation of gene expression

The sometimes large differences between species and even isolates and the varying results from the use of different experimental techniques have been largely ignored in this presentation to give a hopefully more comprehensive picture. While this is certainly incorrect in some cases and the unique mechanisms of each organism are of great importance, by and large this modus operandi has proven useful in biology. A mechanism in one of the trypanosomatids can initially be assumed to be reasonably similar in the others. In time, evidence to the opposite should accumulate via inconsistencies in results that rely on these presumptions.

While it is reasonable to leave gene expression at the stage of finished protein, regulation of protein stability and degradation can achieve changes in concentrations as much as regulation of protein expression, albeit at a longer time scale. The actually discussed mechanisms also span a wide spectrum of time scales, which can sometimes be essential for parasite survival.

Protein modification, such as the acetylation, farnesylation and phosphorylation brings much to the ability to fine tune the metabolome and interactome of the parasites, also directly in response to environmental cues without dramatic changes in protein translation rates. Some processes are compartmentalised in the trypanosomatids, notably the glycolysis to the glycosome, which may simplify metabolic control via transporters and metabolite concentrations.

[2]See e.g. Wellcome News 42 where Dr. Mike Fergusson and Dr. Sara Melville describe their experiences.

It is probably sufficient that a key set of proteins is carefully regulated, to achieve orchestration of the cell cycle, differentiation and interaction with the hosts.

Many of the post-transcriptional regulation mechanisms are shared, at least in part, with other eukaryotes, including human, although there transcriptional regulation is the most important. While this conservation may preclude some possibilities for drug candidates, the trypanosomatids can be put to use as model organisms.

Acknowledgements

I acknowledge that this thesis would not have come about had it not been for a great number of persons. Most of you, I have never met. I am grateful for being a part of this world for a fleeting second. However, I will single out a few of you who have had an immediate input on this work.

Elin, my better half, wisely suggests that I should briefly describe my results here, since this will probably be the most well-read section. This chapter is however already long, since I've been fortunate to have many collaborators, so please turn to the abstract now and browse that before continuing.

I am grateful, in particular, to

Björn Andersson, my supervisor, for believing in my ability, for having and sharing confidence in that everything will be all-right (as long as no measure of effort is spared ;), for enjoyable challenges and interesting questions, for hard earned funding and for skill in scientific writing.

Lena Åslund, for sharing her passion for parasitology research and a portion of her vast knowledge.

Mats Gustafsson, for giving me the opportunity to do interesting research while still an undergraduate, and in so shaping much of my conceptions of the world, in particular about feature extraction and pattern recognition, and obviously for realising that the *T cruzi* genome project was perfect for me.

Martti Tammi, for your intense, far-reaching and peculiar ideas and arguments. Thanks for the many fascinating challenges and ideas. Know that they were rewarding (but I still look forward to those crates of coke, beer and whatever bribes promised). I hope we'll work together again at some point - there is still a fundamental question or two we haven't answered..

Erik Arner, my room-mate since around 2000 sometime. Its been a while, and its been good. Thanks for all the music, the laughs and a keen and inquisitive mind. Although I consider myself curios, you often asked more and deeper.

Johan Elf - the next time you want an informatics collaborator for a High Impact Paper - count me in! Thanks for making it happen and for being Elfish.

Johan Normark - hard rock all the way.

The rest of Genome Analysis crew, past and present - **Alan**, **Anh-Nhi**, **Carole**, **Christina**, **Daryoush**, **Delal**, **Ellen**, **Esteban**, **Hamid**, **Kim**, **Marcela**, **Mariana**, **Shane**, **Sindy**, **Staffan**, **Stephen**, **Tarik** and **Yumi**. We've had a lot of fun so far, guys, and it is not over yet!

Students and projectworkers who worked with me; **Jenny E**, **Johan G**, **Fang W**, **Henrik P**, **Nicolas E**, **Daniel E**. Also, my students on the tdb 10p research projects, UU neural network and bioinformatics courses, KTH and KI bioinformatics courses. I hope you learned something - at the very least you taught me a lot!

Our collaborators at the other TriTryp sequencing centers, in particular:

TIGR - **Najib**, **Daniella** (my first experience with driving in the US comes to mind), **Elodie** (I vividly recall reverse-engineering those broken ABI files), **Gaëlle**, **Scott**, **Neil**, **Lis**, **Art**, **Joshua**, **Gustavo**.

SBRI - **Peter**, **Ken**, **Liz** and **Gautam** (though I probably misunderstood at least half of what the two of you said ;)

WTSI - **Al**, **Hubert** (the books, the baroque, the coffee and the science), **Matt**, **Christiane**, **Chris**, **Martin**.

My collaborators at the EMBL; **Christine Clayton**, **Lys Guilbride** and **Corinna Benz**.

The parasitology experts I interacted with during the *T. cruzi* annotation process: **Carlos Frasch**, **Jose-Luiz Ramirez**, **Jeremy Mottram**, **Marilyn Parsons** and **Shulamit Michaeli**.

All the other members of the T cruzi genome initiative and TriTryp consortium who made the meetings so fun and the conference calls almost bearable, for taking the time to talk to a clueless PhD student; especially **Rick Tarleton**, **John Kelly**, **John Donelson**, **Frederic Bringaud**, **Marc Oulette**, **Barbara Papadopoulou**, **Wim Degrave**, **Mike Gottlieb**, **Jennie Blackwell**, **Sarah Melville**, **Alan Fairlamb**, **Fred Opperdoes**, **Jean Lobry**, **Samson Obado**, **Martin Taylor**, **Steve Beverly** and **David Schwartz**.

Co-authors and collaborators on other studies during these years, who shared their ideas with me and must have invested a lot of effort into understanding what I was trying to tell them; in particular **Erland Ljunggren**, **Jens Mattsson**, **Måns Ehrenberg**, **Mats Wahlgren**, **Ulf Ribacke**, **QiJun Chen**, **Gerhard Wagner**, **Nicolas Joanin**. And all the rest of the co-authors, who I didn't think would read this thesis - shame on me - but thank you for the cooperation; I hope the next thing we do together will be as enjoyable!

All the other great people at the Rudbeck-laboratory, home to the group until november 2001, in particular **Ulf Landegren, Gyllensten and Petterson**, **Per-Ivan**, **Anders Isaksson**, **Tomas**, **Kalle N**, **Pernilla**, **Kicki**, **Maja** and **Patricia**.

The people at the CGB (and GB of CMB) past and present, in particular the bioinformatics unit for all the seminars and questions; notably and in no particular order **Erik S**, **Mark**, **Wynand**, **Albin**, **Alistair**, **Boris**, **Abiman**, **Pär**, **Markus**, **Lukas**, **Daniel**, **Fredrik**, **Carsten**, **Timo**, **Anna**, **Christian**, **Jennifer**, **Volker**, **Andrey**, **David** (for testing my martial skills. I still think that human transcript editing project would have been a killer..), and the rest of you folks - I can't list you all, but I have to thank **Rickard**, **Emily**, **Hagit**, **Geert**, **Nobutaka**, **Joacim**, **Yosuke**, **Zdravko**, **Bent**, **Pierre**, **Vivianne**, **Gitt**, **Elsebrit**, **Christine Jansson**, **Margaret Uhlander**, **Matti Nikkola** (for giving Management by Perkele and Brutally efficient productions a face),

and naturally the chairmen **Claes Wahlestedt**, **Christer Höög** and since the CMB merger **Tomas Perlman** for recruiting this amazing crowd, and sharing your inspiration.

The work was supported by grants from TFR, VR and NIAID.

On the practical side, thanks Zoegas, for Mollbergs blandning, Linus Torvalds, Richard Stallman, Larry Wall and all the others for Linux, emacs, perl and all the other open source software. The SF ISI Web of Science, Google and NCBI's pubmed considerably simplified literature access for the project. You know, the first few years I actually regularly had to walk down to the library. Actually, I had to do that a few times for the thesis writing as well.

I am grateful also to all my other friends who were not directly involved in the thesis work, and not only for all those non-scientific reasons. I've certainly discussed Nature and science, also in detail, with many of you, so thank you!

I owe much early development to **Mattias**, **Per** and **Andreas**; although I didn't understand biology at the time, the physics, maths and cs were a ride! Thanks also to the members of ÅKG, Datorakademin, ION, N3, and the teachers in Nybro and Uppsala. Studying in Uppsala would not have been the same without the "goldfish" gang, KTF and the SFINX folks. The "theater" people - thank you for many a good moment over the last years! Soshite, arigatou, Uppsala BudoKlubb, mata ne!

Fore-fathers and fore-mothers - thank you! Especially my grandmother Siv; it's said - with a hint of irony towards the world of science - that literacy, average intellectual talent and great curiosity is what it takes to become a scientist. You have been a great role model and inspiration to me in all of these areas!

Mina föräldrar, för allt stöd, all kärlek och er lugna, sköna inställning.

Elin, min älskade vän; tack!

Liten, tack för perspektivet och buffarna. Vi ses snart!

"Bernard of Chartres used to say that we are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness on sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size."

– John of Salisbury, *Metalogicon*, 1159

Bibliography

[1] N. Mattock and R. Pink. *Seventeenth Programme Report. Making health research work for poor people. Progress 2003-2004. Tropical Disease Research.* UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases, 2005.

[2] J. Lukes, H. Hashimi, and A. Zikova. Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Curr Genet*, 48(5):277–99, November 2005.

[3] F. R. Opperdoes and P. Borst. Localization of nine glycolytic enzymes in a microbody-like organelle in Trypanosoma brucei: the glycosome. *FEBS Lett*, 80(2):360–4, Aug 15 1977.

[4] B. M. Bakker, F. I. Mensonides, B. Teusink, P. van Hoek, P. A. Michels, and H. V. Westerhoff. Compartmentation protects trypanosomes from the dangerous design of glycolysis. *Proc Natl Acad Sci U S A*, 97(5):2087–92, Feb 29 2000.

[5] B. Andersson, L. Aslund, M. Tammi, A. N. Tran, J. D. Hoheisel, and U. Pettersson. Complete sequence of a 93.4-kb contig from chromosome 3 of Trypanosoma cruzi containing a strand-switch region. *Genome Res*, 8(8):809–16, August 1998.

[6] P. J. Myler, L. Audleman, T. deVos, G. Hixson, P. Kiser, C. Lemley, C. Magness, E. Rickel, E. Sisk, S. Sunkin, S. Swartzell, T. Westlake, P. Bastien, G. Fu, A. Ivens, and K. Stuart. Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc Natl Acad Sci U S A*, 96(6):2902–6, Mar 16 1999.

[7] X. H. Liang, A. Haritan, S. Uliel, and S. Michaeli. trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell*, 2(5):830–40, October 2003.

[8] C. E. Clayton. Life without transcriptional control? From fly to man and back again. *EMBO J*, 21(8):1881–8, Apr 15 2002.

[9] S. Martinez-Calvillo, S. Yan, D. Nguyen, M. Fox, K. Stuart, and P. J. Myler. Transcription of Leishmania major Friedlin chromosome 1 initiates

in both directions within a single region. *Mol Cell*, 11(5):1291–9, May 2003.

[10] S. Martinez-Calvillo, D. Nguyen, K. Stuart, and P. J. Myler. Transcription initiation and termination on Leishmania major chromosome 3. *Eukaryot Cell*, 3(2):506–17, April 2004.

[11] D. A. Campbell, S. Thomas, and N. R. Sturm. Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect*, 5(13):1231–40, November 2003.

[12] E. Pays, L. Vanhamme, and D. Perez-Morga. Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries. *Curr Opin Microbiol*, 7(4):369–74, August 2004.

[13] J. L. Krakow, D. Hereld, J. D. Bangs, G. W. Hart, and P. T. Englund. Identification of a glycolipid precursor of the Trypanosoma brucei variant surface glycoprotein. *J Biol Chem*, 261(26):12147–53, Sep 15 1986.

[14] A. G. Simpson, J. R. Stevens, and J. Lukes. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol*, Feb 24 2006.

[15] M. A. Miles, M. D. Feliciangeli, and A. R. de Arias. American trypanosomiasis (Chagas' disease) and the role of molecular epidemiology in guiding control strategies. *BMJ*, 326(7404):1444–8, Jun 28 2003.

[16] A. C. Frasch. Functional diversity in the trans-sialidase and mucin families in Trypanosoma cruzi. *Parasitol Today*, 16(7):282–6, July 2000.

[17] C. A. Buscaglia, V. A. Campo, J. M. Di Noia, A. C. Torrecilhas, C. R. De Marchi, M. A. Ferguson, A. C. Frasch, and I. C. Almeida. The surface coat of the mammal-dwelling infective trypomastigote stage of Trypanosoma cruzi is formed by highly diverse immunogenic mucins. *J Biol Chem*, 279(16):15860–9, Apr 16 2004.

[18] C. Yao, J. E. Donelson, and M. E. Wilson. The major surface protease (MSP or GP63) of Leishmania sp. Biosynthesis, regulation of expression, and function. *Mol Biochem Parasitol*, 132(1):1–16, November 2003.

[19] J. M. Di Noia, I. D'Orso, L. Aslund, D. O. Sanchez, and A. C. Frasch. The Trypanosoma cruzi mucin family is transcribed from hundreds of genes having hypervariable regions. *J Biol Chem*, 273(18):10843–50, May 1 1998.

[20] S. Vaena de Avalos, I. J. Blader, M. Fisher, J. C. Boothroyd, and B. A. Burleigh. Immediate/early response to Trypanosoma cruzi infection involves minimal modulation of host cell transcription. *J Biol Chem*, 277(1):639–44, Jan 4 2002.

[21] S. M. Teixeira, D. G. Russell, L. V. Kirchhoff, and J. E. Donelson. A differentially expressed gene family encoding "amastin," a surface protein of Trypanosoma cruzi amastigotes. *J Biol Chem*, 269(32):20509–16, Aug 12 1994.

[22] M. Almeida de Faria, E. Freymuller, W. Colli, and M. J. Alves. Trypanosoma cruzi: characterization of an intracellular epimastigote-like form. *Exp Parasitol*, 92(4):263–74, August 1999.

[23] V. Campo, J. M. Di Noia, C. A. Buscaglia, F. Aguero, D. O. Sanchez, and A. C. Frasch. Differential accumulation of mutations localized in particular domains of the mucin genes expressed in the vertebrate host stage of Trypanosoma cruzi. *Mol Biochem Parasitol*, 133(1):81–91, January 2004.

[24] M. A. Miles, A. Souza, M. Povoa, J. J. Shaw, R. Lainson, and P. J. Toye. Isozymic heterogeneity of Trypanosoma cruzi in the first autochthonous patients with Chagas' disease in Amazonian Brazil. *Nature*, 272(5656):819–21, Apr 27 1978.

[25] *Recommendations from a satellite meeting.*, volume 94 Suppl 1. Mem Inst Oswaldo Cruz, 1999.

[26] S. Brisse, C. Barnabe, and M. Tibayrenc. Identification of six Trypanosoma cruzi phylogenetic lineages by random amplified polymorphic DNA and multilocus enzyme electrophoresis. *Int J Parasitol*, 30(1):35–44, January 2000.

[27] M. W. Gaunt, M. Yeo, I. A. Frame, J. R. Stothard, H. J. Carrasco, M. C. Taylor, S. S. Mena, P. Veazey, G. A. Miles, N. Acosta, A. R. de Arias, and M. A. Miles. Mechanism of genetic exchange in American trypanosomes. *Nature*, 421(6926):936–9, Feb 27 2003.

[28] N. R. Sturm, N. S. Vargas, S. J. Westenberger, B. Zingales, and D. A. Campbell. Evidence for multiple hybrid groups in Trypanosoma cruzi. *Int J Parasitol*, 33(3):269–79, March 2003.

[29] J. R. Seed and M. A. Wenck. Role of the long slender to short stumpy transition in the life cycle of the african trypanosomes. *Kinetoplastid Biol Dis*, 2(1):3, Jun 25 2003.

[30] M. Berriman, E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D. C. Bartholomeu, N. J. Lennard, E. Caler, N. E. Hamlin, B. Haas, U. Bohme, L. Hannick, M. A. Aslett, J. Shallom, L. Marcello, L. Hou, B. Wickstead, U. C. Alsmark, C. Arrowsmith, R. J. Atkin, A. J. Barron, F. Bringaud, K. Brooks, M. Carrington, I. Cherevach, T. J. Chillingworth, C. Churcher, L. N. Clark, C. H. Corton, A. Cronin, R. M. Davies, J. Doggett, A. Djikeng, T. Feldblyum, M. C. Field, A. Fraser, I. Goodhead, Z. Hance, D. Harper, B. R. Harris, H. Hauser, J. Hostetler, A. Ivens, K. Jagels, D. Johnson, J. Johnson, K. Jones, A. X. Kerhornou, H. Koo,

N. Larke, S. Landfear, C. Larkin, V. Leech, A. Line, A. Lord, A. Macleod, P. J. Mooney, S. Moule, D. M. Martin, G. W. Morgan, K. Mungall, H. Norbertczak, D. Ormond, G. Pai, C. S. Peacock, J. Peterson, M. A. Quail, E. Rabbinowitsch, M. A. Rajandream, C. Reitter, S. L. Salzberg, M. Sanders, S. Schobel, S. Sharp, M. Simmonds, A. J. Simpson, L. Tallon, C. M. Turner, A. Tait, A. R. Tivey, S. Van Aken, D. Walker, D. Wanless, S. Wang, B. White, O. White, S. Whitehead, J. Woodward, J. Wortman, M. D. Adams, T. M. Embley, K. Gull, E. Ullu, J. D. Barry, A. H. Fairlamb, F. Opperdoes, B. G. Barrell, J. E. Donelson, N. Hall, C. M. Fraser, S. E. Melville, and N. M. El-Sayed. The genome of the African trypanosome Trypanosoma brucei. *Science*, 309(5733):416–22, Jul 15 2005.

[31] L. E. Bingle, J. L. Eastlake, M. Bailey, and W. C. Gibson. A novel GFP approach for the analysis of genetic exchange in trypanosomes allowing the in situ detection of mating events. *Microbiology*, 147(Pt 12):3231–40, December 2001.

[32] D. M. Mosser and A. Brittingham. Leishmania, macrophages and complement: a tale of subversion and exploitation. *Parasitology*, 115 Suppl:S9–23, 1997.

[33] S. L. Croft, M. P. Barrett, and J. A. Urbina. Chemotherapy of trypanosomiases and leishmaniasis. *Trends Parasitol*, 21(11):508–12, November 2005.

[34] J. Pepin, F. Milord, A. N. Khonde, T. Niyonsenga, L. Loko, B. Mpia, and P. De Wals. Risk factors for encephalopathy and mortality during melarsoprol treatment of Trypanosoma brucei gambiense sleeping sickness. *Trans R Soc Trop Med Hyg*, 89(1):92–7, Jan-Feb 1995.

[35] R. MacDonald and G. Yamey. The cost to global health of drug company profits. *West J Med*, 174(5):302–3, May 2001.

[36] S. K. Bhattacharya, T. K. Jha, S. Sundar, C. P. Thakur, J. Engel, H. Sindermann, K. Junge, J. Karbwang, A. D. Bryceson, and J. D. Berman. Efficacy and tolerability of miltefosine for childhood visceral leishmaniasis in India. *Clin Infect Dis*, 38(2):217–21, Jan 15 2004.

[37] R. L. Tarleton. New approaches in vaccine development for parasitic infections. *Cell Microbiol*, 7(10):1379–86, October 2005.

[38] L. Sigman, V. M. Sanchez, and A. G. Turjanski. Characterization of the farnesyl pyrophosphate synthase of Trypanosoma cruzi by homology modeling and molecular dynamics. *J Mol Graph Model*, Mar 13 2006.

[39] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–7, Apr 8 1976.

[40] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–95, Feb 24 1977.

[41] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74(2):560–4, February 1977.

[42] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–7, December 1977.

[43] F. Sanger, A. R. Coulson, B. G. Barrell, A. J. Smith, and B. A. Roe. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J Mol Biol*, 143(2):161–78, Oct 25 1980.

[44] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–65, Apr 9 1981.

[45] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol*, 162(4):729–73, Dec 25 1982.

[46] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda,

T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, and Y. J. Chen. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 15 2001.

[47] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert,

S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 16 2001.

[48] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick and. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, Jul 28 1995.

[49] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, , and J. C. Venter. The minimal gene complement of Mycoplasma genitalium. *Science*, 270(5235):397–403, Oct 20 1995.

[50] A. Goffeau. 1996: a vintage year for yeast and Yeast. *Yeast*, 12(16):1603–5, December 1996.

[51] E. D. Green. Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet*, 2(8):573–83, August 2001.

[52] A. Edwards and C. T. Caskey. Closure Strategies for Random DNA Sequencing. *METHODS: A companion to Methods in Enzymology*, 3(1):41–47, August 1990.

[53] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8(3):186–94, March 1998.

[54] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3):175–85, March 1998.

[55] M. T. Tammi, E. Arner, T. Britton, and B. Andersson. Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. *Bioinformatics*, 18(3):379–88, March 2002.

[56] W. Fiers, R. Contreras, G. Haegemann, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of SV40 DNA. *Nature*, 273(5658):113–20, May 11 1978.

[57] C. M. Fraser-Liggett. Insights on biology and evolution from microbial genome sequencing. *Genome Res*, 15(12):1603–10, December 2005.

[58] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, Oct 3 2002.

[59] J. M. Carlton, S. V. Angiuoli, B. B. Suh, T. W. Kooij, M. Pertea, J. C. Silva, M. D. Ermolaeva, J. E. Allen, J. D. Selengut, H. L. Koo, J. D. Peterson, M. Pop, D. S. Kosack, M. F. Shumway, S. L. Bidwell, S. J. Shallom, S. E. van Aken, S. B. Riedmuller, T. V. Feldblyum, J. K. Cho, J. Quackenbush, M. Sedegah, A. Shoaibi, L. M. Cummings, L. Florens, J. R. Yates, J. D. Raine, R. E. Sinden, M. A. Harris, D. A. Cunningham, P. R. Preiser, L. W. Bergman, A. B. Vaidya, L. H. van Lin, C. J. Janse, A. P. Waters, H. O. Smith, O. R. White, S. L. Salzberg, J. C. Venter, C. M. Fraser, S. L. Hoffman, M. J. Gardner, and D. J. Carucci. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, 419(6906):512–9, Oct 3 2002.

[60] P. Xu, G. Widmer, Y. Wang, L. S. Ozaki, J. M. Alves, M. G. Serrano, D. Puiu, P. Manque, D. Akiyoshi, A. J. Mackey, W. R. Pearson, P. H.

Dear, A. T. Bankier, D. L. Peterson, M. S. Abrahamsen, V. Kapur, S. Tzipori, and G. A. Buck. The genome of Cryptosporidium hominis. *Nature*, 431(7012):1107–12, Oct 28 2004.

[61] M. S. Abrahamsen, T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu, C. A. Lancto, M. Deng, C. Liu, G. Widmer, S. Tzipori, G. A. Buck, P. Xu, A. T. Bankier, P. H. Dear, B. A. Konfortov, H. F. Spriggs, L. Iyer, V. Anantharaman, L. Aravind, and V. Kapur. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science*, 304(5669):441–5, Apr 16 2004.

[62] M. J. Gardner, R. Bishop, T. Shah, E. P. de Villiers, J. M. Carlton, N. Hall, Q. Ren, I. T. Paulsen, A. Pain, M. Berriman, R. J. Wilson, S. Sato, S. A. Ralph, D. J. Mann, Z. Xiong, S. J. Shallom, J. Weidman, L. Jiang, J. Lynn, B. Weaver, A. Shoaibi, A. R. Domingo, D. Wasawo, J. Crabtree, J. R. Wortman, B. Haas, S. V. Angiuoli, T. H. Creasy, C. Lu, B. Suh, J. C. Silva, T. R. Utterback, T. V. Feldblyum, M. Pertea, J. Allen, W. C. Nierman, E. L. Taracha, S. L. Salzberg, O. R. White, H. A. Fitzhugh, S. Morzaria, J. C. Venter, C. M. Fraser, and V. Nene. Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. *Science*, 309(5731):134–7, Jul 1 2005.

[63] A. Pain, H. Renauld, M. Berriman, L. Murphy, C. A. Yeats, W. Weir, A. Kerhornou, M. Aslett, R. Bishop, C. Bouchier, M. Cochet, R. M. Coulson, A. Cronin, E. P. de Villiers, A. Fraser, N. Fosker, M. Gardner, A. Goble, S. Griffiths-Jones, D. E. Harris, F. Katzer, N. Larke, A. Lord, P. Maser, S. McKellar, P. Mooney, F. Morton, V. Nene, S. O'Neil, C. Price, M. A. Quail, E. Rabbinowitsch, N. D. Rawlings, S. Rutter, D. Saunders, K. Seeger, T. Shah, R. Squares, S. Squares, A. Tivey, A. R. Walker, J. Woodward, D. A. Dobbelaere, G. Langsley, M. A. Rajandream, D. McKeever, B. Shiels, A. Tait, B. Barrell, and N. Hall. Genome of the host-cell transforming parasite Theileria annulata compared with T. parva. *Science*, 309(5731):131–3, Jul 1 2005.

[64] B. J. Loftus, E. Fung, P. Roncaglia, D. Rowley, P. Amedeo, D. Bruno, J. Vamathevan, M. Miranda, I. J. Anderson, J. A. Fraser, J. E. Allen, I. E. Bosdet, M. R. Brent, R. Chiu, T. L. Doering, M. J. Donlin, C. A. D'Souza, D. S. Fox, V. Grinberg, J. Fu, M. Fukushima, B. J. Haas, J. C. Huang, G. Janbon, S. J. Jones, H. L. Koo, M. I. Krzywinski, J. K. Kwon-Chung, K. B. Lengeler, R. Maiti, M. A. Marra, R. E. Marra, C. A. Mathewson, T. G. Mitchell, M. Pertea, F. R. Riggs, S. L. Salzberg, J. E. Schein, A. Shvartsbeyn, H. Shin, M. Shumway, C. A. Specht, B. B. Suh, A. Tenney, T. R. Utterback, B. L. Wickes, J. R. Wortman, N. H. Wye, J. W. Kronstad, J. K. Lodge, J. Heitman, R. W. Davis, C. M. Fraser, and R. W. Hyman. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science*, 307(5713):1321–4, Feb 25 2005.

[65] B. Loftus, I. Anderson, R. Davies, U. C. Alsmark, J. Samuelson, P. Amedeo, P. Roncaglia, M. Berriman, R. P. Hirt, B. J. Mann, T. Nozaki, B. Suh, M. Pop, M. Duchene, J. Ackers, E. Tannich, M. Leippe, M. Hofer, I. Bruchhaus, U. Willhoeft, A. Bhattacharya, T. Chillingworth, C. Churcher, Z. Hance, B. Harris, D. Harris, K. Jagels, S. Moule, K. Mungall, D. Ormond, R. Squares, S. Whitehead, M. A. Quail, E. Rabbinowitsch, H. Norbertczak, C. Price, Z. Wang, N. Guillen, C. Gilchrist, S. E. Stroup, S. Bhattacharya, A. Lohia, P. G. Foster, T. Sicheritz-Ponten, C. Weber, U. Singh, C. Mukherjee, N. M. El-Sayed, W. A. Petri, Jr, C. G. Clark, T. M. Embley, B. Barrell, C. M. Fraser, and N. Hall. The genome of the protist parasite Entamoeba histolytica. *Nature*, 433(7028):865–8, Feb 24 2005.

[66] E. A. Worthey and P. J. Myler. Protozoan genomes: gene identification and annotation. *Int J Parasitol*, 35(5):495–512, Apr 30 2005.

[67] W. Degrave, M. J. Levin, J. F. da Silveira, and C. M. Morel. Parasite genome projects and the Trypanosoma cruzi genome initiative. *Mem Inst Oswaldo Cruz*, 92(6):859–62, Nov-Dec 1997.

[68] Alberto C. C. Frasch. The Trypanosoma cruzi genome initiative. *Parasitol Today*, 13(1):16–22, January 1997.

[69] B. Zingales, M. E. Pereira, K. A. Almeida, E. S. Umezawa, N. S. Nehme, R. P. Oliveira, A. Macedo, and R. P. Souto. Biological parameters and molecular markers of clone CL Brener–the reference organism of the Trypanosoma cruzi genome project. *Mem Inst Oswaldo Cruz*, 92(6):811–4, Nov-Dec 1997.

[70] B. Zingales, M. E. Pereira, R. P. Oliveira, K. A. Almeida, E. S. Umezawa, R. P. Souto, N. Vargas, M. I. Cano, J. F. da Silveira, N. S. Nehme, C. M. Morel, Z. Brener, and A. Macedo. Trypanosoma cruzi genome project: biological characteristics and molecular typing of clone CL Brener. *Acta Trop*, 68(2):159–73, November 1997.

[71] D. C. Schwartz and C. R. Cantor. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, 37(1):67–75, May 1984.

[72] W. C. Gibson and M. A. Miles. The karyotype and ploidy of Trypanosoma cruzi. *EMBO J*, 5(6):1299–305, June 1986.

[73] J. Henriksson, U. Pettersson, and A. Solari. Trypanosoma cruzi: correlation between karyotype variability and isoenzyme classification. *Exp Parasitol*, 77(3):334–48, November 1993.

[74] J. Henriksson, B. Porcel, M. Rydaker, A. Ruiz, V. Sabaj, N. Galanti, J. J. Cazzulo, A. C. Frasch, and U. Pettersson. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size

variation in Trypanosoma cruzi. *Mol Biochem Parasitol*, 73(1-2):63–74, July 1995.

[75] M. R. Santos, M. I. Cano, A. Schijman, H. Lorenzi, M. Vazquez, M. J. Levin, J. L. Ramirez, A. Brandao, W. M. Degrave, and J. F. da Silveira. The Trypanosoma cruzi genome project: nuclear karyotype and gene mapping of clone CL Brener. *Mem Inst Oswaldo Cruz*, 92(6):821–8, Nov-Dec 1997.

[76] P. E. Porcile, M. R. Santos, R. T. Souza, N. V. Verbisck, A. Brandao, T. Urmenyi, R. Silva, E. Rondinelli, H. Lorenzi, M. J. Levin, W. Degrave, and J. Franco da Silveira. A refined molecular karyotype for the reference strain of the Trypanosoma cruzi genome project (clone CL Brener) by assignment of chromosome markers. *Gene*, 308:53–65, Apr 10 2003.

[77] C. Branche, S. Ochaya, L. Aslund, and B. Andersson. Comparative karyotyping as a tool for genome structure analysis of Trypanosoma cruzi. *Mol Biochem Parasitol*, Jan 31 2006.

[78] M. I. Cano, A. Gruber, M. Vazquez, A. Cortes, M. J. Levin, A. Gonzalez, W. Degrave, E. Rondinelli, B. Zingales, and J. L. Ramirez and. Molecular karyotype of clone CL Brener chosen for the Trypanosoma cruzi genome project. *Mol Biochem Parasitol*, 71(2):273–8, May 1995.

[79] J. Hanke, D. O. Sanchez, J. Henriksson, L. Aslund, U. Pettersson, A. C. Frasch, and J. D. Hoheisel. Mapping the Trypanosoma cruzi genome: analyses of representative cosmid libraries. *Biotechniques*, 21(4):686–8, 690–3, October 1996.

[80] M. A. Duhagon, B. Dallagiovanna, and B. Garat. Unusual features of poly[dT-dG].[dC-dA] stretches in CDS-flanking regions of Trypanosoma cruzi genome. *Biochem Biophys Res Commun*, 287(1):98–103, Sep 14 2001.

[81] A. Brandao, T. Urmenyi, E. Rondinelli, A. Gonzalez, A. B. de Miranda, and W. Degrave. Identification of transcribed sequences (ESTs) in the Trypanosoma cruzi genome project. *Mem Inst Oswaldo Cruz*, 92(6):863–6, Nov-Dec 1997.

[82] R. E. Verdun, N. Di Paolo, T. P. Urmenyi, E. Rondinelli, A. C. Frasch, and D. O. Sanchez. Gene discovery through expressed sequence Tag sequencing in Trypanosoma cruzi. *Infect Immun*, 66(11):5393–8, November 1998.

[83] B. M. Porcel, A. N. Tran, M. Tammi, Z. Nyarady, M. Rydaker, T. P. Urmenyi, E. Rondinelli, U. Pettersson, B. Andersson, and L. Aslund. Gene survey of the pathogenic protozoan Trypanosoma cruzi. *Genome Res*, 10(8):1103–7, August 2000.

[84] F. Aguero, R. E. Verdun, A. C. Frasch, and D. O. Sanchez. A random sequencing approach for the analysis of the Trypanosoma cruzi genome: general structure, large gene and repetitive DNA families, and gene discovery. *Genome Res*, 10(12):1996–2005, December 2000.

[85] F. Bringaud, C. Vedrenne, A. Cuvillier, D. Parzy, D. Baltz, E. Tetaud, E. Pays, J. Venegas, G. Merlin, and T. Baltz. Conserved organization of genes in trypanosomatids. *Mol Biochem Parasitol*, 94(2):249–64, Aug 1 1998.

[86] E. Ghedin, F. Bringaud, J. Peterson, P. Myler, M. Berriman, A. Ivens, B. Andersson, E. Bontempi, J. Eisen, S. Angiuoli, D. Wanless, A. Von Arx, L. Murphy, N. Lennard, S. Salzberg, M. D. Adams, O. White, N. Hall, K. Stuart, C. M. Fraser, and N. M. El-Sayed. Gene synteny and evolution of genome architecture in trypanosomatids. *Mol Biochem Parasitol*, 134(2):183–91, April 2004.

[87] B. J. Haas, A. L. Delcher, J. R. Wortman, and S. L. Salzberg. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–6, Dec 12 2004.

[88] W. Degrave, A. B. de Miranda, A. Amorim, A. Brandao, M. Aslett, and M. Vandeyar. TcruziDB, an integrated database, and the WWW information server for the Trypanosoma cruzi genome project. *Mem Inst Oswaldo Cruz*, 92(6):805–9, Nov-Dec 1997.

[89] C. Hertz-Fowler, C. S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, J. Parkhill, A. C. Ivens, M. A. Rajandream, and B. Barrell. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*, 32(Database issue):D339–43, Jan 1 2004.

[90] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Res*, 34(Database issue):D16–20, Jan 1 2006.

[91] M. Luchtan, C. Warade, D. B. Weatherly, W. M. Degrave, R. L. Tarleton, and J. C. Kissinger. TcruziDB: an integrated Trypanosoma cruzi genome resource. *Nucleic Acids Res*, 32(Database issue):D344–6, Jan 1 2004.

[92] T. F. Smith. The history of the genetic sequence databases. *Genomics*, 6(4):701–7, April 1990.

[93] M. Ehrenberg, J. Elf, E. Aurell, R. Sandberg, and J. Tegner. Systems biology is taking off. *Genome Res*, 13(11):2377–80, November 2003.

[94] J. W. Fickett and C. S. Tung. Assessment of protein coding measures. *Nucleic Acids Res*, 20(24):6441–50, Dec 25 1992.

[95] J. W. Fickett. The gene identification problem: an overview for developers. *Computers Chem*, 20(1):103–118, 1996.

[96] J. W. Fickett. Finding genes by computer: the state of the art. *Trends Genet*, 12(8):316–20, August 1996.

[97] C. B. Burge and S. Karlin. Finding the genes in genomic DNA. *Curr Opin Struct Biol*, 8(3):346–54, June 1998.

[98] G. Mair, H. Shi, H. Li, A. Djikeng, H. O. Aviles, J. R. Bishop, F. H. Falcone, C. Gavrilescu, J. L. Montgomery, M. I. Santori, L. S. Stern, Z. Wang, E. Ullu, and C. Tschudi. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA*, 6(2):163–9, February 2000.

[99] A. C. Ivens, C. S. Peacock, E. A. Worthey, L. Murphy, G. Aggarwal, M. Berriman, E. Sisk, M. A. Rajandream, E. Adlem, R. Aert, A. Anupama, Z. Apostolou, P. Attipoe, N. Bason, C. Bauser, A. Beck, S. M. Beverley, G. Bianchettin, K. Borzym, G. Bothe, C. V. Bruschi, M. Collins, E. Cadag, L. Ciarloni, C. Clayton, R. M. Coulson, A. Cronin, A. K. Cruz, R. M. Davies, J. De Gaudenzi, D. E. Dobson, A. Duesterhoeft, G. Fazelina, N. Fosker, A. C. Frasch, A. Fraser, M. Fuchs, C. Gabel, A. Goble, A. Goffeau, D. Harris, C. Hertz-Fowler, H. Hilbert, D. Horn, Y. Huang, S. Klages, A. Knights, M. Kube, N. Larke, L. Litvin, A. Lord, T. Louie, M. Marra, D. Masuy, K. Matthews, S. Michaeli, J. C. Mottram, S. Muller-Auer, H. Munden, S. Nelson, H. Norbertczak, K. Oliver, S. O'neil, M. Pentony, T. M. Pohl, C. Price, B. Purnelle, M. A. Quail, E. Rabbinowitsch, R. Reinhardt, M. Rieger, J. Rinta, J. Robben, L. Robertson, J. C. Ruiz, S. Rutter, D. Saunders, M. Schafer, J. Schein, D. C. Schwartz, K. Seeger, A. Seyler, S. Sharp, H. Shin, D. Sivam, R. Squares, S. Squares, V. Tosato, C. Vogt, G. Volckaert, R. Wambutt, T. Warren, H. Wedler, J. Woodward, S. Zhou, W. Zimmermann, D. F. Smith, J. M. Blackwell, K. D. Stuart, B. Barrell, and P. J. Myler. The genome of the kinetoplastid parasite, Leishmania major. *Science*, 309(5733):436–42, Jul 15 2005.

[100] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, March 1970.

[101] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, Mar 25 1981.

[102] W. R. Pearson. Comparison of methods for searching protein sequence databases. *Protein Sci*, 4(6):1145–60, June 1995.

[103] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9, Nov 15 1992.

[104] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–8, April 1988.

[105] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 5 1990.

[106] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1 1997.

[107] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–8, March 1990.

[108] S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci U S A*, 90(12):5873–7, Jun 15 1993.

[109] M. S. Gelfand, A. A. Mironov, and P. A. Pevzner. Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci U S A*, 93(17):9061–6, Aug 20 1996.

[110] A. Krogh, I. S. Mian, and D. Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res*, 22(22):4768–78, Nov 11 1994.

[111] C. J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3(3):265–74, September 2002.

[112] R. Staden and A. D. McLachlan. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res*, 10(1):141–56, Jan 11 1982.

[113] Y. Nakamura, T. Gojobori, and T. Ikemura. Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic Acids Res*, 27(1):292, Jan 1 1999.

[114] J. W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*, 10(17):5303–18, Sep 11 1982.

[115] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput Appl Biosci*, 13(3):263–70, June 1997.

[116] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*, 26(2):544–8, Jan 15 1998.

[117] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 27(23):4636–41, Dec 1 1999.

[118] S. L. Salzberg, M. Pertea, A. L. Delcher, M. J. Gardner, and H. Tettelin. Interpolated Markov models for eukaryotic gene finding. *Genomics*, 59(1):24–31, Jul 1 1999.

[119] K. Murakami and T. Takagi. Gene recognition by combination of several gene-finding programs. *Bioinformatics*, 14(8):665–75, 1998.

[120] G. Aggarwal, E. A. Worthey, P. D. McDonagh, and P. J. Myler. Importing statistical measures into Artemis enhances gene identification in the Leishmania genome project. *BMC Bioinformatics*, 4:23, Jun 7 2003.

[121] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25(5):955–64, Mar 1 1997.

[122] T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283(5405):1168–71, Feb 19 1999.

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

[124] P. Flicek, E. Keibler, P. Hu, I. Korf, and M. R. Brent. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res*, 13(1):46–54, January 2003.

[125] M. Berriman, M. Aslett, N. Hall, and A. Ivens. Parasites are GO. *Trends Parasitol*, 17(10):463–4, October 2001.

[126] M. A. Norrgard, Y. Ivarsson, K. Tars, and B. Mannervik. Alternative mutations of a positively selected residue elicit gain or loss of functionalities in enzyme evolution. *Proc Natl Acad Sci U S A*, 103(13):4876–81, Mar 28 2006.

[127] S. Dalal, S. Balasubramanian, and L. Regan. Protein alchemy: changing beta-sheet into alpha-helix. *Nat Struct Biol*, 4(7):548–52, July 1997.

[128] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–5, October 2000.

[129] T. J. Carver, K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill. ACT: the Artemis Comparison Tool. *Bioinformatics*, 21(16):3422–3, Aug 15 2005.

[130] D. R. Forsdyke and J. R. Mortimer. Chargaff's legacy. *Gene*, 261(1):127–37, Dec 30 2000.

[131] E. CHARGAFF, R. LIPSHITZ, C. GREEN, and M. E. HODES. The composition of the deoxyribonucleic acid of salmon sperm. *J Biol Chem*, 192(1):223–30, September 1951.

[132] J. D. WATSON and F. H. CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, Apr 25 1953.

[133] F. R. Blattner, , C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–74, Sep 5 1997.

[134] N. Sueoka. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol*, 40(3):318–25, March 1995.

[135] J. R. Lobry. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol*, 40(3):326–30, March 1995.

[136] J. R. Lobry. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, 13(5):660–5, May 1996.

[137] A. Grigoriev. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res*, 26(10):2286–90, May 15 1998.

[138] I. J. Fijalkowska, P. Jonczyk, M. M. Tkaczyk, M. Bialoskorska, and R. M. Schaaper. Unequal fidelity of leading strand and lagging strand DNA replication on the Escherichia coli chromosome. *Proc Natl Acad Sci U S A*, 95(17):10020–5, Aug 18 1998.

[139] C. I. Wu and N. Maeda. Inequality in mutation rates of the two strands of DNA. *Nature*, 327(6118):169–70, May 14-20 1987.

[140] I. Mellon and P. C. Hanawalt. Induction of the Escherichia coli lactose operon selectively increases repair of its transcribed DNA strand. *Nature*, 342(6245):95–8, Nov 2 1989.

[141] M. P. Francino and H. Ochman. Strand asymmetries in DNA evolution. *Trends Genet*, 13(6):240–5, June 1997.

[142] B. J. Brewer. When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. *Cell*, 53(5):679–86, Jun 3 1988.

[143] C. Shioiri and N. Takahata. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J Mol Evol*, 53(4-5):364–76, Oct-Nov 2001.

[144] D. K. Niu, K. Lin, and D. Y. Zhang. Strand compositional asymmetries of nuclear DNA in eukaryotes. *J Mol Evol*, 57(3):325–34, September 2003.

[145] P. D. McDonagh, P. J. Myler, and K. Stuart. The unusual gene organization of Leishmania major chromosome 1 may reflect novel transcription processes. *Nucleic Acids Res*, 28(14):2800–3, Jul 15 2000.

[146] J. B. Palenchar and V. Bellofatto. Gene transcription in trypanosomes. *Mol Biochem Parasitol*, 146(2):135–41, April 2006.

[147] A. Gunzl, T. Bruderer, G. Laufer, B. Schimanski, L. C. Tu, H. M. Chung, P. T. Lee, and M. G. Lee. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei. *Eukaryot Cell*, 2(3):542–51, June 2003.

[148] G. Gilinger and V. Bellofatto. Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms. *Nucleic Acids Res*, 29(7):1556–64, Apr 1 2001.

[149] M. Dossin Fde and S. Schenkman. Actively transcribing RNA polymerase II concentrates on spliced leader genes in the nucleus of Trypanosoma cruzi. *Eukaryot Cell*, 4(5):960–70, May 2005.

[150] M. Navarro and K. Gull. A pol I transcriptional body associated with VSG mono-allelic expression in Trypanosoma brucei. *Nature*, 414(6865):759–63, Dec 13 2001.

[151] P. Borst and S. Ulbert. Control of VSG gene expression sites. *Mol Biochem Parasitol*, 114(1):17–27, Apr 25 2001.

[152] S. O. Obado, M. C. Taylor, S. R. Wilkinson, E. V. Bromley, and J. M. Kelly. Functional mapping of a trypanosome centromere by chromosome fragmentation identifies a 16-kb GC-rich transcriptional "strand-switch" domain as a major feature. *Genome Res*, 15(1):36–43, January 2005.

[153] V. Tosato, L. Ciarloni, A. C. Ivens, M. A. Rajandream, B. G. Barrell, and C. V. Bruschi. Secondary DNA structure analysis of the coding strand switch regions of five Leishmania major Friedlin chromosomes. *Curr Genet*, 40(3):186–94, October 2001.

[154] S. Monnerat, S. Martinez-Calvillo, E. Worthey, P. J. Myler, K. D. Stuart, and N. Fasel. Genomic organization and gene expression in a chromosomal region of Leishmania major. *Mol Biochem Parasitol*, 134(2):233–43, April 2004.

[155] J. A. Atwood, D. B. Weatherly, T. A. Minning, B. Bundy, C. Cavola, F. R. Opperdoes, R. Orlando, and R. L. Tarleton. The Trypanosoma cruzi proteome. *Science*, 309(5733):473–6, Jul 15 2005.

[156] M. Kozak. Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, 299(1-2):1–34, Oct 16 2002.

[157] D. R. Morris and A. P. Geballe. Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol*, 20(23):8635–42, December 2000.

[158] K. L. Perry, K. P. Watkins, and N. Agabian. Trypanosome mRNAs have unusual "cap 4" structures acquired by addition of a spliced leader. *Proc Natl Acad Sci U S A*, 84(23):8190–4, December 1987.

[159] M. Milhausen, R. G. Nelson, S. Sather, M. Selkirk, and N. Agabian. Identification of a small RNA containing the trypanosome spliced leader: a donor of shared 5' sequences of trypanosomatid mRNAs? *Cell*, 38(3):721–9, October 1984.

[160] G. M. Zeiner, N. R. Sturm, and D. A. Campbell. Exportin 1 mediates nuclear export of the kinetoplastid spliced leader RNA. *Eukaryot Cell*, 2(2):222–30, April 2003.

[161] G. M. Zeiner, N. R. Sturm, and D. A. Campbell. The Leishmania tarentolae spliced leader contains determinants for association with polysomes. *J Biol Chem*, 278(40):38269–75, Oct 3 2003.

[162] M. Parsons, R. G. Nelson, K. P. Watkins, and N. Agabian. Trypanosome mRNAs share a common 5' spliced leader sequence. *Cell*, 38(1):309–16, August 1984.

[163] R. G. Nelson, M. Parsons, M. Selkirk, G. Newport, P. J. Barr, and N. Agabian. Sequences homologous to variant antigen mRNA spliced leader in Trypanosomatidae which do not undergo antigenic variation. *Nature*, 308(5960):665–7, Apr 12-18 1984.

[164] S. Sather and N. Agabian. 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*, 82(17):5695–9, September 1985.

[165] R. E. Sutton and J. C. Boothroyd. Evidence for trans splicing in trypanosomes. *Cell*, 47(4):527–35, Nov 21 1986.

[166] W. J. Murphy, K. P. Watkins, and N. Agabian. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing. *Cell*, 47(4):517–25, Nov 21 1986.

[167] J. Huang and L. H. Van der Ploeg. Requirement of a polypyrimidine tract for trans-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3' splice acceptor site. *EMBO J*, 10(12):3877–85, December 1991.

[168] T. N. Siegel, K. S. Tan, and G. A. Cross. Systematic study of sequence motifs for RNA trans splicing in Trypanosoma brucei. *Mol Cell Biol*, 25(21):9586–94, November 2005.

[169] E. Vassella, R. Braun, and I. Roditi. 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*, 22(8):1359–64, Apr 25 1994.

[170] R. Manning-Cela, A. Gonzalez, and J. Swindle. Alternative splicing of LYT1 transcripts in Trypanosoma cruzi. *Infect Immun*, 70(8):4726–8, August 2002.

[171] J. H. LeBowitz, H. Q. Smith, L. Rusche, and S. M. Beverley. Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes Dev*, 7(6):996–1007, June 1993.

[172] E. Ullu and C. Tschudi. 2'-O-methyl RNA oligonucleotides identify two functional elements in the trypanosome spliced leader ribonucleoprotein particle. *J Biol Chem*, 268(18):13068–73, Jun 25 1993.

[173] K. R. Matthews, C. Tschudi, and E. Ullu. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev*, 8(4):491–501, Feb 15 1994.

[174] N. Schurch, A. Hehl, E. Vassella, R. Braun, and I. Roditi. Accurate polyadenylation of procyclin mRNAs in Trypanosoma brucei is determined by pyrimidine-rich elements in the intergenic regions. *Mol Cell Biol*, 14(6):3668–75, June 1994.

[175] S. Ruepp, A. Furger, U. Kurath, C. K. Renggli, A. Hemphill, R. Brun, and I. Roditi. Survival of Trypanosoma brucei in the tsetse fly is enhanced by the expression of specific forms of procyclin. *J Cell Biol*, 137(6):1369–79, Jun 16 1997.

[176] I. Roditi, H. Schwarz, T. W. Pearson, R. P. Beecroft, M. K. Liu, J. P. Richardson, H. J. Buhring, J. Pleiss, R. Bulow, and R. O. Williams and. Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of Trypanosoma brucei. *J Cell Biol*, 108(2):737–46, February 1989.

[177] S. Biebinger, S. Rettenmaier, J. Flaspohler, C. Hartmann, J. Pena-Diaz, L. E. Wirtz, H. R. Hotz, J. D. Barry, and C. Clayton. The PARP promoter of Trypanosoma brucei is developmentally regulated in a chromosomal context. *Nucleic Acids Res*, 24(7):1202–11, Apr 1 1996.

[178] H. R. Hotz, S. Biebinger, J. Flaspohler, and C. Clayton. PARP gene expression: control at many levels. *Mol Biochem Parasitol*, 91(1):131–43, Mar 1 1998.

[179] A. Furger, N. Schurch, U. Kurath, and I. Roditi. 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*, 17(8):4372–80, August 1997.

[180] H. R. Hotz, C. Hartmann, K. Huober, M. Hug, and C. Clayton. Mechanisms of developmental regulation in Trypanosoma brucei: a polypyrimidine tract in the 3'-untranslated region of a surface protein mRNA affects

RNA abundance and translation. *Nucleic Acids Res*, 25(15):3017–26, Aug 1 1997.

[181] B. C. Coughlin, S. M. Teixeira, L. V. Kirchhoff, and J. E. Donelson. Amastin mRNA abundance in Trypanosoma cruzi is controlled by a 3'-untranslated region position-dependent cis-element and an untranslated region-binding protein. *J Biol Chem*, 275(16):12051–60, Apr 21 2000.

[182] I. D'Orso and A. C. Frasch. Functionally different AU- and G-rich cis-elements confer developmentally regulated mRNA stability in Trypanosoma cruzi by interaction with specific RNA-binding proteins. *J Biol Chem*, 276(19):15783–93, May 11 2001.

[183] I. D'Orso and A. C. Frasch. TcUBP-1, a developmentally regulated U-rich RNA-binding protein involved in selective mRNA destabilization in trypanosomes. *J Biol Chem*, 276(37):34801–9, Sep 14 2001.

[184] I. D'Orso and A. C. Frasch. TcUBP-1, an mRNA destabilizing factor from trypanosomes, homodimerizes and interacts with novel AU-rich element- and Poly(A)-binding proteins forming a ribonucleoprotein complex. *J Biol Chem*, 277(52):50520–8, Dec 27 2002.

[185] R. Duncan. DNA microarray analysis of protozoan parasite gene expression: outcomes correlate with mechanisms of regulation. *Trends Parasitol*, 20(5):211–5, May 2004.

[186] A. Saxena, E. A. Worthey, S. Yan, A. Leland, K. D. Stuart, and P. J. Myler. Evaluation of differential gene expression in Leishmania major Friedlin procyclics and metacyclics using DNA microarray analysis. *Mol Biochem Parasitol*, 129(1):103–14, June 2003.

[187] T. A. Minning, J. Bua, G. A. Garcia, R. A. McGraw, and R. L. Tarleton. Microarray profiling of gene expression during trypomastigote to amastigote transition in Trypanosoma cruzi. *Mol Biochem Parasitol*, 131(1):55–64, September 2003.

[188] M. Gale, Jr, V. Carter, and M. Parsons. Translational control mediates the developmental regulation of the Trypanosoma brucei Nrk protein kinase. *J Biol Chem*, 269(50):31659–65, Dec 16 1994.

[189] F. Alvarez, C. Robello, and M. Vignali. Evolution of codon usage and base contents in kinetoplastid protozoans. *Mol Biol Evol*, 11(5):790–802, September 1994.

[190] P. P. Mueller and A. G. Hinnebusch. Multiple upstream AUG codons mediate translational control of GCN4. *Cell*, 45(2):201–7, Apr 25 1986.

[191] I. B. Rogozin, A. V. Kochetov, F. A. Kondrashov, E. V. Koonin, and L. Milanesi. Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics*, 17(10):890–900, October 2001.

[192] Alexander Churbanov, Igor B. Rogozin, Vladimir N. Babenko, Hesham Ali, and Eugene V. Koonin. Evolutionary conservation suggests a regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes . *Nucl. Acids Res.*, 33(17):5512–5520, 2005.

[193] S. M. Teixeira, L. V. Kirchhoff, and J. E. Donelson. Trypanosoma cruzi: suppression of tuzin gene expression by its 5'-UTR and spliced leader addition site. *Exp Parasitol*, 93(3):143–51, November 1999.

[194] S. Gopal, G. A. Cross, and T. Gaasterland. An organism-specific method to rank predicted coding regions in Trypanosoma brucei. *Nucleic Acids Res*, 31(20):5877–85, Oct 15 2003.

[195] A. Carbone, A. Zinovyev, and F. Kepes. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16):2005–15, Nov 1 2003.

[196] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–7, June 2000.

[197] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle., 2005.

[198] A. Buschiazzo, M. F. Amaya, M. L. Cremona, A. C. Frasch, and P. M. Alzari. The crystal structure and mode of action of trans-sialidase, a key enzyme in Trypanosoma cruzi pathogenesis. *Mol Cell*, 10(4):757–68, October 2002.

[199] M. L. Cremona, D. O. Sanchez, A. C. Frasch, and O. Campetella. A single tyrosine differentiates active and inactive Trypanosoma cruzi trans-sialidases. *Gene*, 160(1):123–8, Jul 4 1995.

[200] S. Eddy. Hmmer version 2.3.2. Distributed by the author. Department of Genetics, Washington University School of Medicine in St. Louis, USA., 1998.

[201] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–41, Jan 1 2004.

[202] M. Parsons, E. A. Worthey, P. N. Ward, and J. C. Mottram. Comparative analysis of the kinomes of three pathogenic trypanosomatids: Leishmania major, Trypanosoma brucei and Trypanosoma cruzi. *BMC Genomics*, 6:127, Sep 15 2005.

[203] S. K. Hanks and T. Hunter. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J*, 9(8):576–96, May 1995.

[204] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–34, Dec 6 2002.

[205] C. Naula, M. Parsons, and J. C. Mottram. Protein kinases as drug targets in trypanosomes and Leishmania. *Biochim Biophys Acta*, 1754(1-2):151–9, Dec 30 2005.

[206] J. G. De Gaudenzi, I. D'Orso, and A. C. Frasch. RNA recognition motif-type RNA-binding proteins in Trypanosoma cruzi form a family involved in the interaction with specific transcripts in vivo. *J Biol Chem*, 278(21):18884–94, May 23 2003.

[207] J. De Gaudenzi, A. C. Frasch, and C. Clayton. RNA-binding domain proteins in Kinetoplastids: a comparative analysis. *Eukaryot Cell*, 4(12):2106–14, December 2005.

[208] M. A. Chiurillo, I. Cano, J. F. Da Silveira, and J. L. Ramirez. Organization of telomeric and sub-telomeric regions of chromosomes from the protozoan parasite Trypanosoma cruzi. *Mol Biochem Parasitol*, 100(2):173–83, May 25 1999.

[209] D. Kim, M. A. Chiurillo, N. El-Sayed, K. Jones, M. R. Santos, P. E. Porcile, B. Andersson, P. Myler, J. F. da Silveira, and J. L. Ramirez. Telomere and subtelomere of Trypanosoma cruzi chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of T. cruzi telomeres. *Gene*, 346:153–61, Feb 14 2005.

[210] I. D'Orso, J. G. De Gaudenzi, and A. C. Frasch. RNA-binding proteins and mRNA turnover in trypanosomes. *Trends Parasitol*, 19(4):151–5, April 2003.

Part II

Reports