5. SUMMARY
The *E. histolytica* ribosomal RNA genes (rDNA) are present in 200 copies per genome and are carried exclusively on circular DNA molecules. This was reported independently by two groups (Bhattacharya et al., 1989; Huber et al., 1989). The nucleotide sequence of one rDNA circle belonging to the strain HM-1:IMSS, has been determined (Sehgal et al., 1994). It measures 24.5 kb (24522 bp) and carries two copies of the ribosomal RNA transcription unit arranged as inverted repeats. It encodes the small, large and 5.8S rRNAs but not the 5S rRNA. This molecule is named as EhR1 and the complete map with sequence details has been worked out.

The main features of EhR1 are: The two transcription units (rDNA), are each 5889 bp long. They account for half of the DNA content of EhR1. A stretch of 361 bp immediately upstream of the two rDNAs have complete sequence identity. The region downstream of the two rDNAs is 3543 bp in length and contains the *Dra* I and *Sca* I repeat families. The region upstream of the rDNA is 9261 bp and also contains various families of short tandem repeats, some of which share partial or complete homology with repeats in downstream region, e.g., *Pvu*I repeats are very closely related to *Sca*I repeats and *Hinfl* repeats share large stretches of common sequence with *Dra*I repeats. Among the repeat sequences found exclusively in the upstream region are the *Avall* repeats and the 74 bp repeats. In addition to the various repeats, the upstream region contains a stretch of DNA corresponding to a polyadenylated transcript has been detected by northern blot analysis. This is designated as Tr. It encodes a 0.7 kb RNA of unknown function. From nucleotide sequence, this RNA lacks an ORF. It is also not edited because the cDNA and genomic sequences are identical (Burch et al., 1991).

The work described in this thesis mainly involved the study of different aspects of transcription of rRNA genes in EhR1. The areas covered under this umbrella were wide-ranging. Experiments were done to gain an understanding of the extent of sequence variation in the highly repetitive rRNA genes, and especially the two rDNA units in EhR1 because this may have a bearing on transcription. The existence of rRNA sequences in novel RNA molecules was investigated. Attempts were made to map the transcription start sites of the two rDNA units in EhR1 to see if both units had their own promoters. The upstream and downstream sequences of the rDNA circle of *E. moshkovskii* was studied to look for consensus regulatory sequences. The results obtained can be summarised as:
1) EhR1 contains two rDNA transcription units which were designated as rDNA I (rightward unit) and rDNA II (leftward unit). The nucleotide sequence of rDNA I was already determined. To determine if the two rDNA units had any significant sequence differences, the nucleotide sequence of rDNA II was determined in selected regions. The sequence of Internal transcribed spacers was determined first, as the ITSs are known to be highly polymorphic in all systems. Both the units were shown to have complete sequence identity in all the regions sequenced including the ITSs.

A further study of the variation levels in the ITS-1, 5.8S and ITS-2 nucleotide sequences of rDNA units from different EhR1 molecules was done by generating a library of fragments containing this region by PCR amplification of genomic DNA from HM-1:IMSS using specific primers. Complete nucleotide sequence of ten such clones, again confirmed that they had an identical sequence. We could thus conclude that rRNA genes in HM-1:IMSS are fully conserved and all the genes would have an equal potential of being transcribed.

A comparison of the ITS-1, 5.8S and ITS-2 region from other strains of *E. histolytica* like, Rahman and HK-9 and other species like, *E. dispar* (strain CDC 0784 and isolate 43), *E. moshkovskii* (strain laredo) and *E. invadens* (strain IP-1) was used for the phylogenetic placement of these organisms. The results were as expected from comparison of other gene sequences. The 5.8S sequence of *E. histolytica* strains HM-1:IMSS and Rahman was fully identical in terms of both nucleotide sequence and length and strain HK-9 showed only one nucleotide change. The isolate 43 and CDC 0784 strain of *E. dispar* showed differences in 9 and 5 positions respectively, with respect to HM-1:IMSS. *E. moshkovskii* strain laredo and *E. invadens* strain IP-1 were quite divergent showing differences of 29 and 93 in the nucleotide sequence respectively. The 5.8S of *E. invadens* (IP-1) was only 106 bp long, while the 5.8S of laredo was one nucleotide shorter than that of HM-1:IMSS. As expected, the ITS sequences were quite different, with ITS-2 being much more variable than ITS-1. These differences can be used for diagnostic purposes. The species and strains included in the analysis thus separated into three major groups consisting of *E. histolytica*-type where the strain Rahman is overall closer to HM-1:IMSS than the strain HK-9, *E. dispar*-type consisting of isolate 43 and CDC 0784 which branched very close to *E. histolytica*, and other *Entamoeba*, *E. moshkovskii* and *E. invadens* which are quite divergent from *E. histolytica*.
2) The expressed sequence tag (EST) method, used for the identification of novel as well as housekeeping genes of *E. histolytica* by 5' end sequencing of cDNA clones from a representative library showed that 5% of the clones had extensive homology to the rDNA, both from 18S and 28S regions. It was decided to explore whether the rRNA-like ESTs arose due to an artifact of cDNA preparation or were bona fide products of novel mRNA species. To limit the study, only the clones homologous to the 18S region were selected. The rRNA-like clones were polyadenylated with the poly (A) tail ranging in length from 18-40 nu and surprisingly they were fully identical to the 18S nucleotide sequence of *E. dispar*. However, by extensive analysis using oligos specific to the *E. dispar* sequence on northern blots of *E. histolytica* RNA, Southern blot of *E. histolytica* genomic DNA, PCR and RT-PCR, it was shown that no *E. dispar*-like rRNA sequences could be detected in *E. histolytica* cells, *in vivo*. These sequences arising in the cDNA library were therefore artifactual.

Interestingly, 18S rRNA probes when hybridized to northern blots containing total RNA from HM-1:IMSS, gave strong positive signals with at least three other species besides the expected 18S rRNA. These do not appear to be degradation products as they gave discrete bands which hybridized preferentially with the 3'- end of 18S rRNA. Similar situations reported in other systems like yeast, show that rRNA sequences are interspersed within protein coding regions. These molecules are shown to play regulatory roles. The small nucleolar RNAs (sno RNAs) encoded in the nucleolus by RNA pol II, also contain rRNA-like sequences which play a role in rRNA processing and splicing events. The rRNA-like species observed by us could fall into any of these categories.

3) In all known systems, the three rRNA genes (18S, 5.8S and 28S) are transcribed as a single pre-rRNA which is processed to give rise to the individual rRNA species. It was already shown in the strain HK-9 of *E. histolytica* that the transcription start site of the rDNA unit was 2447 bp upstream of the 5' end of mature 18S rRNA (Michel et al., 1995). HK-9, unlike HM-1:IMSS contains only one rDNA unit in its extrachromosomal rDNA circle. Moreover the upstream region of this rDNA unit is exactly identical to the upstream region of rDNA I. The sequence upstream of rDNA II is missing in HK-9. The start site for transcription of rDNA I in HM-1:IMSS would likely be the same as for HK-9 but it needed direct proof due to the different topologies in the two strains. Primer
extension reactions were carried out with primers homologous to different regions within and upstream of the 18S rRNA.

A primer homologous to the immediate 5'- end of mature 18S extended a product of 345 nu and not the expected 2447 bp. The 345 nu product could result from the processing of a larger pre-rRNA transcript. Alternatively, it could be a new transcription start site. The status of the start site identified in HK-9 was established with two primers, which extended products of sizes corresponding to the same start site identified earlier, thus proving that a similar start site exists in HM-1:IMSS also. However, it remains to be determined whether this is the only initiation site of rRNA gene transcription in HM-1:IMSS or whether the 345 nu product could be a second start site of transcription. Transfection studies and 5'- capping experiments are planned to probe this further.

Primers from different regions upstream to rDNA II were used to map the start site of this rRNA and also the Tr transcript in this region. The primers extended products of various sizes. There were four different end points within the Tr region and one end point 30 bp upstream to the Tr. It remains to be seen as to which of these end points correspond with true initiation events and which are due to processing. Tr and rDNA II could be transcribed from a common promoter or they might have individual promoters.

4) The pre-rRNA processing pathway in eukaryotic systems follows a certain pattern of discrete cleavages which results in the formation of the mature 18S, 5.8S and 28S species. The sequential cuts can be summarised as: 1. first cut made in 5'- ETS which is about 300- 400 bases upstream to the end of the mature 18S species. 2. Second cut made in the ITS-1 to generate an intermediate (24S) that carries additional nucleotides at both the 5'- and 3'- ends. The mature ends are generated later by a series of exonucleolytic reactions. 3. Third cut is made in the ITS-2 to generate the immediate precursor of the 5.8S species and the 28S species. The exact mechanism by which these transcripts are rendered mature is yet to be determined.

The location of cleavage sites in the 5'- ETS, ITS-1 and ITS-2 during processing of the pre-rRNA transcript was identified by using primer extension with specific oligos and determining the exact size of the extended product An oligo located 200 bp downstream to the 5'- end of the mature 18S rRNA extended an expected product of 200 bases corresponding to the mature end of the 18S rRNA and another product of 227 bp which may be the site of the endonucleolytic cut made in the 5'- ETS before action by
exonucleases to generate the mature product. The cleavage site in ITS-1 was mapped by using another oligo located in the immediate 5' end of 5.8S region. It extended a product that mapped 23 bp downstream to the end of the mature 18S rRNA. The oligo used for identifying the cleavage made in the ITS-2, was homologous to the immediate 5'- end of 28S and extended a product 31 bp downstream to the mature 3'- end of the 5.8S.

In addition to the cleavage site in the 5' ETS 29 bp upstream to the 5' end of the mature 18S another site mapped 345 bp upstream to the 5' end of mature 18S and that could be a processing site or could be a site for transcription initiation. The cleavage within ITS-1 corresponded to the data for ITS-1 processing in eukaryotes. The first cleavage in ITS-2 prior to mature termini formation is consistent with the rRNA processing pathway reported for *S. cerevisiae* and other eukaryotes.

5) A comparative study of non-pathogenic *E. moshkovskii* with *E. histolytica* was undertaken to study the regulatory sequences which are located in the upstream or downstream regions of rDNA transcription units and to arrive at some consensus feature among the rRNA plasmids of *Entamoeba*. The extrachromosomal rDNA circle of *E. moshkovskii*, strain laredo is smaller in size, being only 19.5 kb, and it contains only one rDNA transcription unit. Most of the upstream spacer was contained in a 8.0 kb *EcoR I* - *Hind III* fragment and the downstream spacer was contained in a 6.4 kb *EcoR I* fragment. These fragments were cloned and some parts were sequenced. A sequence comparison of the rRNA coding region (~ 400 bp from the 3' end of 28S coding region) showed that this region was well conserved. A BLAST analysis of the spacer sequence failed to detect any homology with any other sequence in the database. Unlike HM-1:IMSS, no significant repeat families could be identified in the downstream spacer region of laredo sequenced so far. Further sequencing of upstream and downstream regions is underway.