Computational analysis of Entamoeba histolytica genome- Exploring junk DNA for hidden messages

Manish Kushwaha

"Computational analysis of *Entamoeba histolytica* genome- Exploring junk DNA for hidden messages"

Dissertation submitted to

School of Life Sciences
Jawaharlal Nehru University

in partial fulfilment of the requirements for the award of the degree of
Master of Science

Submitted
by
MANISH KUSHWAHA

SCHOOL OF LIFE SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI 110067, INDIA

April 21, 2006

ABSTRACT

Increasingly large numbers of microRNAs are being discovered everyday and their role in gene regulation being better understood. A number of computational methods for the prediction of microRNAs are homology based and can miss out on any unique microRNA. A part of this work was dedicated to the development of an effective *ab initio* pipeline (CIDmiRNA) for microRNA prediction. This pipeline showed a sensitivity of 91.7% for Human microRNA prediction.

Entamoeba histolytica is the causative agent of human invasive amoebiasis, a common health problem in developing countries. Despite a lot of available information about the biology of this organism, no microRNAs have yet been reported. In this work, 6 putative microRNA genes were discovered in *E. histolytica* using CIDmiRNA, a computational prediction pipeline.

Also, the intergenic regions of *E. histolytica* were analysed, exploring for novel genes using computational methods. We were able to predict at least 2 highly likely candidates for being novel genes missed out in the current annotation.

ACKNOWLEDGEMENTS

I thankfully acknowledge the following people for their help and support during my stay in the MSc Life Sciences program and the successful execution of this project.

1) The Dean, Prof R N K Bamezai, for allowing me the opportunity to undertake this project;
2) My MSc Project supervisor, Prof Alok Bhattacharya, for his constant supervision and guidance during the project, the incisive scientific discussions that have helped me hone my analytical aptitude; and also, for the tea and dinner parties, the music concerts and the deliberations on books, theatre, society and religion;
3) Dr Rohini Muthuswami, for her support and encouragement all through the MSc program, and for being an excellent human being;
4) Sonika, for being a good friend; for the wonderful time I have had in her company; and, for the tea-breaks, the shopping, and the plays we have together enjoyed;
5) Kamal Rawal, for kindly providing the LINES_SINES dataset for use in this project;
6) The teachers at both the School of Life Sciences (SLS) and the Center for Computational Biology and Bioinformatics (CCBB) who have helped advance my knowledge and skills;
7) The other staff at SLS and CCBB who have been very helpful at all times;
8) Chetna, Deepak, Dipti, Priyank, Meera, Sangeeta, Sayantan- the close knit “family of friends”- for staying together through all the thick and thin;
9) Chetan and Prem, my room mates, for silently putting up with all the inconveniences that have had their origin in me;
10) Mudit, for his perpetual concern, his profound tolerance; and, for the emotional backup; and,
11) My family, for their love, care, and support, and for never letting me feel away from home.

-Manish Kushwaha

CONTENTS

1. INTRODUCTION

1.1.1. MicroRNAs: the discovery

It was reported in 1993 that *lin-4*, a gene known to exercise temporal regulation over *C. elegans* postembryonic larval development by negative control of the LIN-14 protein, did not code for a protein [1]. Instead, it produced two RNA molecules; one was approximately 22 nt (*lin-4*S) and the other approximately 61 nt (*lin-4*L) in length. These RNA molecules were found to contain two short blocks of sequence complementary to an element repeated seven times on the *lin-14* mRNA. It is noteworthy that these target elements lay within the 3'UTR of the target mRNA, previously proposed to mediate the repression of *lin-14* by the *lin-4* gene product [2].

Although examples of anti-sense mechanisms that affect mRNA stability had been reported earlier [3, 4, 5, 6], the *lin4-lin14* interaction seemed to suggest a novel mechanism of translational control mediated through interaction with the 3'UTR in the target mRNA. Since then, this suggestion has been backed by results from many studies- both computational and experimental.

1.1.2. MicroRNAs: history, development and current research

Following the suggestions of an RNA-mediated novel mechanism of translational regulation, many more tiny RNA molecules have been discovered. Second in the order was *let-7*, also in *C. elegans* [7]. These RNA molecules were then being dubbed small temporal RNAs (stRNAs) because they were thought to be exclusively involved in temporal patterning during development.

This was followed by a spurt of discovery of about 100 new tiny RNAs by three independent teams in 2001. Now, the tiny RNAs were re-christened as 'microRNAs' [8, 9, 10] or, in short, miRNAs. Fourteen new miRNAs were discovered in *Drosophila* embryo and nineteen new miRNAs in HeLa cells [11]; all these were experimentally verified. Fifty new miRNAs were discovered in mixed-stage *C. elegans* and expression of twenty of the twenty

two tested could be verified [9]. Fifteen new *C. elegans* miRNAs were cloned and their expression verified [10]; of these, ten were common with those identified with Bartel's team.

The extent and significance of miRNA-mediated gene regulation are becoming more evident with the increasingly large number of miRNAs being discovered at a rapid pace. miRNAs have already been shown to be involved in the control of cell division, cell proliferation, cell death, and fat metabolism [12, 13, 14], neuronal patterning [15], modulation of hematopoietic lineage differentiation [16], control of leaf and flower development [17, 18, 19, 20], and even playing both suppressor and enhancer roles in cancer development [21, 22, 23].

The vast implications of miRNA-controlled gene regulation have compelled the development of a number of computational approaches for their prediction [24, 25, 26, 27,28]. Also, a number of methods are being developed for the prediction of their target mRNAs [29, 30]. Alongside, work is on towards the elucidation of the finer details of miRNA biology.

1.2.1. MicroRNAs: gene organisation and expression patterns

A majority of miRNA genes are known to be standalone, coming from intergenic regions far away from known genes, suggesting that they are transcribed from independent transcription units [8, 9, 10]. However, a good fraction of miRNAs are known to be produced by processing of the spliced-out intronic regions of pre-mRNAs [31,32]. Yet other miRNAs appear to be clustered in the genome suggesting some kind of a poly-cistronic expression [8, 9]. As an example, orthologs of the *C. elegans lin-4* and *let-7* are seen to be clustered in flies and humans [31, 33, 34]. This may also suggest that functionally related miRNAs are clustered together.

MicroRNAs control the expression of other genes, but they themselves are also under controlled expression. This can be inferred from their spatial, temporal and numerical patterns of distribution. While miR-1 is primarily expressed in the mammalian heart [10, 35], miR-122 localises primarily in the liver [35]. The *C. elegans lin-4* and *let-7* miRNAs are expressed in

the first and the fourth larval stages of embryonic development, respectively [36, 37]. While more than 50,000 molecules of some miRNAs like miR-2, miR-52 and miR-58 are present per adult *C. elegans* cell, miR-124 is present in the same cells at around 800 molecules per cell.

1.2.2. MicroRNAs: gene transcription

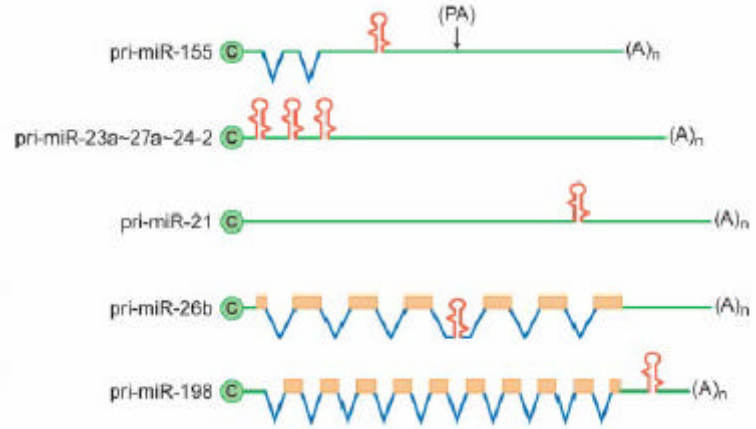

Figure 1

Schematic of the structure of five human pri-miRNAs. miRNA stems are shown in red, noncoding sequences in green, introns in blue and coding exons in beige. This figure is not to scale. PA, alternate poly-A-site.

(Source: reviewed by Cullen, 2004)

While it is implicit that miRNAs derived from the intronic regions would be transcribed by RNA pol II and share their precursors and regulatory elements with the pre-mRNA of the host gene, the role of RNA pol II in standalone miRNA transcription has only recently become clearer. Prior to their maturation into miRNAs (~22 nt), the precursors of these molecules fold into a stem-loop structure (~80 nt, called pre-miRNA) that itself is transcribed as a part of a larger precursor, several hundred nucleotide long, termed pri-miRNA. Analysis of 15 standalone miRNAs in humans has revealed that all are derived from pri-miRNA precursors that bear a 5' 7-methyl guanosine cap and a 3' poly-A tail [39, 40]. These findings suggest that pri-miRNAs may well be the structural analogues of protein coding messengers.

1.2.3. MicroRNAs: maturation

The first step in miRNA maturation involves the nuclear cleavage of the pri-miRNA, liberating the miRNA stem-loop intermediate or the pre-miRNA (precursor miRNA) [41, 42]. In animals, this step is mediated by an RNAse III endonuclease called Drosha which cleaves the pri-miRNA close to the base of the stem-loop structure [43]. Drosha cleaves the dsRNA with a staggered cut and results in a 5'-phosphate and a 2 nt 3' overhang [43, 44]. The pre-miRNA is then transported from the nucleus to the cytoplasm by Ran-GTP and the export receptor Exportin-5 [45, 46]. Cytoplasmic Dicer enzyme recognises the double stranded region of the pre-miRNA, particularly the 5'phosphate and the 3' overhang at the stem-loop base, and cleaves around two helical turns (~22 nt) away from the base [43]. This step results in an miRNA/miRNA* duplex with 2 nt overhangs at the 3' ends, miRNA being the effector molecule.

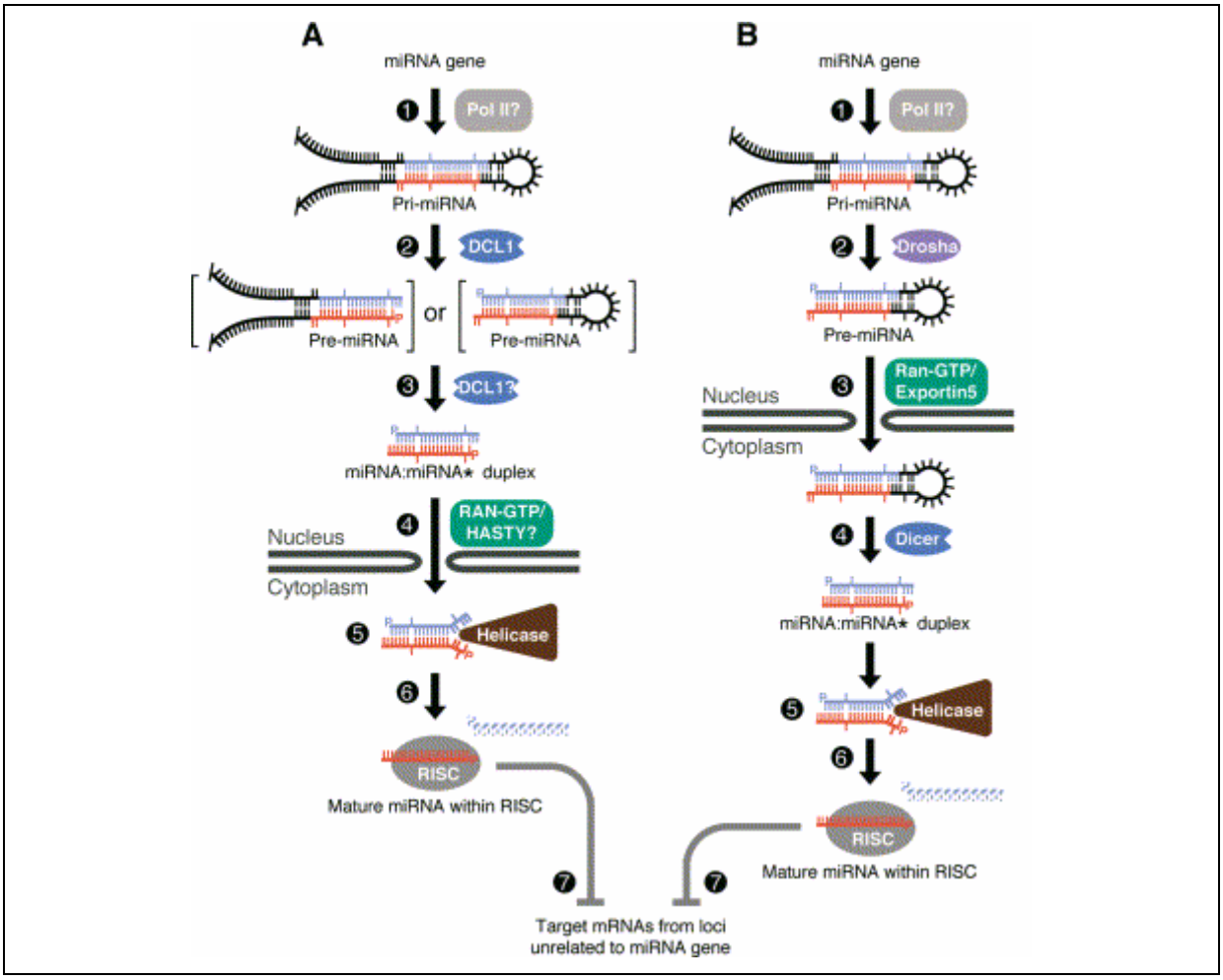

Figure 2

(A) The biogenesis of a plant miRNA (steps 1–6) and its hetero-silencing of loci unrelated to that from which it originated (step 7). The pre-miRNA intermediates (bracketed), thought to be very short-lived, have not been isolated in plants. The miRNA (red) is incorporated into the RISC (step 6), whereas the miRNA* (blue) is degraded (hatched segment). A monophosphate (P) marks the 5' terminus of each fragment.

(B) The biogenesis of a metazoan miRNA (steps 1–6) and its hetero-silencing of loci unrelated to that from which it originated (step 7).

(Source: reviewed by Bartel, 2004)

In plants, the maturation of miRNA from the pri-miRNA involves a Dicer-like homologue called DCL1 [47]. The pre-mRNA so formed is then processed by DCL1 (or another unidentified enzyme), in a manner analogous to the Dicer function in animals, further into the miRNA/miRNA* duplex. While Exportin 5 exports the pre-miRNA to the cytoplasm in animals, its homologue, HASTY, is proposed to export the miRNA/miRNA* duplex [48].

1.2.4. MicroRNAs: the silencing complex

Once the mature miRNA has been formed, the pathways in both plants and animals appear to be indistinguishable. They resemble the RNA silencing pathways called PTGS in plants and RNAi in animals. The mature miRNA becomes incorporated as single-stranded RNAs into a ribonucleoprotein complex, known as the RNA-induced silencing complex (RISC) [11, 49, 50, 51, 52, 53]. The RISC can recognise the target mRNAs by partial or perfect complementarity with the miRNA.

1.2.5. MicroRNAs: target recognition

The study of the miRNA sequences and the miRNA/mRNA duplexes has revealed the following details about the interaction: (1) The residues 2 to 8 of many invertebrate miRNAs show perfect complementarity to the 3'UTR elements [54], (2) the region 2-8 of the miRNAs is most conserved among homologous animal miRNAs [24, 55], (3) the elements of mRNAs that are complementary to the 2-8 region of the cognate miRNAs are perfectly conserved in the orthologous messages of other species [56].

Such asymmetry in the complementary base pairing across the miRNA is probably because the RISC presents only this region to initiate pairing to the mRNAs [48].

1.2.6. MicroRNAs: mechanisms of action

MicroRNAs are known to downregulate gene expression by either of two different mechanisms: mRNA cleavage and translational repression. The choice of the mechanism used is thought to be dependent on the extent of complementarity between the miRNA and the mRNA. Cleavage happens if the two strands are nearly completely complementary, whereas repression is followed if the complementarity is far from absolute [42, 57, 58, 59].

During an miRNA-mediated mRNA cleavage, the cut applied on the mRNA by the RISC is precisely between the pairing residues 10 and 11 of the miRNA [50, 57, 60, 61].

The prevailing belief is that most animal miRNAs downregulate by repression while most plant miRNAs do so by cleavage [48]. The downregulation is often brought about by a cumulative action of multiple RISCs acting at the multiple target sites within the 3'UTRs [59].

1.3. The miRBase [62, 63, 64]

miRBase is the repository of published microRNA data. It is located on the World Wide Web at http://microrna.sanger.ac.uk/

Initially started as "miRNA registry", the miRBase now has three major components:

(1) miRBase Sequences includes a searchable database of published miRNA sequences and annotation. This data was previously provided by the "miRNA registry",

(2) miRBase Registry provides a confidential service for pre-publication assigning of official names for novel miRNA genes, and

(3) miRBase Targets is a database of predicted miRNA target genes.

Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences

are available for searching using BLAST and SSEARCH, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.

The miRBase has expanded from 506 miRNAs from six organisms in 2004 (Release 2.0) to include 3518 miRNAs from primates, rodents, birds, fish, worms, flies, plants and viruses (Release 8.0).

1.4. MicroRNA prediction by computational means

As we have seen in Section 1.2.1, different microRNAs have different spatial, temporal and numerical patterns of distribution. Experimental discovery may therefore be very difficult or simply impossible for the rare miRNAs, as well as, for those that are expressed in specialised stages in the life of the organism. Therefore, computational methods have been used to accompany experimental methods to miRNA gene identification.

Homology searches have revealed orthologues and paralogues of known miRNA genes [8, 9, 10, 65]. Since many miRNA genes are known to be present in clusters, another approach has been to analyse the regions in the vicinity of known miRNA genes for other stem-loop forming structures that may represent additional genes of a genomic cluster [9, 31, 66, 67]. However, homology based methods cannot be used to detect unique and rare miRNAs; this is where the need for *ab initio* methods of miRNA prediction is felt.

Till date, five major computational methods have been applied for miRNA prediction in nematode, vertebrate, insect and plant candidates. These approaches have been briefly discussed below.

1.4.1. MiRscan [24]

MiRscan was used on sequences conserved between *C. elegans* and *C. briggsae* that folded into hairpin-loop structures, about 36,000 in number. The previously published 50 miRNAs served as training set for the program, which was then used to assign scores to each of these 36,000 hairpins, evaluating them on the basis of similarity to the training set with

respect to the a number of structural and sequence parameters. The distribution of MiRscan scores for the ~36,000 hairpins showed that the 50 genes of the training set mostly lay in the high scoring region.

The use of this method helped detect 30 additional miRNA genes in *C. elegans*. The method could predict miRNAs with a sensitivity of 50% and a specificity of 70%.

1.4.2. miRseeker [25]

miRseeker was used to identify miRNA genes by analysing intronic and intergenic regions conserved between *Drosophila melanogaster* and *D. pseudoobscura* and folding into evolutionarily constrained hairpin structures with features characteristic of known miRNAs. A number of parameters like nucleotide divergence, selective pressure on the precise sequence of the non-miRNA-coding-arm, preferential divergence within the loop structure were derived from the conservation analysis of the training set and then used for miRNA prediction.

This strategy could identify 48 novel miRNA candidates within *Drosophila* whose existence was strongly supported by conservation in other insect, nematode and vertebrate genomes. The sensitivity of the procedure was demonstrated to be 75%.

1.4.3. MIRFINDER [26]

MIRFINDER was a genome-wide approach used to detect miRNA genes in *Arabidopsis thaliana* genome. This method relied on the conservation of short sequences between the genomes of *Arabidopsis* and rice (*Oryza sativa*) and on properties of secondary structure of the miRNA precursor. Major rules were derived from the reference set, and later used in MIRFINDER. These rules included features of structure and sequence and conservation.

The method identified 91 potential miRNA genes, of which 58 had at least one nearly perfect match with and *Arabidopsis* mRNA.

1.4.4. ProMiR [27]

ProMiR uses a probabilistic co-learning method based on the paired hidden Markov model (HMM) to implement a general a general miRNA prediction method to identify close homologues as well as distant homologues.

When applied to genome screening for novel miRNAs on human chromosomes 16, 17, 18 and 19, ProMiR effectively searched distantly homologous patterns over diverse pre-miRNAs, detecting at least 23 novel miRNA gene candidates. Importantly, the miRNA gene candidates do not demonstrate clear sequence similarity to the known miRNA genes. 9 of the 23 predicted miRNA genes were experimentally confirmed, indicating that ProMiR may successfully predict miRNA genes with at least 40% accuracy.

On 5-fold cross-validation with 136 referenced human datasets, the efficiency of classification shows 73% sensitivity and 96% specificity.

1.4.5. miRAlign [28]

miRAlign uses both sequence and structure alignment to predict novel miRNAs. This method is able to predict distant homologues and considers more properties of miRNA structure conservation than previous tools. The method first does a conservation analysis and then sequence and structure alignment. Using this tool 59 new miRNAs were detected in *Anopheles gambiae* and it achieved a sensitivity of 70%.

1.5. Formal Grammars in Bioinformatics

Fundamental tasks in bioinformatics are mathematical analyses and comparisons of macromolecular sequences to determine common or consensus patterns among a family of sequences, produce a multiple sequence alignment, discriminate members of the family from nonmembers, and discover new members of the family.

Recently, the benefits of viewing the biological sequences representing DNA, RNA, or protein as sentences derived from a formal grammar has been argued [68]. When we view DNA, RNA, or protein sequences just as strings or formal languages on alphabets of four

nucleotides A, C, G, T or A, C, G, U or 20 amino acids, respectively, a grammatical representation and a grammatical inference method can be applied to various problems for biological sequence analyses. Especially, the increasing numbers of yielded DNA, RNA, and protein sequences highlight a growing need for developing grammatical system in bioinformatics, especially stochastic grammars such as Hidden Markov Models (HMMs) and Stochastic Context-Free Grammars (SCFGs) [69, 70, 71].

1.6. The SCFG_Based_Scorer [72]

An effective method for learning and building Stochastic Context-Free Grammars (SCFGs) to model a family of RNA sequences has been recently provided [71]. In general, the folding of an RNA sequence into a functional molecule is largely governed by the formation of the standard Watson-Crick base pairs A-U and G-C as well as the wobble pair G·U. Such base pairs constitute the so-called biological palindromes in genome and can be clearly described by a context-free grammar. In particular, productions of the forms X ? **A** Y **U** , X ? **U** Y **A**, X ? **G** Y **C**, and X ? **C** Y **G** describe a structure in RNA due to Watson-Crick base pairing. Using productions of this type, a CFG can specify a language of biological palindromes.

The SCFG_Based_Scorer [72] uses a Stochastic Context–Free Grammar model to decide if a given sequence can fold into the pre-miRNA structure. This is a statistical model constructed using base-pairing rules from human microRNA structures.

The rules used to generate the grammar are:

(1) The structure consists of a single long stem with a loop at its distal end.

(2) The paired region of the stem is continuous in only for short intervals; it is interrupted by bulges.

(3) The paired region of the stem at the proximal end is at least 3 base pairs long.

(4) The paired region of the stem at any other location is at least 2 base pairs long.

(5) The bulges interrupting the paired region of the stem can be symmetric or asymmetric.

(6) The symmetric bulge consists of equal number of unpaired bases on both the strands, not exceeding 4 in number on either.

(7) The asymmetric bulge has unequal number of unpaired bases on both sides, not exceeding 4 in number on one side and 7 in number on the other. The differerence in the number of unpaired bases on the two strands does not exceed 4.

(8) Allowed base pairs are A-U, G-C and G-U.

(9) The terminal loop at the distal end of the structure varies from 4 to 12 bases.

These rules have been encoded in the form of a grammar consisting of 5 sets of rewriting rules (Figure 3) to define the various elements of the miRNA structure and can be used to derive parse trees that correspond to miRNA structural elements. The primary sequence of the RNA structure can then be derived by tracing the terminal symbols (a, u, g, c) on the outside of the tree from one end to another (Figure 4).

Start:
S → STEM

Set I:
STEM → A1 STEM1 U1
STEM → U1 STEM1 A1
STEM → G1 STEM1 C1
STEM → C1 STEM1 G1
STEM → G1 STEM1 U1
STEM → U1 STEM1 G1
STEM1 → A1 STEM2 U1
STEM1 → U1 STEM2 A1
STEM1 → G1 STEM2 C1
STEM1 → C1 STEM2 G1
STEM1 → G1 STEM2 U1
STEM1 → U1 STEM G1
STEM2 → A1 STEM2 U1
STEM2 → U1 STEM2 A1
STEM2 → G1 STEM2 C1
STEM2 → C1 STEM2 G1
STEM2 → G1 STEM2 U1
STEM2 → U1 STEM2 G1
STEM2 → S_BULGE
STEM2 → A_BULGE
STEM2 → LOOP

Set II:
S_BULGE → T STEM T
S_BULGE → TT STEM TT
S_BULGE → TTT STEM TTT
S_BULGE → TTTT STEM TTTT

Set III:
A_BULGE → STEM T
A_BULGE → STEM TT
A_BULGE → STEM TTT
A_BULGE → STEM TTTT
A_BULGE → T STEM
A_BULGE → T STEM TT
A_BULGE → T STEM TTT
A_BULGE → T STEM TTTT
A_BULGE → T STEM TTTTT
A_BULGE → TT STEM
A_BULGE → TT STEM T
A_BULGE → TT STEM TTT
A_BULGE → TT STEM TTTT
A_BULGE → TT STEM TTTTT
A_BULGE → TT STEM TTTTTT
A_BULGE → TTTTT STEM TTT
A_BULGE → TTT STEM
A_BULGE → TTT STEM T
A_BULGE → TTT STEM TT
A_BULGE → TTT STEM TTTT
A_BULGE → TTT STEM TTTTT
A_BULGE → TTT STEM TTTTTT
A_BULGE → TTT STEM TTTTTTT
A_BULGE → TTTT STEM
A_BULGE → TTTTSTEM T
A_BULGE → TTTT STEM TT
A_BULGE → TTTT STEM TTT
A_BULGE → TTTTT STEM TT
A_BULGE → TTTTTT STEM TT
A_BULGE → TTTTTTT STEM TTT

Set IV:
LOOP → TTTT
LOOP → TTTTT
LOOP → TTTTTT
LOOP → TTTTTTT
LOOP → TTTTTTTT
LOOP → TTTTTTTTT
LOOP → TTTTTTTTTT
LOOP → TTTTTTTTTTT
LOOP → TTTTTTTTTTTT

Set V:
T → a
T → u
T → g
T → c
A1 → a
U1 → u
G1 → g
C1 → c

Figure 4

SCFG rewriting rules to represent secondary structure of human pre-miRNAs

The Scorer is first trained on a positive dataset using the inside-outside algorithm [73]. Then, the Cocke-Younger-Kasami [74] method is used for calculating the probability of a given sequence being generated using the defined grammar and then generating the most probable parse tree for that sequence. If the sequence cannot be generated using the defined grammar a negative infinite score is returned by the scorer, otherwise the best log-likelihood of generating the sequence is returned as the score (a negative value).

The SCFG_Based_Scorer is therefore an effective core program that can be used to identify the pre-miRNA stem-loop structures and hence to predict miRNAs.

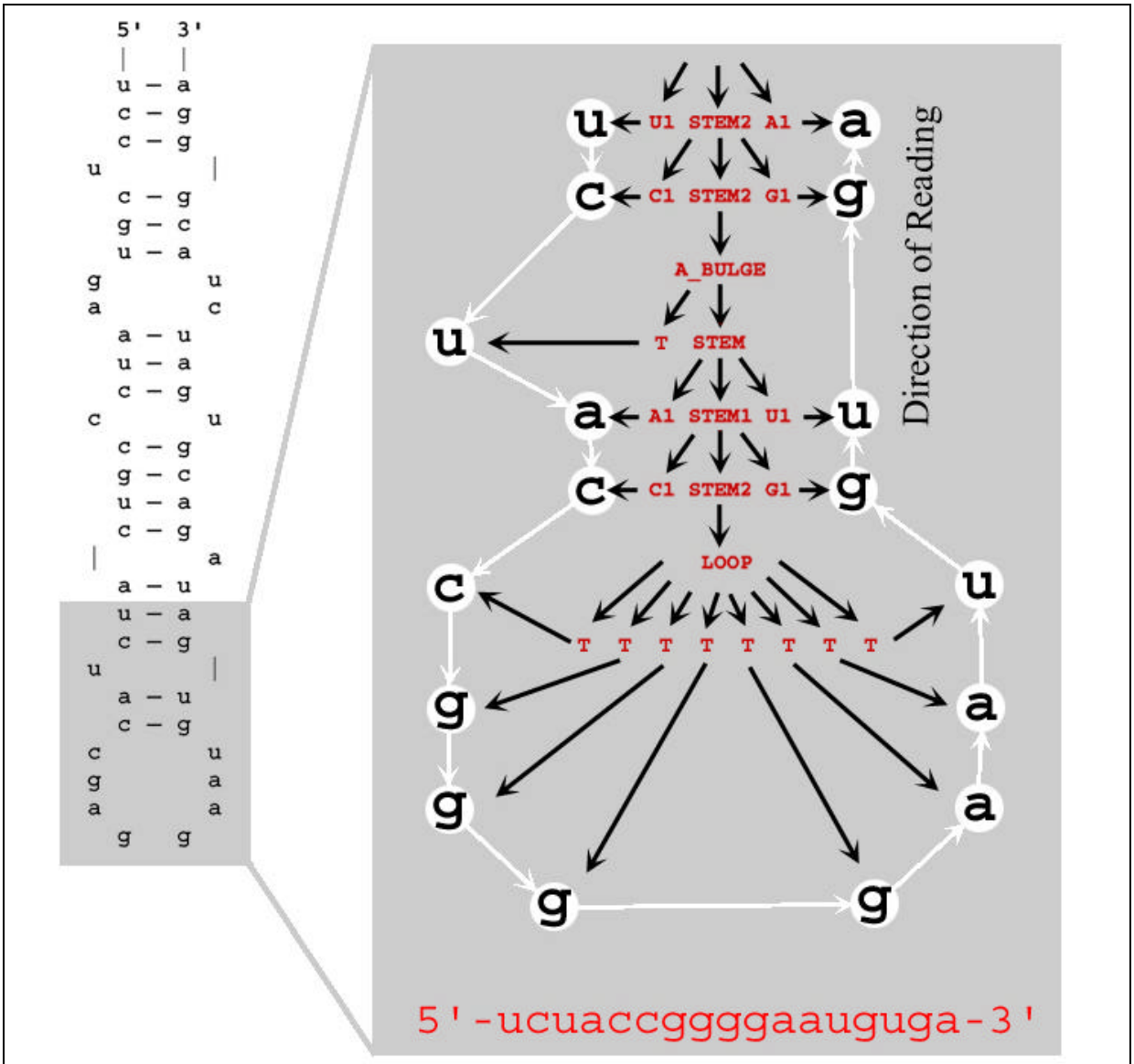

Figure 4

A parse tree for a part of a pre-miRNA derived using the SCFG rewriting rules

1.7.1. *Entamoeba histolytica*: the parasite

The organism, first identified by Fedor Losch in 1875, was initially named "Amoeba coli". However, the formal taxonomic classification was put forward by Fritz Schaudinn who also named it "*Entamoeba histolytica*". In 1925, Emile Brumpt proposed the concept of two morphologically identical species, the *E. histolytica* and *E. dispar*, the former responsible for an invasive disease and the latter surviving as a commensal.

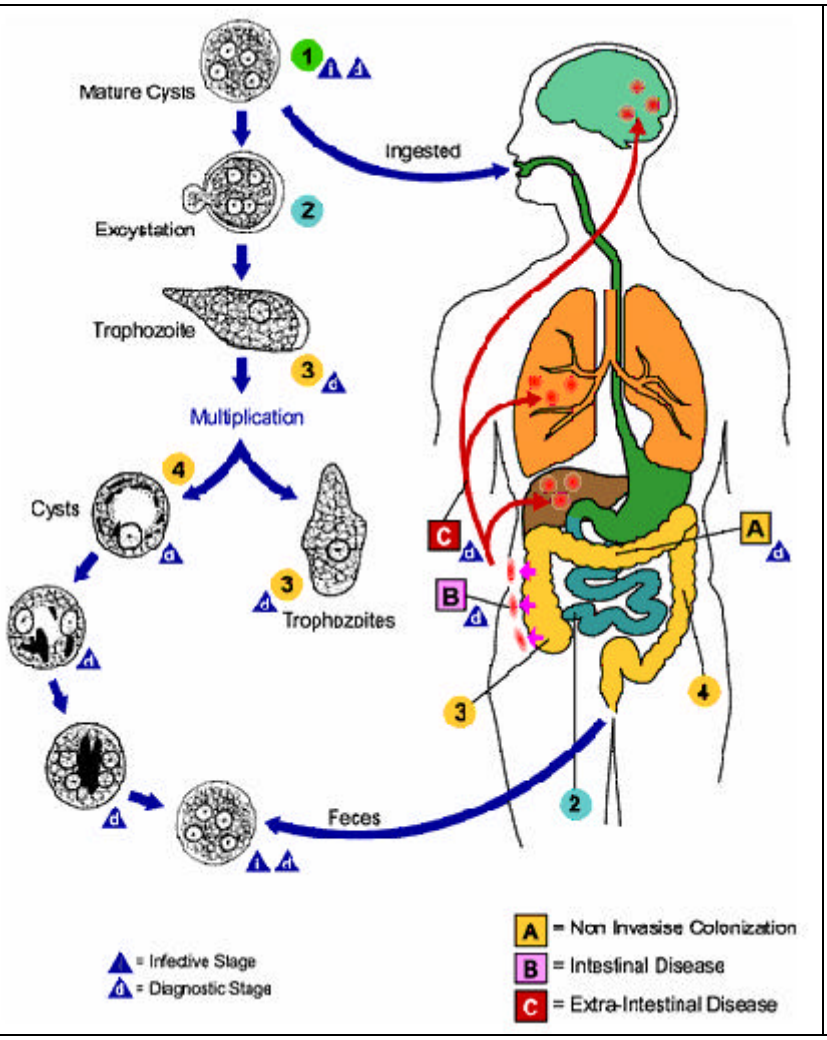

Figure 5
Life Cycle of *Entamoeba histolytica*

The parasitic protozoa *E. histolytica* is the causative agent of human invasive amoebiasis, a common health problem in some developing countries. Amoebiasis is characterized by potentially fatal intestinal and liver lesions. Worldwide, about 30-50 million people are infected and as many as 40,000-100,000 die, directly or indirectly, due to *E. histolytica* infection per annum. Despite the availability of the potent drug metronidazole, *E. histolytica* is a major cause of parasitic deaths worldwide. The motile form of the parasite, the

trophozoite, usually lives as a harmless commensal in the lumen of the distal colon, where it consumes bacteria as its nutrient source. In the colonic lumen it differentiates into cysts, the resistant form, responsible for the transmission of the infection, which has four nuclei and a chitin wall. As a commensal, *E. histolytica* induces no signs or symptoms in this condition. Normally, it causes the disease condition in about 10% of the infected individuals; virulent amoebas invade the intestinal mucosa and produce dysentery or amoeboma shedding fewer infectious cysts and more noninfectious trophozoites. Occasionally, it can spread through blood giving rise to extra intestinal lesions, mainly liver abscess. The main reservoir of *E. histolytica* is man, although morphologically similar *amoeba*s may be found in primates, dogs and cats. Human susceptibility to infection appears to be general, but most individuals harboring the parasite do not develop disease. Some host properties that favor invasive amoebiasis are intestinal microflora, certain nutritional habits, deficient immune response, alcoholism and pregnancy [75].

1.7.2. *Entamoeba histolytica*: genome and gene structure

The total DNA content in *E. histolytica* is ~0.5 pg DNA per cell, all of which is present in the nucleus and there are no other DNA-containing organelles [76]. The haploid genome size of *E. histolytica* was found to be 20 Mb. There are 14 linkage groups with ploidy level of four in *E. histolytica* [77]. The *Entamoeba* genome is low (~22.4%) in [G+C] content. The genome of *Entamoeba* consists of both linear chromosomes and plasmid-like circular DNA molecules; among the latter class are the rDNA circles [78]. In *E. histolytica*, in contrast to other unicellular eukaryotes, rDNA is present exclusively as extrachromosomal circular molecules [79]. The rDNA plasmid EhR1 of *E. histolytica* strain HM1:IMSS is 24.5 kb in size, present in 200 copies, has two transcription units arranged as inverted repeats along with several short tandem repeats upstream and downstream of the transcription units. The size of the plasmid is different in different strains and species with two major categories: one with two RNA transcription units and the other with two rRNA transcription units [80]. The

ribosomal plasmid contains multiple potential origins of replication, which are utilized in a differential manner. Only a few (3%) *Entamoeba* genes have introns in them [81]. However, sequencing of genomic contigs indicates that introns are more common than previously reported [82]. Introns are small in size (~100bp) and show typical eukaryotic features. Genes are tightly packed with short intergenic regions about 0.4-2.3 kb. Analysis of several promoters in *E. histolytica* has shown divergence in the core promoter with a non-consensus TATA and Initiation region (Inr). These semi-conserved motifs in the core promoter [83] are: (1) GTATTTAAA(G/C), the putative TATA element at -30; (2) GAACT, the GAAC box, location variable in the core promoter, between –30 to -14 and (3) AAAATTCA, the putative initiator at the start of transcription. Apart from core promoter four positive (URE1, URE2, URE4 and URE5) and one negative (URE3) regulatory elements has been localized upstream of the lectin gene hgl5 [83]. The conserved $TA^{A}/_{T}TT$ in the untranslated 3' region might function as transcription termination signal. Downstream of these motifs and just upstream of the poly-adenylation signal, a stretch of sequence of up to 12 pyrimidine residues is observed (Bruchhaus I *et al.*, 1993). RNA polymerase II inhibitor α-amanitin does not inhibit transcription of protein coding genes, indicating that the RNA Polymerase II is different from other eukaryotes [84], making it similar to *T. vaginalis* transcription machinery.

1.7.3. *Entamoeba histolytica*: the genome projects

The *E. histolytica* genome projects are simultaneously underway at The Institute for Genomic Research (TIGR) and the Sanger Institute Pathogen Sequencing Unit. The goal of the collaborative efforts is to determine 99% of the genomic sequence of *E. histolytica* strain HM1:IMSS, analyze and annotate the data and provide ready equal access to the sequence information and analysis. The *E. histolytica* genome is ~24 Mb in size and split into 14 chromosomes. A whole genome shotgun sequence (8X coverage) has been produced by Sanger Institute in collaboration with The Institute of Genome Research and assembled into 888 scaffolds. This draft sequence has been published [85] and finishing is now in progress.

2. AIMS AND OBJECTIVES

The aims and objectives of the project titled “Computational analysis of *Entamoeba histolytica* genome: Exploring junk DNA for hidden messages” were as follows:

1. To develop an effective pipeline for computational identification of microRNAs
2. To use this pipeline on conserved intergenic regions of *Entamoeba histolytica* to find novel miRNA genes
3. To find novel genes within the intergenic regions of *Entamoeba histolytica*

3. MATERIALS AND METHODS

3.1. MicroRNA reference sets and other sequences

A dataset of 227 human miRNAs, their precursor sequences (pre-miRNAs), and their structures were downloaded from the MicroRNA Registry [62, 63] release 6.0 (http://microrna.sanger.ac.uk/cgi-bin/sequences/browse.pl). Of these, only the experimentally verified miRNAs (121 in number) were retained, and the rest discarded. A Training_Dataset of 60 pre-miRNAs was then prepared by clustering the experimentally verified pre-miRNAs and selecting only the distantly related species. A dataset of the available 230 and 116 pre-miRNAs from *Mus musculus* and *Caenorhabditis elegans*, respectively, were also downloaded from the same release.

500 rRNA sequences from heterogeneous sources were obtained from GenBank following a keyword based search for rRNAs. These were designated as the 500_rRNA set.

Genomic sequences of five viruses were used in this work, including Simian Virus 40 (NC_001669), Kaposi sarcoma-associated herpesvirus (U75698), Mouse gamma herpesvirus 68 (U97553), Epstein-Barr virus (NC_001345), and Human cytomegalovirus (NC_001347). These genomes were obtained from Genbank; accession numbers have been mentioned within parentheses. The GenBank can be accessed at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide

Human chromosome 22 assembly (Release 3) was downloaded from Sanger Institute's Website (http://www.sanger.ac.uk/HGP/Chr22/cwa_archive/Release_3_14-09-2001/). Annotations from Release 3.1b (March 5, 2002) were used to mask all coding regions before the sequences were used further.

Entamoeba histolytica strain HM1:IMSS genome sequences and annotations, and protein database were downloaded from the Sanger Institute FTP server (ftp://ftp.sanger.ac.uk/pub/pathogens/Entamoeba/histolytica/). Entamoeba dispar genome was

downloaded form the Sanger Institute FTP server (ftp://ftp.sanger.ac.uk/pub/pathogens/Entamoeba/dispar/).

A set of Human Chromosome 22 microarray tiling data was downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/projects/geo/).

3.2. Training of the "SCFG_Based_Scorer"

The Training_Dataset (60 pre-miRNAs) was used to train the SCFG_Based_Scorer [72] such that it could analyse a profile of true miRNAs. Once trained, the Scorer was used to score input sequences. A lower absolute value of the score returned by the SCFG_Based_Scorer suggests a better likelihood of the sequence being a pre-miRNA.

3.3. Determining the cutoff score for effective pre-miRNA prediction

The SCFG_Based_Scorer is designed to score any input sequence for its probability of folding into a pre-miRNA structure. It is therefore necessary to fix a threshold value that will help make a binary decision of whether the input sequence can fold into a pre-miRNA structure or not. To help decide the threshold (1) a Positive_Dataset containing all the 121 experimentally verified human miRNAs, and (2) a Negative_Dataset containing regions from the 500_rRNA set with a duplex stability of at least -13.559 kcal/mol was prepared.

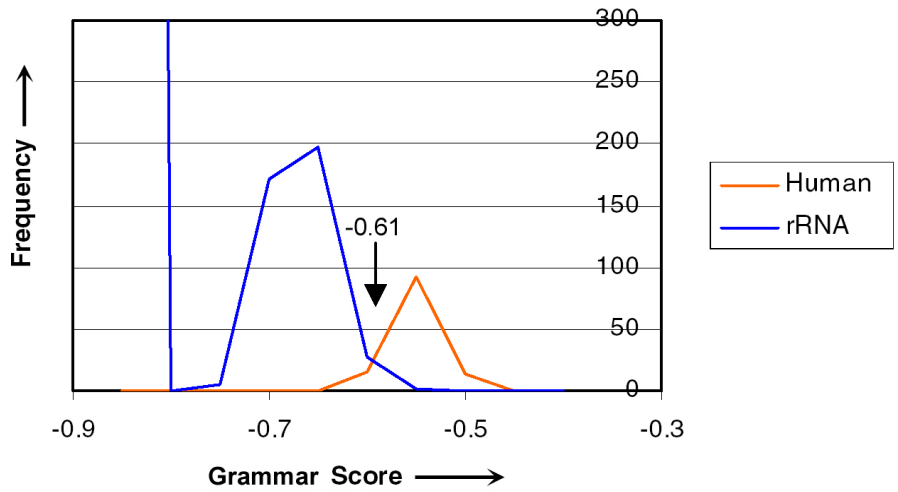

Figure 6

Frequency distribution of Normalised Scores (from the SCFG_Based_Scorer) for the Positive_Dataset and the Negative_Dataset

To determine the cutoff score for effective pre-miRNA prediction, the scores returned by the SCFG_Based_Scorer for the Positive_Dataset and the Negative_Dataset were analysed, and a cutoff value suitably segregating the two sets was chosen (Figure 6).

3.4 Pre-processing in the microRNA prediction pipeline: the Mfold Filter

Although the SCFG_Based_Scorer returns scores that identify pre-miRNAs with good accuracy, this requires that all the possible sub-windows from any given genomic region be supplied to the Scorer one at a time, thus increasing the time taken for pre-miRNA search exponentially. To cut down the prediction time, an optinal Mfold filter was introduced.

Mfold is a popular Minimum Free Energy (MFE) based RNA folding program [86]. It was used to fold all the sequences in the Positive_Dataset. The dG values obtained were in the range of -46.4 kcal/mol to -15.4 kcal/mol, with a Mean value of -31.856 kcal/mol and a Standard Deviation (S.D.) of 6.0990 kcal/mol. A wide range of Mean±3S.D. (-50.153 kcal/mol and -13.559 kcal/mol) was decided for filtering input sequences.

To further reduce the time of search, the flexible scanning window shifting by 1 nucleotide at a time was replaced by a fixed scanning window (length 125 chosen such that all possible miRNA lengths could be sufficiently covered) shifting by 10 nucleotides at a time. The effectiveness of such a window was confirmed by embedding each sequence of the Positive_Dataset in five different randomly generated backgrounds of length 125 each. The sequences so generated were then made to pass through the Mfold filter with a flexible scanning window shifting by 1 nucleotide. 93.468% of all sub-windows analysed could still pass through the filter, suggesting that the 125 length windows will still allow most pre-miRNAs embedded within the window to pass the filter. Any potential sub-windows excluded would be taken care of by the small distance (10 nucleotides) slid at a time.

3.5. Pre-processing in the microRNA prediction pipeline: the Structure Filter

Although the SCFG_Based_Scorer independently identifies pre-miRNAs with good sensitivity (91.7 %), it is relatively poor in its specificity. This is because the Scorer's context-free nature does not allow it to have a near memory. This means that the relative position of any structural element (e.g. S_BULGE, A_BULGE) will not affect the score as long as their length does not extend far enough to interfere with another structural element. The absence of near memory may result in inappropriate structures getting predicted. Therefore, a previously described Structure scoring system [87] was used to assign Structure scores to each of the sequences scored positive by the SCFG_Based_Scorer. Structures were obtained by using RNAfold [88] and any multiple loops at the terminus were opened up to conform to the structures depicted in the MicroRNA Registry. Opening of the multiple loops at the pre-miRNA hairpin heads for depiction in the Registry was confirmed by communication with Sam Griffiths-Jones.

A score of +1 was assigned to each set of paired bases, +0.5 to each unpaired base in symmetrical bulges of length up to 2. A score of –2 was assigned to each base in symmetrical bulges of length greater than 2 and a score of –2 was assigned to each unpaired base in asymmetrical bulges.

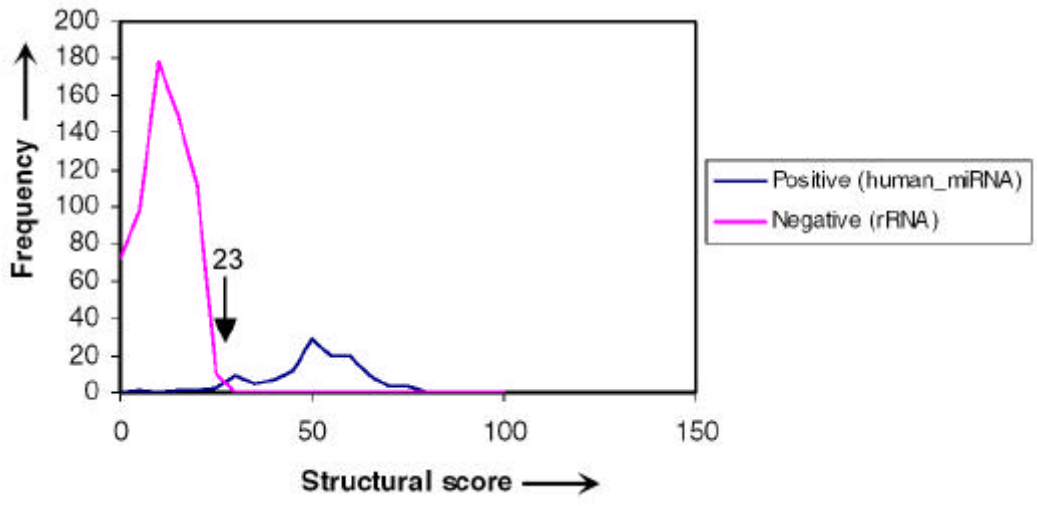

Figure 7

Frequency distribution of Structure Scores for the Positive_Dataset and the Negative_Dataset

The cutoff score was determined by analysing the scores returned by applying the Structure scoring system on the Positive_Dataset and the Negative_Dataset, and a cutoff value suitably segregating the two sets was chosen (Figure 7).

3.6. Obtaining the conserved intergenic regions from *E. histolytica* genome for further analysis

Conserved intergenic regions were obtained from the *E. histolytica* genome in a stepwise manner. The details of the steps follow.

1) Intergenic regions of the genome were extracted using the current annotation (11,522 continuous sequences)
2) The intergenic sequences were BLASTed against the *E. dispar* genome (with E-value cutoff of 1e-9). Sequences at least 50 nt in length and showing at least 92% identity with *E. dispar* were extracted for further processing (78,085 continuous sequences)
3) The sequences extracted in Step 2 were BLASTed against a LINES_SINES dataset (with E-value cutoff of 1e-9; dataset courtesy Kamal Rawal) and any matching sequences were excised out (63,267 continuous sequences remained)
4) To remove any duplicated regions, the sequences obtained from Step 3 were BLASTed against the *E. histolytica* genome (with E-value cutoff of 1e-9) and any sequence that returned more than one hit was removed (972 sequences remained)
5) To remove regions duplicated in *E. dispar*, the sequences obtained from Step 3 were BLASTed against the *E. dispar* genome (with E-value cutoff of 1e-9) and any sequence that returned more than one hit was removed (600 sequences remained)
6) The sequences obtained from Step 5 were BLASTed against the *E. histolytica* protein dababase (using BLASTX with E-value cutoff of 1e-9) and sequences showing a match were segregated into the Putative_Orthologues dataset; the remaining sequences were clubbed together as Eh_Unique_Conserved_IG (27 sequences went to Putative_Orthologues, 562 went to Eh_Unique_Conserved_IG)

3.7. Obtaining the 3'UTRs from *E. histolytica* genome for further analysis

To obtain the 3'UTRs from *E. histolytica* genome, 100 nucleotides downstream from the 3'end of the each gene were extracted and added to the dataset Eh_3'UTRs. This resulted in 9732 UTRs.

4. RESULTS AND DISCUSSION

4.1. Web-based pipeline for prediction of pre-microRNAs

An effective pre-miRNA prediction pipeline called CIDmiRNA (for Computational Identification of microRNA) was developed, with the SCFG_Based_Scorer (unpublished) at its heart. Major features of the pipeline are depicted in Figure 8 below.

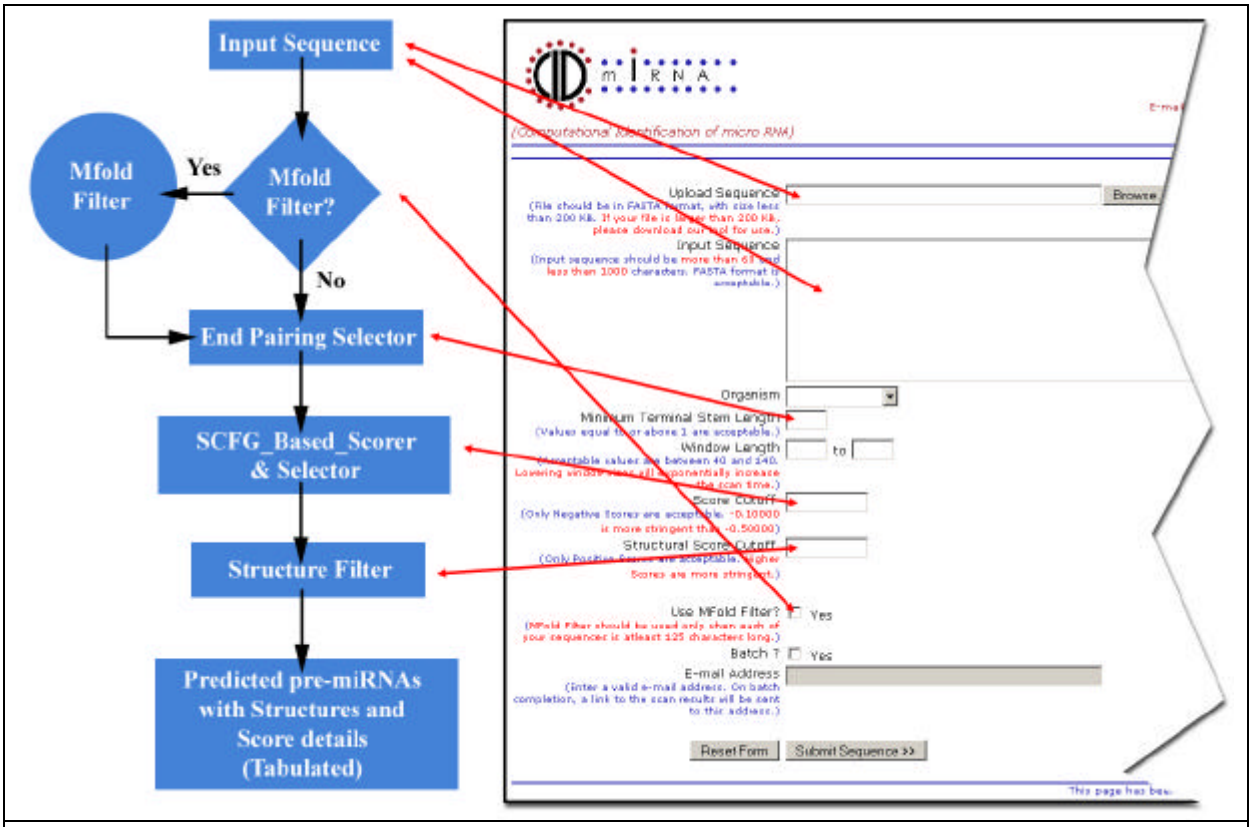

Figure 8

The diagram shows the correlation between the pipeline steps and the options available through the Web-based tool CIDmiRNA. The “Input Sequence” can either be directly pasted (between 60 and 1000 characters long) or the sequence file uploaded (up to 200 KB large). The choice of using the “Mfold Filter” can be exercised by checking/ unchecking the corresponding box. The “End Pairing Selector” selects only those sequences that pair at the ends, with the minimum length of terminal pairing decided by the value entered in the field “Minimum Terminal Stem Length”. The “SCFG_Based_Scorer & Selector” selects sequences that qualify the cutoff value entered in the filed “Score Cutoff”. Similarly, the “Structure Filter” filters out sequences that do not qualify the value entered in the filed “Structural Score Cutoff”. There is also an option of “Batch” processing which allows the user to specify an email address for receiving a link to the results once the batch is completed. A drop down

menu has been provided to choose the optimal parameter values for a given "Organism"; currently, default values for only *Homo sapiens* are available.

4.2. Testing of the pipeline on *Homo sapiens, Mus musculus, Caenorhabditis elegans* reported miRNAs

121 experimentally verified pre-miRNAs of *Homo sapiens*, and 230 and 116 pre-miRNAs (both experimentally verified and homology predicted) from *Mus musculus* and *Caenorhabditis elegans*, respectively, were used as a Test_Set for the CIDmiRNA pipeline. Parameters default for *Homo sapiens* were used on the entire set; Mfold filter was excluded. The sensitivties of prediction for the three species were 91.7%, 50.9% and 12.1% (Table 1). Sensitivity has been calculated using the relation SN = TP/ (TP+FN), where SN stands for sensitivity, TP for true positives and FN for false negatives [89].

Table 1

S. No.	Organism	Number of Input Sequences	Total Length (nt)	Post-SCFG_Based_Scorer & Selector	Post-Structure Filter
1	*Homo sapiens*	121*	9,945	113	111
2	*Mus musculus*	230#	11,410	119	117
3	*Caenorhabditis elegans*	116#	19,184	16	14

* Experimentally verified miRNAs

Experimentally verified or homology predicted miRNAs

These results suggest that CIDmiRNA can distinguish finer details of pre-miRNA structures from one species to another. Since the SCFG_Based_Scorer was trained on a human training set, the results were most promising for *Homo sapiens*. The sensitivity of mouse pre-miRNA prediction is worse than that for human, but better than that for the worm. The sensitivities of prediction appear to be related to the evolutionary distance between the organisms, suggesting that finer details of pre-miRNA structures have undergone divergent evolution.

4.3. Application on human and viral genomes

The CIDmiRNA pipeline was also applied on human chromosome (except contigs 2 and 3) and five viral genomes, including Simian Virus 40, Kaposi sarcoma-associated herpesvirus, Mouse gamma herpesvirus 68, Epstein-Barr virus, and Human cytomegalovirus. Parameters default for *Homo sapiens* were used on all of them; Mfold filter was excluded for the viral genomes. The results of the analysis have been represented in Table 2.

Table 2

S. No.	Chromosome/ Virus	Number of Input Sequences	Total Length (nt)	Post-SCFG_Based_Scorer & Selector	Post-Structure Filter[#]
1	Human Chromosome 22[*]	25,488,164	10,342,054	5,199	3,176 (1/2)[$]
2	Simian Virus 40	12,586	5,243	5	3 (2/2)
3	Kaposi sarcoma-associated herpesvirus	294,731	137,508	94	55 (6/12)
4	Mouse gamma herpesvirus 68	275,205	119,450	98	71 (2/9)
5	Epstein-Barr virus	6,200,339	172,281	173	134 (4/5)
6	Human cytomegalovirus	7,625,286	230,287	246	190 (6/9)

[*] Contigs 2 and 3 excluded; Mfold filter used

[#] Numbers within the parentheses represent the number of known miRNAs predicted/ the total number of known miRNAs in that region or virus

[$] When the regions containing another 3 reported miRNAs (in Contigs 2 and 3 of the Chromosome) were separately supplied, all the 3 miRNAs could also be detected

The 3,176 putative miRNAs predicted by the CIDmiRNA pipeline were used to search the intergenic tiling microarray data for human chromosome 22 from the GEO database (http://www.ncbi.nlm.nih.gov/projects/geo/). 731 of the 3,176 predicted miRNAs were found

in the microarray data, suggesting strongly that these sequences are indeed expressed and hence are strong contenders of being miRNAs.

The idea of using this pipeline on human viruses stemmed from the observation that viruses within a human cell use the human cellular machinery and should therefore result in similar structures of pre-miRNAs as the ones shown by human pre-miRNAs. This idea seems to stand validated because prediction for all human viruses shows at least 50% sensitivity of prediction. However, prediction for Mouse gamma herpesvirus 68 shows only 22.2% sensitivity.

4.4. Application of the pipeline on the conserved intergenic regions of *Entamoeba histolytica*

Following a BLAST against *E. dispar* genome to re-calculate their conservation, the Eh_Unique_Conserved_IG dataset prepared earlier was segregated into two subsets: Eh_Unique_Conserved_IG_Small (41 sequences) and Eh_Unique_Conserved_IG_Large (521 sequences). Eh_Unique_Conserved_IG_Small was obtained by picking up sequences that were no more than 200 in length and had at least 95% identity with *E. dispar*. The remaining sequences were set to the Eh_Unique_Conserved_IG_Large dataset.

The Eh_Unique_Conserved_IG_Large dataset was further processed as follows. (1) The 521 sequences were BLASTed against *E. histolytica* genome (using TBLASTX with E-value cutoff of 1e-9) and all continuous alignments (without stop codons) were discarded (347 sequences retained). (2) The 347 sequences were then BLASTed against themselves (using TBLASTX with E-value cutoff of 1e-9) and any sequence giving more than one hit was discarded (100 sequences retained). (3) Of the 100 sequences so obtained, all sequences larger than 200 nt or less than 95% identical to *E. dispar* were discarded. (93 sequences retained)

The 41 sequences from Eh_Unique_Conserved_IG_Small and the 93 sequences from the processed Eh_Unique_Conserved_IG_Large were then pooled together and BLASTed

against the est_others database at the NCBI Nucleotide-Nucleotide BLAST server (with results limited by an Entrez query string “Entamoeba[organism]”). Any sequences that gave a significant hit (>=98% identity) were discarded (a total of 85 sequences retained).

When the CIDmiRNA pipeline was used (with Minimum Terminal Stem Length 1, Window Length 40 to 140, Score Cutoff -0.9, Structural Score Cutoff 10) on the 85 remaining sequences, 6 putative pre-miRNAs were predicted (details in Table 3, page 37). A flexible window (12 to 25 in length) sliding by 1 nucleotide at a time was used to scan each of these predicted miRNAs and supplied to miRanda with default parameters [29] for a target search within the previously prepared Eh_3’UTRs dataset. This resulted in 4,040 hits that reduced to 1,653 on overlap removal. A distribution of the scores of the 4,040 hits across the length of each of these pre-miRNAs was observed using target-hit scores returned by miRanda. For each position on the pre-miRNA, the scores of all the target-hits of which the position was a part were considered. The distribution of total scores at each position for all the 6 pre-miRNAs was plotted with all pre-miRNAs aligned at their 5’end (Figure 9). This distribution shows a perceptible skew in the distribution of the scores, suggesting that the regions with greater scores in the distribution are more likely to be parts of the mature miRNA sequence. The distribution for the predicted AAFB01001645_miRNA_hit suggests that this pre-miRNA possibly matures into two miRNAs. The positions corresponding to the highest cumulative score have also been highlighted in Table 3 (page 37).

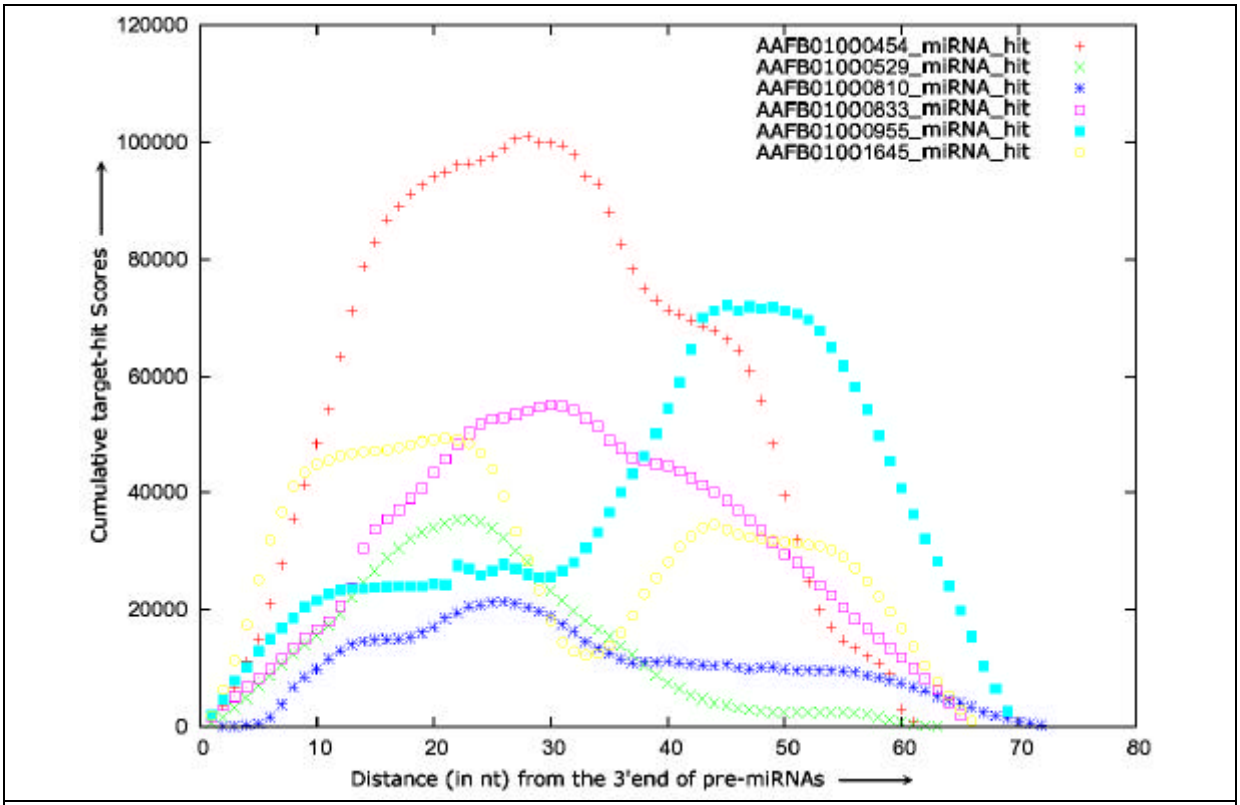

Figure 9

Distribution of the total target-hit scores for each position of the predicted pre-miRNAs

Relaxed parameters were chosen for pre-miRNA prediction in *E. histolytica* because its large evolutionary distance from humans could result in very low sensitivities (observed in Section 2 above). Relaxing the parameters is an effective dodge to catch any pre-miRNAs that have their basic structure conserved, yet the finer details different.

4.5. Further analysis of the conserved intergenic regions of *Entamoeba histolytica*

The Eh_Unique_Conserved_IG dataset was also used for another parallel analysis, with an aim to locate any genes not yet annotated. (1) The 562 sequences were BLASTed against self (using TBLASTX with E-value cutoff of 1e-9) and all continuous alignments (without stop codons) were extracted (227 sequences retained; their range of length was 21 to 867 nt). (2) Of these 227 sequences, sequences smaller than 100 nt were discarded (75 sequences retained).

Each of the 75 sequences obtained above, along with 100 nt flanking regions on both sides, was then subjected to a threefold analysis. (1) The sequences were BLASTed against

the est_others database at the NCBI Nucleotide-Nucleotide BLAST server (with results limited by an Entrez query string “Entamoeba[organism]”) and the nucleotide lengths of any significant hits (>=98% identity) with *E. histolytica* were recorded. (2) The sequences were also BLASTed against the nr PROTEIN database at the NCBI translating BLAST server and the %identities and the nucleotide lengths of any significant hits were recorded. (3) The sequences were supplied to Genscan using default parameters, and the predicted peptide length and the gene features were recorded [90]. Sequences that passed on one or more of these criteria have been recorded in Table 4 (page 40).

Of the 22 sequences recorded in the table, 2 passed all the three criteria, 1 passed only two criteria, whereas the rest passed on only one of the three criteria.

The sequences that gave an EST hit are very likely candidates of being genes, because their presence in the database already suggests that they are being transcribed. Candidates AAFB01000175_(6899-7048) and AAFB01001401_(885-310) can be interesting for further analysis because they are not only confirmed to be transcribed and show homology with known proteins, but also seem to be coding for some specific feature of a gene. While the former is predicted to be an internal exon, the latter is predicted to be an initial exon. Further analysis of regions upstream and downstream of these regions can give us more information.

The sequences that show homology with known proteins need to be compared with the entire homologous protein. It is possible that they are essentially small, non-functional parts of those proteins acquired long ago in evolution.

Although Genscan has been shown to be an effective gene prediction tool, some caution must be exercised while interpreting the results of Genscan. The default settings of Genscan use a Vertebrate training to predict genes. *E. histolytica* being an early branching eukaryote may not faithfully represent all the features of the vertebrate genes.

Table 3

Pre-miRNA Accession/ Locus	Pre-miRNA Length	Pre-miRNA %Conservation in *E. dispar*	Normalised Grammar Score	Structural Score	Non-overlapping hits in *E. histolytica* 3'UTRs[#]	RNAfold[*] dG (kcal/mol) -Folding without constraints
Structure[$]				Pre-miRNA Sequence[$]		
AAFB01000454 (7861-7920)	60	96%	-0.610221	34	304	-10.37
5′- u u --- gg aa u uuua cau uuug auaguaa gag c |||| ||| |||| ||||||| ||| aggu gua agac uauuguu cuu a 3′- a u auu uu c- u				5'-ttttatcattttgggatagtaaaagagtcatttccttgttatttcagattaatgttggaa-3'		
AAFB01000529 (3178-3241)	64	95%	-0.613559	12	174	-12.14
---- au ccac- u 5′- aaugaac uuuaauc caa guugu a ||||||| ||||||| ||| ||||| 3′- uuacuug aaauuag guu uaaca u ugau c- auaau u				5'-aatgaactttaatcatcaaccacgttgttattacaattaatattgcgattaaatagtgttcatt-3'		

AAFB01000810 (7010-7080)	71	97%	-0.619276	44	171	-14.24
- a ca ---- **u**a uc 5′- guuaa uuag uu uuuuuau ggaa gau c ||||| |||| || |||||| |||| ||| 3′- caauu aguc aa aaagaua ucuu cua a u a ag aaaa ua uu				5’-gttaattagattcatttttatggaa**t**agattccattatcatttctaaaaatagaaagaaaactgatttaac-3’		
AAFB01000833 (8776-8839)	64	96%	-0.620081	15	374	-8.70
a uu- a ucaau a u 5′- uca uu gauu guuuu aaua **cg** u ||| || |||| ||||| |||| || 3′- agu ag uuaa cgaaa uuau gu a a uau c uuuu- a u				5’-tcaattttgattagtttttcaataataa**cg**ttattgatatttttttaaagccaatttatgaatga-3’		
AAFB01000995 (12298-12365)	68	95%	-0.628116	16	366	-17.10
5′- u a ---- uuccu- u gu cau ugauucu uguuauauaa u || ||| ||||||| |||||||||| a ca gua acuagga acaauguauu g 3′- a a agau uua**u**cu a				5’-tgtacattgattctttccttgttatataattagattatgtaacatc**t**attaggatcatagaatgaaca-3’		

AAFB01001645 (1155-1219)	65	95%	-0.622005	29	264	-13.29

```
           uu   u uuaaua    u          a
5′- uuaaugg  gug g      guau uuuaauacu a
    |||||||  ||| |      |||| |||||||||| u
3′- aguuauc  uac c      caua aaauuguga c
           c-   - ------    u          c
```

5’-ttaatggttgtgtgttaatagtattttaatactaatccagtgttaaatatacccatcctattga-3’

***** Using RNAfold from *http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi* (Default parameters used. Lowest of the calculated dGs chosen.)

Using miRanda Version 1.9 downloaded from *http://www.microrna.org/miranda_new.html* (100 bases downstream of each *E. histolytica* gene end were extracted as 3’UTR)

$ Base/s with the highest cumulative score across all target-hits have been highlighted in **RED**

Table 4				
S. No.	Putative new gene (Accession/ Locus)*	BLASTn against *Entamoeba* ESTs	BLASTx against the "nr-PROTEIN database"	Genscan Analysis#
1.	AAFB01000031 (51989-51807)	337nt	–	–
2.	AAFB01000073 (15312-15971)	–	–	Init_277aa
3.	AAFB01000073 (16164-17030)	–	–	Sngl_PlyA_290aa
4.	AAFB01000087 (59599-59748)	–	100%id_99nt	–
5.	AAFB01000134 (13179-12982)	–	100%id_96nt	–
6.	AAFB01000157 (20046-19636)	–	45%id_579nt	–
7.	AAFB01000175 (6899-7048)	256nt	61%id_213nt	Intr_63aa
8.	AAFB01000200 (30858-30730)	–	100%id_96nt	–
9.	AAFB01000225 (13629-13483)	231nt	–	–
10.	AAFB01000240 (43466-43353)	163nt	–	–
11.	AAFB01000629 (4888-5004)	–	100%id_99nt	–
12.	AAFB01000663 (3085-2810)	–	–	Sngl_PlyA_105aa
13.	AAFB01000663 (4093-3815)	–	–	Init_119aa
14.	AAFB01000781 (13825-13721)	–	100%id_99nt	–
15.	AAFB01000855 (16777-17256)	–	–	Init_126aa
16.	AAFB01000862 (8330-8740)	–	–	Intr_138aa
17.	AAFB01000885 (2569-2156)	–	–	Init_177aa
18.	AAFB01000906 (3827-4510)	191_nt	–	–
19.	AAFB01000968 (11381-11280)	149nt	–	–
20.	AAFB01001020 (11175-10831)	418_nt	–	Intr_67aa
21.	AAFB01001311 (3143-3003)	180nt	–	–

22.	AAFB01001401 (885-310)	616_nt	33%id_603nt	Init_195aa

* The loci mentioned are those of the conserved regions; however, 100 nt flanking regions were also supplied for the three analyses

Genscan was with default parameters used from http://genes.mit.edu/GENSCAN.html

An explanation of the gene features (from the Genscan output page): Init = Initial exon (ATG to 5' splice site); Intr = Internal exon (3' splice site to 5' splice site); Sngl = Single-exon gene (ATG to stop); PlyA = poly-A signal (consensus: AATAAA)

REFERENCES

1. Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–854.
2. Wightman, B., Burglin, T.R., Gatto, J., Arasu, P., and Ruvkun, G. (1991). Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development. Genes Dev. 5, 1813–1824.
3. Kimelman, D., and Kirschner, M. W. (1989). An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in *Xenopus* oocytes. Cell 59, 687–696.
4. Hildebrandt, M., and Nellen,W. (1992). Differential antisense transcription from the *Dictyostelium* EB4 gene locus: implications on antisense mediated regulation of mRNA stability. Cell 69, 197–204.
5. Simons, R. W. (1988). Naturally occurring antisense RNA control: a brief review. Gene 72, 35–44.
6. Eguchi, Y., Itoh, T., and Tomizawa, J. (1991). Antisense RNA. Annu. Rev. Biochem. 60, 631–652.
7. Reinhart, B.J., Slack, F.J., Basson, M., Bettinger, J.C., Pasquinelli, A.E., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21 nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403, 901–906.
8. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science 294, 853–858.
9. Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science 294, 858–862.
10. Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science 294, 862–864.

11. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001b). Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. EMBO J. 20, 6877–6888.
12. Hatfield SD, Shcherbata HR, Fischer KA, Nakahara K, Carthew RW, Ruohola-Baker H. (2005). Stem cell division is regulated by the microRNA pathway. Nature, 435(7044):974-8.
13. Brennecke, J. et al. (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. Cell 113, 25–36.
14. Xu, P., Vernooy, S.Y., Guo, M., and Hay, B.A. (2003). The *Drosophila* microRNA mir-14 suppresses cell death and is required for normal fat metabolism. Curr. Biol. 13, 790–795.
15. Johnston, R.J., and Hobert, O. (2003). A microRNA controlling left/ right neuronal asymmetry in *Caenorhabditis elegans*. Nature 426, 845–849.
16. Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. Science 303, 83–86.
17. Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell 10, 10.
18. Chen, X. (2003). A MicroRNA as a Translational Repressor of APETALA2 in *Arabidopsis* Flower Development. Science. Published online September 11, 2003. 10.1126/science.1088060.
19. Emery, J.F., Floyd, S.K., Alvarez, J., Eshed, Y., Hawker, N.P., Izhaki, A., Baum, S.F., and Bowman, J.L. (2003). Radial patterning of arabidopsis shoots by class III HD-ZIP and KANADI genes. Curr. Biol.13, 1768–1774.

20. Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by C.B. (2003). microRNAs. Nature 425, 257–263. 787–798.

21. Jun Lu, Gad Getz, Eric A. Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L. Ebert, Raymond H. Mak, Adolfo A. Ferrando, James R. Downing, Tyler Jacks, H. Robert Horvitz and Todd R. Golub (2005). MicroRNA expression profiles classify human cancers. Nature 435, 834-838.

22. Kathryn A. O'Donnell, Erik A. Wentzel, Karen I. Zeller, Chi V. Dang and Joshua T. Mendell (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature 435, 839-843.

23. Lin He, J. Michael Thomson, Michael T. Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W. Lowe, Gregory J. Hannon and Scott M. Hammond (2005). A microRNA polycistron as a potential human oncogene. Nature 435, 828-833.

24. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003a). The microRNAs of *Caenorhabditis elegans*. Genes Dev. 17, 991–1008.

25. Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. (2003). Computational identification of *Drosophila* microRNA genes. Genome Biol 4:R42, 1–20.

26. Bonnet E, Wuyts J, Rouze P, Van de Peer Y: Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. Proc Natl Acad Sci USA 2004, 101(31):11511-11516.

27. Jin-Wu Nam, Ki-Roo Shin, Jinju Han, Yoontae Lee, V. Narry Kim and Byoung-Tak Zhang. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research. 33:3570-3581.

28. Xiaowo Wang, Jing Zhang, Fei Li, Jin Gu, Tao He, Xuegong Zhang and Yanda Li. (2005). MicroRNA identification based on sequence and structure alignment. Bioinformatics.21:3610-3614.

29. A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D.S. Marks; (2003) MicroRNA targets in *Drosophila*; Genome Biology 5(1):R1.

30. Ola Sætrom, Ola Snove Jr, and Pal Sætrom. (2005) Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms.RNA. 11:995–1003.

31. Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during *Drosophila melanogaster* development. Dev. Cell 5, 337–350.

32. Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. RNA 9, 175–179.

33. Bashirullah, A., Pasquinelli, A.E., Kiger, A.A., Perrimon, N., Ruvkun, G., and Thummel, C.S. (2003). Coordinate regulation of small temporal RNAs at the onset of *Drosophila* metamorphosis. Dev. Biol. 259, 1–8.

34. Sempere, L.F., Sokol, N.S., Dubrovsky, E.B., Berger, E.M., and Ambros, V. (2003). Temporal regulation of microRNA expression in *Drosophila melanogaster* mediated by hormonal signals and broad-Complex gene activity. Dev. Biol. 259, 9–18.

35. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr. Biol. 12, 735–739.

36. Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. Cell 75, 855–862.

37. Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R., and Ruvkun, G. (2000). The lin-41 RBCC gene acts in the *C. elegans* heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. Mol. Cell 5, 659–669.

38. Cullen, B.R. (2004). Transcription and Processing of Human microRNA Precursors. Mol. Cell 16, 861-865.

39. Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10, 1957–1966.

40. Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Back, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 23, 4051–4060.

41. Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 21, 4663–4670.

42. Zeng, Y., and Cullen, B.R. (2003). Sequence requirements for microRNA processing and function in human cells. RNA 9, 112–123.

43. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415–419.

44. Basyuk, E., Suavet, F., Doglio, A., Bordonne, R., and Bertrand, E. (2003). Human let-7 stem-loop precursors harbor features of RNase III cleavage products. Nucleic Acids Res. 31, 6593–6597.

45. Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 17, 3011–3016.

46. Lund, E., Gu¨ ttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science 303, 95–98.

47. Kurihara, Y. and Watanabe, Y. (2004) Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proc. Natl. Acad. Sci. USA 101, 12753–12758.

48. Bartel, D.P. (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell. 116, 281-297.

49. Hammond, S.C., Bernstein, E., Beach, D., and Hannon, G.J. (2000). An RNA-directed nuclease mediates posttranscriptional gene silencing in *Drosophila* cells. Nature 404, 293–296.

50. Elbashir, S.M., Leneckel, W., and Tuschl, T. (2001a). RNA interference is mediated by 21- and 22- nucleotide RNAs. Genes Dev. 15, 188–200.

51. Nykänen, A., Haley, B., and Zamore, P.D. (2001). ATP requirements and small interfering RNA structure in the RNA interference pathway. Cell 107, 309–321.

52. Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell 110, 563–574.

53. Schwarz, D.S., Hutvágner, G., Haley, B., and Zamore, P.D. (2002). Evidence that siRNAs function as guides, not primers, in the *Drosophila* and human RNAi pathways. Mol. Cell 10, 537–548.

54. Lai, E.C. (2002). MicroRNAs are complementary to 3'-UTR motifs that mediate negative post-transcriptional regulation. Nat. Genet. 30, 363–364.

55. Lewis, B.P., Shih, I., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell 115, 787–798.

56. Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of Drosophila microRNA targets. PLOS Biol. 1, E60.

57. Hutvágner, G., and Zamore, P.D. (2002). A microRNA in a multiple turnover RNAi enzyme complex. Science 297, 2056–2060.

58. Zeng, Y., Wagner, E.J., and Cullen, B.R. (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol. Cell 9, 1327–1333.

59. Doench, J.G., Peterson, C.P., and Sharp, P.A. (2003). siRNAs can function as miRNAs. Genes Dev. 17, 438–442.

60. Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002b). Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. Science 297, 2053–2056.

61. Xie, Z., Kasschau, K.D., and Carrington, J.C. (2003). Negative feedback regulation of Dicer-like1 in *Arabidopsis* by microRNA-guided mRNA degradation. Curr. Biol. 13, 784–789.

62. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T. (2003) A uniform system for microRNA annotation. RNA, 9(3), 277-279.

63. Griffiths-Jones S. (2004) The microRNA Registry. NAR, 32, Database Issue, D109-D111

64. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. (2006). miRBase: microRNA sequences, targets and gene nomenclature. NAR, 34, Database Issue, D140-D144.

65. Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M., Maller, B., Srinivasan, A., Fishman, M., Hayward, D., Ball, E., et al. (2000). Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide let-7 heterochronic regulatory RNA. Nature 408, 86–89.

66. Seitz, H., Youngson, N., Lin, S.P., Dalbert, S., Paulsen, M., Bachellerie, J.P., Ferguson-Smith, A.C., and Cavaille, J. (2003). Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat. Genet. 34, 261–262.

67. Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., and Burge, C.B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA.

68. D. Searls, “The Language of Genes,” Nature, vol. 420, pp. 211-217, Nov. 2002.

69. R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis. Cambridge Univ. Press, 1998.

70. A. Krogh, M. Brown, I.S. Mian, K. Sjo¨lander, and D. Haussler, “Hidden Markov Models in Computational Biology: Applications to Protein Modeling,” J. Molecular Biology, vol. 235, pp. 1501-1531, 1994.

71. Y. Sakakibara, M. Brown, R. Hughey, I.S. Mian, K. Sjolander, R.C. Underwood, and D. Haussler, “Stochastic Context-Free Grammars for tRNA Modeling,” Nucleic Acids Research, vol. 22, pp. 5112-5120, 1994.

72. Tyagi, S. et al. (unpublished)

73. K. Lari and S.J. Young. (1990) “The Estimation of Stochastic Context-Free Grammars Using the Inside-Outside Algorithm,” Computer Speech and Language, vol. 4, pp. 35-56.

74. A.V. Aho and J.D. Ullman, The Theory of Parsing, Translation and Compiling, vol. I: Parsing, Prentice Hall, 1972.

75. Soh, C.T. (1988). Factors influencing the pathogenicity of *Entamoeba histolytica*. Yonsei Medical Journal; 29(1):1-10.

76. Byers TJ. (1986) Molecular biology of DNA in *Acanthamoeba*, *Amoeba*, *Entamoeba*, and *Naegleria*. Int Rev Cytol. 1986;99:311-41.

77. Willhoeft U, Tannich E. (1999). The electrophoretic karyotype of *Entamoeba histolytica*. Mol Biochem Parasitol.;99(1):41-53.

78. Dhar SK, Choudhury NR, Bhattacharaya A, Bhattacharya S. (1995). A multitude of circular DNAs exist in the nucleus of *Entamoeba histolytica*. Mol Biochem Parasitol.;70(1-2):203-6.

79. Bhattacharya S, Bhattacharya A, Diamond LS, Soldo AT. (1989). Circular DNA of *Entamoeba histolytica* encodes ribosomal RNA. J Protozool.;36(5):455-8.

80. Ghosh S, Zaki M, Clark CG, Bhattacharya S. (2001). Recombinational loss of a ribosomal DNA unit from the circular episome of *Entamoeba histolytica* HM-1:IMSS. Mol Biochem Parasitol.;116(1):105-8.

81. Plaimauer B, Ortner S, Wiedermann G, Scheiner O, Duchene M. (1993). Molecular characterization of the cDNA coding for translation elongation factor-2 of pathogenic *Entamoeba histolytica*. DNA Cell Biol.;12(1):89-96.

82. Willhoeft U, Buss H, Tannich E. (2000). Genetic differences between Entamoeba histolytica and *Entamoeba dispar*. Arch Med Res.;31(4 Suppl):S254.

83. Purdy JE, Pho LT, Mann BJ, Petri WA Jr. (1996). Upstream regulatory elements controlling expression of the *Entamoeba histolytica* lectin. Mol Biochem Parasitol.;78(1-2):91-103.

84. Lioutas C and Tannich E. (1995). Transcription of protein-coding genes in Entamoeba histolytica is insensitive to high concentrations of alpha-amanitin. Mol Biochem Parasitol.;73(1-2):259-61.

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

86. Zuker M, Stiegler P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research.9(1):133-48.

87. Sullivan CS, Grundhoff AT, Tevethia S, Pipas JM, Ganem D, (2005) SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. Nature. 435(7042):682-6.

88. Hofacker IL, Fontana W, Bonhoeffer S, Stadler PF, (1994) Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie, 125:167-188.

89. Burset M, Guigo R.(1996) Evaluation of gene structure prediction programs.Genomics. 34(3):353-67.

90. Burge C and Karlin S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol. 268(1):78-94.