Mining of Expressed Sequence Tags for Identification of ATP Binding Cassette Transporters in Human Brain

Sandeep Kumar Vishwakarma[1, 2], Syed Ameer Basha Paspala[1, 2] and Aleem Ahmed Khan[1, 2]*

[1]Centre for Liver Research and Diagnostics, Deccan College of Medical Sciences, Kanchanbagh, Hyderabad, Andhra Pradesh-5000 58, INDIA.
[2]PAN Research Foundation, Narayanguda, Hyderabad, Andhra Pradesh, INDIA.

**Corresponding author: Aleem Ahmed Khan*, *aleem_a_khan@rediffmail.com*

Abstract

The general quality of expressed sequence tag (EST) sequences and the general absence of positive selection in these sequences make ESTs an attractive tool for the study of evolutionary relationship and sequence searches. Due to emerging scope of ABC transporters in human brain diseases treatment investigations and drug resistance, here we investigate the ABC transporter ESTs present in human brain. 15 ESTs were identified and used for cluster generation to see the gaps and subjected to phylogenetic analysis to identify the evolutionary relationship and conserved structures. Investigation and clustering of human ESTs from brain allows extension of data sampling from outside of the genome project.

Keywords: Expressed sequence tag, ABC transporters, cluster generation, phylogenetic analysis.

Expressed sequence tag (EST) is short sub-sequence of a cDNA molecule and is used to identify gene sequences, transcripts analysis and gene discovery (Adams *et al.,* 1991). Due to its wide applications in genomics identification of various genes, investigation of ESTs has developed quickly, with 74,186,692 ESTs entered in public databases for all organisms. Rapid identification of novel markers and evolutionary relationship among human ESTs searches has improved since last decades and now the highest dbEST is available for human on NCBI approximately 8,704,790 (GenBank, 1st January 2013).

ESTs are generated from individual clones of cDNA library and results in a relatively low quality fragment with limited length to approximately 500 to 800 nucleotides. Because these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. The current understanding of the human set of genes includes the existence of thousands of genes based solely on

EST evidence. In this respect, ESTs have become a tool to refine the predicted transcripts for those genes, which leads to the prediction of their protein products and ultimately their function (Christoffels *et al.,* 2001; Skrabanek *et al.,* 2001).

High-throughput analyses of ESTs often encounter several challenges such as poor data management which makes it difficult to write programs that can unambiguously determine that two EST libraries sequenced from the same tissue (Campagne *et al.,* 2006). Similarly, disease conditions for the tissue are not annotated in a computationally friendly manner which needs further analysis for tissue specific gene expression for several genes and transcription factors.

The present study describes identification of ATP-binding cassette (ABC) transporters ESTs in human brain. This will assist to get an idea of the possible sources for EST data resources and the bottlenecks in EST analysis. We also provide significant tools to analyze their proven utility in different application areas, the general methods and protocols being followed for EST analysis obtained from dbEST on NCBI.

Materials and Methods

dbEST, TIGR and UniGene are the most useful resources for the identification of raw and cluster data of ESTs of many organisms. dbEST is one of the largest repository of EST database on NCBI. To initiate *in silico* analysis, the EST database was searched for ABC transporter gene clusters expressed in human brain tissues using dbEST on NCBI. The ESTs of all 49 known ABC transporter gene entries were downloaded and only those originating from human brain tissues were considered for analysis (Nagaraj *et al.,* 2007).

EST Pre-Processing

Automatically generated ESTs remains in low quality containing higher error rates with vector sequence contamination; hence, these undesired sequences were deleted from ESTs to reduce the overall redundancy and to improve the efficacy of further analysis. Using dbEST in our analysis is obvious as each cluster is generated by combined information from GenBank mRNA database, UniGene and electronically spliced genomic DNA which are clustered and cleaned from vector sequence contamination.

EST Clustering

A total of 15 ESTs expressing in human brain tissues were identified from seven reported ABC transporter gene families. Clustering of 15 identified ESTs for ABC transporters was performed to collect the overlapped sequences from the same transcript of a gene into a unique cluster to reduce the redundancy. All EST cluster scores were identified and converted into bar graph to know the gaps and false

overlapping. Multiple sequence alignment was performed and consensus sequences along with a quality value for each base were computed. Human brain tissue-based ESTs from 15 ABC transporter genes were subjected to cluster analysis. Distance matrix for all ABC transporters ESTs from human brain was computed GeneBee server and converted into graphical representation using Microsoft Excel 2007.

Database Similarity Searches

The consensus sequences of putative genes obtained from clustering were identified using freely available tools like BLASTN and BLASTX. For sequence analysis, all 15 ESTs of ABC transporter genes were aligned to the genome sequence of the human using specialized program like GeneBee server to assist genome mapping and gene discovery.

Phylogenetic tree construction

15 ABC transporters ESTs identified from human brain tissue were aligned and subjected to phylogenetic tree construction in PHYLIP format. Complete EST sequences were used for both cluster and topological algorithms with bootstrap values. Phylogenetic tree was represented in slanted form for better homology and evolutionary analysis.

Results and Discussion

Screening and identification of ESTs

All 48 members from 7 families of ABC transporter genes, 15 ESTs from human brain were identified from dbEST on NCBI.

Sequence similarity in ESTs of ABC transporters in human brain

The database similarity search by querying 15 ESTs on GeneBee against human genome revealed that all the ABC transporter contigs were showing high degree of sequence homology. Refined alignment with power value of 54.46 was found to be highly significant and showed 8.1% of sequence homology.

ESTs cluster generation

From several successful and widely accessed EST databases, the UniGene database was selected as it uses mRNA and other coding sequence data of GenBank as reference sequence for cluster generation. After search for human ABC transporter genes in dbEST, the EST sequences of human brain were selected due to their association and greater functionality in the onset and progression of human brain tumors such as gliomas. Distance matrix was calculated from the generated cluster database and converted into graphical representation showing high degree of

sequence homology and less distance with low matrix value (Figure-1).

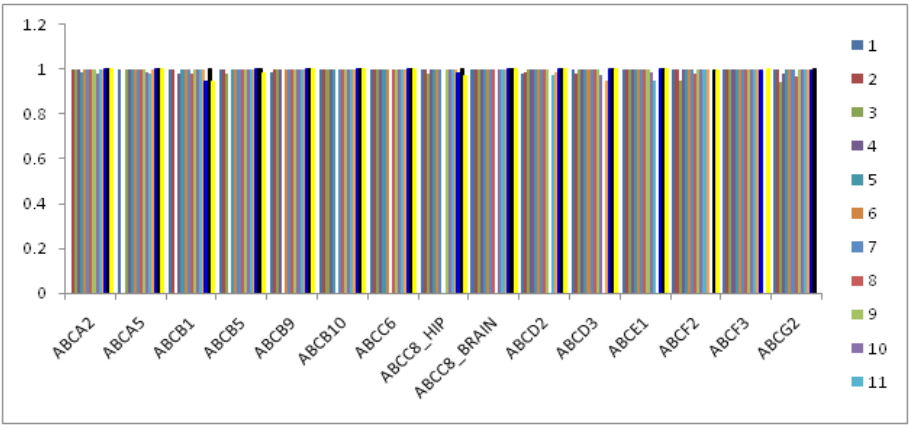
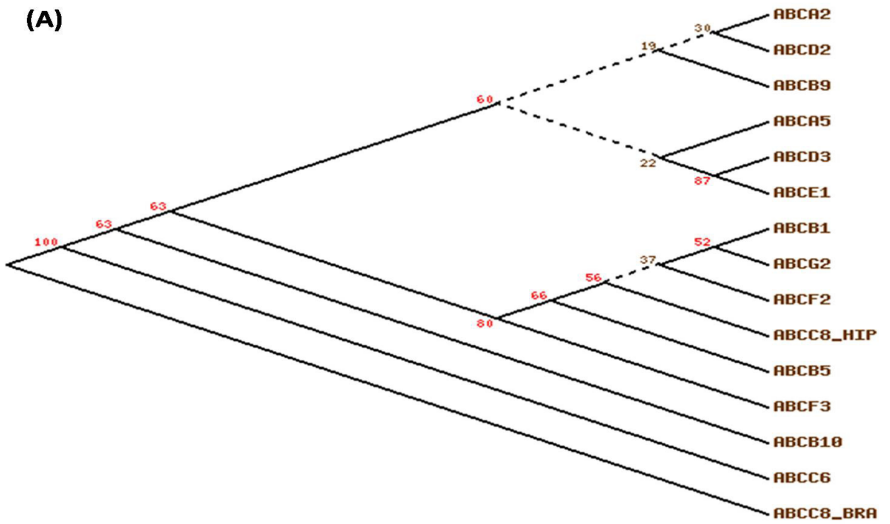


Fig. 1: Graphical representation of distance matrix generated from ABC transporters ESTs cluster database

Data deposition and phylogenetic analysis

Cluster algorithms of all identified 15 ESTs of ABC transporters from human brain were transformed into graphical representation with boot strap values. Slanted phylogenetic tree showed close relationship with high boot strap value for each EST due to the same origin of the tissue (Figure-2A).



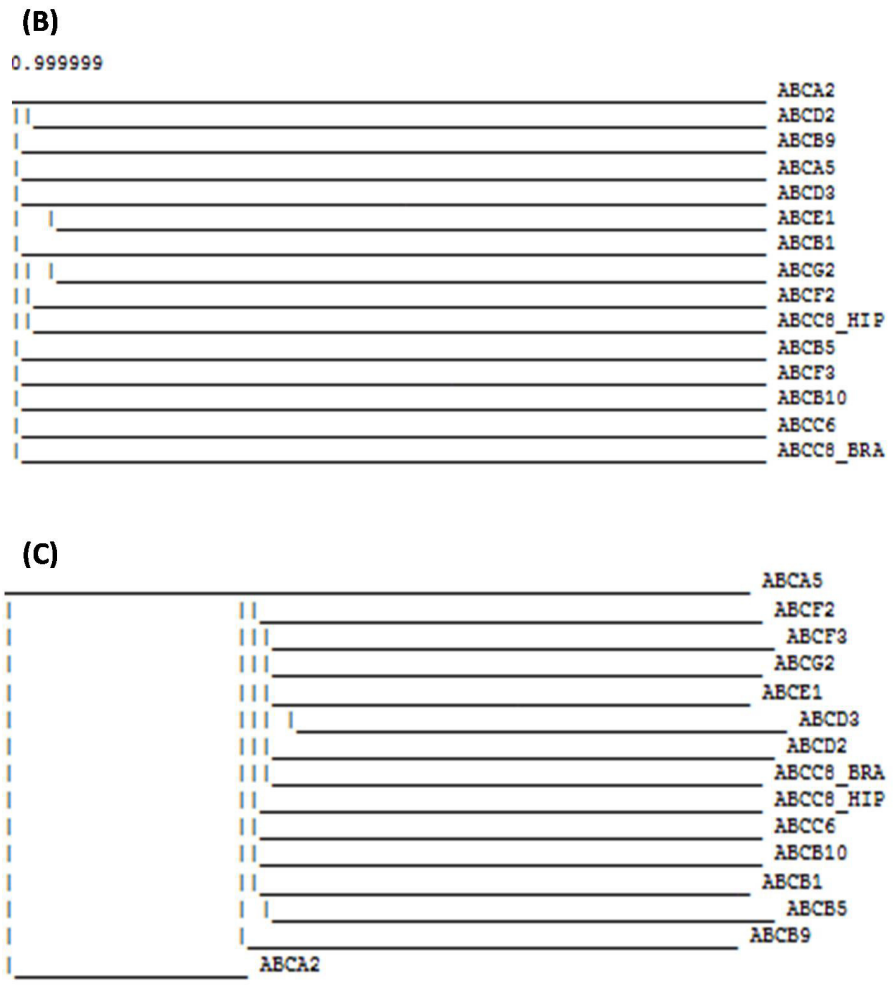


Fig. 2: (A-B) Cluster Algorithm of phylogenetic tree analysis (Phylip format) showing more similar sequences and evolution for ABC transporter ESTs except ABCA2 (C) Topological Algorithm analysis of phylogenetic tree (Phylip format) showing high distance among ABCA2 and other identified ESTs.

Further analysis of evolutionary relationship between ESTs from human brain using cluster algorithm showed 0.999999 value showing highly similar sequences for the ESTs (Figure-2B). Whereas, topological algorithmic analysis showed significant dissimilarity between ABCA2 EST sequence and other 14 identified ABC transporter EST sequences (Figure-2C).

Conclusion

In case of high-throughput EST analysis there is a need for integrated, automated approaches enabling EST data mining for the biologically useful information across disciplinary boundaries. Moreover, ESTs have diverse applications, and the question being addressed will determine the choice of methods or pipelines to be used. As the objective of individual methods and tools can vary substantially, it is difficult to evaluate all of them using a common platform and choose the most appropriate ones for individual projects.

As the data is limited for ABC transporters genes expressed in human, identification of ESTs from various tissue sources need to be identified (Sutcliffe *et al.,* 1982). The correlation is needed now to investigate the difference between the nucleotide/ protein sequences which ultimately results in its absolute function. The BLAST and alignment results have shown not much significant differences in all identified sequences.

While complete genomes are the ultimate data sets for resolving phylogenetic and evolutionary issues of different kinds, the production of human brain ESTs for ABC transporters are still at a level that precludes a dense taxonomic sampling among higher organisms. Therefore, there is need to establish methods that can be used to predict the sequence homology and of general interest for phylogenetic studies. Production and investigation of new ESTs from various tissue sources in human and other higher eukaryotes will gain more attention in future.

References

Adams MD, Kelley JM, Gocayne JD, *et al.,* 1991. "Complementary DNA sequencing: expressed sequence tags and human genome project". *Science.,* **252**(5013): 1651–1656.

Campagne F, Skrabanek L 2006. "Mining expressed sequence tags identifies cancer markers of clinical interest". *BMC Bioinformatics.,* **7**: 481.

Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W 2001. "STACK: Sequence Tag Alignment and Consensus Knowledgebase". *Nucleic Acids Res.,* **29**(1): 234–238.

Nagaraj SH, Gasser RB, Ranganathan S 2007. "A hitchhiker's guide to expressed sequence tag (EST) analysis". *Brief. Bioinformatics.,* **8**(1): 6–21.

Skrabanek L, Campagne F 2001. "TissueInfo: high-throughput identification of tissue expression profiles and specificity". *Nucleic Acids Res.,* **29**(21): E102–2.

Sutcliffe JG, Milner RJ, Bloom FE, Lerner RA 1982. "Common 82-nucleotide sequence unique to brain RNA". *Proc Natl Acad Sci U S A.,* **79**(16): 4942–4946.