# STUDYING THE SECONDARY STRUCTURE OF ACCESSION NUMBER USING CETD MATRIX

**Anamika Dutta**
Department of Statistics, Gauhati University, Guwahati-781014, Assam, India
anamika.dut268@gmail.com
**Kishore K.Das**
Department of Statistics, Gauhati University, Guwahati-781014, Assam, India
daskkishore@gmail.com

Professional Paper
doi:10.5937/jouproman4-11570

**Abstract:** This paper, we have tried to analyze about the Secondary Structure of nucleotide sequences of rice. The data have been collected from NCBI (National Centre for Biotechnology Information) using Nucleotide as data base. All the programs were developed using R programming language using "sequinr" package. Here, we have used CETD matrix method to study the prediction. The conclusions are drawn accordingly.

***Key words:*** *Secondary Structure, ATD Matrix, RTD Matrix, CETD Matrix*

## 1. Introduction:

Proteomics as the name suggest is a very vast field of research. Here we have tried to study the secondary structure of accession number for rice (Molina et al., 2011) using CETD matrix (Kuppuswami et al., 2015). We have considered 24 accession numbers for rice from the paper Cho et al., where the accession number were already been used for other studies (Cho et al., 2000).

The Accession Numbers for rice (Cho et al., 2000) are: D17586, M36469, X58877, Z11920, X07515, U12171, U33175, X64619, D78609, D78506, D30794, L10346, D63901, U40708, M29259, X65183, U08404, D14000, U31771, X53596, U37133, D16221, U49113, L37528

Primary structure of protein is defined as first level of protein i.e., the sequence of 20 amino acids is basically known as the primary structure of protein. The primary structure of protein has different characteristics which is dependent on hydrogen bonding and is known as secondary structure of protein. In other words, secondary structure is the second level of primary structure (Technical Brief 2009).

Here we only secondary structure of the nucleotide sequence of rice has been considered. Out of the secondary structures we have selected only six particles viz. Helix Residue, Beta Sheet Residue, Beta Turn Residue, Helix Region, Beta Sheet Region and Beta Turn Region.

The objective of the paper is:

1. To first split the accession numbers into some groups using lottery method.
2. To find which is the most conservative accession number group using CETD matrix.

### 1.1 Method and materials:

We have used NCBI (Pruitt et al., 2007 and Altschul et al., 1990) nucleotide data base to find the sequences of rice. Later, using Chaufasman Algorithm and "sequinr" package in R programming language, the value of six parts of secondary structure of protein for 24 accession numbers have been found.

Firstly, the Initial Raw Data matrix has to be created (Porchelvi, S. R. and Vanitha, R. (2011), Narayanamoorthy et al., 2013 and Kuppuswami et al., 2015) which can be constructed by taking the accession numbers in rows and the secondary particles in columns. Next we contrive Average Time Dependent (ATD) Matrix (Porchelvi, S. R. and Vanitha, R. (2011), Narayanamoorthy et al., 2013 and Kuppuswami et al., 2015) which is obtained by dividing the entries of Initial Raw Data matrix by the difference between the upper range and the lower range of accession number. Following the mean ($\mu_j$) and standard deviation ($\sigma_j$) of each column has been found separately for ATD matrix. Now using the mean and standard deviation for each column and choosing a parameter $\alpha$ lying between [0,1] and construct a Refined Time Dependent Matrix (RTD Matrix) using the formulae (Porchelvi, S. R. and Vanitha, R. (2011), Narayanamoorthy et al., 2013 and Kuppuswami et al., 2015):

- If $a_{ij} \leq (\mu_j - \alpha\, \sigma_j)$ then $e_{ij} = -1$
- If $a_{ij} \in (\mu_j - \alpha\, \sigma_j, \mu_j + \alpha\, \sigma_j)$ then $e_{ij} = 0$
- If $a_{ij} \geq (\mu_j + \alpha\, \sigma_j)$ then $e_{ij} = 1$

$a_{ij}$ are the elements of ATD matrix and $e_{ij}$ are the elements of RTD matrix (Porchelvi, S. R. and Vanitha, R.(2011)). The RTD matrix consists of only -1, 0 and 1 elements.

Succeeding, each row of the RTD matrix has to be added. The maximum sum gives the group of accession number which is giving a better secondary prediction. Lastly, we combine all the RTD matrices and get a Combined Effective Time Dependent Data Matrix (CETD matrix) Porchelvi, S. R. and Vanitha, R. (2011), Narayanamoorthy et al., 2013 and Kuppuswami et al., 2015). The row sum is obtained for CETD matrix and conclusions are made using the CETD matrix. Subsequently, all the matrices are represented using line diagram.

## 2. Estimation:

The grouping of the accession numbers into six parts has been done using lottery method. We have numbered the accession numbers like:

1 as D17586, 2 as D16221, 3 as M36469, 4 as D78609, 5 as X58877, 6 as Z11920, 7 as D14000, 8 as L37528, 9 as X07515, 10 as U12171, 11 as U33175, 12 as D30794, 13 as X64619, 14 as D78506, 15 as L10346, 16 as D63901, 17 as U40708, 18 as M29259, 19 as X65183, 20 as U08404, 21 as U31771, 22 as X53596, 23 as U37133 and 24 as U49113.

The numbering has been done so that we can group the accession number unbiasedly using Lottery Method; a method used in simple random sampling. The numbering of accession number has been done accordingly.

Subsequently, the 24 accession number has been divided into 6 groups are 1-4, 5-8, 9-12, 13-16, 17-20 and 21-24. With this our first objective gets fulfilled.

We define,

$X_1$: Helix Residue
$X_2$: Beta Sheet Residue
$X_3$: Beta Turn Residue
$X_4$: Helix Region
$X_5$: Beta Sheet Region
$X_6$: Beta Turn Region

Here, CETD matrix has been used in order to check which group of accession number is giving a conservative region. The raw data of 24 accession number has been converted in the form of matrix, known as Initial Raw Data matrix.

**Table 1: Initial Raw Data Matrix**

| Accession Number | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| 1-4 | 420 | 411 | 150 | 48 | 65 | 24 |
| 5-8 | 909 | 514 | 159 | 68 | 74 | 29 |
| 9-12 | 502 | 392 | 80 | 52 | 50 | 12 |
| 13-16 | 287 | 600 | 185 | 32 | 69 | 29 |
| 17-20 | 393 | 260 | 90 | 40 | 41 | 17 |
| 21-24 | 415 | 495 | 155 | 40 | 73 | 22 |

**Table 2: The ATD Matrix**

| Accession No. | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| 1-4 | 105 | 102.75 | 37.5 | 12 | 16.25 | 6 |
| 5-8 | 227.25 | 128.5 | 39.75 | 17 | 18.5 | 7.25 |
| 9-12 | 125.5 | 98 | 20 | 13 | 12.5 | 3 |
| 13-16 | 71.75 | 150 | 46.25 | 8 | 17.25 | 7.25 |
| 17-20 | 98.25 | 65 | 22.5 | 10 | 10.25 | 4.25 |
| 21-24 | 103.75 | 123.75 | 38.75 | 10 | 18.25 | 5.5 |

**Table 3: The Mean and Standard Deviation of the above ATD Matrix**

| Mean (μ) | 121.92 | 111.33 | 34.13 | 11.67 | 15.5 | 5.54 |
|---|---|---|---|---|---|---|
| Standard Deviation (σ) | 49.67 | 26.90 | 9.54 | 2.87 | 3.07 | 1.54 |

It has already been mentioned that the parameter α for the construction in RTD matrix lies in the range [0,1], so three arbitrary values of α has been selected viz. 0.1, 0.15 and 0.3.

**Table 4: Values of $(\mu_j - \alpha\,\sigma_j)$ and $(\mu_j + \alpha\,\sigma_j)$ for different values of α**

| | | | | | | |
|---|---|---|---|---|---|---|
| Interval at α = 0.1 | 116.95 | 108.64 | 33.18 | 11.38 | 15.19 | 5.39 |
| | 126.89 | 114.02 | 35.08 | 11.96 | 15.81 | 5.70 |
| Interval at α = 0.15 | 114.47 | 107.30 | 32.70 | 11.24 | 15.04 | 5.31 |
| | 129.37 | 115.37 | 35.56 | 12.10 | 15.96 | 5.77 |
| Interval at α = 0.3 | 107.02 | 103.26 | 31.27 | 10.81 | 14.58 | 5.08 |
| | 136.82 | 119.4 | 36.99 | 12.53 | 16.42 | 6.00 |

Let us create the RTD matrices for various values of α.

The RTD matrix for α = 0.1 The row sum matrix

$$\begin{pmatrix} -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 2 \\ 6 \\ -3 \\ 2 \\ -6 \\ 1 \end{pmatrix}$$

The RTD matrix for α = 0.15 The row sum matrix

$$\begin{pmatrix} -1 & -1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1 \\ 6 \\ -3 \\ 2 \\ -6 \\ 1 \end{pmatrix}$$

The RTD matrix for α = 0.3 The row sum matrix

$$\begin{pmatrix} -1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} -1 \\ 6 \\ -3 \\ 2 \\ -6 \\ 1 \end{pmatrix}$$

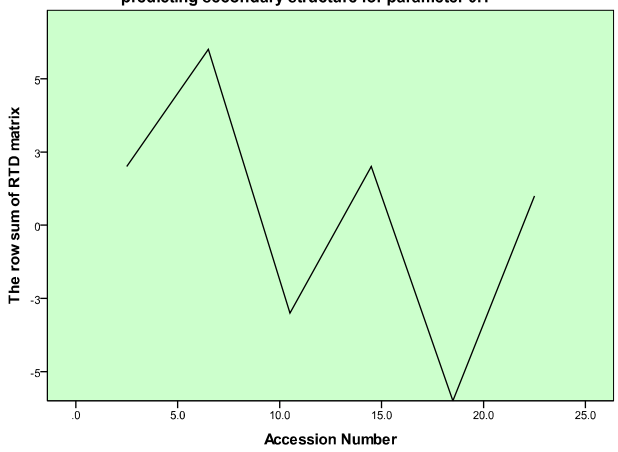
Adding the elements of the three RTD matrices, we get the CETD matrix, which is given below:

The CETD matrix The row sum of CETD matrix

$$\begin{pmatrix} -3 & -3 & 3 & 1 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 0 & -3 & -3 & 3 & -3 & -3 \\ -3 & 3 & 3 & -3 & 3 & 3 \\ -3 & -3 & -3 & -3 & -3 & -3 \\ -3 & 3 & 3 & -3 & 3 & 0 \end{pmatrix} \qquad \begin{pmatrix} 2 \\ 18 \\ -9 \\ 6 \\ -18 \\ 3 \end{pmatrix}$$

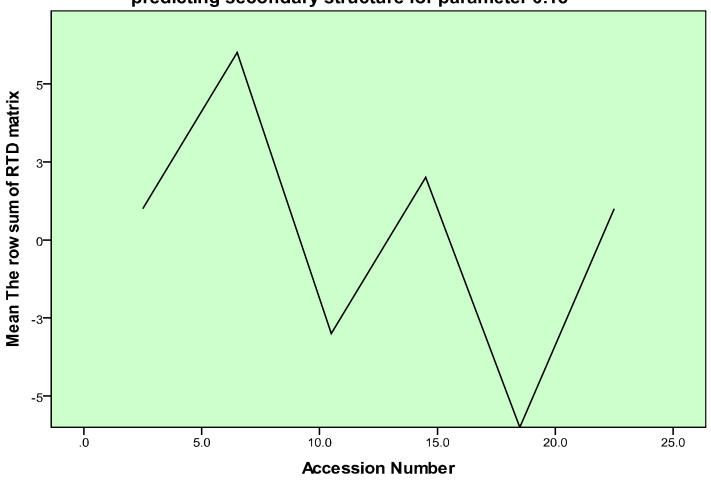
Title

**Figure 1: The graph depicting the maximum group of accession number for predicting secondary structure for parameter 0.1**



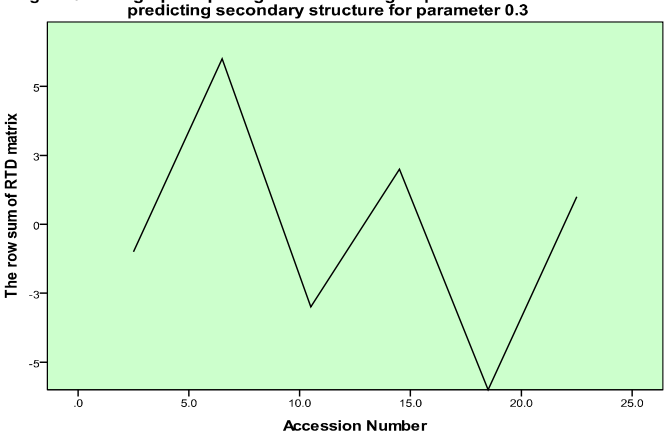

Title

**Figure 2 : The graph depicting the maximum group of accession number for predicting secondary structure for parameter 0.15**



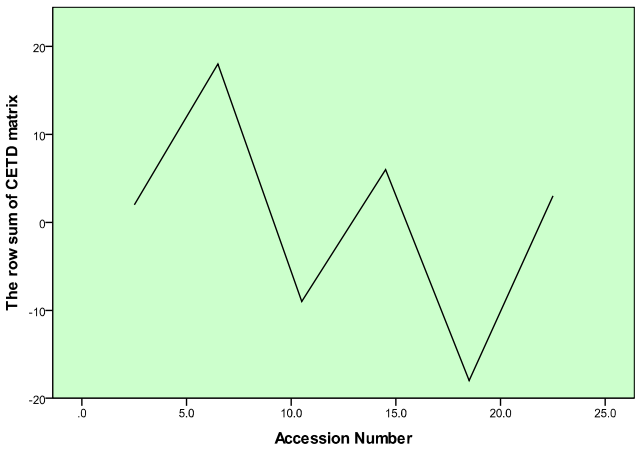

Title

**Figure 3 : The graph depicting the maximum group of accession number for predicting secondary structure for parameter 0.3**




Title

**Figure 4 : The graph predicting the maximum group of accession number for predicting secondary structure in a CETD matrix**



From figure 4, that is the line diagram above, it can be seen that the highest peak is in the second group i.e., in the group 5-8. Which implies the region from 5 to 8 is conservative. The second objective fulfills here.

## 3. Conclusion:

From the CETD matrix, it is seen that the row sum of CETD matrix is highest in the interval 5-8 which means that the group of accession numbers viz. X58877, Z11920, D14000 and L37528 are giving a better secondary prediction than the other intervals of accession numbers, which means that the region is conservative and the sequence of the accession numbers are identical. Also it has been seen that the lowest negative is in the interval 17-20 which means that the group of accession numbers viz. U40708, M29259, X65183 and U08404 are giving the least secondary prediction among the other intervals. It concludes that the sequences in these accession numbers are least alike.

## 4. Acknowledgement:

The author Miss. Anamika Dutta thank to Department of Science and Technology (DST), India for providing financial assistance for carrying out this work as an INSPIRE Fellow.

## 5. Reference:

1. Cho, Y. G., Ishii, T., Temnykh, S., Chen, X., Lipovich, L., McCouch, R. S., Park, D. W., Ayres, N., and Cartinhour, S. (2000). Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice.(*Oryza sativa L.*) *Theor Appl Genet*, Vol. 100, pages: 713-722. *Springer-Verlag*

2. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35, pages: D61–D65

3. Molina, J., Sikora, M., Garud, N., Flowers, M. J., Rubinstein, S., Reynolds, A., Huang, P., Jackson, S., Schaal, A. B., Bustamante, D. C., Boyko, R. A., and Purugganan, D. M. (2011). Molecular evidence for a single evolutionary origin of domesticated rice. *PNAS*, Vol. 108, No. 20, pages: 8351-8356

4. Kuppuswami, G., Sujatha, R., Vasantha Kandaswamy, W.B. (2015). Study of traffic flow using CETD Matrix. *Indian Journal of Science of Technology, Vol 8(24):* 1:5

5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J.(1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410

6. Porchelvi, S. R. and Vanitha, R. (2011). On studying a health problem using fuzzy matrices. *International Review of Fuzzy Mathematics, Vol. 6, No. 1, page: 15-20*

7. Narayanamoorthy, S., V. M. S. and Sivakamasundari, K. (2013). Fuzzy cetd matrix to estimate the maximum age group victims of pesticide endosulfan problems faced in Kerala. *International Journal of Mathematics and Computer, Vol. 3, Issue 2, page: 227-232*

8. Particle Sciences Drug Development Services, *Technical Brief 2009, Vol 8*