Thailand Statistician
January 2022; 20(1): 177-184
http://statassoc.or.th
Contributed paper

# Boolean Alignment Matrix and Quasi Binomial Distribution: A Case Study Using DNA Sequence Data

**Anamika Dutta* and Kishore Kumar Das**
Department of Statistics, Gauhati University, Guwahati, Assam, India.
*Corresponding author; e-mail: anamika.dut268@gmail.com

Received: 23 October 2019
Revised: 2 May 2020
Accepted: 3 May 2020

**Abstract**

A new matrix has been proposed here in this paper which is known as Boolean alignment matrix with variables 1 and 0, constructed for DNA's viz. 'A', 'C', 'G', 'T' along with few gap characters. Quasi binomial distribution of type I with parameters $n$, $p_1$, and $\phi$ has been used in this paper to check whether the Boolean alignment matrix and quasi binomial distribution of type I are inter-related utilizing Kolmogorov-Smirnov goodness of fit test. The highest number of counts from the proposed matrix and the highest value obtained from Kolmogorov-Smirnov goodness of fit test has been found to be DNA 'C'. Consequently, in this paper, it has been found that the proposed matrix has a sharp perceptive with quasi binomial distribution of type I executing Kolmogorov-Smirnov goodness of fit test.

**Keywords**: Quasi binomial type I, alignment matrix, Boolean variable, nucleotide sequence, Abel polynomial.

## 1. Introduction

The values of a variable viz., 1 and 0, is the branch of algebra known as Boolean algebra. It is used where the values of a variable present are denoted by 1, otherwise, it is denoted by 0. Boolean algebra is also known as binary algebra or logical algebra. In Boolean algebra, instead of primary values of the variable, the numbers 1 and 0 are used for notations (Boole 2003). Using this concept, DNA's have been converted to the Boolean alignment matrix.

Boolean alignment matrix proposed in this paper to deal with digits 0 and 1. It is a matrix of 1 and 0, i.e., true or false. For any query or any survey, if the answer is true, it is considered as 1. Otherwise, it is 0. For DNA sequences, i.e., sequences which contain only 'A', 'C', 'G', and 'T'; where A = Adenine, C = Cytosine, G = Guanine, and T = Thymine are the building blocks of amino acids. Now for constructing the Boolean alignment matrix for DNA sequences, all the sequences should be aligned one after another using ClustalX software (Dutta and Das 2018). To construct a Boolean alignment matrix, the first alignment matrix has to be constructed. Later from the alignment matrix (Dutta and Das 2018), the Boolean alignment matrix can be constructed using 0 and 1. Column wise DNA's are examined, if suppose DNA 'A' is present in the first column 3 times, then 3 is written against DNA 'A' in alignment matrix and 1 is written against DNA 'A' in the Boolean alignment

matrix. To construct a Boolean alignment matrix number of count of DNA is ignored, only the presence or absence of DNA is considered.

It is seen that in some experiments, fitting of the binomial model sometimes under-estimate or over-estimate the data. In such a situation quasi binomial distribution of type I with parameters $n$, $p_1$, and $\phi$ gives a better result. Consul (1974) proposed quasi binomial distribution of type I, its applications is seen in medical studies, biology, bioinformatics, medicines, etc. Basically, quasi binomial distribution extends over a very wide range of discrete distributions (Mishra et al. 1992; Das 1993 and Nandi and Das 1994). Quasi binomial distribution of type I have been adopted in this study since the parameter $\phi$ can deal with larger data and DNA particles are count data containing countable infinite particles of DNA or amino acids in any living being or protein molecule (e.g. virus).

In order to explore the DNA count data, following the work of Dutta and Das (2018), we propose to construct the Boolean alignment matrix using Boolean variables and then investigate the matter using quasi binomial distribution of type I. Finally, goodness of fit test using Kolmogorov-Smirnov test has been conducted to oversee the result. The rest of this paper is organized as follows. Section 2 comprises about the data used in this paper and Section 3 describes about how the data has been arranged. All the methods along with an example have been explained in Section 4. Section 5 displays the results and discussion, which is the most important part of our paper. The findings have been summarized in Section 6.

## 2. Description of Data

All the accession numbers have been collected from NCBI in FASTA format using nucleotide sequence as database. R software with the help of 'sequinr' package has been used to arrange and count the DNA sequences.

There are two kinds of sequence alignment. They are global alignment and local alignment (Stormo 2009). Global alignment means given any two sequences but they are not of equal length. Say one sequence is large and the other is small, now to match the maximum identities in both the sequence, gaps has to be introduced between the two sequences so that all possible matches between them can be identified. Whereas local alignment means given any two sequences, both are of equal size, it is not necessary to introduce any gaps because the sequences are already of equal size.

Now let us switch to multiple alignments (Rosenberg 2009). In general, multiple sequence alignment means to align up not only two sequences but also more than two sequences. Multiple sequence alignment is based on dynamic programming (Chenna et al. 2003; Rosenberg 2009; Stormo 2009) which means as soon as the number of sequences changes, the alignment also changes.

## 3. Arrangement of Data

The 12 accession numbers of rice have been collected from a chapter of a book by Dutta and Das (2018). The accession numbers are:

D17586, D16221, M36469, D78609, X58877, Z11920, D14000, L37528, X07515, U12171, U33175 and D30794.

All these accession numbers have been converted into alignment matrix with 5 rows and 7,473 columns. Through the method of scoring and the method of intersection (Dutta and Das 2018) only two parts have been taken into consideration (Dutta and Das 2018; Dutta 2018). Those two parts have been taken into account for Boolean alignment matrix and fitting of quasi binomial distribution of type I.

## 4. Methodology

Taking an example, the procedure of Boolean alignment matrix can be explained in a proper way. As mentioned above, the alignment is always been done by ClustalX software. Let us take some random DNA sequences of different lengths,

C G A A G T C C T A T C
A G A T G T A C A T C G
T C A G G A T A C G T A
G C A C G T A C C A T A.

Boolean alignment matrix has been constructed for the above sequences, which are shown in Table 1.

**Table 1** Boolean alignment matrix

$$\begin{array}{c} A \\ C \\ G \\ T \end{array}\begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

The explanation of how the above matrix has been formed is explained below. Four small sequences have been taken as an example. These sequences are aligned horizontally one after another. Later, observe the DNA's column-wise. In the first column, it can be seen that the entire four DNA's viz. 'A', 'C', 'G' and 'T' are present, so in the Boolean alignment matrix, 1 has appeared against DNA's 'A', 'C', 'G' and 'T'. In the next column DNA 'A' and 'T' are not present; hence in the Boolean alignment matrix in the second column against DNA 'A' and 'T' 0 has appeared and against 'C' and 'G' 1 has appeared. Similarly for the third column, it is capable to see that only DNA "A" is present and hence in the alignment matrix against DNA 'A' 1 has appeared and against the other DNA's i.e., DNA 'C', 'G' and 'T' 0 has appeared. The rest columns of the Boolean alignment matrix have been calculated in the same manner. In this way, the Boolean alignment matrix has been formed. Boolean random variable (Nandi and Das, 1994) $\alpha_{ni},\ i = 1,2,...,n$ where $\alpha_{ni}$ takes only values 0 and 1 and n is defined as the number of trials in any experiment.

Let us consider the random variables in terms of sequence of arrays.

$$\alpha_{11}$$
$$\alpha_{21}\ \alpha_{22}$$
$$\alpha_{31}\ \alpha_{32}\ \alpha_{33}$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\alpha_{n1}\ \alpha_{32}\ \alpha_{n3}\qquad \alpha_{nn}$$

Taking the values 0 and 1. Let $\mu_n = \sum_{i=1}^{n} \alpha_{ni}$. Let us introduce Abel polynomials in terms of Boolean algebra (Nandi and Das, 1994). The Abel Boolean polynomial is defined as

$$\alpha_{\mu_n}\left(a, \sum_{i=1}^{n} z_i\right) = a\left(a + \sum_{i=1}^{n} \alpha_{ni} z_i\right)^{\left(\sum_{i=1}^{n} \alpha_{ni} - 1\right)} \tag{1}$$

where $\sum_{i=1}^{n} z_i = z.$

Then the probability mass function of the Acyclic quasi binomial distribution is given by (Nandi and Das 1994)

$$p(\mu_n) = a\left(a+\sum_{i=1}^{n}\alpha_{ni}z_i\right)^{\left(\sum_{i=1}^{n}\alpha_{ni}-1\right)}\left(b+\sum_{i=1}^{n}\overline{\alpha}_{ni}z_i\right)^{\left(\sum_{i=1}^{n}\overline{\alpha}_{ni}\right)}/\left(a+b+\sum_{i=1}^{n}z_i\right)^{n}, \tag{2}$$

where $\alpha_{ni}$ are stated in the above sequence of arrays and

$$\left(a+b+\sum_{i=1}^{n}z_i\right)^{n} = \sum_{i=1}^{n}a\left(a+\alpha_{ni}z_i\right)^{\left(\sum_{i=1}^{n}\alpha_{ni}-1\right)}\left(b+\sum_{i=1}^{n}\overline{\alpha}_{ni}z_i\right)^{\left(\sum_{i=1}^{n}\overline{\alpha}_{ni}\right)}. \tag{3}$$

Also,

$$\overline{\alpha}_{ni} = 1-\alpha_{ni},\ \ i=1,2,...,n, \tag{4}$$

$$\sum_{i=1}^{n}\alpha_{ni}+\sum_{i=1}^{n}\overline{\alpha}_{ni} = n. \tag{5}$$

For $n_{C_k}$ unordered collections $i_1, i_2, ..., i_k$ of $i = 1,2,...,n$ such that $\alpha_{nij} = 1,\ \ j = 1,2,...,k$ and the remaining α's are zeroes, we get the quasi binomial distribution of type I (Consul 1974; Nandi and Das 1994; Dutta 2018) and is given by

$$p(k) = n_{C_k}\,p_1\left(p_1+k\phi\right)^{(k-1)}\left(p_2+(n-k)\phi\right)^{(n-k)}, k = 0,1,2,...,n, \tag{6}$$

where $0 < p_1 < 1$ and $-p_1/n < \phi < (1-p_1)/n.$ Also, $p_1 + p_2 + n\phi = 1.$

Then the Equation (6) changes to

$$p(k) = n_{C_k}\,p_1\left(p_1+k\phi\right)^{(k-1)}\left(1-p_1-k\phi\right)^{(n-k)}, k = 0,1,2,...,n, \tag{7}$$

where $p_1$ and $\phi$ are the parameters to be estimated. And the estimated values of $p_1$ and $\phi$ are

$$\hat{p}_1 = 1-\left(f_0/N\right)^{(1/n)}, \tag{8}$$

$$\hat{\phi} = 1-p_1-\left[\left(f_1/N\right)/np_1\right]^{(1/n-1)}, \tag{9}$$

where $f_0/n$ is the proportion of 0, $f_1/N$ is the proportion of 1, and $N = \sum_i f_i$ .

## 5. Results and Discussion

The 12 accession numbers of rice have been collected from a chapter of a book by Dutta and Das (2018). The accession numbers are:

D17586, D16221, M36469, D78609, X58877, Z11920, D14000, L37528, X07515, U12171, U33175 and D30794.

All these accession numbers have been converted into alignment matrix with 5 rows and 7473 columns. Through the method of scoring and the method of intersection (Hertz and Stormo 1999; Shu et al. 2012; Dutta and Das 2018) only two parts have been taken into consideration (Dutta and Das 2018; Dutta 2018).

The application of Boolean algebra and Quasi Binomial distribution of type I has been applied to those two matrices which have already been mentioned in the article by Dutta and Das (2018). The first matrix under study (Dutta and Das 2018; Dutta 2018) is shown in Table 2.

**Table 2** Alignment matrix for the first region under study

$$\begin{array}{c} - \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 2 & 2 & 5 & 0 & 3 & 4 \\ 6 & 1 & 1 & 3 & 1 & 10 & 5 & 1 \\ 3 & 7 & 6 & 6 & 5 & 0 & 2 & 7 \\ 1 & 0 & 3 & 1 & 1 & 2 & 2 & 0 \end{bmatrix}$$

The above matrix has been which has been converted to Boolean alignment matrix is shown in Table 3.

**Table 3** Boolean alignment matrix for the first region under study

$$\begin{array}{c} - \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

From Table 3, it can be seen from the matrix that no 0 is present against DNA 'C', which implies that the highest count is DNA 'C' in the matrix. This indicates that DNA 'C' is present more or less in all the accession numbers.

Let us now try to find the expected frequencies by fitting the quasi binomial distribution with observed frequencies taken from Table 2. Then Kolmogorov-Smirnov goodness of fit test would be applied to test the significance of the observed and expected frequencies. Though the Kolmogorov-Smirnov goodness of fit test is intended for continuous distributions; the test may also be used for discrete distributions based on the empirical process in discrete time (see Arnold and Emerson 2011; Khmaladze 2013). The chi-square goodness of fit test cannot be applied as our sample is less and much information would get eliminated through grouping or pooling. The Kolmogorov-Smirnov test statistics is given by (Mukhopadhyay 2009)

$$D = \max\left|F_0(X) - S_N(X)\right|,$$

where $F_0(X)$ is the theoretical cumulative distribution under $H_0$, $S_N(X)$ is the observed cumulative probability distribution of a random sample of $N$ observations. The null hypothesis to be tested is

$H_0$ : The data of observed and expected frequency follows the same distribution

$H_1$ : At least one value does not match the specified distribution.

The observed frequencies from Table 2 are: 1, 21, 28, 36 and 10. All the values of expected frequencies (using Equation (7)) and the difference to find the value of $D$ have been tabulated below.

**Table 4** Kolmogorov-Smirnov goodness of fit test for the first region

| $O$ | $E$ | $F_0(X)-S_N(X)$ |
|---|---|---|
| 1 | 1 | 0 |
| 21 | 21 | 0 |
| 28 | 53 | 0.2604 |
| 36 | 20 | 0.0938 |
| 10 | 1 | 0 |

From the Table 4, it can be seen that the maximum difference is 0.2604. Therefore, $D = 0.2604$. Tabulated value of $D$ for $n = 4$ at 5% level of significance is 0.62394. And since calculated value of $D$ is less than the tabulated value of $D$ at 5% level of significance, we have no reason to reject the null hypothesis and infer that the data of observed and expected frequencies follows the same distribution. The second matrix under study (Dutta 2018) is shown in Table 5.

**Table 5** Alignment matrix for the second region under study

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| – | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 2 | 2 | 5 | 1 | 3 | 5 | 0 | 1 | 0 | 3 | 1 | 6 | 2 | 5 | 5 | 1 | 1 | 6 | 1 | 3 | 7 | 1 | 6 | 0 |
| C | 2 | 2 | 1 | 3 | 2 | 1 | 10 | 1 | 3 | 7 | 6 | 1 | 5 | 2 | 4 | 5 | 4 | 2 | 6 | 1 | 1 | 9 | 0 | 4 |
| G | 7 | 5 | 2 | 3 | 2 | 7 | 1 | 3 | 0 | 1 | 1 | 2 | 2 | 5 | 2 | 0 | 4 | 3 | 1 | 7 | 2 | 1 | 6 | 2 |
| T | 1 | 3 | 3 | 4 | 3 | 0 | 1 | 7 | 9 | 1 | 3 | 3 | 3 | 0 | 1 | 6 | 3 | 1 | 4 | 1 | 2 | 1 | 0 | 6 |

The second matrix (Dutta and Das 2018; Dutta 2018) which has been converted into 0 and 1 are as follows

**Table 6** Boolean alignment matrix for the second region under study

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| – | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| C | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 11 |
| G | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| T | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |

From Table 6, it can be seen from the matrix that 0 is present only once against DNA 'C', which implies that the highest count is DNA 'C' in the matrix. This indicates that DNA 'C' is present more or less in all the accession numbers. The null hypothesis to be tested is

$H_0$ : The data of observed and expected frequency follows the same distribution.

$H_1$ : At least one value does not match the specified distribution.

The observed frequencies from Table 5 are: 5, 67, 82, 69 and 66. All the values of expected frequencies (using Equation (7)) and the difference to find the value of $D$ have been tabulated below:

**Table 7** Kolmogorov-Smirnov goodness of fit test for the second region

| $O$ | $E$ | $F_0(X)-S_N(X)$ |
|---|---|---|
| 5 | 5 | 0 |
| 67 | 67 | 0 |
| 82 | 147 | 0.2249 |
| 69 | 65 | 0.2111 |
| 66 | 5 | 0 |

From the Table 7, it can be seen that the maximum difference is 0.2249. Therefore, $D = 0.2249$. Tabulated value of $D$ for $n = 4$ at 5% level of significance is 0.62394. And since calculated value of $D$ is less than the tabulated value of $D$ at 5% level of significance, we have no reason to reject the null hypothesis and infer that the data of observed and expected frequencies follows the same distribution.

## 6. Conclusions

In Table 3, a 5×8 matrix and in Table 6, a 5×24 matrix have been taken into consideration. And it can be seen that in both the cases the highest count is DNA 'C' out of all the DNA's and gap character, in the first matrix one has no 0 and the second matrix has only one 0, which implies that the repetitions is highest in DNA 'C'. This indicates that DNA 'C' is present more or less in all the accession numbers through the total sum of DNA 'C'.

Next, the conclusion can be stated that both the matrices i.e., 5×8 and 5×24 fits the quasi binomial of type I with Kolmogorov-Smirnov goodness of fit giving the best fit inferring that data of observed and expected frequencies follows the same distribution for both the given regions under study.

It has already been mentioned that DNA 'C' has the highest count in the entire Boolean alignment matrix. Also, it can be seen from Table 4 as well as from Table 7 that the highest value of D falls in the 3rd number which is against DNA 'C' implying that the Boolean alignment matrix and quasi binomial distribution type I are inter-related. Implying that the new proposed matrix i.e., Boolean alignment matrix has a keen connection with quasi binomial type I distribution.

**Acknowledgements**

The first author would like thank Department of Science and Technology (DST), India for providing financial assistance for carrying out this work as an INSPIRE Fellow. Also, the authors are in indebt to all the reviewers and editor who gave their effort and dedication to improve the quality of this paper.

## References

Arnold BT, Emerson WJ. Nonparametric goodness-of-fit tests for discrete null distributions. R J. 2011; 3(2): 34-39.

Boole G. An Investigation of the laws of thought, New York: Prometheus Books; 2003.

Chenna R., Sugawara H, Koike T, Lopez R., Gibson JT, Higgins GD, Thompson DJ. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 2003; 31(13): 3497-3500.

Consul PC. A simple urn model dependent on predetermined strategy. Sankhya.1974; 36: 391-399.

Consul PC. On some properties and applications of quasi-binomial distribution. Commun Stat-Theory Methods. 1990; 19(2): 477-504.

Das KK. Some aspects of a class of quasi binomial distributions. Assam Stat Rev. 1993; 7(1): 33-40.
Dutta A, Das KK. A study on DNA sequence of rice using scoring matrix method and ANOVA technique. Stat Appl. 2018; 244: 15-24.
Dutta A. An empirical study with proteomics datasets using different techniques. PhD [dissertation]. Guwahati, India: Gauhati University; 2018.
Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999; 15(7): 563-577.
Khmaladze E. Note on distribution free testing for discrete distributions. Ann Stat. 2013; 41(6): 2979-2993.
Mishra A, Tiwary D, Singh SK. A class of quasi-binomial Distributions. Sankhya. 1992; 54: 67-76.
Mukhopadhyay P. Mathematical Statistics. Kolkata: Books and Allied; 2009.
Nandi SB, Das KK. A family of the Abel series distributions. Sankhya. 1994; 56(2): 147-164.
Rosenberg SM. Sequence alignment, methods, models, concepts, and strategies. Q Rev Biol. 2009; 84(3): 295-295
Shu JJ, Yong YK, Chang KW. An improved scoring matrix for multiple sequence alignment. Math Probl Eng. 2012; 490649: 1-9.
Stormo GD. An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinform. 2009; 27(1): 3.1.1-3.1.7.