Original Article

TIPS Trends in Phramaceutical Sciences 2016: 2(2): 117-130.

In silico evaluation of rare codons and their positions in the structure of cytosine deaminase and substrate docking studies

Mojtaba Mortazavi[1], Navid Nezafat[2], Manica Negahdaripour[2,3], Ahmad Gholami[2,3], Masoud Torkzadeh-Mahani[1], Safa Lotfi1, Younes Ghasemi[2,3,4]

[1]Department of Biotechnology, Institute of Science and High Technology and Environmental Science, Graduate University of Advanced Technology, Kerman, Iran.
[2]Pharmaceutical Sciences Research Center, Shiraz University of Medical Sciences, Shiraz, Iran.
[3]Department of Pharmaceutical Biotechnology, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.
[4]Department of Medical Biotechnology, School of Advanced Medical Sciences and Technologies, Shiraz University of Medical Sciences, Shiraz, Iran.

Abstract

Cytosine deaminase (CDase) is used with 5-fluorocytosine (5FC) for genetic cancer treatment. This approach has several undesirable features, such as a relatively poor turnover of 5FC and low prodrug conversion activities, thus limiting the overall therapeutic response. We previously reported the molecular cloning of a new type of CDase in *E. coli* AGH09. Here, we describe a hidden layer of information of rare codons in this gene, which can help in problem solving of protein expression. With the help of several web servers, some rare codons in different locations of CDase gene were identified. By in silico modelling of CDase in I-TASSER server, the three rare codons of Arg^{242}, Arg^{286} and Arg^{360} were evaluated. All of these rare codons were located at special positions that seem to have a critical role in proper folding of CDase. In silico docking of the substrate binding site simulation with AutoDock Vina showed that Glu^{218} is involved in substrate binding site, but the others residues were incompatible. Structural analysis showed that the rare codon of Arg^{286} is located adjacent to Glu^{218} in binding site, which may have a critical role in ensuring the correct formation of the binding site structure. Investigation of this hidden information can enhance our understanding of CDase folding. Moreover, studies of these rare codons help to clarify their role in rational design of new and effective drugs.

Keywords: AutoDock, Bioinformatics analysis, CDase, Rare codon.

1. Introduction

Recent studies have demonstrated the effectiveness of cytosine deaminase/5-fluorocytosine (CD/5-FC) in cancer suicide gene therapy (1-3). Cytosine deaminase (CDase) belongs to the family of hydrolases, which act on carbon-nitrogen bonds other than peptide bonds, specifically in cyclic amidines (4). The bacterial CDase gene, which is not expressed in eukaryotic cells, encodes an enzyme capable of converting cytosine and 5-FC to uracil and 5-fluorouracil (5-FU), respectively (5). 5-fluorouracil (5FU) is widely used as a chemotherapeutic drug in tumor-targeted chemotherapy (6). So far, CDase gene has been cloned in bacteria and fungi and used as an antitumor agent (7, 8). However, as previously reported, CDase displays relatively poor turnover of 5FC, thus limiting the overall therapeutic response (9). For practical use of CDase and overcoming this constraints, several approaches such as site-specific mutagenesis have been carried out in order to isolate variants of CDase with improved properties (6, 7, 9). In order to achieve excellent results in this regard,

Corresponding Author: Younes Ghasemi, Department of Pharmaceutical Biotechnology, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.
Email: ghasemiy@sums.ac.ir

Recieved:03/04/2016; Accepted:27/04//2016

some critical points such as situation of rare codons should be considered.

A number of codons in highly expressed genes are used preferentially, while some codons are almost absent, called rare or low usage codons (10). Recent studies suggest that these rare codons can be a serious problem for heterologous protein expression and have a special role in protein activity (11). Because expression systems produce large amounts of the same protein, neglecting substitution of these rare codons can create some problems in protein expression such as translational pausing and premature termination of translation (12). Furthermore, some reports indicate that ribosomal pausing occurs with decrease of tRNAs concentration in rare codons until the rare activated tRNA brings the next amino acid (13, 14). It has been proposed that this pausing may have evolved to ensure the independent folding of some regions of polypeptide chains during their synthesis (15). These rare codons or the hidden layer of information are able to mediate local kinetics of translation (14). Studies of the hidden information in codon sequence, can provide insights into the histories of genes and help in problem solving of protein expression (14).

We previously reported the molecular cloning, purification, and characterization of a new type of CDase in *E. coli* AGH09 (7). As rare codons are functionally important for protein activity (16), we studied, for the first time, the rare codons in CDase gene from *E. coli* AGH09 and identified the location of these rare codons in the structure of CDase. In this study, detection of rare codons were performed using the following servers that detect rare codons or rare codon clusters: ATGme (http://atgme.org/), Rare codon calculator (RaCC) (http://nihserver.mbi.ucla.edu/RACC/), LaTcOm (http://structure.biol.ucy.ac.cy/latcom.html), and Sherlocc program (http://bcb.med.usherbrooke.ca/sherlocc.php) (16). ATGme is a simple DNA sequence optimization tool. Analysis of CDase gene in this server showed the rare and very rare codons (17). RaCC can determine the number of rare codons in a DNA sequence. LaTcOm is a new web tool designed for detecting and visualizing 'rare codon clusters' (RCC) (18). Sherlocc program, using Pfam accession number, detects statistically relevant conserved rare codon clusters (16). By these analyses, some rare codon were identified as Arg (CGA-242), (CGA-286) and (AGG-360). For further understanding of the rare codons role, 3D structure of this enzyme was modeled in the I-TASSER (19) and the situation of these rare codons were visualized and studied using Swiss PDB Viewer software (20) and PyMOL Molecular Graphics System (21). Rare codons can influence translational efficiency, and large clusters will have a greater effect on protein production than an equivalent number of randomly scattered rare codons (22-24). In silico docking simulation was also carried out for identification of substrate binding site and their relationships with rare codon. In this study, the docking protocol was validated by redocking co-crystallized structure of cytosine deaminase in complex with cytosine (PDB ID: 3O7U) (25). The results of this study may help in operational development of this technology and elucidating the enzyme folding mechanism, as well as rational design of new and effective drugs.

2. Materials and Methods

2.1. Detection of rare codon clusters in gene and protein structure of CDase

The protein family (Pfam) accession number of CDase was identified using uniprot database (http://www.uniprot.org/). Pfam is a comprehensive collection of protein domains and families represented as multiple sequence alignments and as profile hidden Markov models (26). Initially, this Pfam ID was analyzed in Sherlocc program and showed that this gene does not have any rare codon cluster. ATGme is a simple DNA sequence optimization tool that provides a simple user-friendly and flexible web-based application for identification of rare codons and proposing several options for codon usage optimization. Analysis of CDase gene in ATGme was done in four steps: (i) Input of the CDase sequence; (ii) Input of the codon usage table of *E. coli* B [gbbct]: 11 CDS's (3771 codons) that was copied from the Codon Usage Database (http://www.kazusa .or.jp/codon/); (iii) starting the process. The rare and highly rare codons are highlighted in orange and red, respectively, in the output sequence. RaCC introduced codons for arginine (AGG, AGA, CGA), leucine

(CTA), isoleucine (ATA), and proline (CCC) with probable problem. Besides, RaCC web server detects the number of tandem rare Arg codon double repeats and triple repeats. LaTcOm is a new web tool designed for detecting and visualizing rare codon clusters (RCC) (18). In this tool, three core RCC detection algorithms are implemented: i) % minimax algorithm, ii) sliding window approach, and iii) a linear-time algorithm named MSS. We used RCC with the following parameters: MSS, Scale: Dong table codon usage (27), cluster length: 7 and transformation: linear+sigmoid. Then the RCC positions were visualized within the submitted sequences.

2.2. Study of rare codons in structure of CDase

To investigate the position of these rare codon clusters in CDase, the 3D structure of this enzyme was modeled by submission of CDase sequence in I-TASSER web server (19). I-TASSER web server was used to generate a total of five most suitable models of target protein. In this web server, 3D models are built based on multiple-threading alignments by LOMETS (Local Meta-Threading-Server) (28) and iterative template fragment assembly simulations. The template used by I-TASSER web server was 3o7uA, 3r0dA, 1qw7A, 3uf9C, and 4jnrB. The models with the best "Confidence Score" and Z-score are chosen by I-TASSER server. The "Confidence Score" indicates the confidence of the predicted template, which is based on a scoring function that takes into account the Z-score of the template, the confidence of the particular server, and the sequence identity between the query and the template. The model which showed the best confidence and Z-score was selected and visualized using swiss PDB viewer (29) and PyMOL molecular graphics system (21). Hydrogen bonds were also calculated by WHAT IF web server (30) and PIC web server (31).

2.3. Molecular docking using AutoDock Vina

AutoDock Vina (version 1.1.2) (32) is used in this project to conduct molecular docking. CDase and cytosine were treated as a receptor and a small molecule ligand, respectively. Three-dimensional structure of CDase was constructed using I-TASSER web server and converted to PDBQT format by means of MGL tools (version 1.5.4) (33). Cytosine in SDF format was obtained from PubChem database (https://pubchem.ncbi.nlm.nih.gov/) and converted to PDB format by Open Babel (version 2.3.1). Since the ligand in PDBQT format is required for performing docking experiments, the ligand was converted from PDB to PDBQT format. Open Babel (version 2.3.1) (34) is used to convert ligand atoms to the PDBQT format. The search space was centered at the similar position of substrate binding site corresponding to the crystal structure of *E. coli* CDase (3O7U) (25). The dimensions of grid box (x=18, y=18, and z=18) were adjusted to include all residues participating in substrate binding. The docking experiments were performed at exhaustiveness value of 20.

3. Results

3.1. Detection of rare codon clusters

The Pfam accession numbers of *E. coli* CDase was identified as PF07969 in the uniprot database (http://www.uniprot.org/). PF07969 ID was studied in the Sherlocc program (16). Based on the result, this program did not identify any rare codon cluster in CDase (Table 1).

Then the nucleotide sequence of CDase was analyzed in ATGme server, which identifies rare codons and gives several options for codon usage optimization. By use of codon usage table of *E. coli* B [gbbct]: 11 CDS's (http://www.kazusa.or.jp/codon/), this gene was analyzed. Figure 1 shows the rare and highly rare codons that were highlighted in orange and red, respectively (Fig. 1). Moreover, the GC and AT contents of this gene were GC%:53.32 and AT%:46.68, calculated by this server.

As this result shows, CDase gene has some rare and highly rare codons. For refine-

Table 1. The result of PF07969 ID analysis in Sherlocc program.

PFAM ID	Pfam Name	Number of rare codon clusters	Rare codon frequency threshold	Size of largest cluster	Number of Sequences	Number of unique organisms
Your query gave 0 match.						

ATG TCG AAT AAC GCT TTA CAA ACA ATT ATT AAC GCC CGG TTG CCA GGC
AAA GAG GGG CTG TGG CAG ATT CAT CTG CAG GAC GGA AAA ATC AGC GCC
ATT GAT GCG CAA TCC GGC GTG ATG CCC ATA ACT GAA AAC AGC CTG GAT
GCC GAA CAA GGT TTA GTT ATA CCG CCG TTT GTG GAG CCA CAT ATT CAC
CTG GAC ACC ACG CAA ACC GCC GGA CAA CCG AAC TGG AAT CAG TCC GGC
ACG CTG TTT GAA GGC ATT GAA CGC TGG GCC GAG CGC AAA GCG TTA TTA
ACC CAT GAC GAT GTG AAA CAA CGC GCA TGG CAA ACG CTG AAA TGG CAG
ATT GCC AAC GGC ATT CAG CAT GTG CGT ACC CAT GTC GAT GTT TCG GAT
GCA ACG CTA ACT GCG CTG AAA GCA ATG CTG GAA GTG AAG CTG GAA GTC
GCG CCG TGG ATT GAT CTG CAA ATC GTC GCC TTC CCT CAG GAA GGG ATT
TTG TCG TAT CCC AAC GGT GAA GCG TTG CTG GAA GAG GCG TTA CGC TTA
GGG GCA GAT GTA GTG GGG GCG ATT CCG CAT TTT GAA TTT ACC CGT GAA
TAC GGC GTG GAG TCG CTG CAT AAA ACC TTC GCC CTG GCG CAA AAA TAC
GAC CGT CTC ATC GAC GTT CAC TGT GAT GAG ATC GAT GAC GAG CAG TCG
CGC TTT GTC GAA ACC GTT GCT GCC CTG GCG CAC CGT GAA GGC ATG GGC
GCG CGA GTC ACC GCC AGC CAC ACC ACG GCA ATG CAC TCT TAT AAC GGG
GCG TAT ACC TCA CGT CTG TTC CGC TTG CTG AAA ATG TCC GGT ATT AAC
TTT GTC GCC AAC CCG CTG GTC AAT ATT CAT CTG CAA GGA CGA TTC GAT
ACG TAT CCA AAA CGT CGC GGC ATC ACG CGC GTT AAA GAG ATG CTG GAG
TCC GGC ATT AAC GTC TGC TTT GGT CAC GAT GAT GTC TTC GAT CCG TGG
TAT CCG CTG GGA ACG GCG AAT ATG CTG CAA GTG CTG CAT ATG GGG CTG
CAT GTT TGC CAG CTG ATG GGC TAT GGG CAG ATT AAC GAT GGC CTG AAT
TTA ATC ACC CAC CAC AGC GCC AGG ACG TTG AAT TTG CAG GAT TAC AGC
ATT GCC GCC GGA AAC AGC GCC AAC CTG ATT ATC CTG CCG GCT GAA AAT
GGA TTT GAT GCG CTG CGC CGT CAG GTT CCG GTA CGT TAT TCG GTA CGT
GGC GGC AAG GTG ATT GCC AGC ACA CAA CCG GCA CAA ACC ACC GTA TAT
CTG GAG CAG CCG GAA GCC ATC GAT TAC AAA CGT

Figure 1. Schematic representation of the codon usage of CDase and position of rare and very rare codons, displayed in orange and red, respectively.

ment of these results, we used the RaCC server. As pointed before, RaCC server introduced the problematic residue codons as Arg, Leu, Ile, and Pro. Results show that CDase gene has three single rare codons of Arg (AGG^{242}-AGA-286 and CGA-360) (fig. 2 red color). Besides, RaCC detected the two single rare codons of Ile, ATA (42-55-codon number -blue color), one single rare codon of Leu, CTA (131-codon number, green color) and two rare codons of Pro, CCC (41-164-codon number , orange color) (Figure 2). This analysis also shows that CDase gene does not have any tandem double or triple repeats of rare Arg codon.

Besides, by use of LaTcOm web tool, rare codon clusters (RCC) were detected and visualized (18). In this web tool, %MINMAX, sliding window and MSS algorithm are employed in detecting the core of RCC. We used Dong table codon usage (27) in all of this algorithm as the reference scale. By use of MSS algorithm, two rare codon clusters were identified (Figure 3A). Likewise, CDase gene analysis in minmax algorithm resulted in the identification of two rare codon clusters (Figure 3B). While sliding_window algorithm in this web serv-

atg tcg aat aac gct tta caa aca att att aac gcc cgg ttg cca ggc aaa gag ggg ctg tgg cag att cat ctg
cag gac gga aaa atc agc gcc att gat gcg caa tcc ggc gtg atg CCC ATA act gaa aac agc ctg gat
gcc gaa caa ggt tta gtt ATA ccg ccg ttt gtg gag cca cat att cac ctg gac acc acg caa acc gcc gga
caa ccg aac tgg aat cag tcc ggc acg ctg ttt gaa ggc att gaa cgc tgg gcc gag cgc aaa gcg tta tta acc
cat gac gat gtg aaa caa cgc gca tgg caa acg ctg aaa tgg cag att gcc aac ggc att cag cat gtg cgt acc
cat gtc gat gtt tcg gat gca acg CTA act gcg ctg aaa gca atg ctg gaa gtg aag ctg gaa gtc gcg ccg
tgg att gat ctg caa atc gtc gcc ttc cct cag gaa ggg att ttg tcg tat CCC aac ggt gaa gcg ttg ctg gaa
gag gcg tta cgc tta ggg gca gat gta gtg ggg gcg att ccg cat ttt gaa ttt acc cgt gaa tac ggc gtg gag
tcg ctg cat aaa acc ttc gcc ctg gcg caa aaa tac gac cgt ctc atc gac gtt cac tgt gat gag atc gat gac
gag cag tcg cgc ttt gtc gaa acc gtt gct gcc ctg gcg cac cgt gaa ggc atg ggc gcg CGA gtc acc gcc
agc cac acc acg gca atg cac tct tat aac ggg gcg tat acc tca cgt ctg ttc cgc ttg ctg aaa atg tcc ggt
att aac ttt gtc gcc aac ccg ctg gtc aat att cat ctg caa gga CGA ttc gat acg tat cca aaa cgt cgc ggc
atc acg cgc gtt aaa gag atg ctg gag tcc ggc att aac gtc tgc ttt ggt cac gat gat gtc ttc gat ccg tgg tat
ccg ctg gga acg gcg aat atg ctg caa gtg ctg cat atg ggg ctg cat gtt tgc cag ctg atg ggc tat ggg cag
att aac gat ggc ctg aat tta atc acc cac cac agc gcc AGG acg ttg aat ttg cag gat tac agc att gcc gcc
gga aac agc gcc aac ctg att atc ctg ccg gct gaa aat gga ttt gat gcg ctg cgc cgt cag gtt ccg gta cgt
tat tcg gta cgt ggc ggc aag gtg att gcc agc aca caa ccg gca caa acc acc gta tat ctg gag cag ccg gaa
gcc atc gat tac aaa cgt

Figure2. Schematic representation of the CDase gene and position of Arg, Leu, Ile, and Pro. These residues have rare codons, displayed in red, blue, green, orange, and red, respectively.

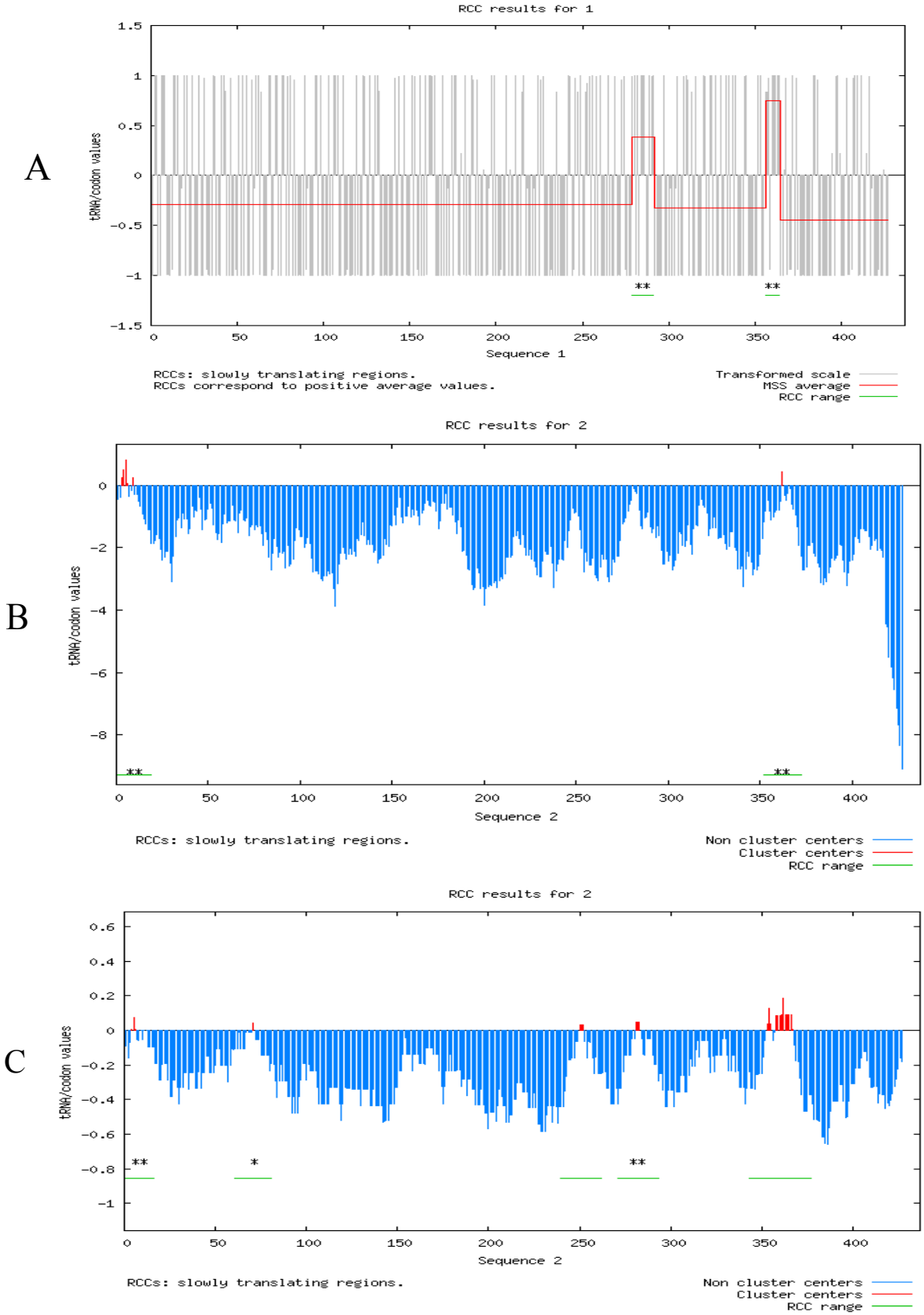

Figure3. The position of rare codon clusters in CDase gene. Detection of RCC using (A) MSS algorithm, (B) minmax algorithm, and (C) sliding-window method.

Table 2. The rare codon clusters characteristics in CDase gene retrieved from LaTcOm web tool.

RCC identification method	Cluster length	Position of clusters	Score (per position)	Expected value	Significance (simulated *p*-value)
MSS	7	279-291	0.384	-0.173	
		356-364	0.745	0.153	
Minmax	21	1-19	-0.002	-0.368	
		352-372	0.182	-0.456	
Sliding_window	21	1-16	0.005	-0.329	**(0.000)
		61-81	0.046	-0.313	* (0.024)
		240-262	-0.019	-0.351	(1.000)
		271-293	-0.062	-0.596	**(0.000) (0.852)
		343-377	-0.158	-0.558	

er identified five rare codon clusters (Figure 3C).

These results show the different features of these algorithms in detection of RCC. As shown, MSS and minmax detected two clusters, and sliding-window method detected five clusters. It is important to note that the cluster length selected for MSS was 7 and by selection of a higher value, this algorithm is not able to detect any RCC. The characteristics of these RCC were reported in Table 2.

These web servers, based on their input initial parameters have identified different parts of CDase gene as rare codon or rare codon cluster. To better understand this analysis and summarize the results, we focused on the common regions that were identified in most of these servers. After their comparison, three clusters located at codon sequence from 240-262, 271-293 and 343-377 were selected. These regions were identified in most of the results taken from various web servers. For further analysis, three Arg codon (CGA-242, CGA-286, and AGG-360) were precisely studied in CDase structure.

3.2. Studying rare codons in CDase structure

As mentioned, three rare codons of Arg were identified in CDase. For detailed study of these rare codons, the 3D model of this enzyme is needed. In this context, the 3D structure of CDase was obtained by I-TASSER Web Server, shown in

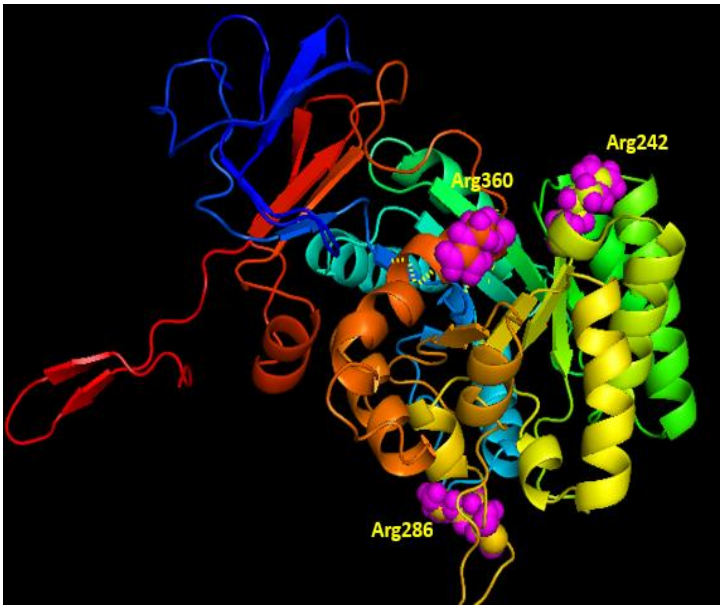

Figure 4. The ribbon diagram of CDase and the location of rare codon residues. The rare codons are shown in pink color.

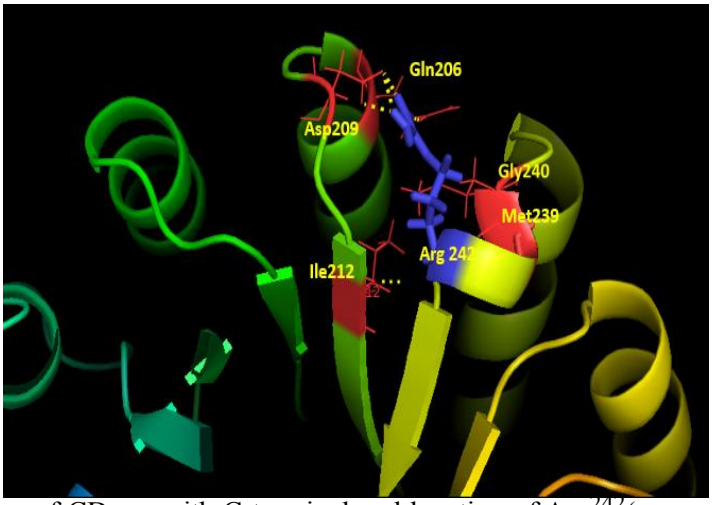

Figure 5. The ribbon diagram of CDase, with C-terminal and location of Arg^{242}(rare codon residues) in blue color. The residues that form hydrogen interaction with Arg^{242} are shown in red color.

Figure 4, along with its three Arg residues.

I-TASSER Web Server generated five models that were visualized with PyMOL. The best model showed 1.75 value of overall C-score and 0.96+-0.05 value of TM-Score and Exp. RMSD was 3.4+-2.4. CDase has 217 residues, in which rare codons of Arg are found in amino acids 242, 286, and 360. Based on the results of modeling, these rare codons are located in the C-terminal region of CDase. Analyzing the 3D model of CDase showed that Arg^{242} residue forms hydrogen bond with Met^{239}, Gly^{240}, Ile^{212}, Asp^{209}, Gly^{240}, and Gln^{206} (Figure 5).

Calculation of non-covalent interactions was done by WHAT IF (30) and PIC (31) web servers, and the results are shown in Table 3.

As shown in figure 6, Arg^{286} residue constitutes hydrogen bonds with His^{282}, Gly^{80}, Gln^{284}, and Ser^{79}.

The results of this analysis are shown in Table 4.

Also, Arg^{360} forms hydrogen bond with His^{356}, His^{357}, Leu^{362}, Asn^{363}, and Asn^{308}. This figure shows the position of these rare codons in CDase protein (Figure 7).

The results of this analysis are shown in Table 5.

An initial review of location of these Arg residues and the large number of hydrogen bonds they formed, show that these residues have a criti-

Table 3. The characteristics of non-covalent interactions of Arg^{242} with other residues. Dd-a=Distance Between Donor and Acceptor, Dh-a=Distance Between Hydrogen and Acceptor, A(d-H-N)=Angle Between Donor-H-N , A(a-O=C)=Angle Between Acceptor-O=C, MO=Multiple Occupancy.

DONOR			ACCEPTOR				PARAMETERS			
POS	RES	ATOM	POS	RES	ATOM	MO	Dd-a	Dh-a	A(d-HN)	A(aO=C)
242	ARG	NH1	206	GLN	OE1	1	3.04	1.99	176.18	999.99
242	ARG	NH1	206	GLN	OE1	2	3.04	3.61	49.44	999.99
242	ARG	NH1	209	ASP	OD2	1	2.76	3.59	33.05	999.99
242	ARG	NH1	209	ASP	OD2	2	2.76	1.82	149.98	999.99
242	ARG	NH2	209	ASP	OD1	1	2.72	3.30	48.41	999.99
242	ARG	NH2	209	ASP	OD1	2	2.72	1.91	133.93	999.99
242	ARG	NH2	209	ASP	OD2	1	3.00	3.93	24.50	999.99
242	ARG	NH2	209	ASP	OD2	2	3.00	2.18	135.52	999.99
212	ILE	N	242	ARG	O		3.01	2.47	167.77	146.99
242	ARG	N	239	MET	O		2.94	2.19	130.72	128.04
242	ARG	N	240	GLY	O		3.24	3.19	84.07	79.19

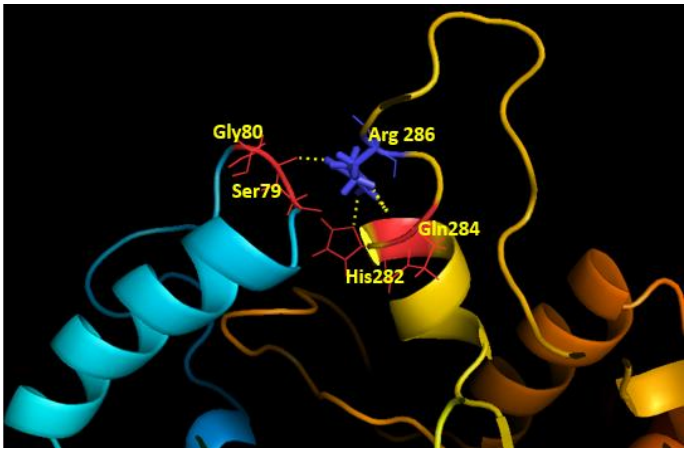

Figure 6. The ribbon diagram of CDase, with the C-terminal and location of Arg^{286} rare codon residue in blue color. The residues that form hydrogen interaction with Arg^{242} are shown in red color.

cal role in proper folding of CDase.

3.3. Enzyme-substrate docking

The computer-simulated docking studies were performed using AutoDock Vina (32). Before docking the cytosine substrate into the CDase binding site, the docking simulation method was checked. For this purpose, the bound phosphonocytosine in the crystal structure of CDase (3O7U) was removed from the active site. Since the PubChem has failed at 3D generation of phosphonocytosine, the cytosine was used and redocked into the binding pocket of CDase. The total interaction modes are shown in Table 6.

In the best mode, the RMSD of all atoms is 0.000 A° (Table 6). The AutoDock binding conformation result is superimposed with the X-ray crystallographic and shown in Figure 8.

This indicated that the docking protocol parameters are reasonable in reproducing the X-ray crystal structure. The ChExV method was used for molecular channel extraction based on the alpha complex representation (35). This method computes geometrically feasible channels, stores both the volume occupied by the channel and its centerline in a unified representation, and reports significant channels (35). The results of this analysis and the situation of cytosine in this channel are shown in Figure 9.

In this study, to determine cytosine binding sites on CDase (*E. coli* AGH09), CDase was treated as the receptor, whereas cytosine as a small molecule ligand. As mentioned in the material and methods part, the search space was designed based on the active site structure of *E. coli* CDase (3O7U). In this box, most involvement residues in substrate binding (Gln^{157}, Glu^{218}, His^{247} and Asp^{314}) were selected. We examined

Table 4. The characteristics of non-covalent interactions of Arg^{286} with other residues.

DONOR			ACCEPTOR				PARAMETERS			
POS	RES	ATOM	POS	RES	ATOM	MO	Dd-a	Dh-a	A(d-HN)	A(aO=C)
286	ARG	NH2	79	SER	O	1	2.82	1.97	135.12	129.25
286	ARG	NH2	79	SER	O	2	2.82	3.16	61.61	129.25
286	ARG	NE	80	GLY	O	-	3.17	3.50	63.34	128.85
286	ARG	NE	282	HIS	O	-	2.83	2.00	134.47	147.36
286	ARG	NH1	282	HIS	O	1	2.76	1.90	137.14	115.95
286	ARG	NH1	282	HIS	O	2	2.76	3.61	31.34	115.95
286	ARG	NH1	282	HIS	O	ND1	1	2.93	2.96	77.83
286	ARG	NH1	282	HIS	O	ND1	2	2.93	2.53	101.31
286	R	N	284	GLN	O		3.41	3.08	100.35	69.75

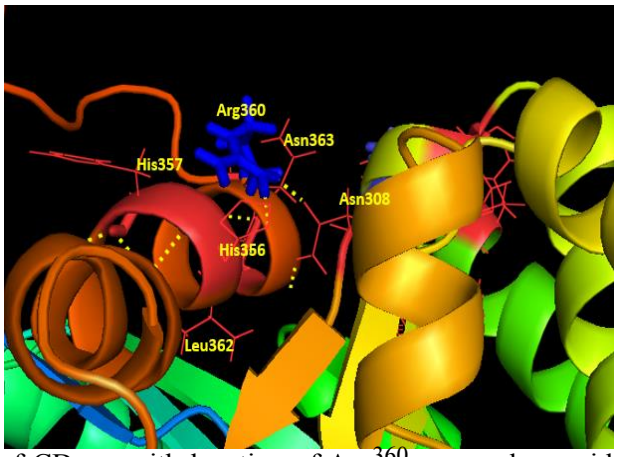

Figure 7. The ribbon diagram of CDase, with location of Arg^{360} rare codon residue in blue color. The residues that form hydrogen interaction with Arg^{242} are shown in red color.

the outcome of molecular docking using different box sizes. The CDase-cytosine complex obtained from docking results is shown in Figure 10. There is a network of diverse nonbonded interactions in the enzyme-substrate complex. As it is clear, two hydrogen bonds can be formed between two residues of CDase (Glu^{218} and Gln^{284}) and cytosine. The ligand also interacts with some other residues of the enzyme (Phe^{83}, Leu^{82}, Ile^{219}, and Phe^{287}) through hydrophobic and van der Waals contacts.

4. Discussion

The preliminary goal of this study was to perform a survey of rare codons in the structure of gene of CDase. Formerly, several in silico studies have been performed on evaluation of different signal peptides for the secretory production of human growth hormone in *E. coli* and vaccine designing (37, 38). In a previous study, we extended this analysis on genome and proteins of HCV and HBV (39). According to our survey, no similar analyses have been reported on CDase. On the other hand, the over-expression of recombinant protein in *E. coli* has very broad applications in life sciences. Therefore, it is very important to recognize everything that can help to overcome a challenge in protein expression such as "rare" codons that are infrequently used by *E. coli.*

In spite of the large amount of studies on CDase, there are a number of unresolved issues regarding the catalytic mechanism of this enzyme (25). Recently, novel CDase gene was cloned from the newly isolated *E. coli* AGH09 (7). The obtained CDase has a suitable thermostability and is able to act optimally at physiological pH, so it can be considered for cancer therapy. However, some routine buffer systems were examined to obtain the maximum activity, but none of them (citrate, phosphate, acetate, and Tris-HCl buffers) were able to promote or maintain the native CDase activity. Some strategies have been investigated for overcoming this problem and better understanding of the catalytic mechanism through site directed mutagenesis. In designing and selection of suitable mutants, considering the structural position of mutations and hidden information in desired codons

Table 5. The characteristics of non-covalent interactions of Arg^{286} with other residues.

DONOR			ACCEPTOR			PARAMETERS				
POS	RES	ATOM	POS	RES	ATOM	Dd-a	Dh-a	A(d-HN)	A(aO=C)	A(aO=C)
360	R	N	356	HIS	O	3.02	2.08	159.97	152.34	999.99
360	R	N	357	HIS	O	3.33	2.86	110.88	108.56	999.99
362	L	N	360	RG	O	3.32	3.27	84.54	76.50	999.99
363	N	N	360	ARG	O	2.98	2.07	148.24	114.69	999.99
360	5R	NE	356	HIS	O	-	3.37	3.21	90.06	999.99
360	R	NH2	308	ASN	OD1	1	2.68	1.90	127.28	999.99

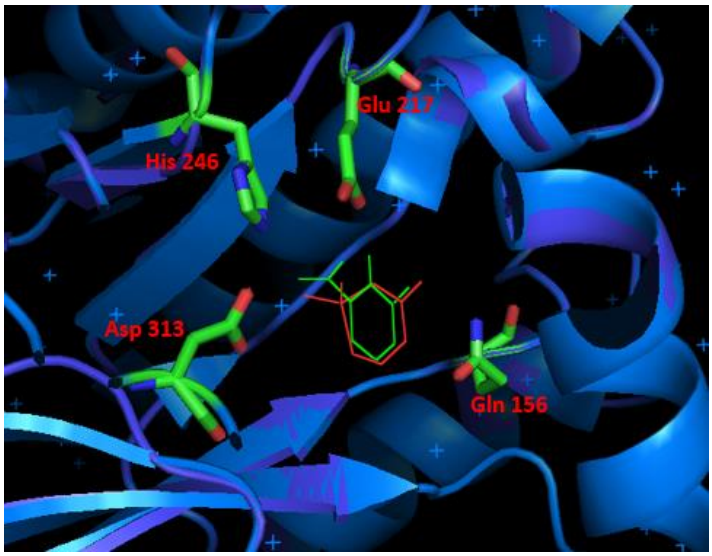

Figure 8. Docked superimposed image of CDase phosphonocytosine (green) and cytosine (red) within the binding site region.

is very important. The suitable mutants are able to determine the roles of specific residues in catalytic function. In this regard, we wanted to reveal the hidden layer of information in the CDase genes and study these rare codons in the enzyme structure. Hence, by means of some web servers, we have tried to identify these rare codons.

Sherlocc program is the primarily web server that was used to study the widespread translational pauses in CDase protein families. As shown in Table 1, Sherlocc program identified no rare codon clusters in the CDase protein family (PF07969). Then by use of ATGme web server, the rare and highly rare codons were identified and also several options for codon usage optimization were given to them. The results showed that CDase had 47 rare codons and 10 very rare codons that may play an essential role in ensuring proper folding of the protein chain. In order to refine the obtained results, the RaCC server that focused on problematic residue codons as Arg, Leu, Ile and Pro was applied. Results showed that CDase gene had three single rare codons of Arg (242, 286, 360 codon number), two single rare codons of Ile (42-55), one single rare codon of Leu (131) and two rare codons of Pro (41-164) (Figure 2). Next by use of LaTcOm web tool, rare codon clusters of CDase gene were also detected (17). Two rare codon clusters were identified via MSS and min-max algorithm, showing that one of these clusters overlapped with the other one. However, in sliding_window algorithm, five rare codon clusters were identified (Figure 3). Aforesaid algorithms have different primary databases; and therefore,

Table 6. Results of analyzing the binding site of CDase (3O7U) with cytosine.

mode	affinity (kcal/mol)	dist from best mode	
		rmsd l.b.	rmsd u.b.
1	-5.9	0.000	0.000
2	-5.2	2.138	2.893
3	-3.3	1.630	2.063
4	-3.0	2.022	3.196
5	-2.5	2.089	2.356
6	-2.1	2.201	2.576
7	-2.0	1.348	1.392
8	-1.9	1.951	2.637
9	-1.2	1.644	2.006

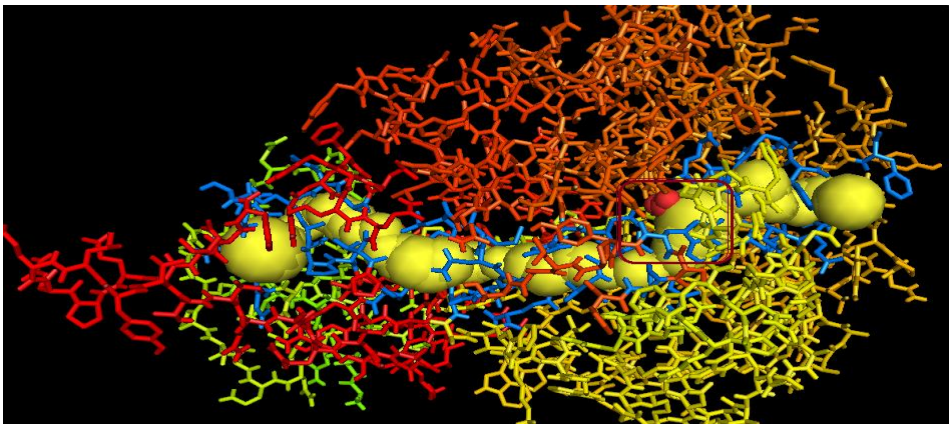

Figure 9. Channel extraction in a CDase. The top ranked pore is shown using the skin surface (yellow) and the position of cytosine is indicated by the brown square.

they have different outputs. It seems difficult to consider all of the rare codons that were obtained in our research. Consequently, for summarization of the results and more accurate analysis, we focused on common and repetitive rare codons.

Results summarization led to selection of three of the rare codons of arginine that had located at codon sequence of 242, 286 and 360. These rare codons were precisely studied in the structure of CDase. For this purpose, the 3D structure of CDase was modeled by I-TASSER Web Server. Structure analyses showed that these rare codons of Arg were located in C-terminal domain of CDase. Results of 3D modeling indicated that Arg^{242} residue forms a hydrogen bond with Met239, Gly240, Ile^{212}, Asp^{209}, Gly^{240} and Gln^{206} (Figure 5). Constitute hydrogen bonds of Arg^{286} residue with His^{282}, Gly^{80}, Gln^{286} and Ser^{79} were shown in Figure 6. Furthermore, Arg^{360} forms hydrogen bond with His^{356}, His^{357}, Leu^{362}, Asn^{363} and Asn^{308} (Figure 5). As shown in Figure 4, these hydrogen bonds have critical roles interactions, which keep together two separately parts of the protein. On the other hand, these residues have relatively formed many hydrogen bonds that involved different parts of the protein. For formation of mentioned bonds, it is pivotal to reduce the rate of protein folding in these residues. It seems that these residues have a very important role in the process of protein folding, which is necessary to slow down the rate of the folding in these positions to grantee the proper folding. Substitutions

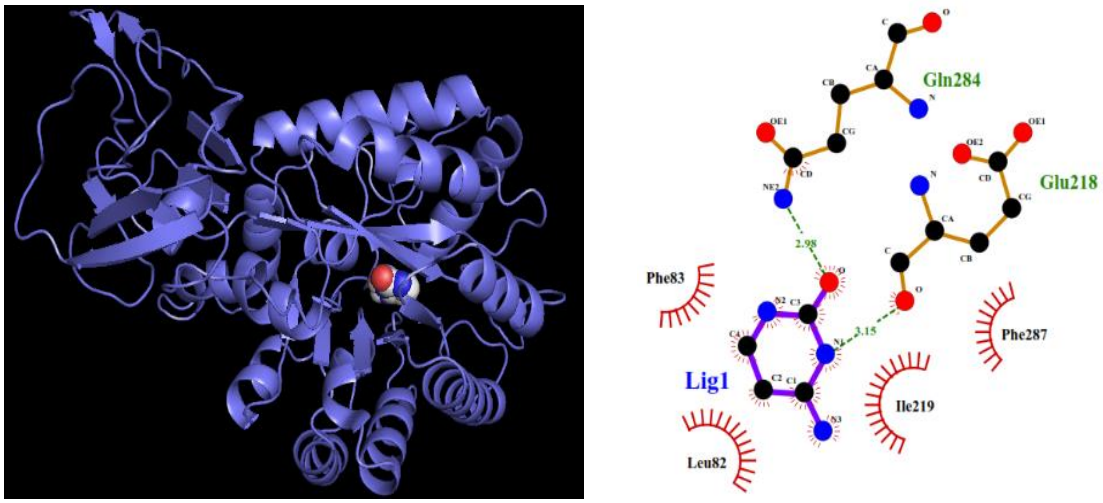

Figure 10. A) Stereo presentation of docking situation of cytosine into CDase created by PyMol (Blue color: CDase structure and Space-filling: cytosine). B) The plot show the diagrams of CDase-cytosine interactions that generated using LIGPLOT program (36).

of these residues with other amino acids lead to elimination of hydrogen bonds, which may disrupt proper folding of the proteins. However, the other rare codons that were identified should undergo further study. Our results showed that mentioned residues may play critical roles in proper folding of CDase protein and disrupting of them may be awful affected on CDase activity. Meanwhile, experimental evidence is required to confirm our theoretical studies.

Several studies have been carried out on CDase protein, which have focused on the role of amino acid and substrate binding site (9). In this context, molecular docking has profound applications in drug discovery and detection of the binding site, which is often a prerequisite for performing structure-based virtual screening (SBVS) (40). To study substrate-binding site in newly sequenced CDase, model1.pdb and cytosine were selected as a macromolecule and ligands, respectively, and were submitted to AutoDock Vina (31). Since CDase from *E. coli* (PDB ID:3O7U) and newly sequenced CDase have high sequence similarity (98%), and also all residues participating in substrate binding in 3O7U exist in the new protein, the grid box was designed based on the substrate binding site of 3O7U. Results of docking conformations were further visualized by the PyMOL and Ligplot (36). Although, the docking experiments were performed with different search space sizes and exhaustiveness values (data not shown), cytosine binding site on the newly sequenced CDase does not match with our predictions. The obtained docking results showed that, only Glu^{218} (corresponding to Glu^{218} in 3O7U) involves in the binding site. Glu^{217} in CDase (3O7U) has a critical role in cytosine binding site and in our docking results, Glu^{218} possesses a similar situation in the substrate binding process. But the other residues that have been reported as substrate-binding sites in previous studies did not involve in the docking results. In fact, the results obtained from docking studies indicated that the zinc ion plays an irreplaceable role in the architecture of the CDase substrate binding site. Since simultaneous docking of a substrate and a metal ion is very difficult and requires a considerable challenge, zinc ion was omitted from our docking experiments. This ion has an inevitable role in proper binding of substrate and enzyme activity; therefore, with the omission of the zinc ion, other residues went away from the binding site. As we know the residues involved in the binding site are critical in enzyme's activity, for this reason, they should be determined accurately. CDase is a zinc metalloprotein, and this ion has an important role in the enzyme activity. Therefore, we tried to introduce the zinc ion in the substrate-docking region to determine the proper cytosine binding site in the further studies.

One of the important results from our study is that the rare codon of Arg^{286} is located adjacent to the Glu^{218} in the binding site, which shows that this rare codon may have a critical role in proper folding and ensuring the correct formation of the binding site structure. As mentioned, CDase displays a relatively poor turnover, which limits the overall therapeutic response. To overcome this hurdle, one of the best methods is introducing mutations in proper sites. Amino acid substitutions should be conducted in positions that do not have critical roles in protein folding. Our study identified some of these residues that may involve in the substrate binding site or proper folding. Our data showed that positions of the rare codon cluster might play a critical role in proper folding and catalytic activity of CDase. This study may also provide new perspectives in drug design for treatment of cancer in future.

Conflict of Interest

None declared.

5. References

1. Negroni L, Samson M, Guigonis J-M, Rossi B, Pierrefite-Carle V, Baudoin C. Treatment of colon cancer cells using the cytosine deaminase/5-fluorocytosine suicide system induces apoptosis, modulation of the proteome, and Hsp90β phosphorylation. *Mol Cancer Ther*. 2007;6:2747-56.

2. Miyagi T, Koshida K, Hori O, Konaka H, Katoh H, Kitagawa Y, *et al*. Gene therapy for prostate cancer using the cytosine deaminase/uracil phosphoribosyltransferase suicide system. *J Gene Med*. 2003;5:30-7.

3. Kaliberov S, Chiz S, Kaliberova L, Krendelchtchikova V, Della Manna D, Zhou T, *et al.* Combination of cytosine deaminase suicide gene expression with DR5 antibody treatment increases cancer cell cytotoxicity. *Cancer Gene Ther.* 2006;13:203-14.
4. Cohen SS, Barner HD. The conversion of 5-methyldeoxycytidine to thymidine *in vitro* and *in vivo*. *J Biol Chem*. 1957;226:631-42.
5. Mullen CA, Kilstrup M, Blaese RM. Transfer of the bacterial gene for cytosine deaminase to mammalian cells confers lethal sensitivity to 5-fluorocytosine: a negative selection system. *Proc Natl Acad Sci U S A*. 1992;89:33-7.
6. Mahan SD, Ireton GC, Knoeber C, Stoddard BL, Black ME. Random mutagenesis and selection of Escherichia coli cytosine deaminase for cancer gene therapy. *Protein Eng Des Sel*. 2004;17:625-33.
7. Gholami A, Shahin S, Mohkam M, Nezafat N, Ghasemi Y. Cloning, Characterization and Bioinformatics Analysis of Novel Cytosine Deaminase from Escherichia coli AGH09. *Int J Pept Res Ther.* 2015;21:365-74.
8. Ireton GC, Black ME, Stoddard BL. The 1.14 Å crystal structure of yeast cytosine deaminase: evolution of nucleotide salvage enzymes and implications for genetic chemotherapy. *Structure.* 2003;11:961-72.
9. Fuchita M, Ardiani A, Zhao L, Serve K, Stoddard BL, Black ME. Bacterial Cytosine Deaminase Mutants Created by Molecular Engineering Show Improved 5-Fluorocytosine–Mediated Cell Killing *In vitro* and *In vivo*. *Cancer Res.* 2009;69:4791-9.
10. Dix DB, Thompson RC. Codon choice and gene expression: synonymous codons differ in translational accuracy. *Proc Natl Acad Sci U S A*. 1989;86:6888-92.
11. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli. Curr Opin Biotechnol*. 1995;6:494-500.
12. Muhlrad D, Parker R. Premature translational termination triggers mRNA decapping. *Nature*. 1994;370:578-81.
13. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. 2000;28:292.
14. Chartier M, Gaudreault F, Najmanovich R. Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics.* 2012;28:1438-45.
15. Widmann M, Clairo M, Dippon J, Pleiss J. Analysis of the distribution of functionally relevant rare codons. *BMC Genomics.* 2008;9:207.
16. Thanaraj T, Argos P. Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci*. 1996;5:1973-83.
17. Goodluck U. ATGme: Open-source web application for rare codon identification and custom DNA sequence optimization. *BMC Bioinformatics*. 2015;16:303.
18. Theodosiou A, Promponas VJ. LaTcOm: a web server for visualizing rare codon clusters in coding sequences. *Bioinformatics*. 2012;28:591-2.
19. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008;9:40.
20. Kaplan W, Littlejohn TG. Swiss-PDB viewer (Deep view). *Brief Bioinform*. 2001;2:195-7.
21. DeLano WL. The PyMOL molecular graphics system. 2002.
22. Clarke TF, Clark PL. Rare codons cluster. *PLoS One.* 2008;3:e3412.
23. Varenne S, Baty D, Verheij H, Shire D, Lazdunski C. The maximum rate of gene expression is dependent in the downstream context of unfavourable codons. *Biochimie.* 1989;71:1221-9.
24. Varenne S, Lazdunski C. Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. *J Theor Biol.* 1986;120:99-110.
25. Hall RS, Fedorov AA, Xu C, Fedorov EV, Almo SC, Raushel FM. Three-dimensional structure and catalytic mechanism of cytosine deaminase. *Biochemistry.* 2011;50(22):5077-85.
26. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins.* 1997;28:405-20.
27. Dong H, Nilsson L, Kurland CG. Co-variation of trna abundance and codon usage inescherichia coliat different growth rates. *J Mol Biol.* 1996;260:649-63.
28. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction.

Nucleic Acids Res. 2007;35:3375-82.
29. Guex N, Peitsch M. Swiss-PdbViewer: a fast and easy-to-use PDB viewer for Macintosh and PC. Protein Data Bank Quaterly Newsletter. 1996;77(7).
30. Vriend G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph*. 1990;8:52-6, 29.
31. Tina K, Bhadra R, Srinivasan N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* 2007;35(Web Server issue):W473-6.
32. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010 ;31:455-61.
33. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, *et al*. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem.* 2009;30:2785-91.
34. OLBoyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform.* 2011;3:33.
35. Masood TB, Sandhya S, Chandra N, Natarajan V. CHEXVIS: a tool for molecular channel extraction and visualization. *BMC Bioinformatics.* 2015;16:119.
36. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*. 1995;8:127-34.
37. Shahbazi M, Haghkhah M, Rahbar MR, Nezafat N, Ghasemi Y. In Silico Sub-unit Hexavalent Peptide Vaccine Against an Staphylococcus aureus Biofilm-Related Infection. *Int J Pept Res Ther*. 2016;22:101-17.
38. Zamani M, Nezafat N, Negahdaripour M, Dabbagh F, Ghasemi Y. *In silico* evaluation of different signal peptides for the secretory production of human growth hormone in *E. coli. Int J Pept Res Ther*. 2015;21:261-8.
39. Fattahi M, Malekpour A, Mortazavi M, Safarpour A, Naseri N. The Characteristics of Rare Codon Clusters in the Genome and Proteins of Hepatitis C Virus; a Bioinformatics Look. *Middle East J Dig Dis.* 2014;6:214-27.
40. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* 2012;14:133-41.