Contextual constraints in the choice of synonymous codons*

A S Kolaskar1, Bharati Joshi & B V B Reddy1
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 problems1-4. 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 sequences9,10, intron exon boundaries11-16, transcription/translation initiation regions17, promoter sequences18,19 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 organism20. Evolutionary drift has also been suggested as one of the factors for the non-random occurrences of codons21. DNA sequence data analysis has shown variations in codon usage from species to species22. 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 recently23-26. 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 Dickerson27,28 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.
1Author for correspondence.
Tel.: 355039/350195; E-mail: kolaskar@bioinfo.ernet.in; FAX: (+91) 212 350087
2Present Address: Centre for Cellular and Molecular Biology, Hyderabad 500 007.
have used carefully chosen cDNA sequence data from certain prokaryotes and Callidine 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 arising 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 +++, ++−, +−+, +++, etc. (k = 1, 27) were counted and normalized to obtain what is called the structural frequency \( j_{sk} \)

\[
\hat{j}_{sk} = \frac{j_{sk}}{J_{sk}}
\]

where \( j_{sk} \) is the frequency of pattern k at position i for synonymous codon j and \( j_{sk} \) 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 \( j_{wk} \) shown below were assigned to normalized frequency values \( j_{sk} \), the indicator of change in B-DNA helix. These ranges of \( j_{sk} \) and corresponding \( j_{wk} \) are given below:

\[
\begin{align*}
j_{sk} > 1.75 & \Rightarrow j_{wk} = 5 \\
1.50 < j_{sk} < 1.75 & \Rightarrow j_{wk} = 4 \\
1.25 < j_{sk} < 1.50 & \Rightarrow j_{wk} = 3
\end{align*}
\]
0.80 ≤ j_{wu} < 1.25 ⇒ j_{wu} = 2
0.64 ≤ j_{wu} < 0.80 ⇒ j_{wu} = 1
0.50 ≤ j_{wu} < 0.64 ⇒ j_{wu} = 0
j_{wu} < 0.50 ⇒ j_{wu} = -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).

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X----Y---Y---R</td>
<td>X---R---R---X</td>
<td>X---R---Y---X</td>
<td>X---Y---R---X</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>+1</td>
<td>-2</td>
<td>+1</td>
<td>+2</td>
</tr>
<tr>
<td>-4</td>
<td>+2</td>
<td>+2</td>
<td>+2</td>
</tr>
</tbody>
</table>

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

A----G----T----A----T----T
X----R----Y----R----Y----X

FRAME 1: 1 -2 1
FRAME 2: 2 -4 2
FRAME 3: 1 -2
SUM: +1 0 -2 0
SIGN: 0 0

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.

-18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1
X----G----T----A----T----T----G----G----A----C----G----C----T----A----T----C----C----A----G
FRAME NO. 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
X----R----Y----R----Y----R----R----Y----R----Y----R----Y----R----Y----Y----R----R----R
Tw in FRAME 1: 1 -2 1 0 0 0 0 0 1 -2 1 1 -2 1 2 -4 2
Tw in FRAME 2: 2 -4 2 2 -4 2 1 -2 1 0 0 0 0 0 0 0 0 0 0
Tw in FRAME 3: 0 1 -2 1 0 0 0 2 -4 2 2 -4 2 0 0 0 1
SUM: 0 -2 0 3 -4 2 1 0 -2 0 3 -3 0 1 2 -4 3
SIGN: 0 - 0 + - + 0 - 0 + - 0 + - +
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,

\[
\sum_{i=-17}^{-1} j_{w_k} + \sum_{i=1}^{17} j_{w_k}
\]

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 \( j_{w_k} \) 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 \( \chi^2 \) values also point out that positions where \( j_{w_k} \) values are different the variability in terms of sequence is also high. However, \( \chi^2 \) values can not provide information regarding patterns of nucleotides or structures being preferred around a particular codon. On the other hand, T, 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 analyze large DNA sequence data to gain an insight in biological problems. Patterns −−−, −−−0, −−−+, −0−0, −0−+, 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 *Anaabaena*, *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 ($k_0$), 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, GTG 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, GTG respectively |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $-0$ $+$ $-0$ $+0$ $-+$ $0+0$ $0-0$ $-0-0$ $+0+0$ $+0-0$ $-0+$ $0-+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ $+0+$ $-0+$ $0+0$ |
Table 3: Fixing of codon in a set of synonymous codon using derived weight matrixes & flanking regions.

Note: High accuracy in fixing codons having different codes in terms of R and Y.

<table>
<thead>
<tr>
<th>REF-CAA</th>
<th>CTS</th>
<th>GLY</th>
<th>HIS</th>
<th>TYR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAC</td>
<td>1062 864</td>
<td>1157 1109</td>
<td>443 153</td>
<td>2363 757</td>
</tr>
<tr>
<td>AAT</td>
<td>974 1493</td>
<td>1231 2344</td>
<td>228 418</td>
<td>1599 1574</td>
</tr>
<tr>
<td>TOT</td>
<td>2638 2157</td>
<td>2478 3513</td>
<td>671 571</td>
<td>4392 2330</td>
</tr>
<tr>
<td>PHE</td>
<td>LYS</td>
<td>ILE</td>
<td>GLY</td>
<td>ALA</td>
</tr>
<tr>
<td>REF-AAA</td>
<td>AAG</td>
<td>ATT</td>
<td>ATT</td>
<td>GCC</td>
</tr>
<tr>
<td>TTC</td>
<td>1819 705</td>
<td>2468 878</td>
<td>828 298</td>
<td>541 335</td>
</tr>
<tr>
<td>TTC</td>
<td>1294</td>
<td>1513 914</td>
<td>88 2148</td>
<td>798 314</td>
</tr>
<tr>
<td>TTC</td>
<td>382 1592</td>
<td>66 52</td>
<td>1732 1223</td>
<td>41 35</td>
</tr>
<tr>
<td>TOT</td>
<td>409 2959</td>
<td>3981 1792</td>
<td>742 3228</td>
<td>2678 1142</td>
</tr>
<tr>
<td>VAL</td>
<td>PRO</td>
<td>SER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REF-GTA</td>
<td>GTG</td>
<td>GTC</td>
<td>GTT</td>
<td>CCA</td>
</tr>
<tr>
<td>GTC</td>
<td>787 1087</td>
<td>31 93</td>
<td>518 155</td>
<td>30 83</td>
</tr>
<tr>
<td>GTG</td>
<td>588 1411</td>
<td>59 144</td>
<td>388 1257</td>
<td>29 99</td>
</tr>
<tr>
<td>GTC</td>
<td>15 39</td>
<td>956 691</td>
<td>66 95</td>
<td>355 216</td>
</tr>
<tr>
<td>GTT</td>
<td>23 37</td>
<td>519 1351</td>
<td>57 132</td>
<td>133 597</td>
</tr>
<tr>
<td>TOT</td>
<td>1393 2554</td>
<td>1595 2285</td>
<td>1031 2239</td>
<td>547 995</td>
</tr>
<tr>
<td>LEU</td>
<td>THR</td>
<td>ARG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REF-CTA</td>
<td>CTG</td>
<td>CTC</td>
<td>CTG</td>
<td>TTA</td>
</tr>
<tr>
<td>TTA</td>
<td>72 665</td>
<td>76 53</td>
<td>561 331</td>
<td>341 43</td>
</tr>
<tr>
<td>TTG</td>
<td>44 442</td>
<td>84 75</td>
<td>242 449</td>
<td>255 999</td>
</tr>
<tr>
<td>CTC</td>
<td>5 50</td>
<td>327 670</td>
<td>6 17</td>
<td>35 1209</td>
</tr>
<tr>
<td>CTC</td>
<td>11 105</td>
<td>535 335</td>
<td>38 26</td>
<td>36 942 820</td>
</tr>
<tr>
<td>CTA</td>
<td>208 998</td>
<td>70 132</td>
<td>224 165</td>
<td>208 998</td>
</tr>
<tr>
<td>CTG</td>
<td>184 1031</td>
<td>85 152</td>
<td>418 304</td>
<td>208 998</td>
</tr>
<tr>
<td>TOT</td>
<td>524 5172</td>
<td>1177 1418</td>
<td>1479 1312</td>
<td>1049 1381</td>
</tr>
</tbody>
</table>
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 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. 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 Xukaryotes 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
3 Hejine Gunnar von (1987), Sequence Analysis in Molecular Biology, Treasure trove or Trivial Pursuit [Academic Press, Inc.]
17 Staden R (1989), CABIOS, 4, 53-60.