Generation of three sample Never Born Protein (NBP) Sequences using the Partitioned Gray code representations of Universal Genetic Code combined with Toeplitz Matrix approach

In order to generate the first sample Never Born Protein sequence, let us choose the 2 x 2 order of the Sample Toeplitz Matrix (STM_1). Now, we have to choose the types of PTM_1, PTM_2, PTM_3 and PTM_4 matrices, which have to be associated with the 48 codon C_{83} and amino acid A_{83} matrices for the first and second rows. Here we choose the PTM_4 type for the first row and the PTM_3 type for the second row. Then we have to choose the number of the codon matrices from the 48 codon C_{83} and amino acid A_{83} matrices for the first and second rows in the form of a Toeplitz matrix to ensure the structural stability of the generated NBP sequence. We choose the numbers of the codon C_{83} matrices as 27 and 17 for the first row and as 17 and 27 for the second row and is given in the following Sample Toeplitz Matrix (STM_1),

$$STM_1 = \begin{bmatrix} 27 & 17 \\ 17 & 27 \end{bmatrix}. \qquad (1)$$

Then the Partitioned Gray code representations of Universal Genetic Code in the form of the codon C_{83} matrices 27 and 17 are retrieved from the PTM_4 type for the first row and 17 and 27 from the PTM_3 type for the second row and are arranged as given below.

27^{th} C_{83} matrix of PTM_4 — 17^{th} C_{83} matrix of PTM_4

gcc	gcu	acu	acc	auu	auc	guc	guu	->	cgc	ugc	uac	cac	uau	cau	cgu	ugu
gca	gcg	acg	aca	aug	aua	gua	gug	->	agc	ggc	gac	aac	gau	aau	agu	ggu
uca	ucg	ccg	cca	cug	cua	uua	uug	->	auc	guc	gcc	acc	gcu	acu	auu	guu
ucc	ucu	ccu	ccc	cuu	cuc	uuc	uuu	->	cuc	uuc	ucc	ccc	ucu	ccu	cuu	uuu
uaa	uag	cag	caa	cgg	cga	uga	ugg	->	aua	gua	gca	aca	gcg	acg	aug	gug
uac	uau	cau	cac	cgu	cgc	ugc	ugu	->	cua	uua	uca	cca	ucg	ccg	cug	uug
gac	gau	aau	aac	agu	agc	ggc	ggu	->	cga	uga	uaa	caa	uag	cag	cgg	ugg
gaa	gag	aag	aaa	agg	aga	gga	ggg	->	aga	gga	gaa	aaa	gag	aag	agg	ggg

cgc	ugc	uac	cac	cgu	ugu	uau	cau	->	gcc	gcu	acu	acc	guc	guu	auu	auc
agc	ggc	gac	aac	agu	ggu	gau	aau	->	gca	gcg	acg	aca	gua	gug	aug	aua
auc	guc	gcc	acc	auu	guu	gcu	acu	->	uca	ucg	ccg	cca	uua	uug	cug	cua
cuc	uuc	ucc	ccc	cuu	uuu	ucu	ccu	->	ucc	ucu	ccu	ccc	uuc	uuu	cuu	cuc
cga	uga	uaa	caa	cgg	ugg	uag	cag	->	gac	gau	aau	aac	ggc	ggu	agu	agc
aga	gga	gaa	aaa	agg	ggg	gag	aag	->	gaa	gag	aag	aaa	gga	ggg	agg	aga
aua	gua	gca	aca	aug	gug	gcg	acg	->	uaa	uag	cag	caa	uga	ugg	cgg	cga
cua	uua	uca	cca	cug	uug	ucg	ccg	->	uac	uau	cau	cac	ugc	ugu	cgu	cgc

17^{th} C_{83} matrix of PTM_3 — 27^{th} C_{83} matrix of PTM_3

The mRNA sequence of the NBP sequence is obtained simply by joining each row of the 4 codon matrices, as indicated by the arrow (->) symbol. The output may be written in FASTA format with the name of the sequence (e.g., NB_mRNA_SEQ_1) and number of bases (4 x 64 x 3 =) 768 in the header line, as given below.

```
>NB_mRNA_SEQ_1 768 bp
gccgcuacuaccauuaucgucguucgcugcuaccacuaucaucguugugcagcgacgaca
augauaguagugagcggcgacaacgauaauagugguucaucgccgccacugcuauuauug
aucgucgccaccgcuacuauuguuuccucuccucccccuucucuucuuucucuucucccc
ucuccucuuuuuuaauagcagcaacggcgaugauggauaguagcaacagcgacgauggug
uacuaucaucaccgucgcugcugucuauuaucaccaucgccgcuguuggacgauaauaac
aguagcggcggucgaugauaacaauagcagcgguggggaagagaagaaaaggagaggaggg
agaggagaaaaagagaagaggggcgcugcuaccaccguuguuaucaugccgcuacuacc
gucguuauuaucagcggcgacaacaguggugauaaugcagcgacgacaguagugaugaua
aucgucgccaccauuguugcuacuucaucgccgccauuauugcugcuacucuucucccc
cuuuuuucuccuuccucuccucccuucuuucuucuccgaugauaacaacgggugguagcag
gacgauaauaacggcgguaguagcagaggagaaaaaaggggggagaaggaagagaagaaa
ggagggaggagaauaguagcaacaaugguggcgacguaauagcagcaaugauggcggcga
cuauuaucaccacuguugucgccguacuaucaucacugcugucgucgc
```

Through transcription of the above sample mRNA sequence, the corresponding sample Never Born DNA sequence is obtained by as simply replacing all `u` with `t`. The output may be written in FASTA format with the name of the sequence (e.g., NB_DNA_SEQ_1) and number of bases in the header line, as given below.

```
>NB_DNA_SEQ_1 768 bp
gccgctactaccattatcgtcgttcgctgctaccactatcatcgttgtgcagcgacgaca
atgatagtagtgagcggcgacaacgataatagtggttcatcgccgccactgctattattg
atcgtcgccaccgctactattgtttcctctcctccccttctcttctttctcttctccccc
tctcctctttttaatagcagcaacggcgatgatggatagtagcaacagcgacgatggtg
tactatcatcaccgtcgctgctgtctattatcaccatcgccgctgttggacgataataac
agtagcggcggtcgatgataacaatagcagcggtgggaagagaagaaaaggagaggaggg
agaggagaaaaagagaagaggggcgctgctaccaccgttgttatcatgccgctactacc
gtcgttattatcagcggcgacaacagtggtgataatgcagcgacgacagtagtgatgata
atcgtcgccaccattgttgctacttcatcgccgccattattgctgctactcttctccccc
ctttttctccttcctctcctcccttctttcttctccgatgataacaacggtggtagcag
gacgataataacggcggtagtagcagaggagaaaaaaggggggagaaggaagagaagaaa
ggagggaggagaatagtagcaacaatggtggcgacgtaatagcagcaatgatggcggcga
ctattatcaccactgttgtcgccgtactatcatcactgctgtcgtcgc
```

From the above sample mRNA sequence, the sample Never Born Protein sequence is obtained by simply translating the triplet codons to their corresponding amino acids using the Universal Genetic Code. Alternatively, the amino acid A_{83} matrices 27 and 17 are retrieved from the *PTM_4*

type for the first row and 17 and 27 from the PTM_3 type for the second row and are arranged as given below.

27^{th} A_{83} matrix of PTM_4									17^{th} A_{83} matrix of PTM_4							
A	A	T	T	I	I	V	V	->	R	C	Y	H	Y	H	R	C
A	A	T	T	M	I	V	V	->	S	G	D	N	D	N	S	G
S	S	P	P	L	L	L	L	->	I	V	A	T	A	T	I	V
S	S	P	P	L	L	F	F	->	L	F	S	P	S	P	L	F
*	*	Q	Q	R	R	*	W	->	I	V	A	T	A	T	M	V
Y	Y	H	H	R	R	C	C	->	L	L	S	P	S	P	L	L
D	D	N	N	S	S	G	G	->	R	*	*	Q	*	Q	R	W
E	E	K	K	R	R	G	G	->	R	G	E	K	E	K	R	G

R	C	Y	H	R	C	Y	H	->	A	A	T	T	V	V	I	I
S	G	D	N	S	G	D	N	->	A	A	T	T	V	V	M	I
I	V	A	T	I	V	A	T	->	S	S	P	P	L	L	L	L
L	F	S	P	L	F	S	P	->	S	S	P	P	F	F	L	L
R	*	*	Q	R	W	*	Q	->	D	D	N	N	G	G	S	S
R	G	E	K	R	G	E	K	->	E	E	K	K	G	G	R	R
I	V	A	T	M	V	A	T	->	*	*	Q	Q	*	W	R	R
L	L	S	P	L	L	S	P	->	Y	Y	H	H	C	C	R	R

17^{th} A_{83} matrix of PTM_3 27^{th} A_{83} matrix of PTM_3

Now, the sample Never Born Protein sequence is obtained simply by joining each row of the 4 amino acid matrices as indicated by the arrow (->) symbol and leaving the STOP codons (*). The output may be written in FASTA format with the name of the sequence (e.g., NBP_SEQ_1) and number of amino acids [(4 x 64) - (4 x 3) = 256 - 12 =] 244 in the header line, as given below.

```
>NBP_SEQ_1 244 aa
AATTIIVVRCYHYHRCAATTMIVVSGDNDNSGSSPPLLLLIVATATIVSSPPLLFFLFSP
SPLFQQRRWIVATATMVYYHHRRCCLLSPSPLLDDNNSSGGRQQRWEEKKRRGGRGEKEK
RGRCYHRCYHAATTVVIISGDNSGDNAATTVVMIIVATIVATSSPPLLLLLFSPLFSPSS
PPFFLLRQRWQDDNNGGSSRGEKRGEKEEKKGGRRIVATMVATQQWRRLLSPLLSPYYHH
CCRR
```

In a similar way, we generate the second sample Never Born Protein sequence using the following Sample Toeplitz Matrix (STM_2) of order 3 x 3, as given by

$$STM_2 = \begin{bmatrix} 27 & 37 & 17 \\ 47 & 27 & 37 \\ 17 & 47 & 27 \end{bmatrix}. \quad (2)$$

We choose the first, second and third rows of the above matrix elements may be associated with the 48 codon C_{83} and amino acid A_{83} matrices of types PTM_4, PTM_3 and PTM_1, respectively. Retrieving the appropriate A_{83} matrices, we generate the resultant sample Never Born Protein sequence (NBP_ SEQ_2) with the number of amino acids [(9 x 64) – (9 x 3) = 576 – 27 =] 549, as given below.

```
>NBP_SEQ_2 549 aa
AATTIIVVGSNDNDGSRCYHYHRCAATTMIVVCRHYHYCRSGDNDNSGSSPPLLLLFLPS
PSFLIVATATIVSSPPLLFFVITATAVILFSPSPLFQQRRWLLPSPSLLIVATATMVYYH
HRRCCVITATAVMLLSPSPLLDDNNSSGGGRKEKEGRRQQRWEEKKRRGGRQQWRRGEKE
KRGGGRREEKKAATTVVIIGSNDGSNDGGSSDDNNAATTVVMICRHYCRHYCCRRYYHHS
SPPLLLLFLPSFLPSWRRQQSSPPFFLLVITAVITAVVIMAATTDDNNGGSSGRKEGRKE
VVIIAATTEEKKGGRRRQWRQFFLLSSPPQQWRRLLPSLLPSLLLLSSPPYYHHCCRRVI
TAVMTARCYHHRCYGGRRKEEKAATTIVVISGDNNSGDGGSSNDDNAATTIVVMIVATTI
VACCRRHYYHSSPPLLLLLFSPPLFSWRRQQSSPPLFFLLLSPPLLSLLLLPSSPYYHHR
CCRRQQRWVVIMTAATDDNNSGGSRGEKKRGEVVIITAATEEKKRGGRIVATTMVAFFLL
PSSPQQRWR
```

Similarly, we generate third sample Never Born Protein sequence using the following Sample Toeplitz Matrix (STM_3) of order 4 x 4, as given by

$$STM_3 = \begin{bmatrix} 45 & 35 & 25 & 15 \\ 15 & 45 & 35 & 25 \\ 25 & 15 & 45 & 35 \\ 35 & 25 & 15 & 45 \end{bmatrix}. \quad (3)$$

We choose the first, second, third and fourth rows of the above matrix elements may be associated with the 48 codon C_{83} and amino acid A_{83} matrices of types PTM_4, PTM_3, PTM_2 and PTM_1, respectively. Retrieving the appropriate A_{83} matrices, we generate the resultant sample Never Born Protein sequence (NBP_ SEQ_3) with the number of amino acids [(16 x 64) – (16 x 3) = 1024 – 48 =] 976, as given below.

```
>NBP_SEQ_3 976 aa
GEKRKRGEAVMTITAVATTAIVVIRRHHYYCCVATMTIVAEGRKRKEGSPPSLFFLRRQQ
WLSPLPLLSWRQRQSPPSLLLLLLPPSSLLWQRQRSLLPLPSLATTAMVVILLPPSSFFF
SPLPLFSYCRHRHYCQQRWRIMTTAAVVCYHRHRCYSFLPLPSFEKKERGGRIITTAAVV
GDNSNSGDAVITITAVDNNDSGGSSSNNDDGGVATITIVADGSNSNDGYHHYRCCRRRKK
```

EEGGRRHHCCYYGEKRGEKRAVMTAVITATTAVIIVRRQQWVATMVATIEGRKEGRKSPP
SFLLFLLPPLLSSLSPLLSPLWRQRQSPPSLLLLLLPPFFSSWQRQRSLLPSLLPATTAV
IMVSSNNGGDDGDNSGDNSAVITAVITDNNDGSSGRRKKGGEEVATIVATIDGSNDGSNY
HHYCRRCIMTTVVAAFSPLFSPLYCRHYCRHQQRRWIITTVVAACYHRCYHRSFLPSFLP
EKKEGRRGATTAIIVVRRHHCYYCGEKREKRGAVMTVITASPPSLLFFRRQQWVATMATI
VEGRKGRKESPPSLLLLLLPPLSSLLSPLSPLLWRQRQATTAIMVVLLPPFSSFWQRQRS
LLPLLPSYHHYRRCCRRKKGEEGVATIATIVDGSNGSNDQQRRWIMTTVAAVFSPLSPLF
YCRHCRHYEKKERRGGIITTVAAVCYHRYHRCSFLPFLPSDNNDSSGGSSNNGDDGGDNS
DNSGAVITVITAAVMTTAVIATTAVVIIRRHHYCCYGEKRRGEKEGRKKEGRSPPSFFLL
RRQQWVATMIVATWRQQRSPPSLLLLLLPPSLLSLSPLLLSPSLLPPSLLATTAVVIMLL
PPSFFSWQRRQSFLPPSFLEKKEGGRRIITTAVVACYHRRCYHAVITTAVIDNNDGGSSS
SNNDGGDGDNSSGDNDGSNNDGSYHHYCCRRRRKKEGGEVATIIVATYCRHHYCRQQWRR
IMTTAVVAFSPLLFSP

Thus, we have described the methodology involved in obtaining three sample NBP sequences using the Partitioned Gray code representations of Universal Genetic Code combined with a novel Toeplitz matrix approach.