CHAPTER 5

SUMMARY
The Genome of an organism is its storehouse of information regarding its functional genes, the proteins it can synthesize and the various functions it can perform vis a vis its biochemistry, physiology, gene regulation and cell biology. Thus it is extremely crucial to make a significant leap in the direction of genome analysis. Genome analysis would in totality encompass karyotype analysis, chromosome mapping, physical mapping, genetic mapping, gene discovery and genomic sequencing allied to ongoing informatics and functional analyses. The physical mapping and sequencing of an organism's complete genome is a tedious and time-consuming task. In order to rapidly discover the genes of an organism numerous approaches have been employed. We have used the expressed sequence tag approach to discover novel genes from the non-pathogenic species *E. dispar*. We have also characterized a repetitive DNA element from *E. histolytica* and *E. dispar*. Our studies show that it is a degenerate non-LTR retrotransposon.

**Expressed sequence tags (ESTs)**
The Expressed Sequence tag approach is widely used for rapid gene discovery. It is much quicker and cost effective than the time taking and expensive genome sequencing projects. ESTs provide information only about the expressed genes of an organism as against whole genome sequences. Recently, the EST approach has been used to identify novel genes in *E. histolytica* (Azam et al., 1996, Tanaka et al., 1997, Willhoeft et al., 1999b). A number of genes hitherto unknown in *Entamoeba* were identified by this approach. During these studies unusual gene expression was also noticed in *E. histolytica*, for example, the translation elongation factor EF-I α was found to express at a very high level. Unlike *E. histolytica*, the information available about *E. dispar* genes is limited and there is a need to identify and characterize many more genes for comparative studies. Our data on the analysis of random ESTs from *E. dispar* shows that a large number of novel genes can be identified by this approach. Of all the ESTs examined, 11% were homologues of known genes from *E. histolytica* and 29% were homologues of genes in other organisms but not reported from *E. histolytica* or *E. dispar*. No identity could be assigned to 59% of the ESTs (Table 4).

The maximum numbers of ESTs (17%) were found to encode ribosomal proteins. This has also been observed with other organisms including *E. histolytica* (Azam et
al., 1996, Tanaka et al., 1997). Out of 16 ESTs that were found to be homologous of already known genes of *E. histolytica*, ten ESTs represented a different gene each and six accounted for two distinct genes only (Table 5A). The maximum number of identical ESTs whose function could be assigned was four and these corresponded to translation elongation factor EF-1α. Besides EF-1α, a number of ESTs corresponding to the translation initiation factor eIF5A were also identified. It is the only protein known which is post-translationally modified by deoxyhypusine synthase. An EST corresponding to a putative deoxyhypusine synthase has also been identified in *E. dispar* (Table 5B), which suggests that this modification also takes place in *Entamoeba* as found in animal cells. The first enzyme in the biosynthesis of methionine is cystathionine-gamma synthase and one of the ESTs is likely to encode this enzyme (Table 5B). There are two ESTs encoding proteasome sub-units DD5 and IOTA. It appears that *Entamoeba* may have the full complement of ubiquitin-dependent proteasome mediated protein degradation pathway reported in other eukaryotes. Cysteine proteinases are thought to be involved in pathogenesis. Recent data suggests that there are six cysteine proteinase genes in *E. histolytica* and four in *E. dispar* and the level of expression is substantially different in the two species (Bruchhaus et al., 1996). The two ESTs in our study that appear to encode cysteine proteinases are found to be homologues of *E. histolytica* cysteine proteinase ACP3 and EhCP6 indicating that these are expressed in the trophozoites of *E. dispar*.

**Presence of polyadenylated untranslated RNAs**

Several clones corresponding to unidentified ESTs were also found. Remarkably, there were ten ESTs corresponding to a single sequence UEEI (unidentified *Entamoeba* EST I) with percent identity at nucleotide level varying from 86 to 98. All the ten ESTs had different 5'-ends suggesting again that the high frequency may not be due to amplification of a few clones. However, a relatively high rate of transcription of UEEI by RT during the preparation of the library cannot be ruled out. A number of UEEI ESTs with slightly different nucleotide sequence suggests that there may be multiple genes encoding UEEI. From the search of the GenBank and dBEST for homologies with UEEI, two ESTs of *E. histolytica* with no known function were also found to be similar to this sequence (Tanaka et al., 1997). The multiple
alignment of a few representative *E. dispar* UEEI sequences and *E. histolytica* homologues is shown in Fig. 5. The sequence identity between the *E. dispar* and *E. histolytica* sequences was much lower than that observed within *E. dispar*. None of the UEEI sequences was found to have an extensive open reading frame (ORF). The average size of the reading frames was about 40 amino acid residues. The functions of these RNAs are not known at present and will be of tremendous interest. Identification of novel genes hitherto unknown could be useful in understanding the biology of *Entamoeba*. Before this study, the nucleotide sequence of only a few genes of *E. dispar* was known. Availability of a much larger number of gene sequences would facilitate research in this area and will be useful to understand the molecular basis of pathogenesis. Our data also emphasizes that EST is a powerful strategy for discovering a large number of expressed novel genes of an organism.

**Cloning and characterization of a retrotransposon–like element from *Entamoeba***

Repetitive DNA sequences are interspersed throughout the genomes of mammals, higher eukaryotes and also prokaryotes. In *E. histolytica* numerous repetitive DNAs have been identified but no retrotransposon-like element has been reported so far. Our repetitive DNA element HMc was originally identified as a non-ribosomal chromosomal DNA element found in several copies across the genome (Mittal et al., 1994). Complete sequencing and homology search recently identified a region with considerable homology to reverse transcriptase. We have further characterized this repeat element, renamed as EhRLE1 (*E. histolytica* retrotransposon-like element1). EhRLE1 has a copy number of 140 and is found distributed all over the genome i.e. present on all chromosomes. The entire EhRLE was found to be repetitive hence we cloned out larger fragments containing this repeat in order to characterize the ends of the repeat and its flanking sequences. Characterization of five EhRLE elements (EhRLE2-6) was done. Two of the elements (EhRLE2 and EhRLE6) were found to contain only a truncated portion of EhRLE flanked by unique sequences. EhRLE6 had a highly truncated repeat region hence was excluded from comparative analysis. The largest repeat unit was found to be 4087 bp long from...
EhRLE5. The repeat region was flanked by unique sequences at both ends hence represented a complete repeat unit. The repeat started with a region homologous to the SOD gene upstream sequence. Downstream of this was a region of homology to the reverse transcriptase gene. A part of the cpn60 gene upstream region was present adjacent to the RT region. This region partly overlaps with the RT coding region. The cpn60 upstream sequence is followed by a sequence with no match in the database. This is followed by region of homology to the IE element. Both ends of the repeat are flanked by 14-mer inverted repeats and 8-mer direct repeats. EhRLE4 also has a large part of the repeat unit but the IE element is absent. At the 5' end both elements have sequence homologies to known genes in the database. At the 3' ends unique sequences without a database match are present. The repeat ends and unique region probes show expected results upon Southern hybridization with EcoR1 digested genomic DNA.

Each of the elements shows sequence variation from each other over the entire stretch of homologous sequence corresponding to RT. Hence we derived a consensus sequence of the element. The majority nucleotide was taken as the consensus but whereever a majority was absent, the nucleotide present in EhRLE3 was taken as this showed maximum BLASTX scores for the RT conserved region. We found that the sequences showed overall variation from 1.9% (EhRLE3) to 4% (EHRLE5) from the consensus sequence. The variation from the consensus was lesser in the RT region 1% to 3% respectively. The consensus sequence also did not show any continuous open reading frame, hence we reconstructed a continuous ORF of 464 amino acids spanning the RT conserved region. This required the insertion of a 'T' at position 1365 of the consensus sequence. The scores increased significantly from 84 (EhRLE3) to 134 (consensus) and the p value lowered from 1e-26 (EhRLE3) to 3e-30 (consensus). We found seven amino acids varying from the RT conserved residues in our element. We found that substitution of 9 nucleotide positions could get the conserved RT signature sequence.

We further analysed our element and found that it did not have any LTRs. A TATA like element was found within the repeat unit upstream of the RT coding region but no other promoter element could be identified. A phylogenetic tree was constructed using EhRLE and several other LTR and non-LTR RT sequences. EhRLE grouped
along with non-LTR retrotransposons. The genescan software detected coding potential in the RT region which improved after reconstruction.

We looked for the presence of this element in other species of *Entamoeba*. It was found to be absent in *E. moshkovskii* but present in *E. dispar* at a low copy number. We used EhRLE1 as probe and fished a copy from the *E. dispar* genomic library. The EdRLE showed about 85% sequence identity to EhRLE. The *E. dispar* sequence showed identity to the cpn60 upstream region and RT region. The RT homology values were much lower probably due to the presence of excessive mutations. The genic regions of *E. histolytica* show about 3% variation on an average. The intergenic regions show approximately 12% average variation. This indicated that EhRLE was present as an intergenic region at the time of divergence of *E. histolytica* and *E. dispar* as separate species.

We have thus characterized a non-LTR retrotransposon from *Entamoeba* which is now degenerate. We cannot rule out the presence of a functional copy. Reconstruction of transposase has been achieved successfully in other systems, hence may be possible for EhRLE RT also. Such an integration vector would help in mutagenesis and functional studies.