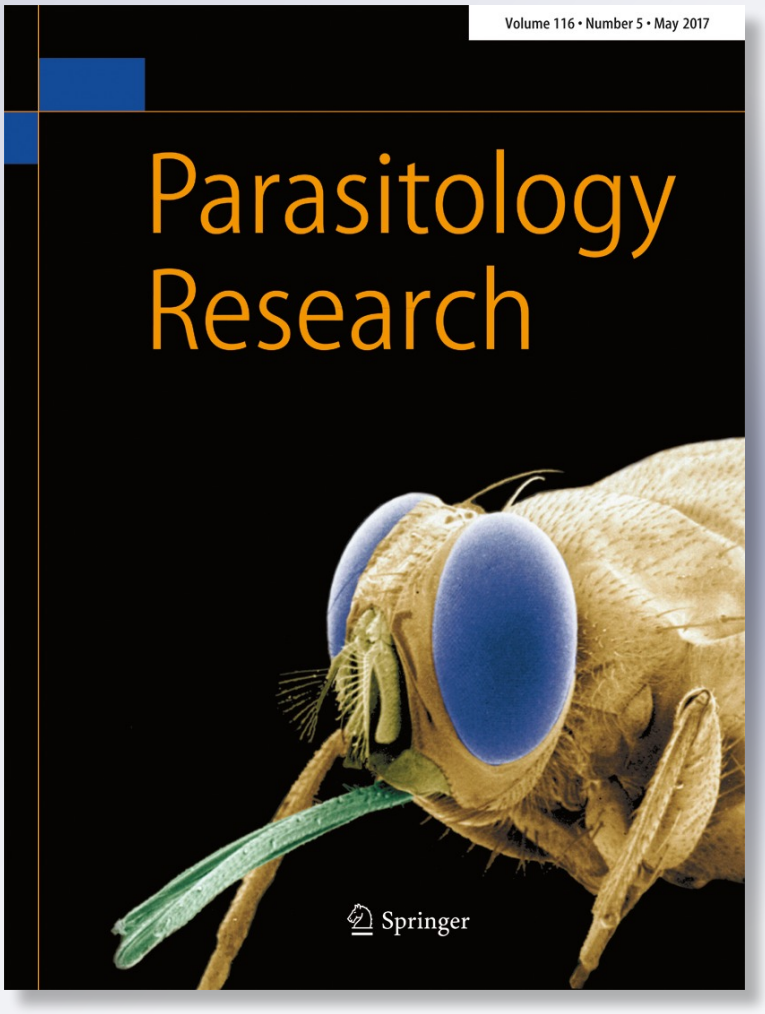
In silico identification and validation of a novel hypothetical protein in Cryptosporidium hominis *and virtual screening of inhibitors as therapeutics*

Arpit Kumar Shrivastava, Subrat Kumar, Priyadarshi Soumyaranjan Sahu & Rajani Kanta Mahapatra

Parasitology Research
Founded as Zeitschrift für Parasitenkunde

ISSN 0932-0113
Volume 116
Number 5

Parasitol Res (2017) 116:1533-1544
DOI 10.1007/s00436-017-5430-1



Springer

Your article is protected by copyright and all rights are held exclusively by Springer-Verlag Berlin Heidelberg. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

Author's personal copy

Parasitol Res (2017) 116:1533–1544
DOI 10.1007/s00436-017-5430-1

ORIGINAL PAPER

In silico identification and validation of a novel hypothetical protein in *Cryptosporidium hominis* and virtual screening of inhibitors as therapeutics

Arpit Kumar Shrivastava[1] · **Subrat Kumar**[1] · **Priyadarshi Soumyaranjan Sahu**[1,2] · **Rajani Kanta Mahapatra**[1]

Received: 17 February 2017 / Accepted: 21 March 2017 / Published online: 7 April 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract Computational approaches to predict structure/function and other biological characteristics of proteins are becoming more common in comparison to the traditional methods in drug discovery. Cryptosporidiosis is a major zoonotic diarrheal disease particularly in children, which is caused primarily by *Cryptosporidium hominis* and *Cryptosporidium parvum*. Currently, there are no vaccines for cryptosporidiosis and recommended drugs are ineffective. With the availability of complete genome sequence of *C. hominis*, new targets have been recognized for the development of effective and better drugs and/or vaccines. We identified a unique hypothetical protein (TU502HP) in the *C. hominis* genome from the CryptoDB database. A three-dimensional model of the protein was generated using the Iterative Threading ASSEmbly Refinement server through an iterative threading method. Functional annotation and phylogenetic study of TU502HP protein revealed similarity with human transportin 3. The model is further subjected to a virtual screening study form the ZINC database compound library using the Dock Blaster server. A docking study through AutoDock software reported *N*-(3-chlorobenzyl)ethane-1,2-diamine as the best inhibitor in terms of docking score and binding energy. The reliability of the binding mode of the inhibitor is confirmed by a complex molecular dynamics simulation study using GROMACS software for 10 ns in the water environment. Furthermore, antigenic determinants of the protein were determined with the help of DNASTAR software. Our findings report a great potential in order to provide insights in the development of new drug(s) or vaccine(s) for treatment and prophylaxis of cryptosporidiosis among humans and animals.

Keywords *C. hominis* · Hypothetical protein · Molecular docking · Molecular dynamics simulation

Introduction

Cryptosporidiosis is a gastrointestinal illness caused by *Cryptosporidium* species. *Cryptosporidium* is an apicomplexan protozoan parasite belongs to the class Conoidasida. The pathogen is responsible for diarrheal disease in both immunocompetent and immunodeficient humans. Cryptosporidiosis has been emerged as the second major cause of diarrheal disease and death in infants (Snelling et al. 2007; Striepen 2013). The most common route of *Cryptosporidium* transmission is through consumption of contaminated water (Efstratiou et al. 2017; Mahon and Doyle 2017). In humans, *Cryptosporidium hominis* and *Cryptosporidium parvum* are two epidemiological important diarrheal pathogens, which cause cryptosporidiosis (Snelling et al. 2007; Berahmat et al. 2017). In children, *Cryptosporidium* infection is also associated with malnutrition and impaired physical fitness (Guerrant et al. 1999; Sharling et al. 2010). Actual quantification of zoonotic and anthroponotic transmission is very difficult in the environment. However, history suggests major epidemic outbreaks occurred due to contaminated water and food; therefore, it poses

Electronic supplementary material The online version of this article (doi:10.1007/s00436-017-5430-1) contains supplementary material, which is available to authorized users.

✉ Priyadarshi Soumyaranjan Sahu
priyadarshi_sahu@yahoo.com

✉ Rajani Kanta Mahapatra
rmahapatra@kiitbiotech.ac.in

[1] School of Biotechnology, KIIT University, Bhubaneswar, Odisha, India

[2] Divisions of Pathology, School of Medicine, International Medical University, 57000 Kuala Lumpur, Malaysia

Author's personal copy

significant challenges for controlling transmission in both developing and developed nations (Yoder and Beach 2010).

Treatment strategies are extremely limited; no vaccine is available for this parasite, and nitazoxanide (NTZ) is the only FDA-approved drug for cryptosporidiosis that shows moderate efficacy in immunocompromised individuals (Abubakar et al. 2007). There is an immediate need for developing new therapeutics for cryptosporidiosis. Continuous long-term maintenance of *Cryptosporidium* in cell culture and frequent genetic modifications in *Cryptosporidium* genome are extremely difficult. Therefore, alternative search strategies like computational approach to exploit the genomics data provide a potential solution for identification and characterization of novel therapeutic candidates.

The sequencing of *Cryptosporidium* genomes provides better knowledge of microbial biology, pathogenicity, evolution, and virulence for this parasite. Pertinent to this context, the CryptoDB (Heiges et al. 2006) database enables the identification of genes based on text, sequence similarity, and motif queries. However, the CryptoDB database contains ~50% of proteins in both *C. hominis* TU502 and *C. parvum* IOWA as hypothetical or unnamed. The abundance of hypothetical proteins in the parasite genome makes their study a difficult task; therefore, more rational approach is required for screening and identification of novel drug/vaccine targets.

In this study, a unique hypothetical protein of *C. hominis* TU502 was identified and characterized, utilizing in silico methodologies with the objective of identification of a novel target for new drug(s) or vaccine candidate for control and prevention of cryptosporidiosis in humans and animals.

Materials and methods

Sequence retrieval

The CryptoDB (Heiges et al. 2006) database is a community bioinformatics resource database of the *Cryptosporidium* species genome, which provides a comprehensive Web interface for mining and visualization of data. *C. hominis* and *C. parvum* are the two major causes of human cryptosporidiosis; therefore, we targeted those sequences, which were present in both *C. hominis* and *C. parvum*. The search reported a total list of 105 proteins. The next step of our approach was based on screening of unique proteins in the *C. hominis* and *C. parvum* genome. For that, the amino acid sequences of all 105 proteins were checked for their uniqueness through BLAST analysis to target as a potential therapeutic candidate. TU502 hypothetical protein which was conserved in *C. hominis* and *C. parvum* and was unique in nature was selected for generation of three-dimensional structure and physiological and biochemical characterization.

Three-dimensional model prediction and functional annotation

General physiological characters like molecular weight, isoelectric point (pI), amino acid composition, molecular extinction coefficient, half-life, instability index, and aliphatic index of the hypothetical protein were predicted using an ExPASy ProtParam tool (Gasteiger et al. 2005). Hydrophobicity and hydrophilicity were analyzed using ProtScale (Gasteiger et al. 2005). We used the I-TASSER (Iterative Threading ASSEmbly Refinement; Yang and Zhang 2015) online server to determine secondary structure elements of uncharacterized hypothetical protein.

I-TASSER is a bioinformatics server for generating three-dimensional protein structure. Out of the five models generated, the best model was selected based on the highest C score and Template Modeling (TM) value. The quality of the best-predicted model was further validated through Ramachandran plot analysis using PROCHECK (Laskowski et al. 1993), ERRAT (Colovos and Yeates 1993), and VERIFY3D (Eisenberg et al. 1997). The ProSA server (Wiederstein and Sippl 2007) was used for calculating native conformation, and PyMOL software (Schrödinger, LLC 2010) was used for target template superimposition. The DALI server (Plewczynski et al. 2004) was used for structural and functional similarity search of TU502HP with other proteins of database. Subcellular localization of the hypothetical protein was predicted by using CELLO v.2.5 (Yu et al. 2006), which is a multi-class support vector classification system. CELLO algorithm determines the localization for a protein if it has a confidence score of 1 for a particular localization. However, a minimum confidence score of 2 was retained for a stringent screening method in this study.

Sequences, alignment, and construction of phylogenetic tree

Phylogenetic analysis is an important step in understanding the ancestral relationship of a set of sequences. Genome analysis of *Cryptosporidium* reported extremely streamlined metabolic pathways and a lack of many cellular structures. Previous studies suggested that *Cryptosporidium* possess an extensive array of transporter protein which enables the import of essential nutrient from the host (Xu et al. 2004; Rider and Zhu 2010; Pain et al. 2005).

Therefore, a phylogenetic tree of TU502HP protein was constructed with other 14 related sequences based on functional annotation study results of TU502HP. Molecular Evolutionary Genetics Analysis (MEGA) v6.0 (Tamura et al. 2013) was used for the creation of phylogenetic tree. The evolutionary history was inferred using the neighbor-joining (NJ) method (Saitou and Nei 1987). NJ is a method for

Author's personal copy

constructing phylogenies from a set of distances between each pair of sequences by successive clustering.

Virtual screening and ADMET analysis

The Dock Blaster server (Irwin and Shoichet 2005) was used for virtual screening of inhibitor molecules. The server performs structure-based virtual screening against the ZINC chemical database (Irwin and Shoichet 2005) by using the docking program DOCK 3.5.54 (Lorber and Shoichet 1998). The modeled hypothetical protein TU502 was subjected to a structure-based virtual screening study of potential lead compound from the ZINC database. Dock Blaster identified 199 lead compounds docked in the active site of target protein. These compounds were filtered and arranged orderly based on docking score and Lipinski's rule of five (Lipinski et al. 2001).

Molecular docking

Molecular docking studies were carried out to understand the binding mode of lead compounds with TU502HP using AutoDock v4.2.6 software (Morris et al. 2009). The grid parameters were arranged at $x = 0.547$, $y = -0.098$, and $z = -1.245$ dimensions. Prior to docking, all the water and solvent atoms of the protein were removed, and the polar hydrogen atoms were added. The protein was kept rigid while the ligand was allowed to rotate and explore more flexible binding pockets. AutoDock ranks the poses using binding free energy (ΔG). The interaction plot representing the H bond interactions of the protein-ligand complex was generated using PyMOL v 1.3 (Schrödinger, Inc.).

Molecular dynamics simulation and energy minimization

A molecular dynamics (MD) simulation study of the best identified compound was performed to confirm the binding mode determined from a molecular docking study using the GROMACS v5.0 software package (Berendsen et al. 1995). GROMOS96 43a1 force field (Scott et al. 1999) was chosen for the simulation. The topology file of the ligand was prepared using the PRODRG server (Schüttelkopf and Van Aalten 2004). One Na+ ion was added as a counter ion to neutralize the charges in the system. The protein was solvated with SPC 60,590 water molecules and centered in a cubic box with dimensions 40 nm × 40 nm × 40 nm. This system was subjected to energy minimization for 884 steps by using the steepest descent algorithm. The MD simulation was performed with periodic boundary conditions to ensure that the atoms stay inside the simulation box. Berendsen temperature coupling was turned on with a temperature set to 300 K. The leapfrog algorithm was used for integrating Newton's equation in MD simulation. Finally, linear constraint solver (LINCS) algorithm (Hess et al. 1997) was used, and equilibrated systems were subjected to MD simulation for 10 ns at 300 K.

Trajectory analysis

The simulation trajectories were analyzed using tools from the GROMACS package and were viewed using PyMOL software. The root-mean-square deviation (RMSD) was calculated using g_rms. The radius of gyration (Rg) of the protein was calculated using g_gyrate. All GROMACS analysis plots were generated using Xmgrace plotting software (Vaught 1996).

Antigenic epitope prediction

Antigenic determinates of the TU502HP were predicted with the help of DNASTAR (Burland 1999) software. Algorithms used in this analysis were hydrophobicity plot flexible regions, antigenic index, and surface accessibility. B cell and T cell linear epitopes were predicted by using BCPred (Saha and Raghava 2004) and CTLPred (Bhasin and Raghava 2004; Saha and Raghava 2004) software. The BCPred server predicts epitopes by using flexibility, hydrophilicity, polarity, and surface properties combined at a threshold of 2.38. CTLPred is a direct method for prediction of CTL epitopes. The methods followed were based on elegant machine learning techniques, viz., artificial neural network (ANN) and support vector machine (SVM). In this method, information or patterns of T cell epitopes instead of MHC binders were used for the development of methods. The cutoff scores were 0.51 and 0.36 for ANN and SVM, respectively, above which peptides were designated to be antigenic.

Source of *Cryptosporidium* DNA and target gene amplification

C. hominis genomic DNA was obtained from a standard laboratory of Christian Medical College, Vellore (India), on request, and *C. parvum* oocysts (Product No. P102C) were purchased from Waterborne, Inc., USA. *C. parvum* genomic DNA was extracted by DNeasy Blood and Tissue extraction from Qiagen (Cat. No. 69504). The TU502HP encoding gene was amplified by standard PCR using a newly designed pair of specific primers (Table 1). The PCR was performed in a total of 25 μl reaction by adding 5 μl (1–10 ng) of the genomic DNA to 20 μl of the reaction mixture containing 10× PCR buffer (200 mM Tris-HCl; 500 mM KCl), 2 mM of $MgCl_2$, 2 mM dNTP, 20 pm of each primer, and 1 unit of Taq DNA polymerase (Invitrogen) and water. The PCR cycling conditions were as follows: initial denaturation at 94 °C for 5 min followed by 35 cycles of 94 °C for 45 s, 61.9 °C for 30 s, and 72 °C for 2.45 min followed by 1 cycle of 10 min at 72 °C.

Author's personal copy

Table 1 Detailed descriptions of primers used in the study

Target sequence	Primer name	Primer sequence (5′ to 3′)	Tm (°C)	Size (bp)
TU502HP	Crypto_X_1_Fw	ATGGAAAGTTATTCGTCTT CTTTAAG	61.9	2220
	Crypto_X_2_Rw	TCAAGTTTTTAGACCATTG TTAATAAATGC		

The PCR product was subjected to 1% agarose gel electrophoresis to confirm amplification.

TA cloning and sequence identification of the TU502HP encoding gene

The TU502HP gene was amplified from genomic DNA using conventional Taq polymerase (Invitrogen) that generates PCR products with 3′ A overhangs. PCR products were then purified using a Promega Gel cleanup kit (Promega, USA) and cloned into a TA cloning vector using a TOPO TA cloning kit (Invitrogen, USA), according to the manufacturer's protocol. The ligation mix was then incubated at 22 °C for 30 min. Two microliters of the ligation mix is then transformed into chemically competent DH5α cells. Transformation was carried out in a 42 °C water bath for 60 s. Then, 400 μl of the SOC medium was added to the mixture and incubated at 37 °C for 1 h at 150 rpm. This was then plated on ampicillin or kanamycin selection plates. Next day, colonies on the plate were subjected to colony PCR using vector-specific primers. Positive colonies showing specific bands were inoculated for plasmid isolation. Next day, plasmids were isolated and again confirmed by PCR. These plasmid clones were sequenced on both strands using a Big Cycle Termination v 3.1 cycle sequencing kit (Applied Biosystems, Foster City, USA). BLAST search was carried out to confirm the sequence similarity with the target sequence.

Results and discussion

Sequence retrieval and physiological and biochemical characteristics of TU502HP

The selected *C. hominis* hypothetical protein sequence (TU502HP) reported 91% sequence identity with *C. parvum* hypothetical protein (CryptoDB: cgd2_2550). Length of the TU502HP was found to be 739 amino acids long and is encoded by a 2220-nucleotide-long ORF sequence. The molecular weight and theoretical isoelectric point was found to be 84 kDa and 6.75, respectively. The pI of the TU502HP is near to the neutral pH. pI is the pH at which the amino acids of protein tolerate no net charge; moreover, proteins become more stable and compact at isoelectric pH. The protein contains Ser (11.1%), Asn (8.1%), and Tyr (5.7%) as major polar amino acid residues and Leu (11.9%), Ile (11.4%), and Phe (6.2%) as major non-polar amino acid residues. The total number of negatively charged residues (Asp + Glu) and positively charged residues (Arg + Lys) is 78 and 77, respectively, which give a stable charge to the protein. The estimated half-life of this protein was found to be 30 h in mammalian reticulocytes and >20 h in yeast. The aliphatic index was found high as 106, and the instability index was found to be 43. The molar extinction coefficient of this protein at 280 nM in water was found to be 79,303 m/cm. A two-dimensional structural analysis of this protein revealed that the majority of protein contains alpha helices. Hydrophobic effect is an important factor for protein folding and architectural stability (Malleshappa Gowder et al. 2014). The grand average of hydrophobicity (GRAVY) of this protein was found to be 0.079.

Tertiary structure prediction and validation of TU502HP

In silico methods are extremely useful in preliminary structure prediction and functional annotation of proteins. Presently, various protein structure prediction methods are available as

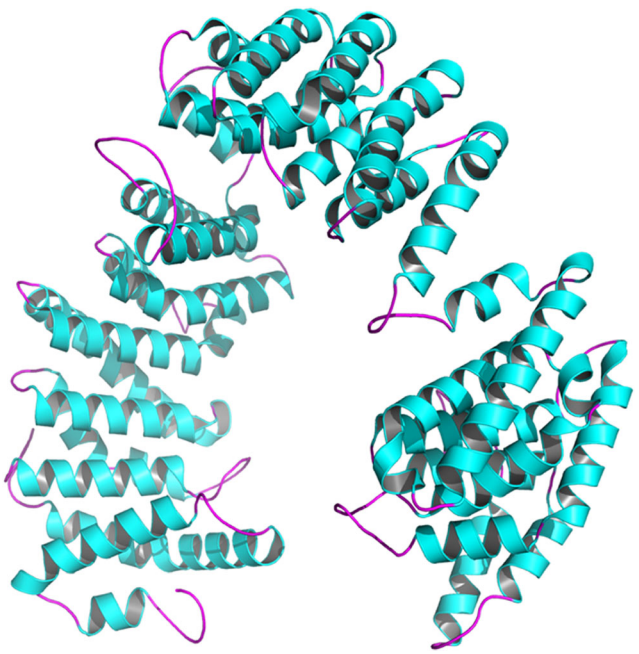
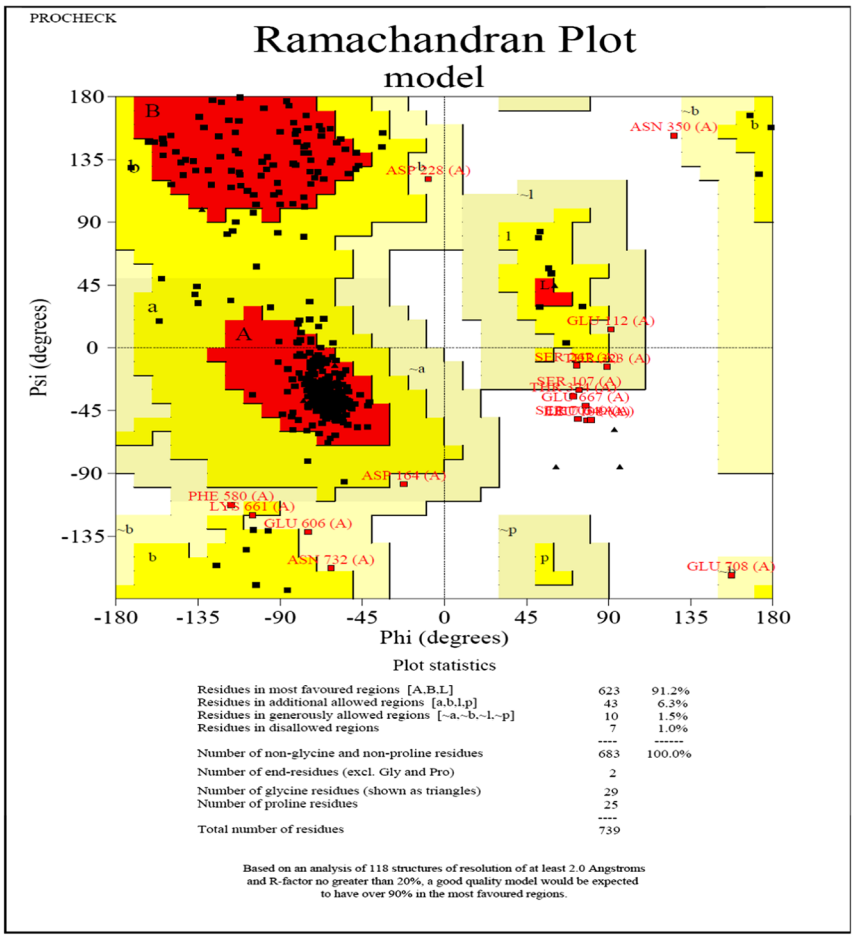
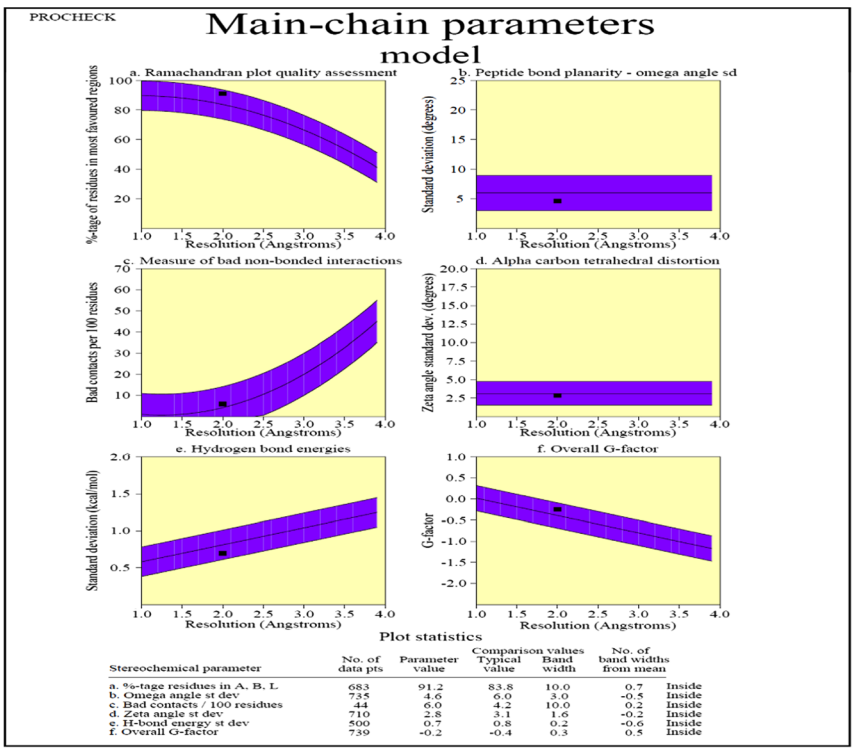


Fig. 1 Three-dimensional model of *C. hominis* hypothetical protein in cartoon representation showing a secondary structure (α-helices in *blue color* and connecting loops in *magenta color*) of hypothetical protein

Author's personal copy



a



b

Fig. 2 Model validation plot of best generated hypothetical protein model. **a** Ramachandran plot. *Dark black band* represent amino acids, and *red zones A*, *B*, and *L* represent the most favored regions. **b** Main chain parameter. All six graphs show how the structure compares with well-refined structures at the similar resolution. The *dark band in each graph* represents the results from the well-refined structures; the *central line* is a least-squares fit to the mean trend as a function of resolution, while the width of the band on either side of it corresponds to a variation of 1 standard deviation about the mean

Author's personal copy

Web servers (Watson et al. 2005). A well-established function prediction approach relies on the analysis of the three-dimensional structure of protein (Adams et al. 2007; Baker 2006). The protein sequence was subjected to BLASTp (Altschul et al. 1997) search against the Protein Data Bank (PDB) database (Bernstein et al. 1977). However, BLASTp against PDB reported no suitable template for three-dimensional structure predictions through a homology modeling study. Therefore, a three-dimensional structure of this hypothetical protein was predicted using the I-TASSER online server. I-TASSER is one of the most successful and freely available tools for protein tertiary structure prediction. This online server detects the structure of templates from PDB by a threading technique and generates five top models for the given protein with a C score ranging from −5 to 2. On the basis of these parameters, we selected the first model (Fig. 1) with the highest C score of −1.59 and TM 0.43 ± 0.14. The RMSD of the structure with the template was computed to be 2.47 Å and with an IDEN score of 0.077. The quality of the model was validated through the Ramachandran plot by the PROCHECK server. This plot revealed 623 residues (91.2%) were in the most favored core region, 43 residues (6.3%) in the

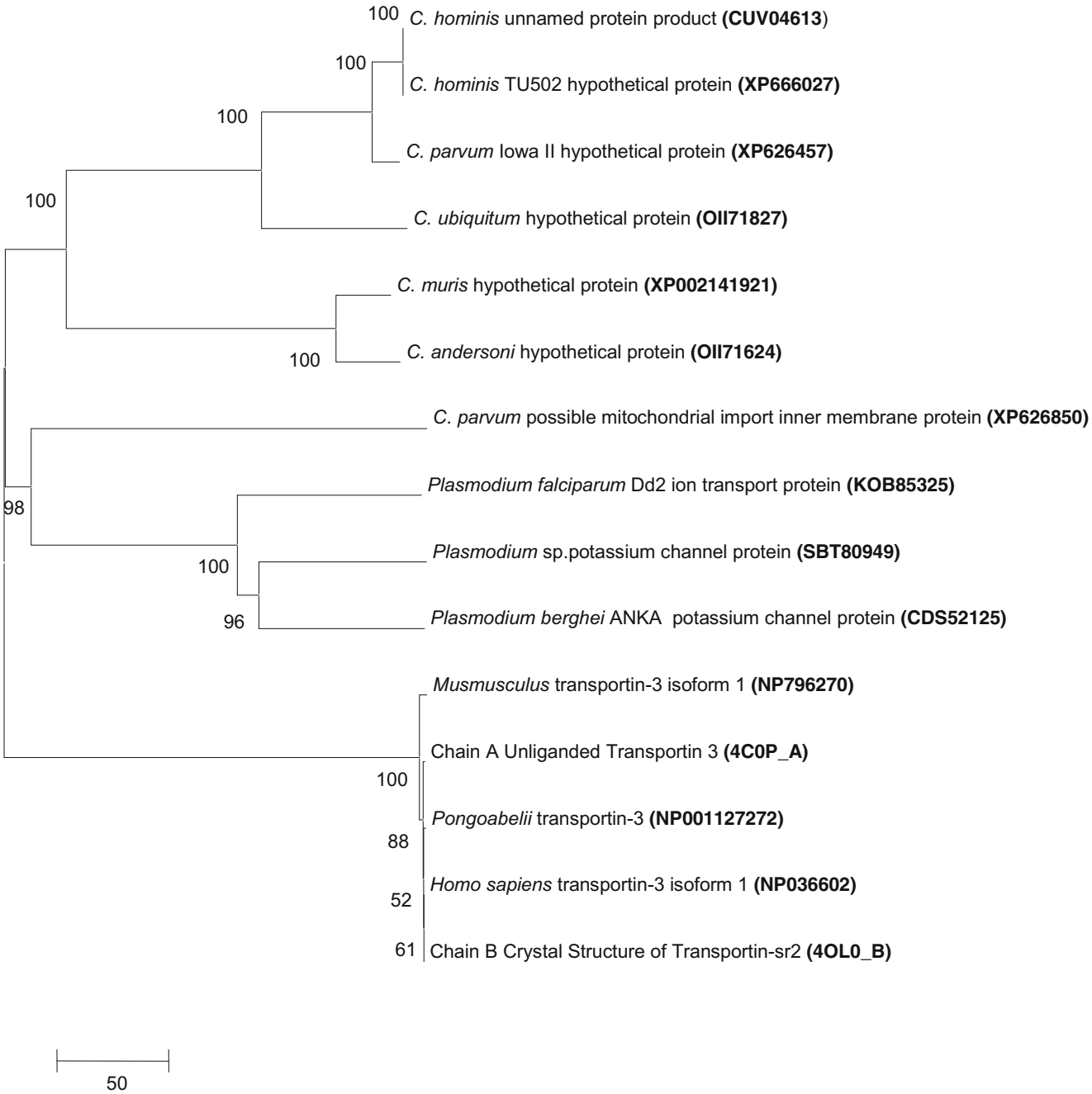


Fig. 3 Phylogenetic tree constructed for *C. hominis* hypothetical protein (TU502HP)

Author's personal copy

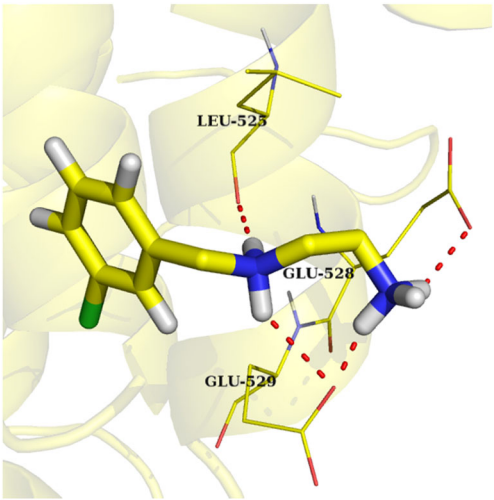


Fig. 4 Binding pose of *C. hominis* hypothetical protein before MD simulation; hydrogen bond interactions are shown in *red dashes*

additional allowed region, 10 residues (1.5%) in the generously allowed region, and 7 residues (1%) in the disallowed region (Fig. 2a). The main chain parameter plots for the model were generated using PROCHECK shown in the (Fig. 2b). The plots are for the Ramachandran plot quality, peptide bond planarity, bad non-bonded interactions, C-alpha tetrahedral distortion, main chain hydrogen bond energy, and the overall G factor. Overall analysis reported the I-TASSER-generated model of hypothetical protein TU502HP is of good quality. The compatibility of three-dimensional model was checked by the VERIFY3D program which reported 36.27% of the residues had an average three-dimensional-one-dimensional score of ≥0.2. The ERRAT server gives the overall quality factor of the protein model as 48.974. We studied the quality of model using the ProSA Web server, and it revealed a Z score value of −6.65 which is in the range of native conformations. So, it can be concluded that several quality assessment and validation parameters computed in this study indicate the reliability of the model for a structure-based drug design approach.

The functional annotation of TU502HP through the DALI server suggested structural similarity with human transportin 3. Transportin 3 is a nuclear import protein that plays an important role in the import of splicing factor and HIV-1 replication (Maertens et al. 2014). In silico prediction of subcellular localization provides a rapid and inexpensive way to find out information regarding protein function. The knowledge of the subcellular localization of a protein plays a very significant role in target identification during the drug discovery process. CELLO v.2.5 was used to determine the subcellular localization. CELLO algorithm determines the localization of protein if it has a confidence score of 1 for a particular location. The software reported a confidence score of 2.064 for the extracellular location. Apart from CELLO, ESLPred (Bhasin and Raghava 2004), BaCelLo (Pierleoni et al. 2006), and EuLoc (Chang et al. 2013) servers, which

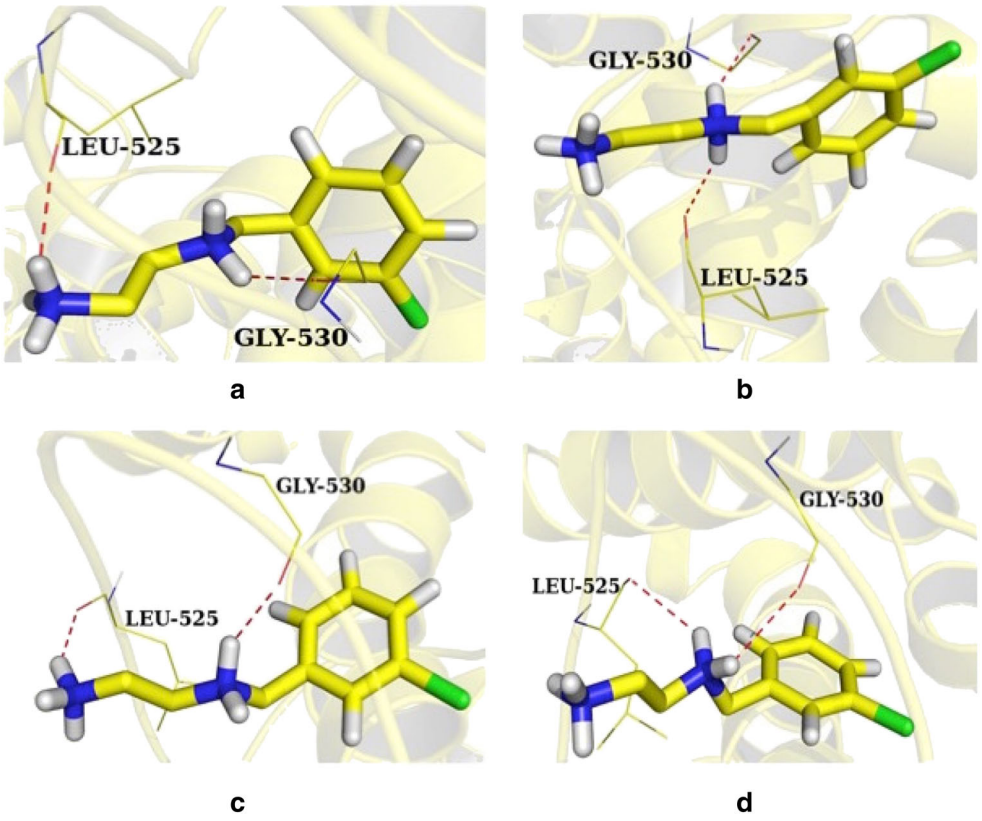


Fig. 5 Structural conformation of protein and compound complex at different time intervals during MD simulation. **a** At 2.5 ns. **b** At 5 ns. **c** At 7.5 ns. **d** At 10 ns

Author's personal copy

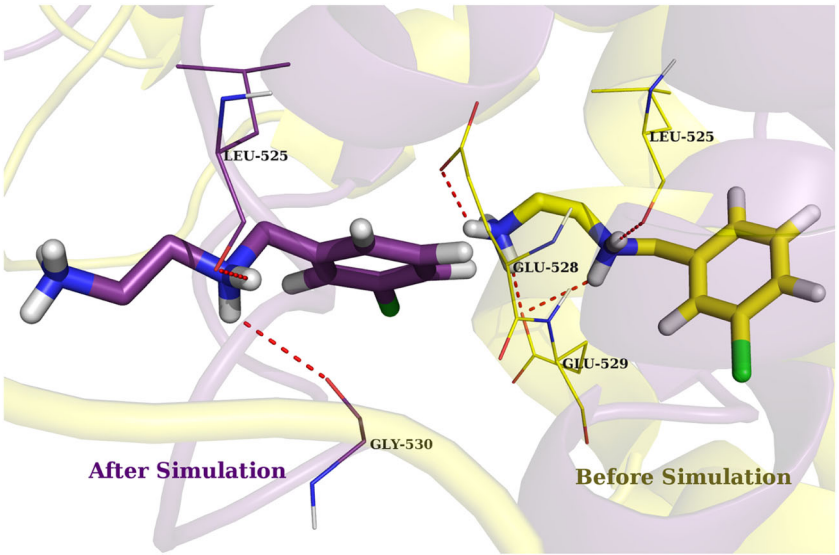


Fig. 6 Binding pose of the *C. hominis* hypothetical protein-ligand complex before and after MD simulation. Hydrogen bond interactions are shown in *red dashes*

specifically used for subcellular localization prediction in eukaryotic protein, were also used to predict the localization. The consensus report revealed that the protein is having a location in the cytoplasm.

Phylogenetic analysis

Phylogenetic analysis of TU502HP using MEGA inferred that this protein is most closely related to other *Cryptosporidium*

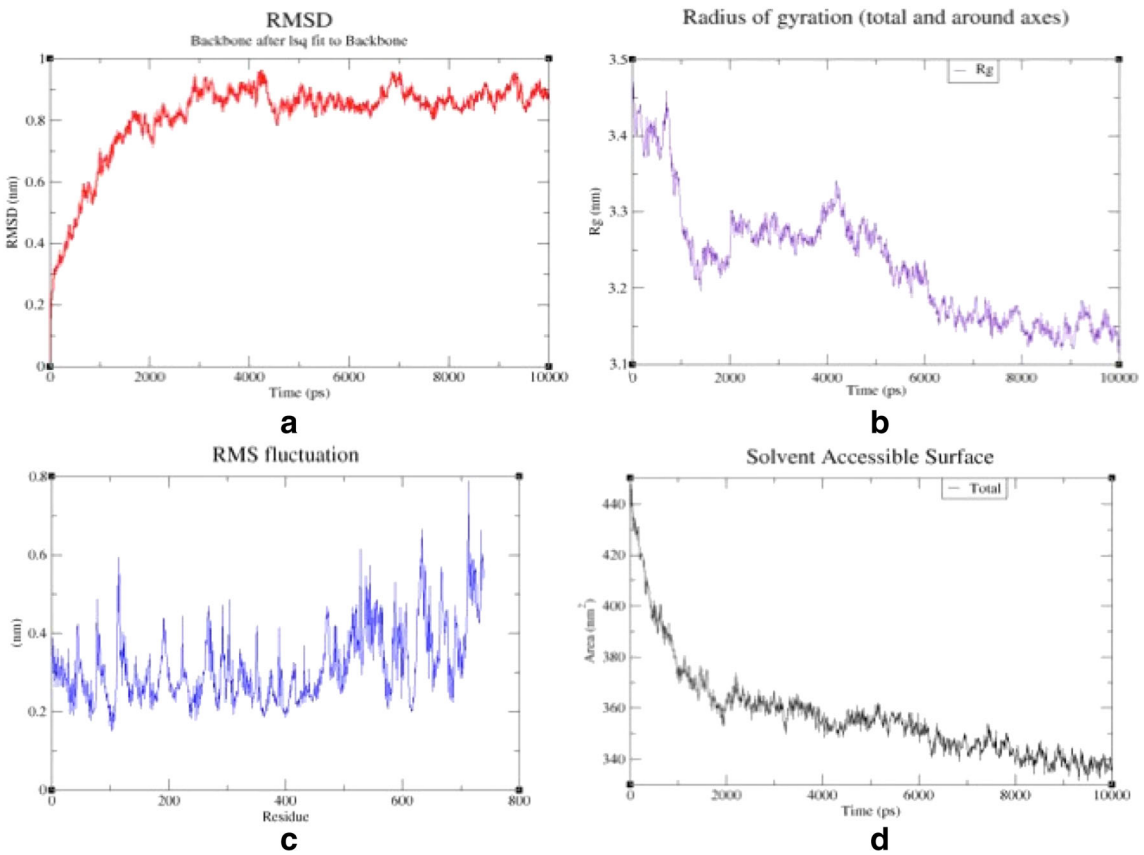


Fig. 7 **a** Backbone root-mean-square deviation (*RMSD*) plot of the protein-ligand complex shown in *red color* at 10 ns. **b** Radius of gyration plot. **c** Root-mean-square fluctuation (RMSF) of protein residues shown in *blue color*. **d** Solvent accessible surface area plot

Author's personal copy

Table 2 Potential B cell epitopes of TU502HP predicted by BCPred

BCPred peptide rank	Start position	Sequence	Score
1	340	FRYKLTIPTTNNKSIVKTKD	0.997
2	479	NLLCWDSNNGTSKHCSYIQV	0.973
3	620	NLPHFIKHNNNPQESGFVFT	0.921
4	14	TKIRMVGKYSGFGGETPVSF	0.893
5	593	PKIKDIIFKPDFSECPIPEI	0.86
6	410	RPGDLPSDSSPQIHLMKRVP	0.812
7	192	TSILDIPVYESDISQATESY	0.805
8	705	KRTEVESAKIVSNQIRSIIP	0.771

hypothetical proteins. *C. hominis* (CUV04613) was most closely related to TU502HP followed by *C. parvum* (XP626457), *Cryptosporidium ubiquitum* (OII71827), *Cryptosporidium muris* (XP002141921), and *Cryptosporidium andersoni* (OII71624) hypothetical proteins, respectively. Among the characterized proteins, it showed maximum similarity with human transportin 3. The nodes in the phylogenetic tree indicate separate evolutionary paths, and the lengths of the branches give an estimate of how distantly these sequences are related. Results clearly predicted among characterized proteins, this protein is closely related with human transportin 3 in comparison to other *Cryptosporidium* transporter proteins (Fig. 3). Genome analysis of *Cryptosporidium* revealed that this parasite possesses hundreds of transporter and transporter-like proteins and most of these transporter proteins help in the scavenging nutrients from the host. In our study, functional annotation and phylogenetic analysis reported that this TU502HP could be an important transporter protein or transporter-like protein.

Structure-based virtual screening

Result from an in silico virtual screening study provides a list of top 199 lead compounds from the ZINC database, which interact with TU502HP three-dimensional structure (Supplementary Table 1). The 199 lead compounds were checked for their physiochemical properties based on Lipinski's rule of five.

Molecular docking study

A docking simulation study was performed to explore the binding interaction of inhibitor to *C. hominis* hypothetical protein (TU502HP). Top 5 compounds from virtual screening analysis were selected for a molecular docking study (Supplementary Table 2). From the overall molecular docking study, *N*-(3-chlorobenzyl)ethane-1,2-diamine was found to have the highest binding energy (ΔG) interaction of −7.9 kcal/mol. Molecular docking studies revealed the binding modes of the ligand with TU502HP, giving insights of the key amino acid residues that are involved during the binding conformation. The ligand (Zinc_32919754) has the highest binding energy, and it forms four hydrogen bonds with residues Leu 525, Ile 526, Glu 528, and Glu 529 (Fig. 4). The next compound (Zinc_79438559) has the second highest binding energy of −7.79 kcal/mol, and it forms three hydrogen bonds with active site residues.

Molecular dynamics simulation

After the docking analysis, the binding stability of the best inhibitor and modeled protein was further evaluated using the 10-ns molecular dynamics (MD) simulation in

Table 3 Potential T cell epitopes of TU502HP predicted by CTLPred

CTLPred peptide rank	Start position	Sequence	Score
1	228	DNSILLNIN	1.000
2	489	VFRYLKSNI	1.000
3	65	SIYKNLAPL	0.990

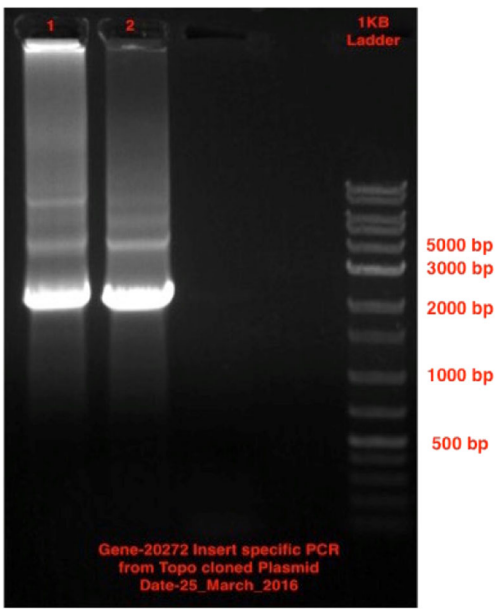


Fig. 8 Representative gel image of the TA clone *C. hominis* hypothetical protein (TU502HP) encoding gene

Author's personal copy

GROMACS within the water solvent system. Trajectories were analyzed at each 2.5-ns time interval, among which the hydrogen bond with Leu 525 (NH–O=C) and Gly 530 is significantly present (100%) in MD trajectories (Fig. 5). The binding pose of protein and compound complex before and after MD simulation is shown in Fig. 6. The RMSD of the protein backbone after 10 ns showed that the protein-ligand complex attains a maximum RMSD value of approximately 0.9 nm between 3000 and 10,000 ps (Fig. 7a). Compactness of the protein can be determined by the Rg of the protein; during the simulation of this protein, a gradual decrease of the Rg was observed after 4 ns to reach a compact structure at around 10 ns (Fig. 7b). The flexibility of structure was calculated through root-mean-square fluctuation (RMSF) from 0- to 10-ns simulation; the plot of protein residues displayed a maximum fluctuation in the region of 0.8 nm (Fig. 7c). Increased fluctuation was predominately observed in the 634th and 708th amino acid residues. Solvent accessible surface area (SASA) was calculated for protein to check for available surface area for ligand interactions and extension of hydrophobicity of the model. The SASA result indicates that after the protein unfolds itself, i.e., after 2 ns, the protein becomes more compact (Fig. 7d).

Most of the hydrogen bonds of the protein-ligand complex were retained in an energy-minimized complex structure at 10-ns MD simulation. The MD simulation study provides substantial evidence of the reliability of molecular docking and binding stability of the protein-ligand complex.

Potential B and T cell epitopes

B cell epitopes are usually recognized by host immune cells and induce a protective immune response against infection. In total, eight potential B cell epitopes were predicted on the basis of variability, surface accessibility, fragment mobility, and secondary structures (Table 2). The amino acid sequences with a high binding score have high possibility to induce an antibody response (Tambunan and Parikesit 2009). Three out of eight epitopes were predicted with a score of ≥0.9. The highest ranking epitope was predicted to be FRYKLTIPTTNNKSIVKTKD starting at position of the 340th residue in the protein. The list of potential T cell epitopes is also shown in Table 3. CTLPred predicted three T cell epitopes of nine residues long with a higher recommended cutoff score of 0.51 for ANN and 0.36 for SVM. The server predicted DNSILLNIN and VFRYLKSNI as the highest ranking epitopes at the 228th and 489th residues of the protein, respectively. The majority of predicted epitopes contain polar amino acids (serine, threonine, tyrosine, and asparagine) similar to the epitopes reported from the envelope glycoprotein E of dengue virus type 3 (Ilyas et al. 2011). The Jameson-Wolf index analysis of TU502HP demonstrated that it had a good antigen index of 1.7 and suggested that it could be a potential candidate to detect *C. hominis* infection. Furthermore, TU502HP also possesses class 1 immunogenicity and antibody epitopes (Supplementary Tables 3 and 4). For developing synthetic peptide vaccines, immunodiagnostics, and antibody production, it is necessary to identify antigenic determinants on target proteins (Chen et al. 2007).

Identification of TA-cloned TU502HP

PCR-cloned product was visualized by electrophoresis; a 2220-bp band was observed on 1% agarose gel (Fig. 8). The sequencing results also revealed 97.2% identity with a TU502HP gene sequence which shows that the particular hypothetical protein exists in the *C. hominis* genome (Supplementary Fig. 1).

Conclusions

A continuous increase in access to the parasite genomic sequence data has provided an opportunity to predict the parasite-borne proteins to be potential drug and/or vaccine candidates. Computational screening of such targets using docking with a dynamics protocol and a drug-like compound library database has enhanced significantly the identification of possible novel therapeutic agents. Currently, in biomedical research, the search for novel drug targets and there characterization by various predictive tools allow a more complete and unbiased interpretation of the data, leading to suggestions for experimental validation of predicted targets and biomarkers for the uncharacterized proteins.

In the present study, the identified *C. hominis* hypothetical protein (TU502HP) is expected to be a drug target for the suggested inhibitor molecule. However, an in vitro or in vivo study to confirm the above proposition is warranted. A further study might facilitate the identification of possible other functional proteins and putative inhibitor molecules in novel drug designing for cryptosporidiosis.

Acknowledgements The authors duly acknowledge Prof. Mrutyunjay Suar, Director of School of Biotechnology, KIIT University, Bhubaneswar, for the institutional support. Especially, we are thankful to the bioinformatics facility of the school for the software and computing resources used in this study. The help from Ms. Subhashree Rout, PhD scholar, Bioinformatics Lab, is also of great importance particularly during the docking and MD simulation analysis. The kind help from Prof. Gagandeep Kang, Christian Medical College, Vellore (India), is greatly acknowledged.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Author's personal copy

References

Abubakar I, Aliyu SH, Arumugam C, Usman NK, Hunter PR (2007) Treatment of cryptosporidiosis in immunocompromised individuals: systematic review and meta-analysis. Br J Clin Pharmacol 63:387–393. doi:10.1111/j.1365-2125.2007.02873.x

Adams MA, Suits MD, Zheng J, Jia Z (2007) Piecing together the structure–function puzzle: experiences in structure based functional annotation of hypothetical proteins. Proteomics 7:2920–2932. doi:10.1002/pmic.200700099

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI BLAST: a new generation protein database search programs. Nucleic Acids Res 25:3389–3402

Baker D (2006) Prediction and design of macromolecular structures and interactions. Philosophical Transactions of the Royal Society of London B: Biological Sciences 361:459–463

Berahmat R, Spotin A, Ahmadpour E, Mahami-Oskouei M, Rezamand A, Aminisani N, Ghojazadeh M, Ghoyounchi R, Mikaeili-Galeh T (2017) Human cryptosporidiosis in Iran: a systematic review and meta-analysis. Parasitol Res 116(4):1111–1128. doi:10.1007/s00436-017-5376-3

Berendsen HJ, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun 91:43–56. doi:10.1016/0010-4655(95)00042-E

Bernstein FC, Koetzle TF, Williams GJ Jr, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 80:319–324. doi:10.1111/j.1432-1033.1977.tb11885.x

Bhasin M, Raghava GPS (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine 22:3195–3204. doi:10.1016/j.vaccine.2004.02.005

Burland TG (1999) DNASTAR's Lasergene sequence analysis software. Methods Mol Biol 132:71–91. doi:10.1385/1-59259-192-2:71

Chang TH, Wu LC, Lee TY, Chen SP, Huang HD, Horng JT (2013) EuLoc: a webserver for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC. J Comput Aided Mol Des 27(1):91–103

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428. doi:10.1007/s00726-006-0485-9

Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 9:1511–1519

Efstratiou A, Ongerth JE, Karanis P (2017) Waterborne transmission of protozoan parasites: review of worldwide outbreaks—an update 2011–2016. Water Res 114:14–22. doi:10.1016/j.watres.2017.01.036

Eisenberg D, Lüthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three dimensional profiles. Methods Enzymol 277:396–404

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A, (2005) Protein identification and analysis tools on the ExPASy server. In: John M. Walker (ed) Proteomics Protoc Handb, Humana Press, Totowa p 571–607. doi:10.1385/1-59259-890-0:571

Guerrant DI, Moore SR, Lima AA, Patrick PD, Schorling JB, Guerrant RL (1999) Association of early childhood diarrhea and cryptosporidiosis with impaired physical fitness and cognitive function four-seven years later in a poor urban community in northeast Brazil. AmJTrop Med Hyg 61:707–713

Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su Y, Miller J, Kraemer E, Kissinger JC (2006) CryptoDB: a Cryptosporidium bioinformatics resource update. Nucleic Acids Res 34:D419–D422. doi:10.1093/nar/gkj078

Hess B, Bekker H, Berendsen HJ, Fraaije JG (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472

Ilyas M, Rahman Z, Shamas S, Alam M, Israr M, Masood K (2011) Bioinformatics analysis of envelope glycoprotein E epitopes of dengue virus type 3. African J Biotechnol 10(18):3528–3533

Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182. doi:10.1021/ci049714+

Laskowski RA, MacArthur MN, Moss DS, Thorton JM (1993) PROCHECK-a program to check the steriochemical quality of protein structures. J Appl Crystallogr 26:283–291

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational apparoaches to estimate solubilitty and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46:3–26. doi:10.1016/S0169-409X(96)00423-1

Lorber D M, Shoichet BK (1998) Flexible ligand docking using conformational ensembles. Protein Sci 7:938–950

Maertens GN, Cook NJ, Wang W, Hare S, Gupta SS, Oztop I, Lee K, Pye VE, Cosnefroy O, Snijders AP, Kewalramani VN, Fassati A, Engelman A, Cherepanov P (2014) Structural basis for nuclear import of splicing factors by human transportin 3. Proc Natl Acad Sci U S A 111:2728–2733. doi:10.1073/pnas.1320755111

Mahon M, Doyle S (2017) Waterborne outbreak of cryptosporidiosis in the South East of Ireland: weighing up the evidence. Ir J Med Sci 1–6. doi: 10.1007/s11845-016-1552-1

Malleshappa Gowder S, Chatterjee J, Chaudhuri T, Paul K (2014) Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. Scientific World Journal. doi:10.1155/2014/971258

Morris GM, Ruth H, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Software news and updates AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791. doi:10.1002/jcc.21256

Pain A, Crossman L, Parkhill J (2005) Comparative apicomplexan genomics. Nat Rev Microbiol 3:454–455. doi:10.1080/10409238.2017.1290043

Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22(14):e408–e416

Plewczynski D, Pas J, Von Grotthuss M, Rychlewski L (2004) Comparison of proteins based on segments structural similarity. Acta Biochim Pol 51:161–172

Rider SD, Zhu G (2010) *Cryptosporidium*: genomic and biochemical features. Exp Parasitol 124:2–9. doi:10.1016/j.exppara.2008.12.014

Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic tree. Mol Bio Evol 4(4):406–425

Saha S, Raghava GPS (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. Artif Immune Syst Third Int Conf 3239:197–204. doi:10.1007/978-3-540-30220-9_16

Schrödinger, LLC (2010) The PyMOL molecular graphics system, version 1.3r1. Schrödinger, LLC, Portland, Oregon

SchuÈttelkopf AW, Van Aalten DM (2004) PRODRG: a tool for high-throughput crystallography of protein–ligand complexes. Acta Crystallogr D Biol Crystallogr 60(8):1355–1363

Scott WRP, Hunenberger PH, Tironi IG et al (1999) The GROMOS bimolecular simulation program package. J Phys Chem A 103(19):3596–3607

Sharling L, Liu X, Gollapalli DR, Maurya SK, Hedstrom L, Striepen B (2010) A screening pipeline for antiparasitic agents targeting Cryptosporidium inosine monophosphate dehydrogenase. PLoS Negl Trop Dis 4(8):e794. doi:10.1371/journal.pntd.0000794

Author's personal copy

Snelling WJ, Xiao L, Ortega-Pierres G, Lowery CJ, Moore JE, Rao JR, Smyth S, Millar BC, Rooney PJ, Matsuda M, Kenny F, Xu J, Dooley JSG (2007) *Cryptosporidiosis* in developing countries. J Infect Dev Ctries 1:242–256. doi:10.3855/jidc.360

Striepen B (2013) Time to tackle cryptosporidiosis. Nature 503:189–191

Tambunan US, Parikesit AA (2009) *In silico* analysis of envelope dengue virus-2 and envelope dengue virus-3 protein as the backbone of dengue virus tetravalent vaccine by using homology modeling method. Online J Biol Sci 9:6–16

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30(12):2725–2729

Vaught A (1996) Graphing with Gnuplot and Xmgr: two graphing packages available under linux. Linux Journal 28es:7

Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 5:275–284. doi:10.1016/j.sbi.2005.04.003

Wiederstein M, Sippl MJ (2007) ProSA-Web: interactive Web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410. doi:10.1093/nar/gkm290

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA (2004) The genome of *Cryptosporidium hominis*. Nature 431:1107–1112

Yang J, Zhang Y (2015) Protein structure and function prediction using I-TASSER. Curr Protoc Bioinformatics 52:5–8. doi:10.1002/0471250953.bi0508s52

Yoder JS, Beach MJ (2010) Cryptosporidium surveillance and risk factors in the United States. Exp Parasitol 124:31–39. doi:10.1016/j.exppara.2009.09.020

Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins 64:643–651. doi:10.1002/prot.21018