Heliyon 8 (2022) e10476

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

BESFA: bioinformatics based evolutionary, structural & functional analysis of prostrate, Placenta, Ovary, Testis, and Embryo (POTE) paralogs

Sahar Qazi [a,e,1], Bimal Prasad Jit [a,1], Abhishek Das [a], Muthukumarasamy Karthikeyan [b], Amit Saxena [c], M.D. Ray [d], Angel Rajan Singh [d], Khalid Raza [e], B. Jayaram [f], Ashok Sharma [a,*]

[a] Department of Biochemistry, All India Institute of Medical Sciences, Delhi 110029, India
[b] National Chemical Laboratory, Council of Scientific and Industrial Research (NCL-CSIR), Pune, Maharashtra, India
[c] Centre for Development of Advanced Computing, Pune, Maharashtra, India
[d] Dr. B.R.A Institute-Rotary Cancer Hospital, All India Institute of Medical Sciences, Delhi 110029, India
[e] Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India
[f] Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology, Delhi, India

ARTICLE INFO

Keywords:
POTE paralogs
Evolution
Homology
Adaptive divergence
Molecular dynamic simulation molecular mechanics/generalized born surface area (MMGBSA)
Molecular docking

ABSTRACT

The POTE family comprises 14 paralogues and is primarily expressed in Prostrate, Placenta, Ovary, Testis, Embryo (POTE), and cancerous cells. The prospective function of the POTE protein family under physiological conditions is less understood. We systematically analyzed their cellular localization and molecular docking analysis to elucidate POTE proteins' structure, function, and Adaptive Divergence. Our results suggest that group three POTE paralogs (POTEE, POTEF, POTEI, POTEJ, and POTEKP (a pseudogene)) exhibits significant variation among other members could be because of their Adaptive Divergence. Furthermore, our molecular docking studies on POTE protein revealed the highest binding affinity with NCI-approved anticancer compounds. Additionally, POTEE, POTEF, POTEI, and POTEJ were subject to an explicit molecular dynamic simulation for 50ns. MM-GBSA and other essential electrostatics were calculated that showcased that only POTEE and POTEF have absolute binding affinities with minimum energy exploitation. Thus, this study's outcomes are expected to drive cancer research to successful utilization of POTE genes family as a new biomarker, which could pave the way for the discovery of new therapies.

1. Introduction

Cancer testis/Cancer Germline (CT/CG) antigens are highly immunogenic and have observed mostly in adult testis/germline tissues and various cancer tumors, can be seen as potential candidates for cancer biomarkers and therapeutics. POTE (Prostrate, Placenta, Ovary, Testis, and Embryo) is a primate-specific class of proteins, first discovered by an *in-silico* screening approach using the Expressed Sequence Tags (EST) database [1]. POTE has 14 family members and is grouped into classes based on their similarities. POTE family members are localized on eight different chromosomes: 2, 8, 13, 14, 15, 18, 21, and 22, respectively [1, 2, 49]. The POTE proteins are composed of mainly three types of repeats, ANK motifs (33 amino acids), cysteine-rich region (CRR; 37 amino acids), and alpha-helical region of varying length in these paralogs. All the paralogs of the POTE family code for a different number of repeats have revealed that during gene evolution, several members of the POTE gene family comprise an actin-retroposon, which is present at the C-terminal of their ancestral paralogs [1, 3].

POTED (POTE-21) is located on chromosome 21. The first-ever POTE gene discovered codes for a protein of 66 kDa composed of three CRR, 5 ANK, and many helical regions [1]. POTEH (POTE-22) and POTE-G (POTE-2C) are two genes highly similar to POTED, the only exception being that the latter does not have alpha-helical motifs in their sequences. POTEH (POTE-22) viz. located on chromosome 22, codes for a 34 kDa protein containing two ANK and four CRR conserved regions (motifs). It has been observed that when these three POTE protein sequences (POTED, POTEH, and POTEG) are aligned, there is approximately 73% similarity showing their high homology [4]. This high homology indicates that these proteins may be similar in their functionality, which would make it easier to illustrate their significance in

* Corresponding author.
E-mail address: ashoksharma1202@gmail.com (A. Sharma).
[1] Equally contributed.

https://doi.org/10.1016/j.heliyon.2022.e10476
Received 22 December 2021; Received in revised form 25 January 2022; Accepted 23 August 2022

many diseases such as cancer. When comparing dexterous studies focused on the functionality and specific regions based on evolutionary protein analysis, POTE paralogs have limited literature recapitulation. Harking back to Bera et al. [2]; Barger et al. (2018), the research group suggests that the POTE family is primate-specific and belongs to a cancer-testis antigen (CTA) family and is likely to play a pivotal role in primate biological dynamics. Their analysis paved the way for looking at the POTE members in the direction of cancer-testis antigens (CTAs). Recently, Barger et al. (2018) discerned that POTE groups 1 and 2 encapsulate that POTEA, POTEB, POTEB2, POTEC, and POTED a normal tissue expression that is relevant to cancer-testis antigens (CTAs). However, group 3 paralogs, POTEE, POTEF, POTEG, POTEH, POTEI, POTEJ, POTEKP, and POTEM, have a function in normal tissues and is not considered as cancer-testis antigens (CTAs). This study also somehow clubs our POTE paralogs based on their specificity. When comparing the two-benchmark studies [2, 5], it is evident that a thin layer of vague specificity surrounds POTE paralogs. Henceforth, it is essential to determine the POTE paralogs' evolutionary spectra, which can help determine their structural and functional aspects.

In this study, we have used in silico modus operandi to compute the POTE protein family members' evolutionary relationship so that a correlation can be deduced between the evolutionary divergence of the POTE proteins and their functionality, accordingly. This study aims to identify whether POTE family members have been exposed to Darwinian selection in the process of evolution. Furthermore, structural predictions, molecular docking of the POTE paralogs to anticancer drugs, and molecular refinement with molecular mechanics/generalized Born surface area (MMGBSA) were computed to understand the POTE stability target receptors. Additionally, these POTE paralogs were subjected to a functionality assessment wherein each paralog's function was deduced in biological, chemical, and molecular aspects.

2. Materials & methods data sources

All the 14 POTE paralogs members have been considered, i.e. POTEA (NM_001002920; Q6S8J7), POTEB (NM_001277304; A0A0A6YYL3), POTEB2 (NM_001277303; H3BUK9), POTEB3 (NM_207355; A0JP26), POTEC (NM_001137671; B2RU33), POTED (NM_174981; Q86YR6), POTEE (NM_001083538; Q6S8J3), POTEF (NM_001099771; A5A3E0), POTEG (NM_001005356; Q6S5H5), POTEH (NM_001136213; Q6S545), POTEI (NM_001277406; P0CG38); POTEJ (NM_001277083; P0CG39), POTEM (NM_001145442; A6NI47) and POTEKP (AY014272; Q9BYX7) respectively. The above protein paralogs FASTA sequences were retrieved from HUGO Gene Nomenclature Committee (HGNC) [18].

2.1. Sequence similarity & alignment

As the protein sequence data of these paralogs is quite large and highly divergent, we have executed our study with leaves of 14 sequences (POTEA, POTEB, POTEB2, POTEB3, POTEC, POTED, POTEE, POTEF, POTEG, POTEH, POTEI, POTEJ, POTEM, and POTEKP) could keep genetic divergence small. The UniProt (https://www.uniprot.org/) accession numbers of the 14 POTE member sequences are shown in Table 1. For deducing the sequence similarity of the paralogs, we employed bioinformatics tools such as Simple Modular Architecture Research Tool – Basic Local Alignment Search Tool (SMART BLAST) (https://blast.ncbi.nlm.nih.gov/smartblast/smartBlast.cgi?CMD=Web), which tends to provide a breviloquent graphical summary of the entire proteins based on their evolutionary tracts [19] and Position-Specific Iterated PSI-BLAST (https://www.ebi.ac.UK/Tools/sss/psiblast/) iteratively searches many protein databases like NR (non-redundant) and utilizes a profile Position-Specific Scoring Matrix (PSSM) for the same [20]. After deriving the sequence similarity, we moved to a sequence alignment of the POTE paralogs that were executed based on the multiple sequence alignment (MSA) analogy [21]. Multiple sequence alignments (MSA) (https://www.ebi.ac.uk/Tools/msa/) are pivotal in various sequence analysis methods and are generally calculated using heuristic methods. The protein sequences were aligned by using CLUSTALW [22]. Multiple Sequence Comparison by Log Expectation (MUSCLE) [23] and CLUSTAL Omega [24] with default settings.

Table 1. POTE gene family and UniProt accession IDs of POTE paralogs.

Symbol	Description	Location	Uni-Prot ID
POTE A	POTE ankyrin domain family member A	8p11.1	Q6S8J7
POTE B1	POTE ankyrin domain family member B	15q11.2	A0A0A6YYL3
POTE B2	POTE ankyrin domain family member B2	15q11.2	H3BUK9
POTE B3	POTE ankyrin domain family member B3	15q11.2	A0JP26
POTE C	POTE ankyrin domain family member C	18p11.21	B2RU33
POTE D	POTE ankyrin domain family member D	21q11.2	Q86YR6
POTE E	POTE ankyrin domain family member E	2q21.1	Q6S8J3
POTE F	POTE ankyrin domain family member F	2q21.1	A5A3E0
POTE G	POTE ankyrin domain family member G	14q11.2	Q6S5H5
POTE H	POTE ankyrin domain family member H	22q21.1	Q6S545
POTE I	POTE ankyrin domain family member I	2q21.1	P0CG38
POTE J	POTE ankyrin domain family member J	2q21.1	P0CG39
POTE KP	POTE ankyrin domain family member KP	2q21.1	Q9BYX7
POTE M	POTE ankyrin domain family member M	14q11.2	A6NI47

2.2. Phylogenetic analysis

The phylogeny of POTE protein sequences was developed based on the multiple alignments of amino acid by employing Un-weighted Pair Group Method with Arithmetic Mean (UPGMA), Neighbour Joining (NJ), Minimum Evolution (ME), Maximum Parsimony (MP) [21], and Maximum Likelihood (ML) in MEGA 5.2 [25]. UPGMA is an agglomerative clustering method that shows the phenotypic similarities between operational taxonomic units (OTU) by showing an ancestral root. It assumes that evolution rates are more or less constant among different lineages [26]. Neighbor-joining (NJ) attempts to correct the UPGMA method for its inappropriate assumption about constant evolutionary rates throughout the lineage; thus, it gives a rootless phylogenetic tree. NJ is similar to the UPGMA method because the distant pairs of nodes are linked, and their common ancestral node is added to the tree, and their nodes are pruned from the tree accordingly [27]. Minimum evolution (ME) is a distance-based phylogenetic method where the trees are calculated from the pair-wise distances between the sequences rather than from the fit of individual nucleotide sites to a tree [28]. The trio, UPGMA, NJ, and ME are distance-based methods of the phylogeny.

Maximum Parsimony (MP) and Maximum Likelihood are two cladistics methods of generating phylogenetic trees for a commonly set species or reproductively isolated populations of a single species. Maximum parsimony (MP) searches for a tree that needs only a few evolutionary changes to explain the differences observed among the OTUs (Biology 1971). Maximum Likelihood (ML) creates all possible trees containing the set of organisms considered and then uses the statistics to evaluate the most likely tree for a small number of populations [29, 30].

Furthermore, we have also discerned each of the POTE paralogs' amino acid compositions and the number of amino acid substitutions in each paralogue. Some informative, conserved, and Mark-Parsim sites have also been identified, which can help POTE family members' structural analysis. The phylogenetic analysis also led to the estimation of POTE's average evolutionary divergence rate, which is essential in deducing its motion of divergence. The disparity index test was also determined using Monte Carlo replications. Supplemental Figure S1 represents the graphical summary of the research work executed in the paper. This paper has been bifurcated in to two sections – *primary and secondary research aspects*. The primary analyses encapsulate – sequence retrieval, sequence similarity, sequence alignment, evolutionary analysis,

structure prediction and quality assessment, molecular dynamic simulation (MDS) and MMGBSA along with molecular docking with selected NCI drug candidates. The secondary analyses on the other hand, talks about the sub-cellular localization, functional enrichment of POTE paralogs and protein-protein network associations.

2.3. Secondary structure prediction

Secondary structures of all the POTE Paralogs were predicted and produced using PSIPRED software. To know the helix, Beta, and loop exact position of amino acid of all POTE paralogs, we had made the PSI-blast-based secondary structure Prediction (PSIPRED) (http://bioinf.cs.ucl.ac.uk/psipred/) [31].

2.4. Homology and threading strategies for structure prediction

Structures of all the POTE paralogs were produced using homology and protein threading approaches of B.I. software such as the Swiss Model, MODELLER, and Phyre2 [32,33,34] respectively. Swiss Model and MODELLER are based on the sequence homology of the proteins, which are template-based. Phyre2 is a remote homology recognition threading strategy that employs Hidden Markov Models (HMMs) or only profiles to build precise tertiary structures of proteins.

2.5. Quality assessment of protein models

The retrieved protein models were then subjected to evaluations using Protein Quality (ProQ), Protein Structure Analysis (ProSA), and RAMPAGE [35, 36]. This was done to get the best optimal and stable tertiary structures, which can be further analyzed using MD simulations. The main criteria for assessment were the z-scores, over-all quality, residual identity and coverage.

2.6. Molecular dynamic simulations (MDS)

We first executed a refinement analysis for all the 14 POTE paralogs at 10 ns using the GalaxyRefine [37] online tool wherein the paralogs are stabilized using the AMBER force field ff94 [38]. After retrieving the refinement results, based on the ROG and RMSF results, we selected group III POTE paralogs- POTEE, POTEF, POTEI, and POTEJ for a detailed molecular dynamic simulation (MDS) using QwikMD toolkit [39] of Visual molecular dynamics (VMD) [40]. NVT dynamics were deployed, which hold an amount of substance (N), volume (V), and temperature (T) constants. The Noose-hover temperature was set to 300 K, and the entire simulation from each selected paralog was executed at 50 ns (a total of 1000 steps for each paralog). The first 10 ns was used to equilibrate the system whereas the remaining 40 ns was used for other electrostatic analyses. CHARMM27 protein-lipid parameter set was used to assign the topology and force field parameters [50]. Generalized Born Molecular Mechanics (GBMM) was deployed to retrieve the approximate results in explicit solvent. The visualization of the refined structures was done in PyMol (https://pymol.org/2/).

2.7. Molecular mechanics/generalized born surface area (MMGBSA) & electrostatics computation

The Molecular Mechanics-/generalized Born surface area (MM-GBSA) approach was also deployed to estimate the binding free energy (delta G) for complexes over simulation time [41]. This was executed using the APBS plugin available in VMD software (https://pymolwiki.org/index.php/APBS_Electrostatics_Plugin). Electrostatics were computed using Blues software [42].

2.8. Sub-cellular localization analysis of POTE proteins

To study POTE paralogs' functions, it is essential to know its sub-cellular localization as a protein can be localized either in the outer membrane, inner membrane, periplasm, extracellular space, or cytoplasm. Hence, before proceeding with the functional analysis, we checked for the sub-cellular localization using four different software, i.e., WoLFPSORT [43], Hum_mPLoc 3.0 [44], DeepLoc [45], and MDLoc [46]. WoLF PSORT software uses amino acid sequence and some sorting signal motifs of targeted protein to predict its subcellular localization of the protein. It displays information about detected sorting signals. Hum_mPloc 3.0 is also based on amino acid sequence and predicts 12 human subcellular localizations. DeepLoc software predicts the sub-cellular localization based on a neural network that processes the entire protein sequence and an attention mechanism identifying protein regions important for the subcellular localization. This online resource can differentiate between different localizations: Nucleus, Cytoplasm, Extracellular, Mitochondrion, Cell membrane, Endoplasmic reticulum, Golgi apparatus, Lysosome/Vacuole, and Peroxisome. MDLoc predicts the multiple locations for proteins using inter-dependencies among locations. It is based on an iterative process and uses the DBMLoc dataset for predicting subcellular localization. We give POTE protein sequences input in all the software, which is retrieved from the UNIPROT database.

2.9. Functional prediction using ProFunc

Functional prediction and analysis were executed by employing ProFunc [47] on the optimal selected tertiary structures on both predicted and refined models obtained after an exhaustive validation. ProFunc provides insight into proteins' functional capacity and various other aspects such as domain and clefts of the proteins, binding capacity, biological, cellular, and metabolic processes.

2.10. Protein-protein interaction (PPI) network

Further, we tried to assess the importance of the most highly connected proteins. For this, we analyzed their clusters using the bioinformatics database STRING [13] were constructed, and thus, protein interaction networks were retrieved. This database derives high throughput experimental data ($\geq$0.700) from a wide range of sources, analyses the co-expression of genes computationally, uses a scoring framework, and outputs a single confidence score per prediction. This confidence score is a measure of the predicted interactions' reliability, and a high score indicates that the predicted interactions are also replicated in the KEGG database [48].

2.11. Molecular docking on POTE proteins against different anticancer drugs

All the molecular docking study's computational procedures were carried out with the Molecular Operating Environment (MOE). POTE protein receptors were initially prepared with the default 3D protonation procedure in MOE [43]. The drug compounds were downloaded from NCI (https://www.cancer.gov/about-cancer/treatment/drugs#D) and then converted from name to 2D structure using ChemAxon tool (http://www.chemaxon.com) followed by 3D structure generation. Docking was performed using all default parameters with Triangle Matcher, Rigid Receptor, initial scoring method London dG retaining 30 poses, and final scoring method used was GBVI/WSA with 5 poses. POTE protein structures were imported into MOE after removing water molecules. All hydrogen atoms were then added to the structure with their standard geometry, followed by their energy minimization using default parameters with Forcefield value Amber 10: EHT and RMS gradient of 0.1

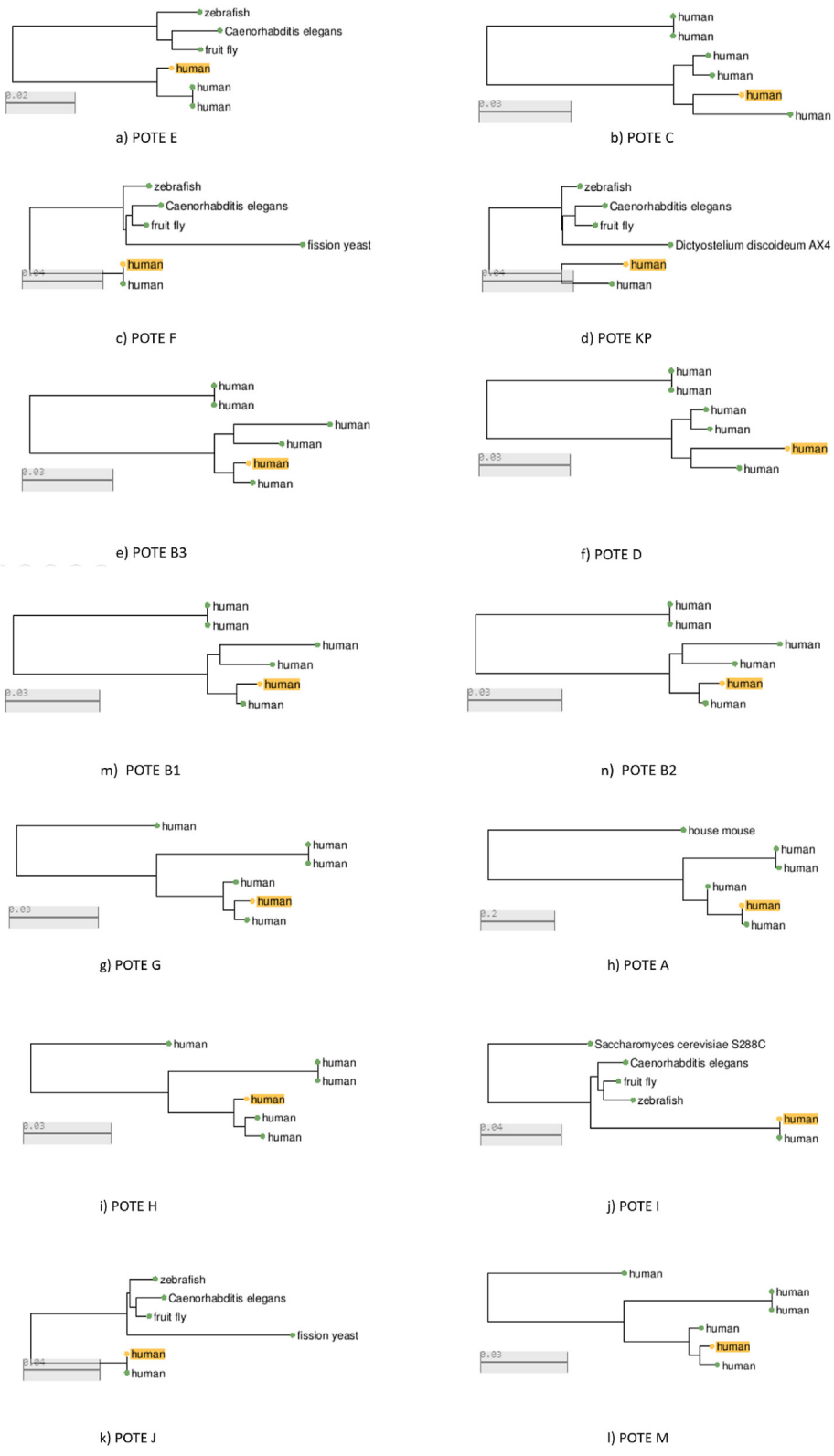


Figure 1. SMART BLAST results of POTE paralogs suggesting homology with other species.

kcal/mol. Each POTE receptor's binding site was identified through the MOE Site Finder program, which uses a geometric approach to calculate putative binding sites in a protein, starting from its tridimensional structure. Active sites were identified, and dummy atoms were created around the resulting alpha sphere centers. The backbone and residues were kept fixed, and energy minimization was performed.

3. Results

3.1. Sequence similarity & alignment

We identified a recent functional divergence in 14 POTE paralogs, as shown in Table 1. Because the duplications were identified through

similarity of full-length protein sequences, this method detected functional proteins instead of non-functional ones, where an exception was present only for POTEKP. Our study has successfully identified the homogeneity of the POTE paralogs, not only within themselves but also with other species. We used SMART BLAST to identify our POTE proteins' similarity on the evolutionary basis and found that POTEE, POTEF, POTEI, and POTEJ are orthologous to *C. elegans*, *D. discoideum* AX4, Thale cress, *S. cerevisiae* S288C. POTEA and POTEH are orthologous to the house mouse. POTEKP is orthologous to zebrafish, C. elegans, fruit fly, *Dictyostelium discoideum* AX4. There are no orthologous genes of POTEB, POTEB2, POTEB3, POTEC, POTED, POTED, POTEM, and POTEG. The pictorial representations of SMART BLAST results have been shown in Figure 1. It lucidly indicates the fact that POTE paralogs may have a recent divergence from other species to humans. This also reiterates that some of the POTE family members are clusters of orthologous groups (COGs) sharing high similarity with another genus. We can hypothesize that if they have sequence similarities with other species, then it is possible that their functions can be derived from these orthologous. It is not an apocryphal tenet that a protein sequence and structure can describe its appropriate functioning and throw light on its dynamic mechanistic pathways, providing insights to many specialized domains that can be fruitful in drug designing developments for various concerned diseases.

PSI-BLAST is a statistically driven protein similarity search method that hunts regions of similarity between the query sequence and landmark database sequences and generates gapped alignments. The PSI-BLAST program is more sensitive than BLAST because it can find distantly related sequences missed in a BLAST search. It can repeatedly search the target landmark databases such as – nr (non-redundant) using, multiple alignments of high-scoring sequences found in each iteration to produce a new PSSM for the next round. The program iterates until no new sequences are found or if the threshold is achieved. The PSI-BLAST results show similar domains in the query sequence and similar sequence hits retrieved by the program (Table 2). The results retrieved by PSI-BLAST indicate the fact that all the 14 POTE paralogs have high similarity with *Bos taurus* (3U4L_A, 2OAN_A, 2BTF_A), *Sus scrofa* (5NW4_V & 5AFT_H), *Drosophila melanogaster* (4JHD_B, 4JHD_A, 4RWT_A, 2HF3_A, 3EKS_A & 4M63_C), *C. elegans* (1D4X_A) and *Limulus polyphemus* (3B63_L) and a total number of amino acids in these species as retrieved by PSI-BLAST shown in Figure 2. The higher the percentage similarity between sequences, the better is the alignment. The similarity score (refer Table 2) depicts a good alignment for all the POTE paralogs with *Bos taurus*, *C elegans* and *L. polyphemus* that show 92% of similarity with the POTE sequences.

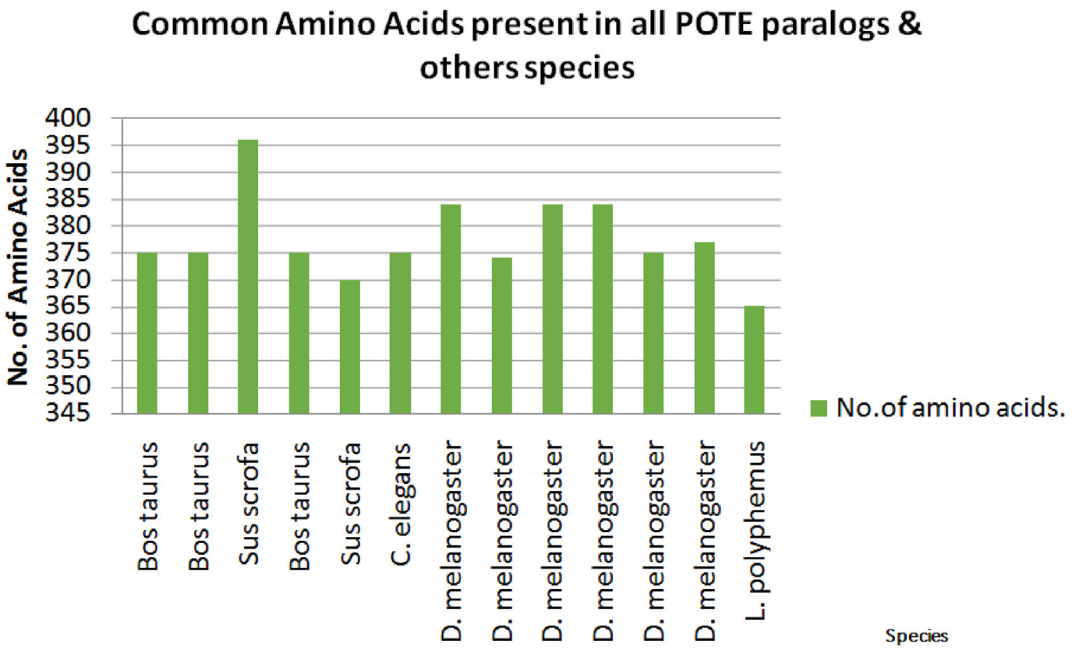
PSI-BLAST may successfully determine some subtle relationships that surpass the standard database in similarity searches but is dependent on the amino acid pattern, viz. conserved within the protein family of interest. We chose only the highest hits retrieved by the PSI-BLAST algorithm, and in each case; it reports a simple but structurally and functionally relevant relationship between humans and other species such as- *D. melanogaster, C. elegans, L. Polyphemus, B. Taurus, and S. scrofa* as these are optimal hits highly significant with an E-value of 0. The alignments suggest that these relationships have clear family members, henceforth hinting for further research analysis on the reliability and correlation of these protein hits with our POTE protein sequences query. POTE proteins, which have been presumed to share an evolutionary relationship, descend from a common root or origin. Thus, a multiple sequence alignment using Omega, Muscle, ClustralW & MEGA 5.02 was executed to infer the sequence homology and phylogenetic analysis. The results discern mutational events such as point mutations occurring at different locations as different characters in a single alignment column and insertion or deletion mutations (indels or gaps), which appear as hyphens in one or more of the sequences alignment. Since POTEB and POTEB2 are the same and equivalent in length, they have been tagged simply as POTEB while keeping POTEB3 separately, which is lengthier. The detailed results have been displayed in supplementary figure S2 (a–c).

Table 2. Highest sequence similarity of POTE paralogs with various organisms other than *Homo sapiens* with accession numbers have been provided by PSI-BLAST having significant expectation values (E-value).

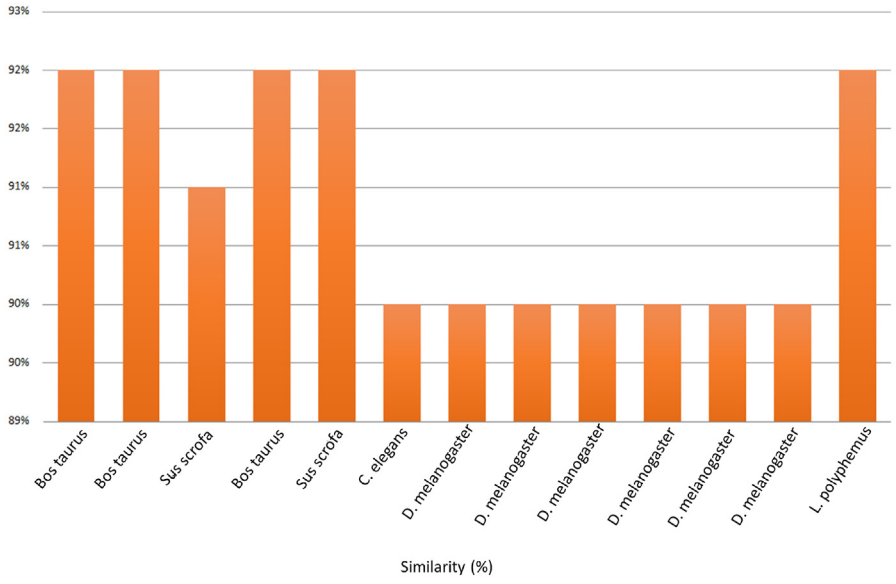
Organism Name	Accession Id	Query Coverage	E-Value	Similarity (%)	No.of amino acids.
Bos taurus	3U4L_A	34%	0.0	92 %	375
Bos taurus	2OAN_A	34%	0.0	92%	375
Bos taurus	2BTF_A	34%	0.0	92%	375
Sus scrofa	5AFT_H	34%	0.0	92%	370
Sus scrofa	5NW4_V	34%	0.0	91%	396
C.elegans	1D4X_A	34%	0.0	90%	375
D. melanogaster	4RWT-A	34%	0.0	90%	384
D. melanogaster	2HF3_A	34%	0.0	90%	374
D. melanogaster	4JHD_A	34%	0.0	90%	384
D. melanogaster	4JHD_B	34%	0.0	90%	384
D. melanogaster	3EKS_A	34%	0.0	90%	375
D. melanogaster	4M63_C	34%	0.0	90%	377
L. polyphemus	3B63_L	34%	0.0	92%	365

ClustalW, along with the alignment, also predicts phylogeny have been shown in Supplemental Figure S2. The phylogenetic analysis executed by CLUSTALW discerns that these proteins have diverged at a slow pace and clubbed into five specific groups. Although POTEA is quite close to POTEB, POTEB3, POTEC, POTED, POTEH, POTEM, and POTEG, it has diverged and separated itself from the rest, hinting at its monoexistence. The reason being that POTEA has a varying structural skeleton; thus, in group I. POTEB, POTEB2, POTEB3, POTEC, and POTED are closer to one another because of highly similar structural framework and hence makes group II. POTEH, POTEM, and POTEG make group III as they are clubbed together because of the same ancestral root and similarity. POTEE, POTEF, POTEI, and POTEJ, however, make a visibly distinct group IV arising from a common root but segregating to sub clubs of POTEE and POTEF and POTEI and POTEJ, referring again to their similar framework. POTEKP is a non-functional protein that rises from a pseudogene and is bereft and an outlier with the lowest similarity with the other paralogs. The cladogram retrieved from CLUSTAL W has been represented in Figure 3.

Phylogenetic analysis using Molecular Evolutionary Genetics Analysis (MEGA) provided a better overview of POTE's alignment and phylogenetic by giving an edge to its slow divergence and conserved regions. Multiple sequence alignment by MEGA showcases the gapped alignments and stringently conserved regions in all the 14 paralogs as shown in Supplemental Figure S3 by marking them in similar color and letter. At the same time, the mismatches have been highlighted with a different color and letter. Phylogenetic tree construction discerns that POTEE, POTEF, POTEI, and POTEJ are clubbed together, hinting at their common divergence from other species during evolution, while POTEB, POTEB2, POTEB3, POTEC, POTED, POTEG, POTEH, and POTEM are placed relatively nearer to each other. POTEKP and POTEA are placed straightforwardly as outliers and nearer to one another, showing that their ancestral origin might be linked. The phylogeny was deduced using UPGMA, Neighbor-Joining, Minimum Evolution, Maximum Parsimony, and Maximum Likelihood algorithms, respectively, which are based on distance and cladistics approaches, respectively. All the five phylogenetic trees were obtained to hint at the concept of adaptive evolution in POTE paralogs. The phylogenetic trees



(a)



(b)

Figure 2. Position specific similarity search for POTE paralogs. a) Graphical representation of the common of amino acids, b) Sequence similarity (%) of various species with POTE paralogs by PSI-BLAST.

of POTE paralogs using a) *UPGMA*, b) *Minimum Evolution*, c) *Neighbor Joining*, d) *Maximum Likelihood* and e) *Maximum Parsimony* have been displayed in Figure 4.

3.2. Secondary structure predictions

The secondary structures of all the POTE paralogs were developed by using PSIPRED software. The secondary structure prediction result is shown in Supplemental Figure S4, which indicates that the percentage of β-strands is much more significant in POTEE, POTEF, POTEJ, POTEI, and POTEKP paralogs than the percentage of β-strands in other paralogs, which makes our evolutionary analysis more subjective as a concept for adaptive divergence.

3.3. Structural predictions & molecular dynamic simulations (MDS)

Based on homology and protein threading strategies, POTE paralogs protein structures were generated using the Swiss Model, MODELLER, and Phyre2. The templates are taken for the Swiss Model, and query coverage, e-value, and sequence identity have been mentioned in

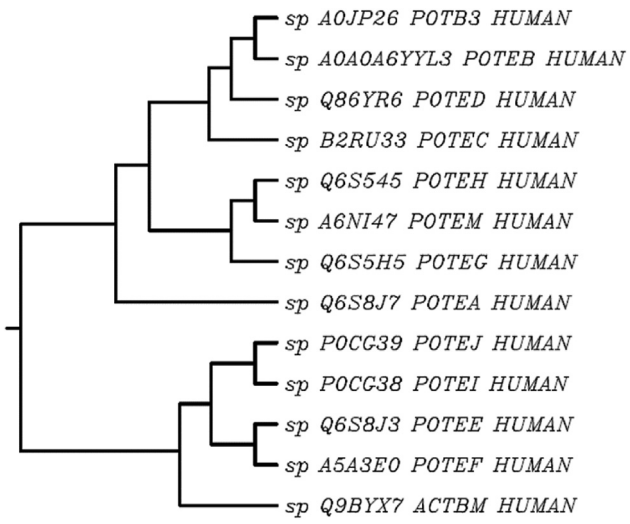


Figure 3. Cladogram of POTE paralogs by CLUSTAL W.

Supplemental Table 1. It was observed that all the protein paralogs comprised mainly of helices and supercoiled regions. Only POTEE, POTEF, POTEJ, and POTEI were the only paralogs, which contained β-strands along with the helices and coils in their tertiary structures, which matches the evolutionary analysis executed. POTEE, POTEF, POTEI, and POTEJ are clubbed together; thus, they have similar structures. POTEKP also contains β-strands. The POTE family's tertiary models and structures were developed using the Swiss Model, Phyre 2, and MODELLER, as shown in Figure 5. We refined the models created using the Swiss Model since these tertiary models had a better overall quality than Phyre2 and MODELLER, predicted by Protein Quality (ProQ), Protein Structure Analysis (ProSA), and RAMPAGE. The resolution change has been represented in Table 3.

Refined structures also predict helices in the predicted structures, as shown in Figure 6. Refinement of the predicted structures was done first at 10 ns to develop an idea about the accessible surface area, fluctuations present in the POTE paralogs' initial structures, and the number of hydrogen bonds formed by each paralog. The obtained results have been portrayed in Supplemental Figure S5. As the models get compact during the refinement procedure, we thus discern that ROG is reduced after stabilization in all the models, as shown in supplemental figure S6. Further, we calculated the number of hydrogen bonds, which subsequently increased, showing that hydrogen bonds' formation is enhanced across the trajectory shown in supplementary figure S7.

By exploring the Accessible Surface Area (ASA), Radius of gyration (ROG), hydrogen bonds, and Energy potential on our preliminary models, stabilized forms of all POTE paralogs were observed. However, after the initial 10 ns refinement, it was noted that only POTEE, POTEF, POTEI, and POTEJ have short regions of remodeled gaps. Furthermore, it was also observed that the ROG of all the four POTEs lay between 40 and 55 cm, and the number of hydrogen bonds formed was less compared to the remaining paralogs. Figure 7 displays the circos plot showing the relationship of each POTE paralog on the basis of evolutionary, structural and function.

Therefore, we selected only these four POTE paralogs, namely- POTEE, POTEF, POTEI, and POTEJ, for a detailed molecular dynamic simulation analysis executed at 50 ns. It is evident that the complexes have been refined to the best potential, and the total energy of the complex has also been stabilized, with all the structures having a good RMSD score with fewer clash scores and are well-fitting the Ramachandran plot criterion. Supplementary figure S8 represents the four simulated POTE paralogs – POTEE, POTEF, POTEI and POTEJ.

Table 4 summarizes the best refined POTE targets, including preliminary criteria such as - RMSD scores, clash core, accuracy score (refinement of the backbone), Ramachandran plot score, and MolProbity scores. Root mean square deviation (RMSD) describes the various hinges present in the structure during the molecular dynamic simulation (MDS) that comprise the refined structure's stability and confirms whether the simulation has been equilibrated. We calculated the RMSD values for each of the paralogs at 10 ns wherein we observed major helices and beta strands present in their tertiary structure. The RMSD scores were recorded to be 0.3, indicating a few minor changes after the POTE paralogs' refinement. The accuracy score defines the improvement of the backbone structure of the initial structure, represents that POTEE (accuracy = 0.9853), POTEF (accuracy = 0.9813), and POTEI (accuracy = 0.9851)

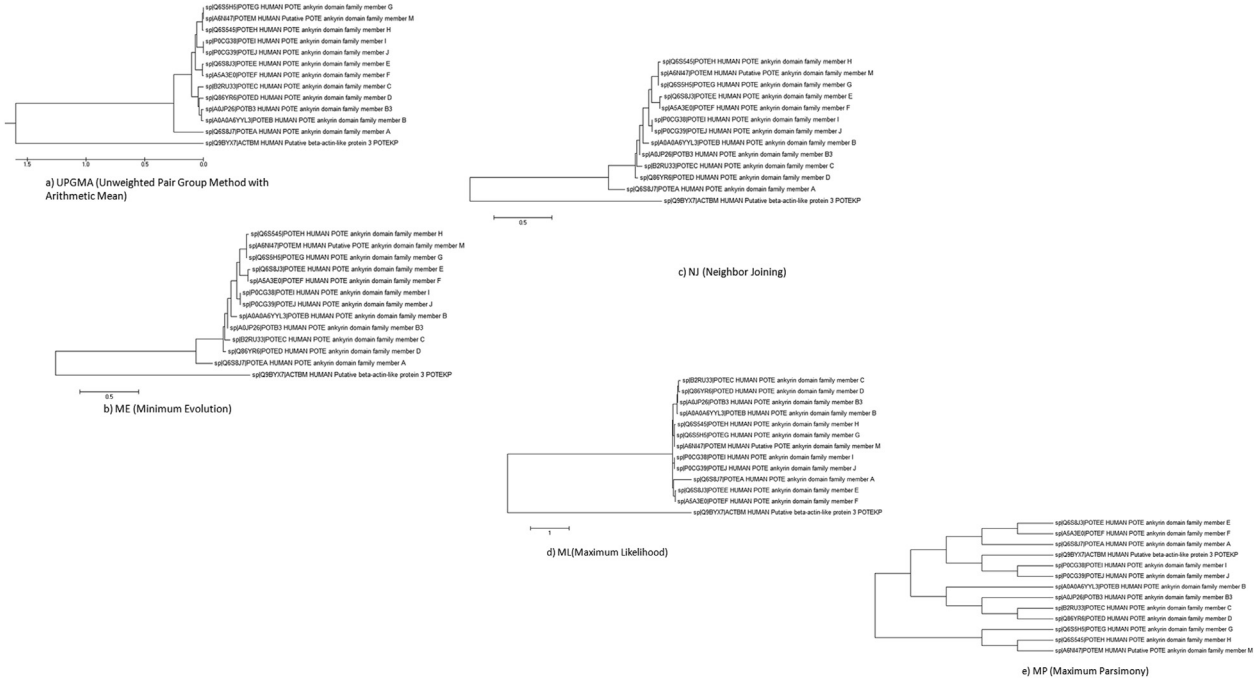
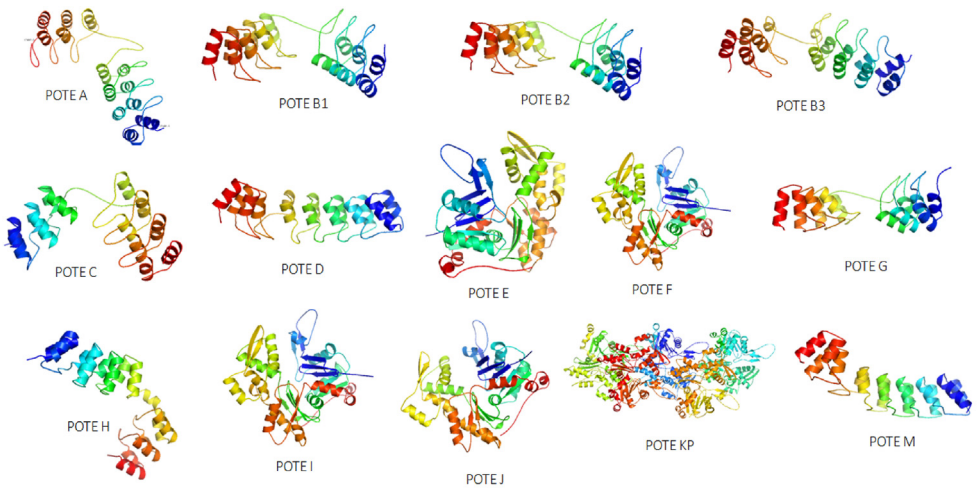
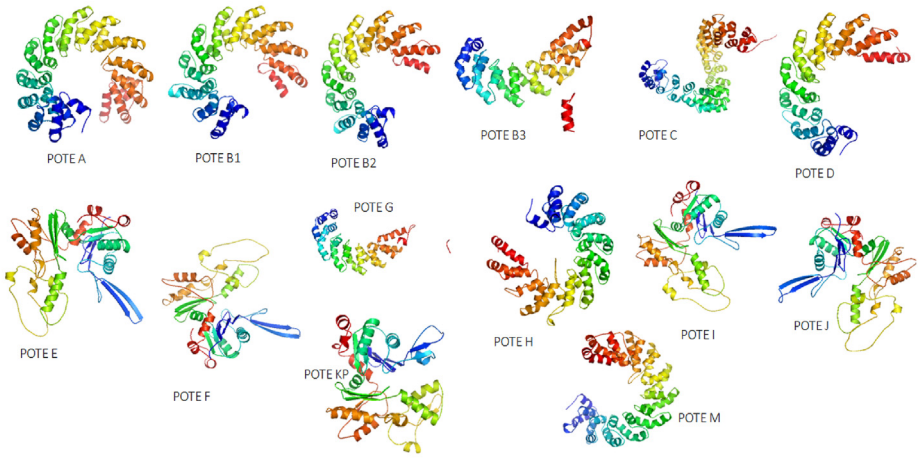


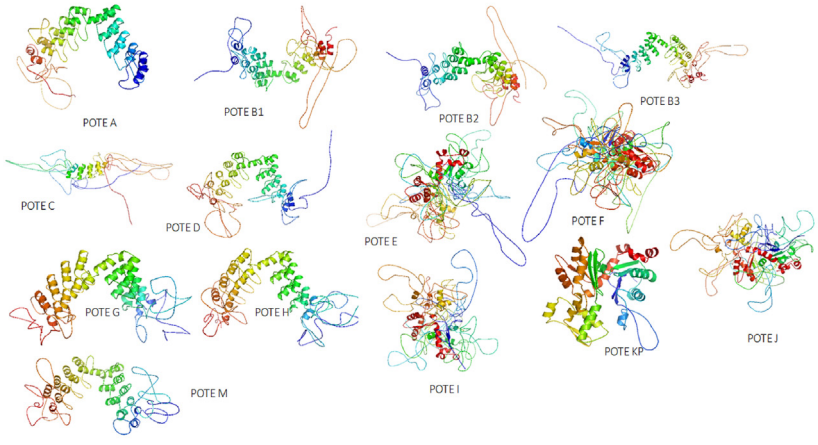
Figure 4. Phylogenetic trees of POTE paralogs using a) UPGMA, b) Minimum Evolution, c) Neighbor Joining, d) Maximum Likelihood and e) Maximum Parsimony.



(a) Swiss Model



(b) Phyre 2



(c)MODELLER

Figure 5. Tertiary models developed by using a). Swiss Model, (b) Phyre2, and c) MODELLER.

Table 3. Root Mean Square Deviation (RMSD) of both Predicted and Refined Structure.

S.No.	MODEL ID	Initial Cα RMSD	GalaxyRefine
1	POTEA	12.28	12.008
2	POTEB	4.8	4.6
3	POTEB2	4.8	4.6
4	POTEB3	9.40	9.077
5	POTEC	10.3	10.163
6	POTED	11.861	11.6
7	POTEE	0.554	0.41
8	POTEF	0.556	0.260
9	POTEG	11.811	11.6
10	POTEH	12.357	12.01
11	POTEI	0.691	0.54
12	POTEJ	11.32	11.01
13	POTKP	1.578	NA
14	POTEM	5.89	5.32

have refined better when compared to POTEJ as its accuracy score is only 0.9739. Out of these four, POTEE is much stable as it has fewer steric hindrances and clashes, and accuracy scores (refer Table 4). The MolProbity score gives the optimal physical correctness of the best refines structure. Typically, MolProbity scores for tertiary structures fall in the range of 1–2 Å (A). Our results showcase that paralogs POTEF and POTEJ (MolProbity score = 1.85) structures have good physical correctness compared to the rest (refer Table 4).

3.4. Molecular mechanics/generalized born surface area (MMGBSA) & electrostatic computation

Group III POTE paralogs, POTEE, POTEF, POTEI, and POTEJ, were subjected to MM-GBSA and various other essential electrostatic calculations that define the overall stability and energy of the thermodynamic system invariant pH environment. Electrostatics is a crucial factor in understanding how biomolecules interact with one another under various molecular environments. The Adaptive Poisson–Boltzmann Solver (APBS) software was developed to solve the equations of

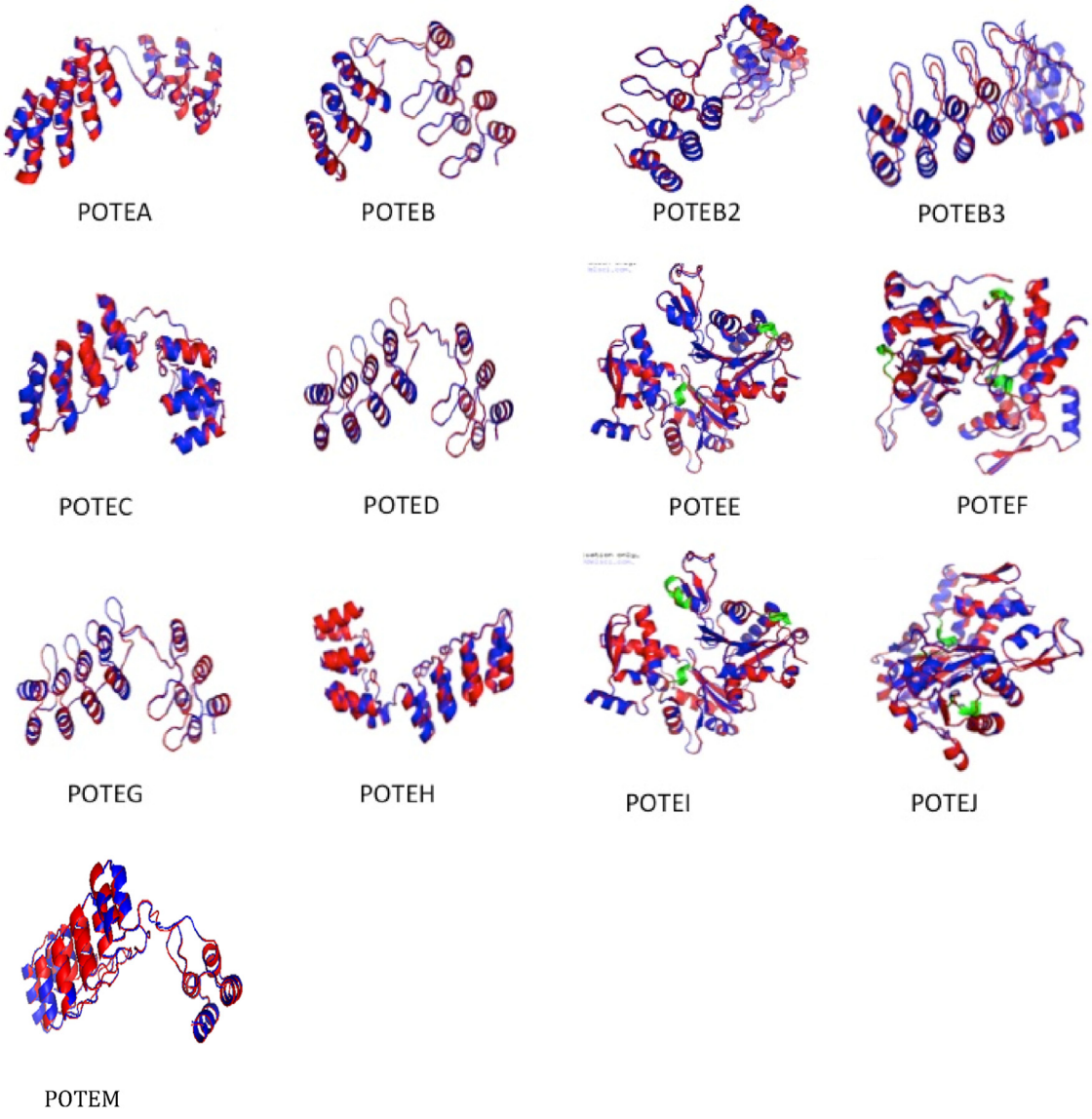


Figure 6. The structures superimposed of pre refinement (red) and post refinement (blue).

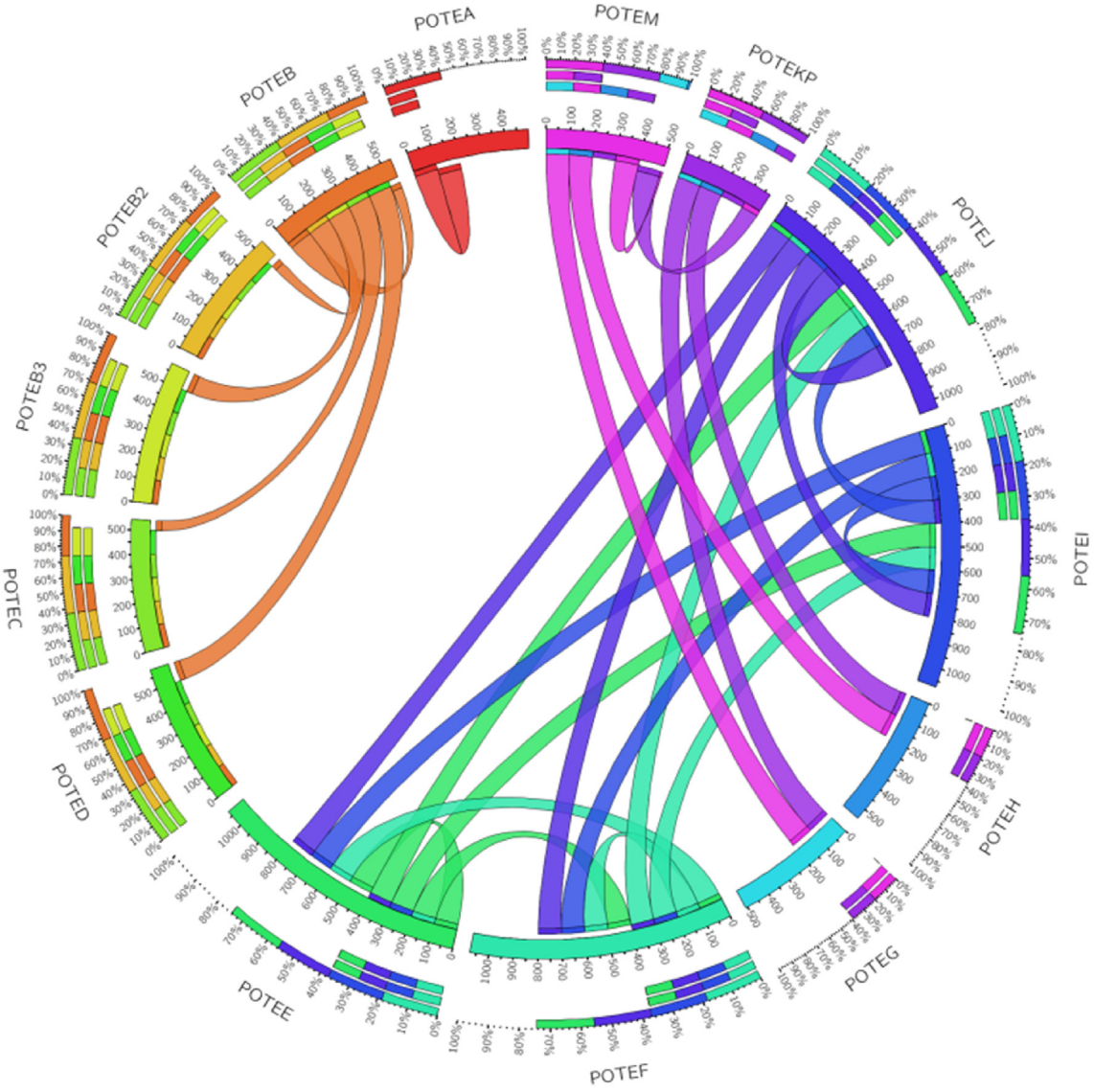


Figure 7. Circos plot: This plot is made using the length of each POTE paralog (mentioned in plot) and the relationship draw on the basis of evolutionary, structural and function. We found that POTE paralog grouped into 4 different groups: POTEA with red color and POTEB, B2, B3, C, D are in orange color and POTEE, F, I, J are shown in blue shades and POTEM, KP, G, H are shown in purple and pink color.

Table 4. Molecular dynamics simulation (50ns) detailed results of POTEE, POTEF, POTEI and POTEJ.

POTE Paralog	Accuracy Score	RMSD Score	MolProbity	Steric Hindrance Score	Ramachandran Favoured (%age)
POTEE	0.9853	0.31	1.80	12.1	97.3
POTEF	0.9813	0.31	1.85	15.3	97.0
POTEI	0.9851	0.30	1.83	13.3	96.8
POTEJ	0.9739	0.32	1.85	14.1	96.8

continuum electrostatics for large biomolecule complexes to understand the chemical, biological, and biomedical applications [6]. It was observed that POTEE had an APBS range in between-590.718 to 508.187, POTEF recorded an APBS range in between-565.402 and 503.512, POTEI, on the other hand, had an APBS range in between-575.287 and 504.991, and POTEJ ranged from −561.651 to 499.492. The molecular mechanics generalized Born surface area continuum solvation (MM-GBSA) calculations suggest that POTEE and POTEF are much more robust and electrostatically stable than POTEI and POTEJ. Figure 8 below displays the MM-GBSA calculations in the form of an APBS map as visualized in PyMol software. Table 5 summarizes the necessary electrostatics computations for the four POTE paralogs. It displays the efficiency of paralogs POTEE and POTEF with a good MMGBSE (POTEE = −15055.071461; POTEF = −14831.959332) and the overall energy values indicating a good stability as a receptor molecule (refer Table 5). The overall stability of POTEE, POTEF, POTEI and POTEJ are depicted in the form of macromolecular energy frustration plots as supplementary figure S9. The green peaks indicate minimal energy fluctuations between residues, while red peaks depict maximum energy fluctuations. The overall stability of the structure that combines the minimal and maximum fluctuations are depicted with black color.

3.5. POTE protein expression and subcellular localization

Subcellular localization is predicted for all POTE paralogs using various computational tools. It was observed that DeepLoc, MDLoc, and Hum_mPLoc software give approximately the same prediction, but WoLF PSORT predicted different localization. WoLF PSORT software predicts

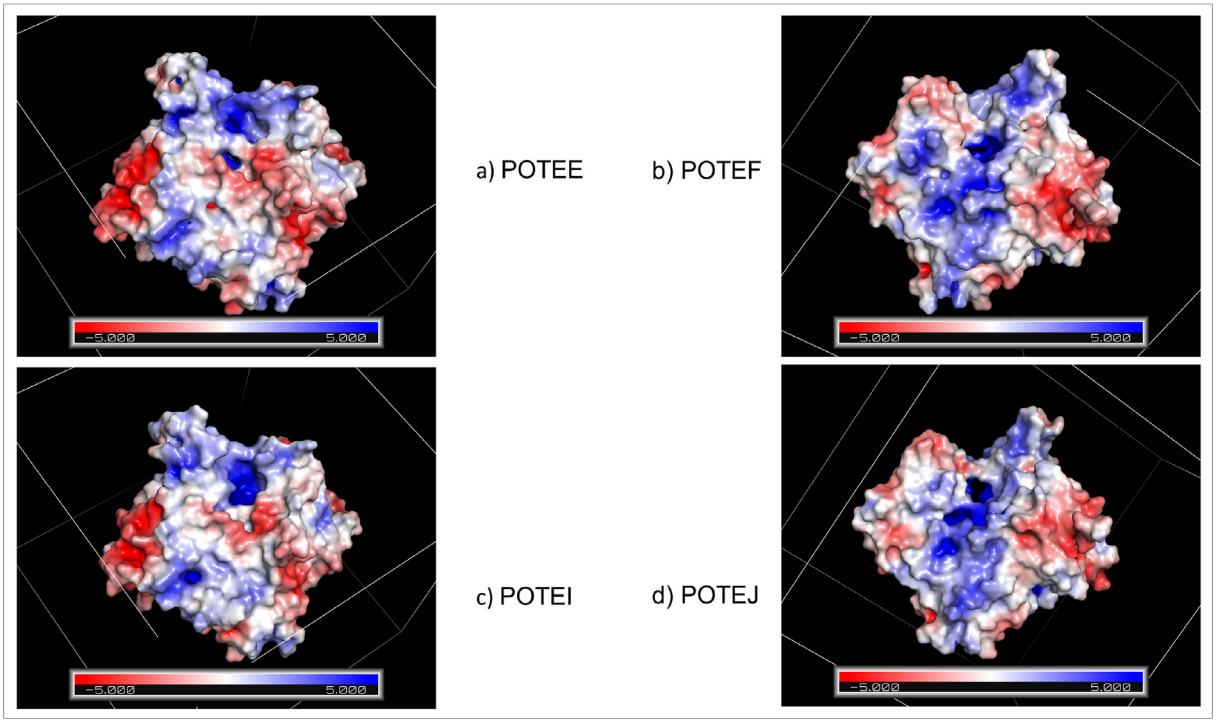


Figure 8. MM-GBSA calculations in the form of an APBS map as visualized in PyMol software. POTEE had an APBS range in between-590.718 to 508.187, POTEF recorded an APBS range in between-565.402 to 503.512, POTEI, on the other hand, had an APBS range in between-575.287 to 504.991, and POTEJ ranged from -561.651 to 499.492.

Table 5. MMGBSA and other essential electrostatics calculated for POTEE, POTEF, POTEI and POTEJ.

POTE paralog	System Surface Area (A^2)	Generalized Born Self Energy (GBSE) (kJ)	Coulomb Energy (kJ)	Electrostatic Solvation Energy (kJ/mol)	Total Energy (kJ/mol)	APBS Potential
POTEE	3.39090e + 04	−15055.071461	−97816.405699	−3693.103361	−99474.970252	−590.718 to 508.187
POTEF	3.32948e + 04	−14831.959332	−97843.869365	−3425.638721	−99271.818874	565.402 to 503.512
POTEI	3.35568e + 04	−14920.890240	−97415.190820	−3566.078112	−98967.859069	−575.287 to 504.991
POTEJ	3.30252e + 04	−14695.354653	−97299.238959	−3328.072766	−98645.797324	−561.651 to 499.492

subcellular localization such as the nucleus, mitochondria, cytosol, plasma membrane, extracellular, peroxisome, Golgi, etc., with competitive accuracy but also provides detailed information relevant to protein localization. The data was analyzed, and the result depicted that POTE paralogs are majorly localized in Cytoplasm and Cell membrane, as shown in Supplemental Tables 2, 3, and 4.

3.6. Interaction network

Using the STRING database, we fined tuned the network association by setting k-means clustering to the POTE paralogs with a high confidence score, i.e., 0.70. We observed that group III paralogs – POTEE, POTEF, POTEI and POTEJ are highly connected and form one cluster. While POTEC and POTED form a separate cluster. POTEB2, POTEB3, POTEM, POTEG, POTE H come out as outliers as they don't fit in any of the clusters. POTEA and POTEKP were not significantly bound to any of the network associators with a very low confidence score, therefore, they didn't show up in k-means clustering. Figure 9 displays the network formed by k-means clustering in STRING webserver.

3.7. Function prediction

ProFunc was employed for predicting POTE paralog's main functionality. The results are astounding as we discern that POTE paralogs have succumbed themselves according to their nearest neighbor with an attributable course of evolutionary time. Similar structures of paralogs have shared functions. This analysis's crux is simple, ProFunc assay suggests that all the POTE paralogs are mainly involved in binding like protein, nucleotide, and ATP binding. Moreover, they are also engaged in enzyme regulation, catalytic activity, transporter, and transferase activity, respectively, as shown in Table 6.

3.8. Molecular docking on POTE proteins against different anticancer drugs

We performed a docking study on 14 POTE paralogs with a list of selected compounds from ovarian, prostate, and testicular anticancer drugs. Doxorubicin showed the highest binding energy of −15.945 kcal/mol for the active site of POTEF among 15 ovarian cancer bioactive

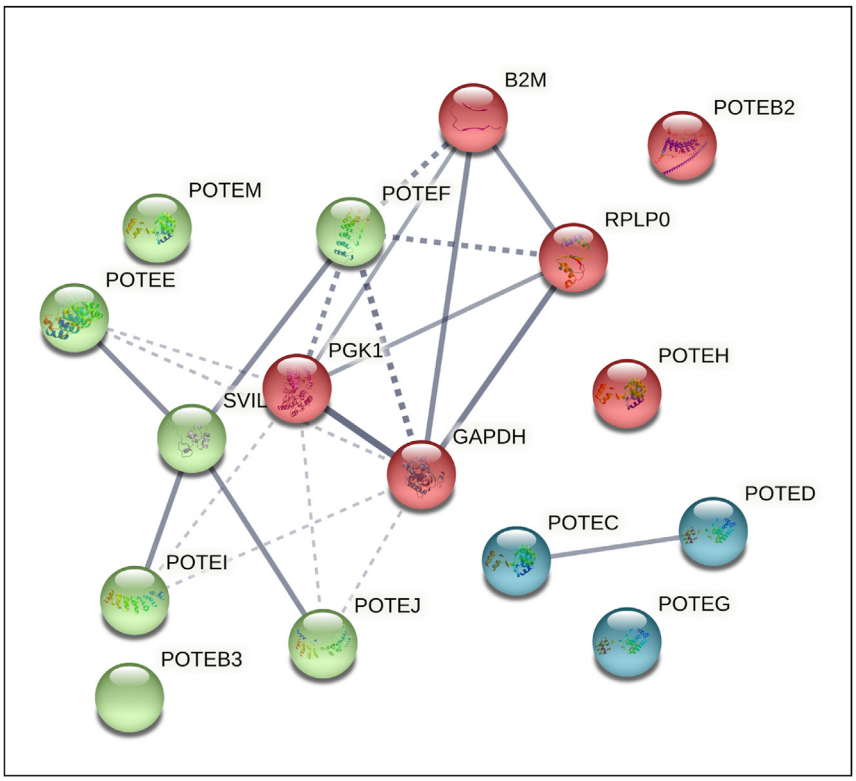


Figure 9. POTE network displayed in three different clusters using k-means clustering algorithm in STRING.

compounds, which were taken for the docking study. Notably, one of the quaternary amines of the ligand sites is well in the POTEF receptors active site where the metal interaction with Asp887 is observed, and the hydroxyl establishes an H-bond with Glu907. In contrast, the arene part of the ligand, which has hydroxyl and acidic groups, sits on the middle of the POTEF receptor, where the cation−π interaction with Arg762 is observed Figure 10a. Mitoxantrone showed the highest binding energy of -15.0988 kcal/mol for the active site of POTEKP. In this protein-ligand interaction, one of the quaternary ammonium ions interacts with the amino acid residue of AspD187 of the POTEKP receptor through acceptor interaction. The hydroxyl group, which is 2 carbon atoms away from the amine group, establishes an H-bond with two amino acid residues Lysb284 and Lysd191, and another hydroxyl group is located on the arene group of the ligand concatenate through H-bond with AsnB280 residue as shown in Figure 10b. Whereas, the docking results on the testicular anticancer compound, etoposide, showed the highest binding energy of −19.831 kcal/mol for the active site of POTEM among all the three classes of anticancer compounds. The maximum binding interactions between amino acid residues with ligand atoms as Lys205 establish H-bond with the phosphate group's oxygen atom, and the oxygen atom of the anisole group appended benzene rings. Arg207 shows multiple H-bond interactions with two oxygen atoms of a phosphate group and anisole's oxygen atom. Simultaneously, the Gln215 interacts through H-bond with two oxygen atoms, one from phosphate and another from the anisole group. The oxygen atom, which is a part of the six-member ring, interacts with POTEM receptors active site where the H-bond interaction is observed with Arg303, and another oxygen atom of the five-member ring establishes an H-bond with Asn269 residue as shown in Figure 10c. Thus, the Etoposide ligand has shown high binding energy with POTEM protein compared with all other anticancer compounds with POTE proteins. Therefore, highest binding affinity were observed in complexes namely – POTEB_ligandID5 (−14.6565 kcal/mol); POTEB2_ ligandID5 (−15.1491 kcal/mol); POTEC_ ligandID5 (−14.5191 kcal/mol); POTEC_ ligandID4 (−15.0707 kcal/mol); POTEE_ligandID15 (−14.222 kcal/mol); POTEE_ligandID3 (−14.2351); POTEF_ligandID5 (−15.9455); POTEG_ligandID3 (−15.8356 kcal/mol); POTEI_ligandID4 (−16.0921 kcal/mol); POTEJ_ligandID4 (−15.6959 kcal/mol); POTEJ_-ligandID15 (−14.1459 kcal/mol); POTEK_ligandID15 (−15.0036 kcal/mol); POTEK_ligandID1 (−15.5574 kcal/mol) and POTEM_ligandID3 (−19.8317 kcal/mol).

A list of bioactive drug compounds from ovarian, prostate, and testicular cancers with the highest binding energy averse to all POTE proteins are shown in supplementary tables 5 and 6.

3. Discussion

In the post-genomic era, protein sequences, which deposited, have increased at an exponential rate. The latest UniProtKB shows ~99,261,416 protein sequence entries in the repository. However, protein structures present in the protein data bank (PDB) (RCSB) are ~125,799. The time, labor, and cost involved in the protein structure determinations are augmenting the sequence-structure gap. Structures are essential for function annotation and pursuing structure-based drug discovery [5, 7, 8]. Computational methods based on ab initio and homology methods can accelerate structure generation and can be used to partly alleviate the dilemma [9]. Hence, we targeted an exhaustive exploratory *in-silico* analysis to know POTE paralogs' evolutionary status and behavior.

POTE paralogs have had a fate of dilatory rate of evolution, hinting at their high conservation of amino acids. Our study reveals that POTE paralogs are very analogous to one another yet very different. They have gone through an adaptive divergence. Most POTE paralogs diverge from many species, not just primates. POTE proteins are orthologous to many different species such as *D. melanogaster, C. elegans*, fission yeast, *D. discoidum*, yeast, and zebrafish, hinting that the POTE sequences might have undergone an adaptive divergence in the attributable course of evolution. Furthermore, this divergence is not in all the POTE paralogs, but only POTEE, POTEF, POTEI, POTEJ, POTEKP, and POTEA, wherein we get to know that POTEA and POTEKP are outliers and do not fall in any of the common clusters which have been formed in our clustering

Table 6. Functional enrichment of POTE paralogs.

S. No	Biochemical Function	Biological Process
POTEA	Binding, Protein Binding, Enzyme Binding, Nucleic acid Binding& Catalytic activity.	Cellular, Metabolic, Regulation of Biological & quality process.
POTEB	Binding, Protein Binding, Metal ion Binding and catalytic activity	Cellular, Regulation of Biological quality & its process
POTE-B2	Binding, Protein Binding, Metal ion Binding and catalytic activity	Cellular, Regulation of Biological quality & its process
POTEB3	Binding, Protein Binding, Metal ion Binding and catalytic activity	Cellular, Biological, Regulation of Biological quality & its process
POTEC	Binding, Protein Binding and catalytic activity	Cellular, Biological, Regulation of Biological quality & its process
POTED	Binding, Protein Binding, Enzyme Binding, Nucleic acid Binding, Transferase and catalytic activity	Cellular, Metabolic, Regulation of Biological & quality process.
POTEE	Binding, Nucleotide Binding, Protein Binding, ATP Binding	Cellular Process, Cellular Component Organization, Organelle Organization, Biological Regulation.
POTEF	Binding, Nucleotide Binding, Protein Binding, ATP Binding	Cellular Process, Cellular Component Organization, Organelle Organization, Biological Regulation & Metabolic Process.
POTEG	Binding, protein binding enzyme binding, metal ion binding	biological regulation, cellular, regulation of biological quality & biological process
POTEH	Binding, Nucleotide Binding, Protein Binding, ATP Binding & Transporter activity	biological regulation, cellular, regulation of biological quality & biological process
POTEI	Binding, Nucleotide Binding, Protein Binding, ATP Binding	Cellular & Metabolic Process, Cellular Component Organization, Organelle Organization,
POTEJ	Binding, Protein Binding, Nucleotide Binding, ATP Binding	Cellular Process, Cellular Component Organization, Metabolic Process, Organelle Organization
POTEKP	Protein, Nucleotide and ATP Binding	Cellular component Organization. Organelle Organization
POTEM	Binding, Protein Binding, Transporter, catalytic & ion channel Activity	Cellular, Biological, Regulation of Biological quality & its process

algorithms. Phylogenetic analysis of the POTE paralogs indicates that POTEE, POTEF, POTEI, POTEJ, POTEA, and POTEKP are clubbed together, hinting at their common divergence from other species during evolution, while POTEB, POTEB2, POTEB3, POTEC, POTED, POTEG, POTEH, POTEM comes from only primates. Since our work highlights the similarities within the three groups of POTE family members, other phylogenetic studies have worked on categorizing the gene structure of the POTE genes among human, great ape, chimpanzee, gorilla and orangutan, macaque, marmoset genomes [49]. Here the researchers suggest that there are four groups in POTE gene family and not three using phylogenetic analyses. They concluded that the POTE gene family is bifurcated as follows: *new world monkey* (NWM) genomes with two copies in marmoset and four paralogs in macaque while in *old world monkey* (OWM) genomes – six in gorilla, seven in both orangutan and chimpanzee, and 14 in human [49]. It is noteworthy to mention here, this study also highlights the "*divergence of POTE family members*" validating our hypothesis of adaptive divergence in POTE paralogs.

The evolutionary analysis, which was done using MEGA 5.02, discerns that all the POTE paralogs' aggregate evolutionary divergence is 0.66, again referring to its dilatory evolution rate. Most of the residues in POTE protein sequences are conserved, referring to a low rate of alterations/mutations. Common amino acids present in all the POTE paralogs are mainly: Serine, Arginine, Lysine, and Valine, respectively, which may vary in their composition in each paralog. Structure developments showcase each paralog's varying structures, but every paralog has a helical and supercoiled structure as its primary protein skeleton. Evaluations of every model were generated to discern that the Swiss Model structures are better than Phyre2 and MODELLER. The refinement provided a successful 0.5 Å of resolution. The graphical study of the radius of gyration, accessible surface area, hydrogen bond, and the energy potential shows every POTE paralog structure's compactness. The explicit MD simulation, MM-GBSA, and APBS electrostatics suggest that only POTEE and POTEF targets have absolute high affinities with minimal energetic exploitation. Both of these paralogs – POTEE and POTEF are highly stable as a system with equilibrated thermodynamic properties and Generalized Born self-energy (GBSE). The MM-GBSA estimates also indicate that POTEE and POTEF have a higher entropic contribution than POTEI and POTEJ.

We also ascertained that the tertiary structures of POTEE, POTEF, POTEI, POTEJ, and POTEKP (pseudogene) have beta strands along with helices and loops. However, POTEA, POTEB, POTEB2, POTEB3, POTEC, POTED, POTEG, POTEH, and POTEM had only helices. This ardently provides evidence to our evolutionary analysis that some of the POTE paralogs have adaptively diverged and differed in sequence, structure, and functioning compared to their counterparts. Adaptive divergence leads to new forms resulting from the adaptation to a new environmental condition [10, 11, 12]. Thus, new forms of POTE paralogs have emerged with time from ancestral origins, giving rise to more robust gene and protein structures. We would also like to correlate our study with [5, 13, 14], where the authors suggest that group 3 POTE members are not cancer-testis antigens functioning only in normal tissues. This differing nature of group 3 POTE paralogs is for sure to ponder upon. Henceforth, we propose that this differing nature of group 3 POTE members POTEE, POTEF, POTEI, POTEJ, and POTEKP (non-functional) is because of their adaptive divergence.

Further, we have identified the sub-cellular localization of POTE paralogs, as predicted by WoLF PSORT, Hum_mPLoc, DeepLoc, discerns that POTEE, POTEF, POTEI, POTEJ, POTEKP, and POTEM are localized in the cytoplasm, while others are located in the extracellular region. Our study on interaction network analysis discerns that POTEE, POTEF, POTEI, and POTEJ have a high confidence score of 0.70 with only one interactor protein, yet again corroborating our phylogenetic studies, which attributable to the advent of subdivisions in already existing group 3.

Function prediction and its analysis suggest that all the POTE paralogs, which are predicted and refined, are mainly involved in biological functions such as protein, nucleotide, and ATP binding. Moreover, they are also engaged in enzyme regulation, catalytic activity, transporter, and transferase activity. POTE family encodes six or seven ankyrin repeats in the middle of the molecule, spectrin-like structure at the carboxy-terminus, and three cysteine-rich repeats at the amino-terminus. Ankyrin repeats have been identified in numerous functionally diverse proteins and are involved in protein-protein interactions in many functional pathways within the cell [15, 16]. In accord, POTE expression in the testis is primarily confined to spermatids, which are caspase-3 positive [17]. Ingenuity pathway analysis of both microarray and RNA-Seq data also suggests that POTEs might be associated with cancer pathogenesis and connected to functional networks involving cancer-relevant pathways such as cellular growth and proliferation.

Further, the molecular docking results show a higher affinity of POTE members for different anticancer drugs. Etoposide and doxorubicin showed the highest binding energy of −19.831 kcal/mol and −15.0988 kcal/mol, respectively, for the active site of POTEM and POTEF proteins. However, the results are preliminary and surely need experimental confirmation, which will be conducted soon via molecular biology studies. However, contemplating all these structural aspects and Glide score, the POTE family might be the first choice that could be exploited to design as an anticancer therapy in the future.

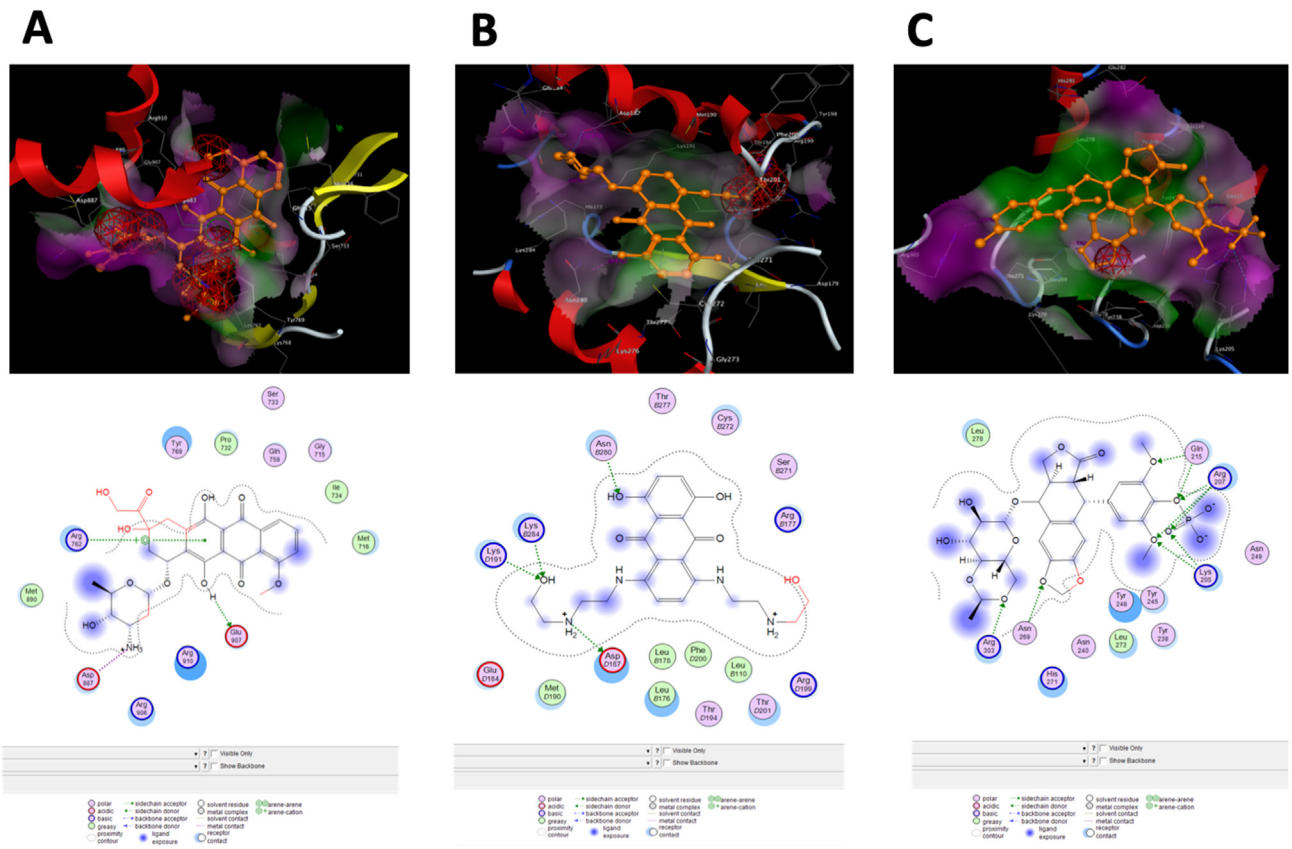


Figure 10. Molecular Docking Analysis of POTE with Anticancer Agents (A). Binding conformation of Doxorubicin in an active pocket of POTEF protein. (B). Binding conformation of Mitoxantrone in an active pocket of POTEK protein. (C). Binding conformation of Etoposide in an active pocket of POTEM protein.

4. Conclusion

Our study reveals that POTE paralogs have gone through an adaptive divergence. Most of the POTE paralogs diverge from many species, and not just primates. The evolutionary analysis discerns that all the POTE paralogs' total evolutionary divergence is 0.66, again referring to its dilatory evolution rate. POTE paralogs are mainly involved in binding like protein, nucleotide, and ATP binding. Moreover, they are also engaged in enzyme regulation, catalytic activity, transporter, and transferase activity. Of interest, cytoplasmic/nuclear/membrane-localized POTEs are more evident in advanced cancer cases. Taken together, further investigations are required to better understand the roles of POTEs members in cancer progression and malignancy and to determine how dysregulation of these genes/proteins may alter responses to therapeutic. Using virtual screening methodology, this study may also help understand and design the novel anticancer compounds concerning POTE proteins.

Declarations

Author contribution statement

Sahar Qazi: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Bimal Prasad Jit: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Abhishek Das: Performed the experiments; Wrote the paper.

Muthukumarasamy Karthikeyan: Performed the experiments.

Amit Saxena: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

M.D. Ray and Angel Rajan Singh: Analyzed and interpreted the data.

Khalid Raza and B. Jayaram: Contributed reagents, materials, analysis tools or data.

Funding statement

Dr Ashok Sharma was supported by Department of Biotechnology [DBT/BT/RLF/Re-entry/51/2013].

Dr Ashok Sharma was supported by Department of Science and Technology [DST/ECR/2016/001740].

Dr Ashok Sharma was supported by Indian Council of Medical Research [R.11012/01/2018-HR].

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e10476.

Acknowledgements

I thank Prof. Adam R Karpf, Professor, Fred & Pamela Buffett Cancer Center, University of Nebraska Medical Center, the USA, for his constructive scientific input. SQ is supported by the DST-INSPIRE provided by the Department of Science and Technology, Govt. of India.

Thanks to Neetu for her contribution in project work. SQ & BJ equally contributed.

References

[1] T.K. Bera, et al., Five POTE paralogs and their splice variants are expressed in human prostate and encode proteins of different lengths, Gene 337 (2004).
[2] T.K. Bera, et al., POTE paralogs are induced and differentially expressed in many cancers, Cancer Res. 66 (2006).
[3] Y. Lee, et al., Evolution and expression of chimeric POTE-actin genes in the human genome, Proc. Natl. Acad. Sci. 103 (2006).
[4] T. Ise, et al., Expression of POTE protein in human testis detected by novel monoclonal antibodies, Biochem. Biophys. Res. Commun. 365 (2008).
[5] P.J. Gane, P.M. Dean, Recent advances in structure-based rational drug design, Curr. Opin. Struct. Biol. 10 (2000).
[6] E. Jurrus, et al., Improvements to the APBS biomolecular solvation software suite, Protein Sci. 27 (2018).
[7] D. Baker, Protein structure prediction and structural genomics, Science (80–.) 294 (2001).
[8] S. Lutz, Beyond directed evolution—semi-rational protein engineering and design, Curr. Opin. Biotechnol. 21 (2010).
[9] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction: progress and new directions in round XI, Proteins Struct. Funct. Bioinforma. 84 (2016).
[10] J.M. Good, C.A. Hayden, T.J. Wheeler, Adaptive protein evolution and regulatory divergence in Drosophila, Mol. Biol. Evol. 23 (2006).
[11] J.-Y. Wu, et al., Adaptive evolution of cry genes in Bacillus thuringiensis: implications for their specificity determination, Genomics. Proteomics Bioinf. 5 (2007).
[12] J.A.M. Raeymaekers, et al., Adaptive and non-adaptive divergence in a common landscape, Nat. Commun. 8 (2017).
[13] D. Szklarczyk, et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res. 47 (2019).
[14] D. Szklarczyk, et al., STRING v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (2015).
[15] J. Li, A. Mahajan, M.-D. Tsai, Ankyrin repeat: a unique motif mediating Protein−Protein interactions †, Biochemistry 45 (2006).
[16] D.A. Voronin, E.V. Kiseleva, Functional role of proteins containing ankyrin repeats, Cell Tissue Biol. 2 (2008).
[17] T.K. Bera, D.A. Walker, R.J. Sherin's, I. Pastan, POTE protein, a cancer-testis antigen, is highly expressed in spermatids in human testis and is associated with apoptotic cells, Biochem. Biophys. Res. Commun. 417 (2012).
[18] B. Yates, et al., Genenames.org: the HGNC and VGNC resources in 2017, Nucleic Acids Res. 45 (2017).
[19] S. Altschul, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997).
[20] F. Madeira, et al., The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic Acids Res. 47 (2019).
[21] S. Kumar, M. Nei, J. Dudley, K. Tamura, MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, Brief. Bioinf. 9 (2008).
[22] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994).
[23] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004).
[24] F. Sievers, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011).
[25] K. Tamura, et al., MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Mol. Biol. Evol. 28 (2011).
[26] Sokal Robert Reuven & Michener Charles Duncan. A_Statistical_Method_for_Evaluating_Syst. Volume 38, Part 2, Issue 22 of University of Kansas Science Bulletin, University of Kansas, 1958.
[27] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. (1987).
[28] A. Rzhetsky, M. Nei, Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference, J. Mol. Evol. 35 (1992).
[29] M. Arenas, A. Sánchez-Cobos, U. Bastolla, Maximum-Likelihood phylogenetic inference with selection on protein folding stability, Mol. Biol. Evol. 32 (2015).
[30] J. Bertl, G. Ewing, C. Kosiol, A. Futschik, Approximate maximum likelihood estimation for population genetic inference, Stat. Appl. Genet. Mol. Biol. 16 (2017).
[31] D.W.A. Buchan, D.T. Jones, The PSIPRED protein analysis workbench: 20 years on, Nucleic Acids Res 47 (2019).
[32] A. Šali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, J. Mol. Biol. 234 (1993).
[33] L.A. Kelley, S. Mezulis, C.M. Yates, M.N. Wass, M.J.E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis, Nat. Protoc. 10 (2015).
[34] A. Waterhouse, et al., SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (2018).
[35] S.C. Lovell, et al., Structure validation by Cα geometry: φ,ψ and Cβ deviation, Proteins Struct. Funct. Bioinf. 50 (2003).
[36] B. Wallner, A. Elofsson, Can correct protein models be identified? Protein Sci. 12 (2003).
[37] L. Heo, H. Park, C. GalaxyRefine Seok, Protein structure refinement driven by side-chain repacking, Nucleic Acids Res. 41 (2013).
[38] V. Hornak, et al., Comparison of multiple Amber force fields and development of improved protein backbone parameters, Proteins Struct. Funct. Bioinf. 65 (2006).
[39] J.V. Ribeiro, et al., QwikMD — integrative molecular dynamics toolkit for novices and experts, Sci. Rep. 6 (2016).
[40] W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics, J. Mol. Graph. 14 (1996).
[41] S. Genheden, U. Ryde, The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities, Expert Opin. Drug Discov. 10 (2015).
[42] I. Walsh, et al., Blues server: electrostatic properties of wild-type and mutated protein structures, Bioinformatics 28 (2012).
[43] J. Sun, et al., A systematic analysis of FDA-approved anticancer drugs, BMC Syst. Biol. 11 (2017).
[44] J.J. Almagro Armenteros, C.K. Sønderby, S.K. Sønderby, H. Nielsen, O. Winther, DeepLoc: prediction of protein subcellular localization using deep learning, Bioinformatics 33 (2017).
[45] R. Simha, S. Briesemeister, O. Kohlbacher, H. Shatkay, Protein (multi-)location prediction: utilizing interdependencies via a generative model, Bioinformatics 31 (2015).
[46] R.A. Laskowski, The ProFunc Function Prediction Server, 2017.
[47] D. Szklarczyk, et al., The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, Nucleic Acids Res. 45 (2017).
[48] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation, Nucleic Acids Res. 44 (2016).
[49] F.A.M. Maggiolini, L. Mercuri, F. Antonacci, et al., Evolutionary dynamics of the POTE gene family in human and nonhuman primates, Genes 11 (2) (2020) 213.
[50] A.D. MacKerell Jr., D. Bashford, et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B 102 (1998) 3586–3616.