Indian Journal of Biochemistry & Biophysics
Vol. 32, December 1995, pp. 417-423

Contextual constraints in the choice of synonymous codons*

A S Kolaskar†, Bharati Joshi & B V B Reddy‡
Bioinformatics, Distributed Information Center, University of Pune, Pune 411 007

Received 3 June 1995; revised 14 September 1995; accepted 18 September 1995

From EMBL Nucleotide Sequence Database, protein coding sequences of all *E. coli* and its DNA phages, were extracted using our computer programme. Same programme has been used to form a database of sequence of oligonucleotides of length 18 nucleotides on both sides of each of the 61 codons. From analysis of this database and study of variations in twist parameter (Tw) values, as an indicator of sequence dependent variations in B-DNA helix, a method is developed to fix the codon among the set of synonymous codons. The accuracy of the method was checked on enlarged data set by adding data from more prokaryotes. Our method assign the codon 85-90% times correctly if the selection has to be made between codons having different sequence in terms of R and Y. The accuracy of the method is somewhat lower when choice of the codon has to be made between codons having same codes in terms of R and Y. This study points out that the major factors which decide the choice of a codon from a set of synonymous codons are contextual constraints arising from flanking regions.

Size of the DNA sequence data is increasing very fast. One of the main reason for such increase in data is the impetus to genome projects and advancements in DNA sequencing techniques. Careful analysis of such data can provide insight into biological problems[1-4]. Various data analysis techniques and tools are being used to obtain useful information and patterns of sequences which are involved in biological functions. These approaches include analysis of data using simple statistical methods to more complex techniques based on artificial intelligence, neural nets, grammars of sublanguages etc.[5-8]. More often a 'null hypothesis' is made and data analysis is carried out to check its validity. DNA sequence data analysis has been used to develop methods to predict protein coding sequences[9,10], intron exon boundaries[11-16], transcription/translation initiation regions[2,17], promoter sequences[18,19] etc., which can be used to study certain specific biological functions or evolution of organisms. One of the problems which requires an indepth analysis is to find the rationale for the choice of a codon among a set of synonymous codons as more than one codon usually codes for a single amino acid and in a given cDNA or mRNA only particular codon is used for coding the amino acid. Further, various studies have pointed out that the occurrence of synonymous codons in protein coding regions is non-random. The non-random occurrence of a codon is suggested to be directly proportional to the percentage of tRNA contents in the organism[20]. Evolutionary drift has also been suggested as one of the factors for the non-random occurrences of codons[21]. DNA sequence data analysis has shown variations in codon usage from species to species[22]. However, implications of structure or neighbouring nucleotides on the choice of synonymous codons is not investigated to the best of our knowledge. The effect of the neighbouring nucleotides on the three dimensional structure of DNA is being understood only recently[23-26]. In other words the dependence of three dimensional structures on the sequence of nucleic acids has not been used to gain an insight into the problem of choice of synonymous codons. It was shown by Calladine and Dickerson[27,28] that the variation in DNA structure are directly related to purine (R) and pyrimidine (Y) sequences in nucleic acids. Similar rules at the individual nucleotide (A, T, G and C) level are not derived as few crystal structures of oligonucleotides, particularly AT containing sequences are available. Therefore, errors in calculated helix base pair parameter values for individual nucleotides are large. In this study we

*This paper is submitted in honour of Prof. D.P. Burma, the inspiring scientist, on the occasion of his 70th Birthday.

†Author for correspondence.
Tel.: 355039/350195; E-mail: kolaskar@bioinfo.ernet.in; FAX: (+91) 212 350087
‡*Present Address*: Centre for Cellular and Molecular Biology, Hyderabad 500 007.

have used carefully chosen cDNA sequence data from certain prokaryotes and Calladine and Dickerson type approach is applied to study sequence dependent variations in B-DNA double helix. The results presented here point out that DNA sequence around the codon under study is such that the helical base pair structural parameter values are different, in several cases, even for codons having same code in terms of R and Y but different code at nucleotide level.The accuracy of the method, developed and described below, to fix codon in a set of synonymous codons, having different codes in terms of R and Y is as high as 85-90%. These results suggest that contextual constraints play an important role in the choice of a codon among a set of synonymous codons.

Method

To pick up protein coding regions of prokaryotes automatically, from the EMBL nucleic acid sequence data, a program was written in C-language. The EMBL data bank was searched for the word 'Prokaryote' in the organism field and the 'CDS' in the feature table. Using information in CDS field, DNA sequence regions were extracted and checks were carried out to confirm the exact position of initiator (ATG, GTG) and terminator (TAA, TAG, TGA) codons. In addition, the non existence of terminators in the reading frame were also checked. Those protein coding cDNA sequences that passed the above mentioned checks were used to prepare 61 files, one file each for one type of codon. In each of these files sequence data of 18 nucleotides flanking a codon at (0, 0, 0) was extracted. Such sequences were translated into R and Y for analysis. Further studies were carried out only on data from *E. coli* and its DNA phages. No other prokaryotic DNA sequence data were used to prepare the weight matrix.

Preparation of weight matrix

The Callidine and Dickerson rules quantitate the changes in base pair parameters needed to relieve constraints in the major and minor groove of the ideal B-DNA helix arosing as a function of DNA sequence in terms of R and Y. We are aware that these are rough rules. Among these rules the twist angle parameter (Tw) values have the least errors according to these authors. Therefore, the change in the twist parameter values, using a *hexanucleotide* as the building block, were studied. The choice of hexanucleotide is based on trial and error and also because most of the known DNA helices do not contain more than twelve nucleotides per turn. It would have been ideal to consider a building block of ten nucleotides, which form one turn of the B-helix, but the possible number of twist parameter values being large, the statistical evaluation of variation in these values will become difficult at the present size of data set. Three overlapping tetranucleotides were assumed to form hexanucleotide and used to calculate Tw parameter values as shown in Table 1b. Twist angle parameter values suggested by Dickerson (Table 1a) were assigned to each of the three base pair doublets in a tetranucleotide. Tw values for each base were added to obtain sum as shown in Table 1b. Values of twist parameter for such overlapping hexanucleotide were calculated for every oligonucleotide in the data file having -18 to $+18$ region. Table 1c shows an example of actual calculations. Numerical values obtained in this fashion around each codon may contain errors and therefore only the signs (+, − and 0) associated with each of these numbers were extracted. These signs indicate whether the twist will be +ve, −ve, or there will be no change in the twist angle compared to the twist per nucleotide in the B-DNA structure. The patterns of three consecutive signs indicate local perturbations in the B-DNA helix per hexanucleotide unit. Therefore, at every position i, from -17 to -1 and $+1$ to $+17$ the occurrence of k patterns $+++$, $++-$, $++0$, etc. (k = 1, 27) were counted and normalized to obtain what is called the structural frequency $j_{S_{ik}}$

$$j_{S_{ik}} = \frac{j_{F_{ik}}}{j_{F'_{ik}}}$$

where $j_{F_{ik}}$ is the frequency of pattern k at position i for synonymous codon j

and $j_{F'_{ik}}$ is the frequency and pattern k at position i for the remaining codons in the set of synonymous codons for that Amino acid.

k varies from 1 to 27 over different sign combinations and i varies from -17 to -1 and $+1$ to $+17$.

In order to further reduce errors due to statistical variations, the weight values $j_{W_{ik}}$ shown below were assigned to normalized frequency values $j_{S_{ik}}$, the indicator of change in B-DNA helix. These ranges of $j_{S_{ik}}$ and corresponding $j_{W_{ik}}$ are given below:

$$j_{S_{ik}} > 1.75 \Rightarrow j_{W_{ik}} = 5$$

$$1.50 < j_{S_{ik}} < 1.75 \Rightarrow j_{W_{ik}} = 4$$

$$1.25 < j_{S_{ik}} < 1.50 \Rightarrow j_{W_{ik}} = 3$$

$0.80 < j_{S_{ik}} < 1.25 \Rightarrow j_{W_{ik}} = 2$

$0.64 < j_{S_{ik}} < 0.80 \Rightarrow j_{W_{ik}} = 1$

$0.50 < j_{S_{ik}} < 0.64 \Rightarrow j_{W_{ik}} = 0$

$j_{S_{ik}} < 0.50 \Rightarrow j_{W_{ik}} = -1$

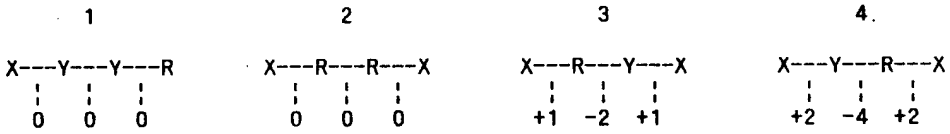
Thus for each codon, one weight matrix of the order (37×27) was prepared. Such 59 weight matrices were obtained. Single codons code for Trp and Met and thus these codons were excluded from the present study.

Fixing of the codon in a set of synonymous codons

The codon being fixed is assumed to be at (0, 0, 0) position. The nucleotide sequence of 18 re-

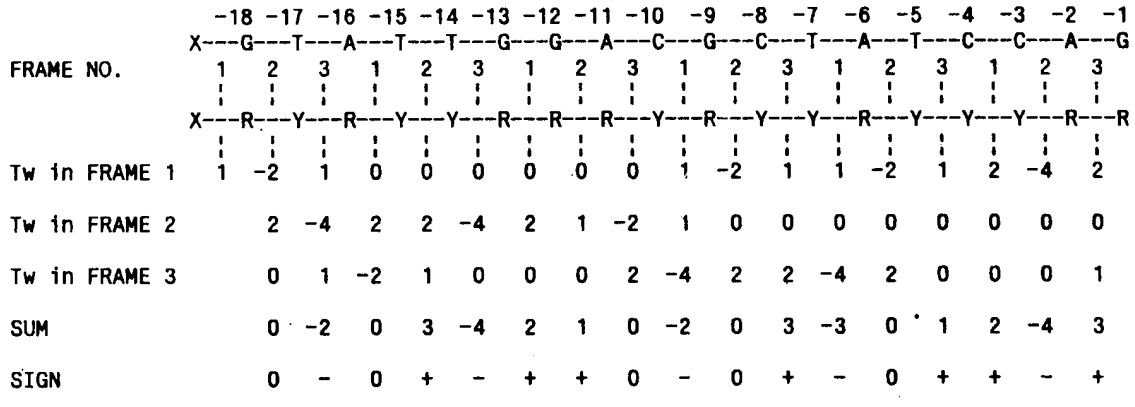
Table 1

1a: Twist parameter values as assigned by Dickerson (1983).



1b: Calculation of change in Twist angle parameter value using Dickerson (1983) rules for hexanucleotide as building block.

1c: Calculation of change in twist angle parameter value for a subsequence of 18 nucleotides by using overlapping hexanucleotides as building blocks. Only 5' - flanking region of the codon is shown.



sidues on both sides of the codon is extracted and translated to R and Y. Using a hexanucleotide as the building block and the procedure described above, the signs associated with twist parameter values were obtained for every bond -17 to -1 and $+1$ to $+17$. A pattern of (+, − and 0) for every overlapping hexanucleotide was obtained. By taking into consideration the hexanucleotide position i and the pattern k weight values were obtained from each of the j type of synonymous codon weight matrix to calculate, j_{W_k}, the algebraic sum of weights,

$$j_{W_k} = \sum_{i=-17}^{-1} j_{W_{ik}} + \sum_{i=1}^{17} j_{W_{ik}}$$

Thus the code for codon at (0, 0, 0) was assumed X-X-X. Among j_{W_k}, the one which has maximum value was picked up and corresponding jth codon was assigned at center.

Results and Discussion

As described in the method, 59 weight matrices were prepared for twist parameter values for each type of codon. It should be noted here that during generation of the weight matrix the central codon sequence (0, 0, 0), in terms of R and Y was considered, as these weight matrices are derived from experimental protein coding cDNA sequence data. Weight values at a specific position in a set of synonymous codons is directly related to local sequence of the hexanucleotide in terms of purine (R) and pyrimidine (Y). Table 2 shows weight values for four synonymous codons GGA, GGG, GGC, GGT respectively for amino acid Glycine. Table 2 points out that though codons such as GGA and GGG are indistinguishable when translated to R and Y, weight values $j_{W_{ik}}$ are different for patterns such as $-0+$, $-++$, $++0$ etc. at certain positions i. It is also clear from this table that weight matrices are quite different for RRY and RRR in a set of synonymous codons. A study of the variations in the occurrence of oligonucleotide patterns in these flanking sequences and measured by χ^2 values also point out that positions where $j_{W_{ik}}$ values are different the variability in terms of sequence is also high. However, χ^2 values can not provide information regarding patterns of nucleotides or structures being preferred around a particular codon. On the other hand, Tw, a twist angle variation parameter, being directly dependent on the local base pair sequence of oligonucleotide, provides a good measure to study the sequence dependent structural change that may occur around a central codon, in a set of synonymous codons. Though several studies are carried out on sequence dependent DNA structure in recent years which are very useful and important in understanding the flexibility in DNA structure, the attempt made here points out that even the weight matrices generated using rough parameters such as those of Calladine and Dickerson, provide resolving power to assign a codon among a set of synonymous codon given flanking nucleotide sequence. This study also points out an urgent need to develop more accurate sequence dependent structural parameters which can be used to analyse large DNA sequence data to gain an insight in biological problems. Patterns $---$, $--0$, $--+$, $-0-$, -00, $0--$, $00-$, $+--$, and $+++$ will not occur in any flanking sequence at any position, as can be seen from rules given in Table 1 and its use in Table 2. To check the usefulness of the weight matrices to assign a codon from a set of synonymous codons, for given flanking sequence, we applied these weight matrices to a large data set. The data set included not only those sequences used to derive the weight matrices, but also, sequences from other prokaryotes such as *Anabaena*, *Aerogenes*, *Typhimurium* etc. It must be mentioned here that, to fix the codon X-X-X is assumed at (0, 0, 0) position, as its sequence is unknown. Results obtained on this enlarged data set are given in Table 3. The reference codons in the table are along horizontal direction, while the codons fixed using our method are given column wise. For example, in the data set considered, Gln amino acid occurs 4707 times. In this data set CAA and CAG occurs 1667 and 3040 times respectively. The algorithm described above could fix CAA correctly 1177 times and assigned CAG 490 times in place of CAA. Similarly, 1824 times CAG was correctly fixed but 1216 times CAA was assigned in place of CAG (see Table 3). Please note that CAA and CAG have the same sequence YRR and still we could resolve CAA with 70% accuracy and CAG with 60% accuracy. These results point out that patterns of R and Y, in flanking region as studied through twist parameter are different for the codons CAA and CAG at critical places. In fact the data given in Table 3 on the fixing of the codons of Gly, Ala, Val, Pro, Thr, which have four synonymous codons point out clearly that the matrices derived have the ability to fix the codon, from a set of synonymous codons, with high accuracy provided the codon sequence in terms of R and Y is different. Results are very similar when there are six synonymous codons for amino acids such as Leu, Ser, Arg or amino acid

Ile having three synonymous codons. Such high accuracy indicates that the major factor in the choice of a codon, of a particular amino acid, is the sequence of flanking region and thus its structure. It may be further mentioned that the structural frequency (js_{ik}), values are normalized only for a particular set of synonymous codons and not for all 59 codons. Because of this normalization procedure, the weight values are consistent only for the particular set and are comparable only within that set. In other words weight matrices of GGA, GGG, GGC, GGT codons for an amino acid Gly, can be compared with each other but its comparison with the weight matrices of codons

TABLE 2 : Weight Matrices of Glycine codons GGA, GGG, GGC, GGT respectively

	- 0 +	- + -	- + 0	- + +	0 - 0	0 - +	0 0 0	0 0 +	0 + -
-17	1 3 2 2	-	-	-	2 1 2 2	-	2 2 1 2	-	-
-16	-	-	-	-	-	-	1 3 1 3	4 2 1 2	-
-15	-	-	-	1 2 2 2	3 2 0 3	-	1 4 0 4	-	-
-14	-	-	-	1 2 2 2	2 2 2 1	-	2 3 0 3	-	-
-13	-	-	-	3 2 1 2	1 2 2 2	1 2 3 2	2 4 2 2	2 2 1 3	-
-12	1 2 2 2	-	-	-	2 2 4 1	2 1 2 2	2 3 1 2	-	-
-11	3 2 2 1	-	-	1 2 1 3	2 4 2 2	1 2 2 2	3 2 2 2	5 2 1 2	-
-10	1 2 2 2	2 1 2 2	-	2 3 2 2	2 2 2 3	2 2 1 2	-	4 2 2 2	4 2 2 2
-9	-	-	-	3 2 2 2	2 4 2 2	-	3 2 2 2	2 2 2 3	1 3 2 2
-8	-	2 1 2 2	-	2 1 2 2	1 2 2 2	3 2 2 2	2 2 3 1	4 2 2 1	1 2 2 2
-7	-	-	-	2 2 1 3	-	2 3 2 2	2 2 3 0	1 1 2 3	2 2 2 2
-6	-	-	-	2 0 3 2	2 3 2 1	2 3 2 2	2 1 3 1	2 3 2 2	2 2 2 2
-5	2 3 2 1	-	1 2 1 4	-	1 3 2 2	-	2 1 3 1	2 3 2 2	-
-4	-	-	-	2 3 2 2	-	-	-	3 0 4 1	-
-3	-	3 2 2 2	-	2 1 2 2	5 -1 -1 4	2 2 4 1	2 2 1 3	0 1 2 2	2 3 1 3
-2	5 -1 1 2	1 1 2 2	3 3 2 1	-1 -1 2 4	0 0 1 5	2 4 2 2	2 2 1 2	-1 3 5 -1	-
-1	1 2 1 3	2 2 0 3	5 5 -1 -1	-1 -1 4 5	-	-	5 5 -1 -1	-1 -1 4 5	5 -1 0 4
1	-1 -1 4 4	0 2 3 2	-1 -1 4 4	2 1 2 2	5 5 -1 -1	5 5 -1 -1	5 5 -1 -1	5 5 -1 -1	5 5 -1 -1
2	4 2 0 2	4 5 -1 2	3 5 1 1	3 5 0 -1	5 5 -1 -1	5 5 -1 -1	5 3 -1 0	5 4 0 0	-1 -1 4 3
3	-	-	2 3 2 2	1 3 2 2	2 1 2 2	2 -1 3 2	5 2 2 1	3 1 2 2	3 2 2 2
4	2 2 2 3	1 2 2 2	2 1 2 2	-	1 2 2 2	1 2 2 2	4 2 2 1	3 2 2 2	-
5	-	-	1 3 2 2	-	1 5 2 2	-	5 2 1 2	2 3 2 1	3 2 2 2
6	1 2 2 2	-	-	2 2 2 1	2 2 3 2	-	5 2 1 2	4 1 2 2	-
7	-	-	-	2 2 1 2	1 0 4 1	-	2 3 -1 4	3 2 2 2	1 2 2 3
8	1 2 2 2	-	2 2 3 1	2 1 2 2	2 1 2 2	-	5 4 -1 1	2 1 1 4	-
9	-	-	-	-	-	-	5 1 1 2	3 5 0 2	2 1 2 3
10	-	-	-	3 2 1 2	3 3 4 -1	-	5 2 1 2	-	2 3 2 1
11	-	-	1 1 2 3	-	-	-	3 2 2 2	3 1 2 2	-
12	2 3 3 1	-	-	-	1 2 2 2	-	-	-	3 2 2 2
13	1 2 2 2	-	-	3 2 2 2	3 -1 2 2	2 2 3 1	-	2 1 2 2	-
14	3 2 2 2	-	2 2 1 2	4 2 2 1	2 3 2 1	-	2 2 1 3	0 3 2 2	-
15	-	-	-	2 1 2 2	-	-	-	2 2 0 3	-
16	-	1 0 3 2	5 5 -1 2	0 -1 3 2	0 -1 4 4	3 2 1 1	-	1 3 0 3	-
17	-1 -1 5 2	3 4 0 2	-1 -1 5 3	5 5 -1 -1	-1 -1 5 2	5 5 -1 0	-1 -1 3 4	5 5 -1 -1	3 3 -1 2

	0 + 0	0 + -	+ - 0	+ - +	+ 0 -	+ 0 0	+ 0 +	+ + -	+ + 0
-17	-	-	3 2 2 1	-	-	-	-	-	-
-16	2 3 1 2	2 3 2 2	-	-	-	2 2 1 3	2 1 1 2	-	3 0 2 2
-15	-	5 2 2 1	-	-	-	3 2 2 1	-	-	2 3 2 1
-14	3 2 1 2	2 1 3 2	1 2 2 2	3 2 2 2	1 2 2 2	-	3 2 2 2	-	2 1 2 3
-13	4 2 2 1	2 1 2 2	-	-	2 1 2 2	-	3 2 2 2	-	-1 2 2 2
-12	2 3 1 2	2 1 3 2	4 2 2 1	-	2 3 2 2	5 0 2 2	3 2 2 2	3 2 2 2	3 3 0 2
-11	2 3 2 2	2 2 1 3	1 2 2 3	-	2 2 1 2	-	-	2 2 2 1	0 2 3 2
-10	2 2 3 1	-	1 2 2 2	-	2 3 2 2	3 1 2 2	1 4 2 2	2 2 1 3	1 0 2 4
-9	3 2 1 2	3 1 3 1	-	-	-	2 2 3 1	1 2 3 2	-	2 4 2 2
-8	1 2 2 2	3 2 2 1	-	-	-	-	2 1 2 2	3 1 2 2	3 2 2 1
-7	-	3 3 1 2	1 1 2 2	-	2 3 2 1	-	-	-	-
-6	1 2 2 2	1 5 2 2	2 3 3 1	-	-	-	-	2 2 1 2	2 2 2 3
-5	-	0 1 3 2	-	-	-	4 2 2 2	-	-	0 1 5 1
-4	1 2 3 1	5 5 1 -1	-	-	2 2 4 1	0 1 2 4	1 2 2 5	1 1 3 2	3 5 2 1
-3	2 0 4 1	2 -1 2 2	5 -1 2 2	-	2 4 2 2	3 4 0 2	1 0 0 2	5 5 1 0	-
-2	-1 -1 5 3	-1 5 4 -1	2 4 0 3	3 3 1 2	-	-	5 5 2 2	2 0 2 2	-1 5 -1 5
-1	-1 5 3 1	5 2 3 -1	2 -1 2 5	2 1 2 2	2 3 2 2	5 5 -1 -1	-1 -1 -1 3	1 1 3 2	-1 3 2 4
1	5 5 -1 -1	-1 -1 -1 -1	-1 -1 2 5	-	5 5 -1 -1	-1 -1 3 4	-1 -1 3 3	-1 -1 3 5	1 1 1 4
2	0 0 2 3	-1 -1 4 5	-1 0 4 2	0 1 2 2	2 -1 2 2	1 -1 3 2	-1 -1 3 5	4 2 2 2	-1 -1 5 3
3	3 2 1 2	2 2 0 3	2 2 2 3	-	1 2 2 2	1 3 2 2	2 3 2 1	0 4 2 2	4 5 2 -1
4	3 2 2 2	1 2 2 2	-	-	2 3 2 2	2 3 2 2	4 3 2 2	-	0 5 1 2
5	1 1 1 5	-1 3 2 2	1 2 2 3	-	-	5 0 2 2	-	-	1 2 3 2
6	2 3 2 2	1 1 3 2	2 3 2 2	-	-	0 3 0 4	1 2 0 3	0 3 2 2	5 -1 1 2
7	3 1 2 2	2 4 1 2	1 2 2 2	-	-	4 2 2 2	-	2 2 3 1	0 3 2 2
8	-	-	2 2 2 1	-	-	2 2 1 3	3 1 1 2	2 3 1 2	3 2 2 2
9	-	2 2 3 1	-	-	2 2 2 1	-	1 3 2 2	-	0 2 3 2
10	3 2 1 2	-	2 2 1 2	-	-	-	2 2 2 1	-	-
11	-	2 1 2 2	1 3 3 1	-	-	2 3 1 2	2 2 1 1	3 2 2 2	2 3 1 2
12	2 1 2 2	3 2 2 2	-	-	2 2 2 1	2 1 2 2	0 1 2 5	2 1 2 2	4 3 2 0
13	1 2 3 1	-	3 2 2 1	-	-	1 3 2 2	-	-	1 2 2 2
14	1 2 2 2	1 5 2 1	-	-	-	2 2 1 3	0 1 1 2	-	-
15	-	-	-	-	-	2 3 1 3	1 2 1 2	-	-
16	1 2 0 5	5 5 -1 1	-1 -1 5 2	3 3 2 2	-	-	2 1 2 3	1 0 3 2	5 4 -1 1
17	-1 -1 5 2	-1 -1 5 -1	-1 -1 5 2	-	5 5 -1 2	-1 -1 6 5	5 5 2 -1	2 1 2 2	-1 -1 5 -1

Note : positions where same weight values are assigned in all four matrices are not shown in the table and indicated by -.

Table 3: Fixing of codon in a set of synonymous codon using derived weight matrices & flanking regions.
Note :High accuracy in fixing codons having different codes in terms of R and Y.

AGN

REF-	AAC	AAT
AAC	1662	864
AAT	974	1493
TOT	2636	2157

ASP

REF-	GAC	GAT
GAC	1157	1169
GAT	1321	2344
TOT	2478	3513

CYS

REF-	TGC	TGT
TGC	443	153
TGT	228	418
TOT	671	571

GLU

REF-	GAA	GAG
GAA	2363	757
GAG	1999	1574
TOT	4362	2330

GLN

REF-	CAA	CAG
CAA	1177	1216
CAG	490	1824
TOT	1667	3040

HIS

REF-	CAC	CAT
CAC	647	396
CAT	448	874
TOT	1095	1270

TYR

REF-	TAT	TAC
TAT	1396	579
TAC	615	937
TOT	2011	1516

PHE

REF-	TTT	TTC
TTT	1619	705
TTC	641	1394
TOT	2360	2099

LYS

REF-	AAA	AAG
AAA	2468	878
AAG	1513	914
TOT	3981	1792

ILE

REF-	ATA	ATT	ATC
ATA	628	298	318
ATT	68	2148	798
ATC	46	782	1562
TOT	742	3228	2678

GLY

REF-	GGA	GGG	GGC	GGT
GGA	728	401	335	368
GGG	314	618	312	315
GGC	68	52	1732	1223
GGT	32	51	846	1260
TOT	1142	1122	3225	3166

ALA

REF-	GCA	GCG	GCC	GCT
GCA	1330	841	79	48
GCG	764	2158	99	170
GCC	41	35	1631	706
GCT	18	67	658	1479
TOT	2153	3101	2467	2403

VAL

REF-	GTA	GTG	GTC	GTT
GTA	787	1067	31	99
GTG	568	1411	59	144
GTC	15	39	956	691
GTT	23	37	519	1351
TOT	1393	2554	1565	2285

PRO

REF-	CCA	CCG	CCC	CCT
CCA	518	755	30	83
CCG	388	1257	29	99
CCC	68	95	355	216
CCT	57	132	133	597
TOT	1031	2239	547	995

SER

REF-	TCT	TCC	TCG	TCA	AGT	AGC
TCT	733	307	39	25	42	48
TCC	377	511	22	19	57	78
TCA	87	80	250	677	78	135
TCG	132	72	505	210	43	46
AGT	53	59	41	47	458	510
AGC	58	41	30	53	237	749
TOT	1440	1080	887	1031	915	1566

LEU

REF-	CTA	CTG	CTC	CTG	TTA	TTG
TTA	72	666	76	53	561	331
TTG	44	442	84	76	242	449
CTT	5	50	327	670	6	17
CTC	11	105	535	335	38	26
CTA	208	996	70	132	224	185
CTG	184	2913	85	152	418	304
TOT	524	5172	1177	1418	1479	1312

THR

REF-	ACA	ACG	ACC	ACT
ACA	741	341	83	63
ACG	255	969	170	113
ACC	27	35	1209	448
ACT	26	36	942	820
TOT	1049	1381	2404	1444

ARG

REF-	AGA	AGG	CGT	CGC	CGG	CGA
AGA	315	71	130	122	39	49
AGG	70	121	96	85	46	25
CGT	4	5	996	730	37	37
CGC	0	4	713	947	44	30
CGG	15	5	475	480	396	88
CGA	7	2	378	378	160	269
TOT	411	208	2788	2742	722	498

for amino acid Ser—TCT, TCC, TCA, TCG, AGT, AGC, will not be proper. Therefore this approach will not be useful to fix a codon among 61 possible codons which code for proteinous amino acids. This is mentioned here to avoid any confusion regarding suitability and applicability of our method to choose any codon. Thus factors such as tRNA contents, and evolutionary drifts will play a role in the selection of the amino acid, but the contextual constraints are probably the single most important factor, which decides the choice of the codon in a set of synonymous codons. Context dependent synonymous codon choice particularly in highly expressed genes is suggested in other studies also[29,30]. The method described here may find its use in carrying out back translation without taking into consideration the codon usage Table of a particular organism. This method will be particularly useful to design oligonucleotide probes for hybridization and other studies, when only the amino acid sequence is known and very little of the genome sequence of the organism is studied.

The analysis carried out and results presented, thus point out a rationale for the choice of a codon from a synonymous codons set. The rationale is contextual constraints arising out of sequence dependent structure of DNA. The study can be extended for Eukaryotes by preparing similar weight matrices. The matrices can be refined as and when more data from single crystal structures of oligonucleotides become available. Studies to prepare helical base pair parameters at individual nucleotide level which can be used in above mentioned studies are in progress. Thus if DNA sequence is metaphorically considered as a language of cell then this study points out the importance of semantics in this DNA language.

Acknowledgement

We acknowledge the financial support from Department of Biotechnology, Govt. of India (New Delhi).

References

1 Nussinov R (1981). *J Biol Chem*, 256, 8458-8462.
2 Staden R & McLachian A D (1982), *Nucleic Acids Res*, 10, 141-156.
3 Heijne Gunnar von (1987), *Sequence Analysis in Molecular Biology*, Treasure trove or Trivial Pursuit (Academic Press, Inc.).
4 Gribskov M & Degvereux J (1991), *Sequence Analysis Primer* (McMillan Publishers, Stockton Press).
5 Stormo G D, Schneider T D, Gold L & Ehrenfeucht A (1982), *Nucleic Acds Res*, 10, 2997-3011.
6 Lapedes A, Barnes C, Burks C, Farber R & Sirotkin K (1990), *Computers and DNA, SFI studies in science of complexity* (Bell G & Marr T, eds.), Vol. 7, pp. 157-182, Addison-Wesley, Reading, Massachusetts.
7 O'Neill M C (1991), *Nucleic Acids Res*, 19, 313-318.
8 Demeler B & Zhou G (1991), *Nucleic Acids Res*, 19, 1593-1599.
9 Fickett J W (1982), *Nucleic Acids Res*, 10, 5303-5318.
10 Kolaskar A S & Reddy B V B (1985), *Nucleic Acids Res*, 13, 185-194.
11 Naora H & Deacon N J (1982), *Proc Natl Acad Sci, USA*, 79, 6196-6200.
12 Senapathy P (1986), *Proc Natl Acad Sci, USA*, 83, 2133-2137.
13 Senapathy P (1988), *Proc Natl Acad Sci, USA*, 85, 1129-1133.
14 Konopka A K, Smythers G W, Owens J & Maizel J V Jr (1987), *Gene Anal Techn*, 4, 63-74.
15 Hawkins J D (1988), *Nucleic Acids Res*, 16, 9893-9905.
16 Reddy B V B & Pandit M W (1995), *J Biomolec Struc & Dyn*, 12, 785-801.
17 Staden R (1989), *CABIOS*, 4, 53-60.
18 Bucher P & & Trifonov E N (1986), *Nucleic Acids Res*, 14, 10009-10026.
19 Nussinov R, Owens J & Maizel J V (1986), *Biochim Biophys Acta*, 866, 109-119.
20 Ikemura T (1985), *Mol Biol Evol*, 2, 13-34.
21 Brenner S (1988), *Nature, London*, 334, 528-530.
22 Steinberger C (1987), *J Theor Biol*, 124, 89-95.
23 Bansal M & Sasisekharan V (1990), *Theoretical Chemistry and Molecular Biophysics* (Beveridge D L & Lavery R, eds.) pp. 329-341, Adenine Press, New York.
24 Bolshoy A, McNamara P, Harrington R E & Trifonov E N (1991), *Proc Natl Acad Sci, USA*, 88, 2312-2316.
25 Zhurkin V B, Gorin A A, Charakchyan A A & Ulyanou N B (1990), *Theoretical Chemistry and Molecular Biophysics* (Beveridge D L & Lavery R, eds.), pp. 411-431, Adenine Press, New York.
26 Lavery R (1994), *Computational Biology*, JAI Press Inc., Vol. 1, 69-145.
27 Calladine C R (1982), *J Mol Biol*, 161, 343-352.
28 Dickerson R E (1983), *J Mol Biol*, 166, 419-441.
29 Yarus M & Folley L S (1985), *J Mol Biol*, 182, 529-540.
30 Folley L S & Yarus M (1989), *J Mol Biol*, 209, 359-378.