*Combinatorial Chemistry & High Throughput Screening*, 2022, 25, 365-16

REVIEW ARTICLE

BENTHAM SCIENCE

# Numerical Characterization of DNA Sequences for Alignment-free Sequence Comparison – A Review

Natarajan Ramanathan[1,*], Jayalakshmi Ramamurthy[2] and Ganapathy Natarajan[3]

[1]*Department of Chemistry, Sri Sarada Niketan College for Women, Karur-639005, Tamil Nadu, India;* [2]*Department of Computer Science, Sri Sarada Niketan College for Women, Karur-639005, Tamil Nadu, India;* [3]*Department of Mechanical Engineering and Industrial Engineering, University of Wisconsin, Platteville, WI 53818, USA*

ARTICLEHISTORY

Received: December 30, 2020
Revised: June 16, 2021
Accepted: June 24, 2021

DOI
10.2174/1386207324666210811181437

**Abstract:** ***Background:*** Biological macromolecules, namely, DNA, RNA, and protein, have their building blocks organized in a particular sequence and the sequential arrangement encodes the evolutionary history of the organism (species). Hence, biological sequences have been used for studying evolutionary relationships among the species. This is usually carried out by Multiple Sequence Algorithms (MSA). Due to certain limitations of MSA, alignment-free sequence comparison methods were developed. The present review is on alignment-free sequence comparison methods carried out using the numerical characterization of DNA sequences.

***Discussion:*** The graphical representation of DNA sequences by chaos game representation and other 2-dimensional and 3-dimensional methods are discussed. The evolution of numerical characterization from the various graphical representations and the application of the DNA invariants thus computed in phylogenetic analysis are presented. The extension of computing molecular descriptors in chemometrics to the calculation of a new set of DNA invariants and their use in alignment-free sequence comparison in an N-dimensional space and construction of phylogenetic trees are also reviewed.

***Conclusion:*** The phylogenetic tress constructed by the alignment-free sequence comparison methods using DNA invariants were found to be better than those constructed using alignment-based tools such as PHLYIP and ClustalW. One of the graphical representation methods is now extended to study viral sequences of infectious diseases for the identification of conserved regions to design peptide-based vaccines by combining numerical characterization and graphical representation.

**Keywords:** Numerical characterization, DNA sequences, alignment-free, sequence comparison, phylogenetic analysis, peptide-based vaccines.

## 1. INTRODUCTION

### 1.1. Evolutionary Studies using Biological Sequence Data

The sequences of biological macromolecules are the product of molecular evolution. The genomic sequences are a treasure of information database created by mother nature through the long process of evolution. When the sequences share a common ancestral sequence, they tend to exhibit similarities in their sequences, structures, and biological functions. Similarities might lead to clues regarding the evolutionary history of the gene or the function of the protein. Genes and proteins are therefore related by evolution and thus, they have an evolutionary history [1]. Information is encoded in the sequential order of the bases (four letters) on the chain of a DNA molecule. This forms the basis for studying evolutionary relationships using biological sequences (DNA, RNA, and protein). Moreover, owing to the availability of a large number of sequence data in the post-genomic era, scientists are using sequence data rather than morphological and other information to study evolutionary relationships. Random mutations may accumulate more easily in non-critical parts of a protein sequence. On the other hand, in parts of a protein sequence that are critical for the function of the protein, hardly any mutations will be accepted; this behavior is due to self-preservation of the protein function. Analysis of evolutionary relationships between protein or gene sequences depends critically on sequence alignments.

*Address correspondence to this author at the Department of Chemistry, Sri Sarada Niketan College for Women, Karur-639005, Tamil Nadu, India; Tel: 09790912713; E-mail: nataraj@lakeheadu.ca

1875-5402/22 $65.00+.00 

## 1.2. Sequence Alignment

The sequence alignment procedure is used in bioinformatics for comparing two or more sequences [2-3]. Identical or similar characters are placed in the same column and non-identical characters can either be placed in the same column as a mismatch or opposite, making a gap in the other sequence. In an optimal alignment, non-identical characters and gaps are introduced in one sequence to bring as many identical or similar characters as possible. To attain optimal alignment, insertion and deletion of some bases are also allowed in the other sequence. Biologically, this corresponds to a mutation event that eliminates a part of a gene or introduces new DNA into a gene. Different alignment algorithms have different basic assumptions and parameterizations. Hence, there is no such thing as the single best alignment.

There are two types of sequence alignment, namely, global and local. In global alignment, an attempt is made to align the entire sequence using as many characters as possible, up to both ends of each sequence. Local alignment searches for segments of the two sequences that match well. This type of alignment favors finding conserved nucleotide patterns, DNA sequences, or amino acid patterns in protein sequences. In local alignment, the two sequences are not forced into an alignment.

Pair-wise sequence alignments are fundamental in nature but have problems in comparing sequences in a database. They become computing-intensive and practically impossible when the sequences are long. An extension of the fundamental principles of pair-wise alignment is multiple sequence alignment (MSA). In multiple sequence alignment, all similar sequences can be compared by aligning sequences on top of each other. In order to perform this, a coordinate system is set up, where each row is for one sequence (DNA, RNA, or protein), and each column is the 'same' position in each sequence. Each column corresponds to a specific residue in the 'template' sequence. In MSA, the gaps introduced in some sequences are shown by the dash '-' or dot '.' character. To construct multiple alignments, one may have to introduce gaps in sequences at positions where there were no gaps in the corresponding pair-wise alignment. This means that multiple alignments contain more gaps than any given pair of aligned sequences.

### 1.3. Dynamic Programming

Dynamic programming is a computational method that is used to align two biological sequences (nucleic acid or protein). This method is very important for sequence analysis because it provides the best or optimal alignment between sequences. The method compares every pair of characters in the two sequences and generates an alignment. This alignment will include matched and mismatched characters and gaps in the two sequences that are positioned such that the number of matches is maximum between identical or related characters. Dynamic programming algorithm provides a reliable computational method for aligning DNA and protein sequences. The method has been proven mathematically to produce the best or optimal alignment between two sequences under a given set of match conditions. Both global and local types of alignments may be made by simple changes in this basic alignment algorithm. Dynamic programming can be performed using a divide-and-conquer strategy. The problem should be divided into smaller sub-problems. The smaller sub-problems can be solved optimally. Using these sub-optimal solutions, a global optimal solution can be constructed for the original problem. A global alignment program is based on the Needleman-Wunsch algorithm [4] and a local alignment program is based on the Smith-Waterman algorithm [5]. Dynamic programming algorithm is guaranteed to produce an optimal alignment that has the highest possible alignment score between two sequences for a given scoring system. The rules for computing the scoring matrix are:

A matrix $S(i, j)$ indexed by residues of each sequence is built recursively,

$$S_{(i,j)} = \max \left\{ \begin{array}{l} S_{(i-1,j-1)} + w_{(x_i, y_j)}, \text{ match/mismatch} \\ S_{(i-1,j)} + w \\ S_{(i,j-1)} + w \end{array} \right.$$

where, $S_{(i,j)}$ = maximum substitution score for residues $i$ and $j$

w = gap penalty

$w_{(xi, yj)}$ = match or mismatch of residues

In global alignment, the gap penalty value is set to zero, and, any positive integer value is set for match or mismatch of the residues.

Local alignments using the Smith-Waterman algorithm are usually more meaningful than the global matches because they include patterns that are conserved in the sequences. The local alignments can be used instead of the Needleman-Wunsch algorithm to match two sequences that may have a matched region that is only a fraction of their lengths that have different lengths that overlap or where one sequence is a fragment or subsequence of the other. The rules for calculating scoring matrix values are slightly different. The scoring system includes negative integers for mismatch.

The above-mentioned algorithms consist of three steps, namely, (i) initiation of the score matrix, (ii) calculation of scores (matrix fill), and (iii) trace back, which give rise to a 2-D matrix representation. The alignments are produced by starting at the highest scoring positions in the scoring matrix and following the trace path from those positions up to a box that scores zero.

Alignment-based bioinformatics tools that made a good impact in the field are:

- Sequence similarity search tools: BLAST [6], FASTA [7],
- MSA multiple sequence alignment tools: ClustalW [8], Muscle [9], MAFFT [10]
- Sequences' profile search programs: PSI-BLAST [6], HMMER/Pfam [11]),
- Whole-genome alignment tools progressive Mauve [12], BLASTZ [13], TBA [14].

### 1.4. Problems with Multiple Sequence Alignment (MSA) Approach

MSA approach is computing-intensive and the computational load escalates as a power function of the length of the sequences. The compute time/load is $2^n$ for un-gapped alignment and somewhat higher for the best-gapped algorithms. Due to this limitation, MSA is not feasible for searching large databases. In addition to the demand for long computational time, they have other problems owing to fundamental basis. In order to infer homology among sequences, MSA converts sequences of unequal length into sequences of equal length by introducing gaps. MSA overlooks well-documented evolutionary events such as insertion and deletion (*indels*), and mutation and does not consider fluidity resulting from recombination with shuffling of conserved segments without loss of function [15-16]. This fact was illustrated by Radomski *et al.* [17] by comparing the genomes of *Aeropyrum pernix* and *Ureoplasma urealiticum*one. Only 3% (114 out of 3499) of protein sequences of *Ureoplasma urealiticum*one could be aligned with those of the *Aeropyrum pernix genome*. Moreover, inferring gaps to align sequences introduces uncertainty and this does not affect the results significantly when orthologous sequences that do not have great divergence between the taxons are considered. The alignment uncertainty significantly affects the results of phylogenetic studies of a large number of alignments. Wong *et al.* [18] showed the effects of uncertainty in alignment using genomic data from seven yeast species. Zielezinski *et al.* [19] discuss situations where the alignment-based sequence comparison method becomes troublesome. Several new alignment-free sequence comparison methods are reported as an alternative to eliminate or minimize the defects in the alignment-based approaches.

## 2. ALIGNMENT-FREE SEQUENCE COMPARISON

### 2.1. Graphical Representation

Graphical representation of DNA sequences was introduced to facilitate visual comparison of similarities or dissimilarities among the lengthy sequences, and they provide a simple way of viewing, sorting, and comparing various gene structures. In a way, they may be regarded as alignment-free sequence comparison methods.

#### *2.1.1. G- and H- curves*

The first attempt towards developing a graphical representation of DNA sequences was of Hamori [20-22] using the *G*- and *H*- curves. *G*-curves are generated in a virtual five-dimensional space whose orthogonal coordinates are assigned to the four DNA nucleotides on a DNA chain. G-curves are constructed by reading the DNA sequence data and directing the *G*-curve along the *a*-axis by one unit if the first nucleotide is adenine (*a*), along the *c*-axis if it is cytosine (*c*), and so on. The position axis is incremented by one unit after each nucleotide is drawn and the procedure is continued until the 3' end is reached. The resulting line drawn is a continuous curve along the position axis in 5-D space. It is very difficult to comprehend 5-D geometry; *G*-curves are useful only conceptually and do not provide any visual representation. To obtain a visual representation of DNA, Hamori and Ruskin [22] placed the four bases at corners of a square assigning them coordinates: $a = (1, -1)$; $c = (-1, 1)$; $g = (-1,-1)$; $t = (1, 1)$ and added the third coordinate as the running index. This way of 3-D graphical representation of DNA was referred by Hamori and Ruskin as *H*-curve [22]. Sophisticated computer graphic tools are needed to generate the *H*-curve.

#### *2.1.2. Two-dimensional Graphical Representations*

Two-dimensional (2-D) graphical representation is the most commonly used approach by many researchers, and it is much simpler than the *H*-curves proposed by Hamori. The 2-D graphical representation of DNA sequence was first proposed by Gates [23]. Nandy [24], and Leong and Morgenthaler [25] independently developed 2-D graphical representation methods. The three methods are similar to one another and only differ in the assignment axis in the Cartesian coordinates. Two dimensional (2-D) graphical representation by any of the three approaches was obtained by associating the four DNA bases with stepwise movement in the Euclidean plane, the Cartesian coordinate system. Successive points obtained by depicting the base distribution in $(x, y)$ plane are connected to obtain the DNA graphs. Applications of graphical representation on the Euclidean plan were studied by several authors [26-34].

Major disadvantages of 2-D representation using rectangular walks on Cartesian coordinates are the formation of overlapping paths due to repetitive sequences, formation of circuits or loops. Though Nandy and Nandy [35] showed that the loss of information in such approaches is not very significant and several alternative systems were suggested to remove the degeneracy. The directed graph method [36] and the condensed graph representation by Randić [37] aimed at overcoming the problem of degeneracy in graphical representations using regular cartesian coordinates. Guo *et al.* [38-39] suggested using four special vectors to represent the four DNA bases to eliminate degeneracy. However, this method results in stochastic degeneracy (accidental). Liu *et al.* [40] showed that complete elimination of degeneracy in the graphical representations generated by rectangular walk in Cartesian coordinates is impossible. Bielińska-Waż *et al.* [41] modified Nandy's approach and used a dynamic model by assigning masses to the points and this removed most of the degeneracy in the Nandy method [24].

Yau *et al.* [42] generated DNA graphs by using the first and fourth quadrant of the Cartesian system. The two DNA bases *t* and *c* (pyrimidine) were plotted in the first quadrant, while *a* and *g* (purine) were plotted in the fourth quadrant. This approach completely eliminated degeneracy in the 2-D graphical representation of DNA sequences. Similar methods were proposed by Liao *et al.* [43]. This method was utilized to analyze similarity/ dissimilarity among the sequences. Huang *et al.* [44] suggested a modification to the method of Yau *et al.* [42]. In their method, the four bases are assigned four different two-component vectors. The 2-D graphical representations thus obtained were called the *H-L* curves by the authors.

He and Wang [45] generated 2-D graphical representations based on the different classifications of the four bases into pyrimidine (*t* and *c*) and purine (*a* and g), amino

(*a* and *c*) and keto (*g* and *t*), and weak (*a* and *t*) and strong (*g* and *c*) hydrogen bond. This method eliminates degeneracy and took into consideration the chemical, structural and hydrogen bonding natures. "Four line" graphical representation [46-48] is another 2-D graphical representation of DNA sequence. In this representation, the four bases in a sequence are represented on four horizontal lines separated by unit distances. The bases constituting the DNA sequence are represented by dots placed along the horizontal lines sequentially at unit distance intervals. The adjacent dots are then connected to obtain a zig-zag curve. The four horizontal lines can be labeled in any order. Hence, the number of possible graphs is 4! Several authors used the 2-D graphs to compare sequences [49-50]. Nandy and his co-workers [51-53] have used the graphical representation to identify conserved domains and developed a method to design peptide-based drugs for the Zika virus. Their approach uses graphical representation and numerical characterization (discussed later in this review) of DNA sequences or parts of them.

#### *2.1.3. 3-D and Higher Order Graphical Representation*

The generalization of 2-D graphical representation has led to 3-D and even 4-D representations. The higher-order representations do not suffer from the arbitrary assignment of nucleic acid bases to choose the directions of the partition coordinate system axis. These methods generally used geometric possibility. Wang *et al.* [54] came up with a symbolic dynamic procedure of graphical representation. This method can be used for visual comparison of DNA sequences in 3-D coordinates and it does not suffer from any loss of information.

#### *2.1.4. Chaos Game Representation (CGR)*

Chaos game representation (CGR) of DNA sequences was introduced by Jeffrey [55]. CGR is a scale-independent, highly compact graphical representation that characterizes the various fractal patterns of DNA sequences. Jeffry's CGR representation is an improvement of the method introduced by Barnsley [56]. In Barnsley's method, the vertices of a polygon are chosen randomly. In Jeffrey's CGR method, the polygon is always a square and the random vertices are defined by the four bases as A: (0, 0), C: (0, 1), G: (1, 1) and T: (1, 0). CGR representation of a DNA sequence is obtained by starting at the center of the square. The first base is plotted halfway between the center of the square and the corner corresponding to the base. The next base is plotted halfway between the point just plotted and the corner corresponding to the next base. Thus, successive bases are plotted halfway way between the previous point and the corner corresponding to the base being plotted. This process continues till the last base in the DNA sequence is reached.

Hill *et al.* [57] used the CGR representation as a map of 16 dinucleotides and compared seven genes of human globin regions (coding regions) and 29 alcohol dehydrogenase genes that are relatively conserved. They used the frequencies of dinucleotides in each CGR map and found that CGRs were similar for the genes of the same or closely related species but were different for the genes from distantly related species. The density of CGR points was used to derive entropy profiles for DNA sequences and was used to study variability; intra and intergenomic variability [58]. Goldman [59] demonstrated the use of CGR map to study the frequencies of mono, di, and trinucleotides. This is an extension of the dinucleotide frequencies by Hill *et al.* [57]. Thus, each position in the CGR map may be considered to represent an oligonucleotide. Deschavanne *et al.* [60] found the oligomer frequency counts using CGR and tried to interpret it in terms of the genomic signature concept. Tino [61] expanded the application of CGR to other biological sequences, including proteins, by generalizing the arbitrary sequence of symbols. The work of Basu *et al.* [62] may be considered a motivation for developing CGRs of proteins.

Goldman [58] stated that the quadrant frequencies of mono-, di-, and tri-nucleotides completely account for the complex pattern of CGRs of DNA sequences and this appeared to limit the application of CGR. On the contrary, Almeida *et al.* [63] showed that one can go beyond trinucleotide and, as a matter-of-fact, oligonucleotide of all lengths [64]. Almeida and Vinga [65] showed that CGR is a special case of Markov chain models where probability tables of all non-integer orders are covered. They also pointed out that the closer the CGR points, the higher the sequence similarity between two sequences. Almeida and Vinga [65] extended the application of CGR in sequence comparison by developing a universal sequence map (USM). Joseph and Sasikumar [66] utilized the distance between CGR points to identify sequence identities of whole-genome sequences and developed an algorithm for alignment-free sequence comparison using CGR maps. Randić [67] used chaos-game representation and came up with a spectrum-like graphical representation of DNA. He utilized the 8 × 8 table of codons in the chaos-game representation to get the spectrum like representation. Hao *et al.* [68] proposed the global visualization method of DNA sequences based on counting and coarse-graining of frequency of appearances of nucleic acid bases in a given length. Similar to chaos-game representation maps, the approach of Hao *et al.* resulted in distinctive patterns for different genomes and revealed fractal patterns of DNA sequences.

## 3. NUMERICAL CHARACTERIZATION

### 3.1. Matrices used to Represent DNA-graphs

Once a connected graph is obtained for a DNA sequence, it can be represented by a matrix. The adjacency matrix is not generally used in the numerical characterization of DNA graphs. However, distance matrix and other matrices such as L/L matrix, M/M matrix, and higher order L/L matrices are used. The *E* matrix [47, 69] is obtained based on the Euclidean distance. It is a symmetric matrix, $E = E^T$, whose $(i, j)$ element is defined as the Euclidean distance between vertices $i$ and $j$ of the zig-zag curve (DNA-graph). This is also referred to as the *d*-matrix. M/M matrix is obtained as the quotient of the Euclidean distance (d) between two vertices of the zig-zag curve and the number of edges (the so-called graph-theoretical distance, D) between the two vertices [68-71]. The M/M matrix is also square symmetric. The L/L matrix is the symmetric matrix whose off-diagonal elements are defined

as a quotient of the Euclidean distance between two vertices of the zig-zag curve and the sum of geometrical lengths of edges between the two vertices [70-78]. By definition, all diagonal entries are zero. The 'higher order' *L/*L matrix is a symmetric matrix whose $(i, j)$ element is $([L/L]_{ij})^k$. In other words, the *L/*L matrix is the product of Hadamard multiplication of the L/L matrix by itself $k$-times. The exponent $k$ can take larger and larger values and as $k$ approaches infinity, the eigenvalues converge to the eigenvalues of the sparse binary matrix L/L obtained from the *L/*L matrix by the substitution of all its elements, whose values are less than one, with zero. The M/M matrix may also serve as a source of the 'higher-order' matrices and it is necessary to normalize the elements of the M/M matrix to secure convergence because they are greater than one. Since E, M/M, and L/L matrices are associated with the Euclidean and geometrical distances, they cannot distinguish curves that are symmetrical about a line in 2-D or 3-D space. In order to eliminate this limitation, Dai *et al.* [79] proposed the use of $C(i, j)$ matrix as an alternative. If $v_i$ and $v_j$ are two points (vertices) in a DNA graph, we consider the center of the circle determined by $v_i$, $v_j$, and $v_0$. The fixed point $v_0$ is chosen according to the plane curve. On choosing the origin (0, 0) as the fixed point $v_0$, the $C(i, j)$ matrix is constructed as follows:

$$C_{ij} = \frac{1}{r} \text{ if } r \neq 0; r \text{ otherwise}$$

where r is the radius of the circle determined by $v_i$, $v_j$, and $v_0$. Elements of $C(i, j)$ matrix are symmetric and the main diagonal elements are zero.

### 3.2. Numerical Descriptors from DNA-graphs

Once a matrix is constructed, its eigenvalues can be computed. The complete set of eigenvalues or any specified value, such as the leading eigenvalue of a matrix, can be used as a numerical descriptor. A DNA sequence that was converted into a 2-D or 3-D graph can thus be represented by a matrix. By using appropriate matrix manipulation, a numerical descriptor is obtained.

Euclidean distance (d) is always equal to or less than the graph distance ($D$). Hence, before computing the eigenvalues for the d/D matrix, the matrix elements are raised to higher powers until all the elements <1 vanish, leaving only the unit ratios. Randić *et al.* [80] proposed a condensed matrix for DNA sequences. The rows and columns of this matrix are associated with the appearances of the four DNA bases. This was extended to the appearances of the codons (triplet) to construct 4 × 4 matrices. These matrices were used to extract the corresponding eigenvalues. In the case of a $C(i, j)$ matrix, if an appropriate fixed point is selected, then the elements of $C(i, j)$ matrix are less than or equal to 1. This allows the computation of $C^*(i, j)$ matrix, the product of Hadamard multiplication of the $C(i, j)$ matrix by itself $k$ times ($k > 1$). Dai *et al.* [79] used the invariants derived from the $C(i, j)$ matrix to construct a phylogenetic tree for the first exons of ten β-globin genes.

Bielińska-Waż *et al.* [81] extended the utility of their 2-D dynamic representations approach by using distribution moments of the 2-D representation as a numerical descriptor. Differences in the plots of a family of genes can be assessed qualitatively. The first-order moments such as $\mu_{1x}$ and $\mu_{1y}$ were obtained. $\mu_1(x)$ and $\mu_1(y)$ are the sum of $x$ and $y$ coordinates values of each point averaged by the total number of points in the distribution. Thus, graph radius can be defined as:

$$\mu_x = \frac{\sum x_i}{N} \quad \mu_y = \frac{\sum y_i}{N}$$

$$gR = \sqrt{\mu_x^2 + \mu_y^2}$$

Where the $x_i$ and $y_i$ represent coordinates of each point on the plot, and N is the total number of bases in the segment. $gR$ can be called the base distribution indices. Graph similarity/dissimilarity index is another invariant and can be defined as $\Delta gR$

$$\Delta gR = \sqrt{(\mu_{1x} - \mu_{2x})^2 + (\mu_{1y} - \mu_{2y})^2}$$

where $\mu_1$ and $\mu_2$ are two different sequences. $gR$ is specifically useful in comparing equal length sequences [81]. Base distribution index ($gR$), $\mu_x$, $\mu_y$ and graph similarity/dissimilarity index ($\Delta gR$) are the descriptors obtained from the graphical method. Ghosh *et al.* [82, 83] and his co-workers improved the numerical characterization by using the coordinates of points in a 20-D plane formed using the amino acids corresponding to the triplets. This has been used by them to study the neuraminidase gene sequences of Influenza virus-A subtypes *H5N1* and *H1N1*.

Zhang *et al.* [84] extended the scope of 3D graphical representation by introducing an invariant to compare DNA sequences. DNA similarity analysis proposed by Li *et al.* [85] approach requires neither the graphical representation nor the conversion of the DNA graphs to invariants. They used the relative frequency of codon usage as given by the formula

$$v_{ij} = \frac{\left|F_{ij} - F'_{ij}\right|}{\frac{1}{n_i}\sum_{m=1}^{n} F_{im}}$$

Where $F_{ij}$ denotes the frequency of occurrence of the jth triplet of $C_i$ (i = 0,1,2 ……..,20), while $F'_{ij}$ is the ideal codon usage frequency of the corresponding triplet, $n_i$ is the size of $C_i$. Todeschini *et al.* [86] used partial orders to find the similarity among DNA sequences of first exons of β-globin genes for eight different species.

### 3.3. Alignment-free Sequence Comparison by a k-string Method

Though there are a large number of papers on graphical representation and their numerical characterization, they did not attract the attention of biologists as alternatives to alignment-based sequence comparison. However, chaos game representation is an exception to this. $k$-String method uses the linguistics of biological sequences. A biological sequence $S$ of length $l$ is a linear succession of $n$ symbols from a finite alphabet, A, of length $r$. For a DNA sequence

A = {a, t, g, c} and $r = 4$, while for an amino acid sequence A = {symbol of 20 amino acids}, and $r = 20$. A segment of $k$ symbols where $k \leq n$ is called a $k$-tuple, $k$-mer, $k$-string or $k$-word. The number of words or strings of length $k$ in a sequence is usually counted by moving a sliding window of width $k$ ($k$-wide) through the sequence from position one to $(n - k + 1)$. For example, in a DNA sequence *ttaggcggcggaacc* of length fifteen, there are eleven strings of $k = 4$ as shown below:

*ttagg taggc aggcg ggcgg gcggc cggcg ggcgg gcgga cggaa ggaac gaacc*

The number of possible combinations of $k$-tuple of the symbols in the alphabet A, $N_k = r^k$. For DNA sequences, $N_k = 4^4$, and $N_k = 20^k$ for amino acid sequences; however, all the possible combinations ($N_k$) may not be present in the sequence $S$ of length $l$. There may be some repetition of certain combinations. In the illustration, there are eleven 5-strings but they belong to ten different types (equivalent classes). From the frequencies of different word-length, one can calculate the probability of finding a specific word. The $k$-tuple frequencies profiles thus obtained can be used to compare sequences without alignment. The first usage of $k$-mer counts for alignment-free sequence comparison was implemented by Blaisdell [87]. Blaisdell advocated the use of Euclidean distance to measure similarity/ dissimilarity using $k$-string frequency profiles. Information-theoretic measures such as the Kullback-Leibler distance (KLD), which is also called 'relative entropy' was introduced by Wu *et al.* [88]. Other methods include geometric measures such as the cosine of the angle between the count vectors [89], and statistical measures such as the correlation coefficient [90]. Standardized Euclidean distance and 'Mahalanobis' distance were suggested [88, 91] as alternative measures to account for the variances of $k$-words in computing the Euclidean distance between the count vectors. Instead of considering counts of all $k$-words, some selected subsets were used to compare sequences [86]. In their approach, prior to computing the cosine of the angle between the $k$-word count vectors of two sequences, they obtained the most important k-words using singular value decomposition. Van Helden [92] used a motif-finding program to identify statistically significant $k$-words and based the similarity score of two regulatory sequences on the counts of these significant $k$-words only. The fixed $k$-word string composition is also known as the Feature Frequency Profile (FFP) method. Generally, FFP method uses "distance" between word frequency profiles of two tests as a measure of dissimilarity between two texts. However, there are no "words" in the long string of bases, and therefore, Sims *et al.* [93, 94] used relative $l$-mer frequencies to calculate distance scores. Wu *et al.* [88] supplemented the FFP method with horizontal gene transfer [HGT] detection techniques and constructed the whole-genome phylogeny for a population of viruses from eleven viral families comprising 142 large dsDNA eukaryote viruses. Frequency information of all profile features ($k$-word) of a given length and the length of resolution of the features are used to obtain optimal range. These methods used to compare two sequences based on $k$-word profile are also referred to as the $D_2$ method; Jing *et al.* [95] suggested statistical measures to improve its application. Qi *et al.* [96, 97] developed a composition vector tree (CVTree) method based on $k$-word composition to construct a phylogenetic tree of prokaryotes. They used the composition vector of $k$-string to construct distance matrix and the trees were drawn using PHYLIP online tool (https://evolution.genetics.washington.edu/phylip.html). Wu *et al.* [98] improved the method using a complete composition vector, i.e. they used strings of variable length instead of a particular length ($k$-string) to form the composition vector. Lu *et al.* [99] used CCV and improved CCV (ICCV) tree methods to improve the CVTree originally proposed by Qi *et al.* [97]

### 3.4. Alignment-free Sequence Comparison using N-dimensional Similarity Space

Jayalakshmi *et al.* [100-103] developed an alignment-free sequence comparison using an N-dimensional similarity space constructed by the principal components extracted from a pool of DNA descriptors. The descriptors used by them comprise of some already known ones and several new descriptors that were developed by them using the ideas prevalent in chemometrics for the numerical characterization of molecules (molecular descriptors). The complete set of descriptors was classified into primary and secondary descriptors. Primary descriptors are those obtained directly by the numerical characterization of a given DNA sequence. For example, information content is a primary descriptor. Secondary descriptors are those obtained from the graphical representation of a DNA sequence. For example, graph radius explained in section 3.2 on Numerical descriptors from DNA graphs is a secondary descriptor. Fig (1) shows a schematic representation of the calculation of primary and secondary descriptors.

#### *3.4.1. Information Content*

Information content (IC) based on nucleotides and the one based on amino acids that would be formed from the coding sequence are similar to the $L$-tuple or $k$-mer approaches. In addition to IC, two new measures of information content, namely, sequence information content ($SIC_L$) and the complementary sequence information content ($CSIC_L$) were also calculated [103] based on the following expressions:

Sequence information content $SIC_L = \dfrac{IC_L}{\log_2 n_L}$

Complementary sequence information content $CSIC_L = \log_2 n_L - IC_L$

These calculations are similar to information contents introduced by Basak *et al.* for chemical molecules [104].

#### *3.4.2. Connectivity Indices for DNA Sequences*

The connectivity type index proposed by Natarajan *et al.*, [105] is an extension of molecular connectivity introduced in the realm of chemical graph theory by Randić [106]. In the connectivity approach, each chemical molecule is depicted as a graph (molecular graph) in which the atoms form the vertices, and the covalent bonds form the edges. In a molecular graph, the degree ($\delta$) of each vertex is the number of vertices directly connected to it. Randić connectivity index is calculated from the degrees ($\delta$) of a vertex using the relation given below:

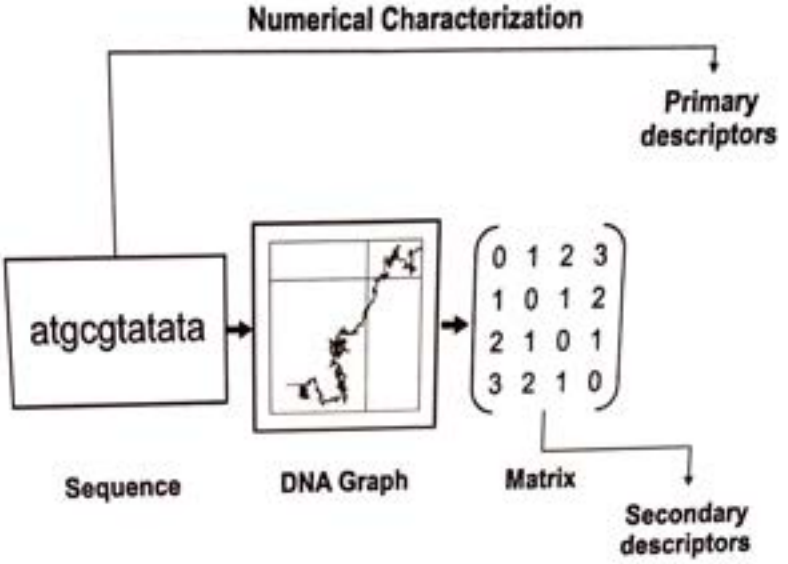



**Fig. (1).** Schematic representation calculation of primary and secondary descriptors.

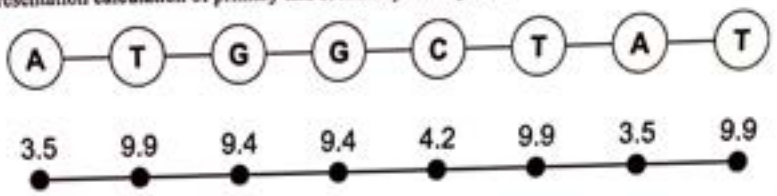



**Fig. (2).** Conversion of a DNA sequence into a line graph using $pK_a$ values of the four bases as the vertex weights. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

$$^{1}\chi = \sum \frac{1}{\sqrt{\delta_i \delta_j}}$$

Where $i$ and $j$ are pairs of non-hydrogen atoms connected by a bond (edge) and the summation is over all the bonds in a molecule, and degree $\delta_i$ of a vertex $i$ is the number of edges incident on the vertex. Kier *et al.* [107] developed a generalized connectivity index $^{h}\chi$ considering paths of type $v_0, v_1, \ldots v_h$ of length $h$ in a molecular graph. It may be noted that the path is the shortest distance, showing the number of edges between any two vertices. In the case of weighted graphs, vertices are assigned weights based on several schemes such as bond-order and valency. A generalized connectivity index $^{h}\chi$ of length $h$ can be calculated from the equation.

$$^{h}\chi = \sum \frac{1}{\sqrt{\delta_i \delta_j \ldots \delta_h}}$$

The approach used in chemical graph is extended to the numerical characterization of DNA sequences. In order to calculate connectivity type indices, DNA sequences need to be represented as graphs similar to molecular graphs containing edges and vertices. A DNA sequence was converted into a graph in which each nitrogenous base represented a vertex while the phosphate and sugar units were suppressed. Each vertex was then assigned a vertex weight based on the dissociation constant ($pK_a$) of the base corresponding to the vertex. Dissociation constants ($pK_a$) at 25°C are: adenine = 3.5, thymine = 9.9, guanine = 9.4, cytosine = 4.2. The conversion of a DNA sequence to a DNA-line graph is shown in Fig. (2).

Connectivity indices for a DNA sequence were calculated based on $pK_a$ values of each of the four nucleic acid bases. Hence, the notation $^{h}\chi^{pK_a}$ is suggested for the new set of connectivity descriptors. For the hypothetical sequence of length fifteen (*cgttgtaataattat*), there are fourteen edges namely, (*cg*) (*gt*) (*tt*) (*tg*) (*gt*) (*ta*) (*aa*) (*at*) (*ta*) (*aa*) (*at*) (*tt*) (*ta*) (*at*). Connectivity index $^{1}\chi$ is then,

$$^{1}\chi = \frac{1}{\sqrt{4.2\times9.4}} + 3\frac{1}{\sqrt{9.4\times9.9}} + 2\frac{1}{\sqrt{9.9\times9.9}} + 6\frac{1}{\sqrt{9.9\times3.5}} + 2\frac{1}{\sqrt{3.5\times3.5}} = 2.2629$$

### *3.4.3. Secondary Descriptors from DNA Graphs*

Of the various methods of graphical representation, the authors preferred to use the methods proposed by 1) Nandy [24] and 2) Yau *et al.* [42]. Though this method suffers from loss of information due to degeneracy and formation of circuits (retracing the path traversed), Nandy and co-workers used it extensively in various applications, including the studies on sequences of H5N1 [82, 83].

A series of new DNA-invariants that could be derived from the DNA-graphs generated by the two methods were proposed and these new descriptors are namely, leading eigenvalue of the distance matrix ($\lambda_D$), Wiener index [108] from the distance matrix ($W_D$), leading eigenvalues of the D/d matrix ($\lambda_{D/d}$), and Wiener index from the D/d matrix ($W_{D/d}$). These invariants were calculated for Nandy and Yau graphical representations of sequences based on the graph from which they are derived. These invariants are denoted as $\lambda_{N\text{-}D}$, $\lambda_{Y\text{-}D}$, $\lambda_{N\text{-}D/d}$, $\lambda_{Y\text{-}D/d}$, $W_{N\text{-}D}$, $W_{Y\text{-}D}$, $W_{N\text{-}D/d}$, and $W_{Y\text{-}D/d}$. In

addition to this, ovality (O) and ovality per nucleotide ($\bar{O}$) were calculated only for plots generated using the method proposed by Yau *et al.* [42]. Ovality is the ratio of length versus width of the smallest rectangular box that could contain the DNA graph. Graph radius [81] was calculated for both Nandy and Yau-type graph ($G_{r\text{-}N}$ and $G_{r\text{-}Y}$).

#### 3.4.3.1. Leading Eigenvalue ($\lambda_D$) of the Graph Distance Matrix (D)

In order to calculate the leading eigenvalue $\lambda_D$, the distance matrix had to be constructed first. The distance matrix is formed by calculating the number of edges traversed in order to move from point (nucleotide) *i* to point *j* on the DNA graph. The number of edges traversed from one point to the other is represented as Graph Distance (*D*). For each point on the graph, *D* is calculated to every other point on the graph resulting in the graph distance matrix. The resulting matrix is symmetrical along the diagonal. The leading eigenvalue ($\lambda_D$) of this matrix is a unique identifier of the sequence and hence a descriptor of that sequence.

#### 3.4.3.2. Leading Eigenvalue of D/d Matrix ($\lambda_{D/d}$)

Construction of the D/d matrix requires the construction of the *d* matrix. The distance *d* from point *i* to point *j* was calculated as the Euclidean distance between the points ($d_{ij}$). It is the shortest distance through space (plane) between the two points. Upon constructing the *d* matrix, every element in the *D* matrix is divided by the corresponding element from the *d* matrix in order to construct the D/d matrix. The leading eigenvalue of the D/d matrix ($\lambda_{D/d}$) is another useful descriptor of the sequence.

#### 3.4.3.3. Wiener Index for DNA Graphs (W)

Wiener index was introduced [108] as path index and was shown to correlate with boiling points of alkanes. Later, it was showed to be the sum of the upper triangle or the lower triangle of the distance matrix. Since the distance matrix is square symmetric containing elements $D_{ij}$ and the diagonal elements are zero (when $i = j$, $D_{ij} = 0$), *W* can be calculated as half the sum of the distance matrix ( $W = \frac{1}{2}\sum d_{ij}$ ). Calculation of Wiener index or Wiener number is extended from molecular graphs to DNA graphs. Wiener number can be calculated for both the *D-matrix* and the *D/d-matrix* of Nandy graph and modified Yau-type graph. Four new DNA descriptors, namely, Wiener index for *D*-matrix ($W_D$) and Wiener index for *D*-matrix and D/d matrix ($W_{D/d}$) for Nandy-type and Yau-type DNA graphs were introduced. The four Wiener numbers are denoted as $W_{N\text{-}D}$, $W_{N\text{-}D/d}$, $W_{Y\text{-}D}$ and $W_{Y\text{-}D/d}$.

The complete set of DNA descriptors used by Jayalaskhmi and co-worker [100-103, 105] is given in Table **2**.

### 3.5. Construction of N-dimensional Space for Sequence Comparison

The scheme for sequence comparison after data reduction by Principal Component Analysis (PCA) is shown in Fig (**3**).

Some of the calculated descriptors are highly inter-correlated and thus encoded redundant information. Perfectly correlated descriptors (correlation coefficient $r = 1$) were grouped and one descriptor from each group was retained and the others from the same group were dropped. In addition to this, descriptors that had the same numerical value for more than 90% of the sequences and those with sparse data (only zeroes and ones) were also dropped.

**Table 2. Symbols and brief definitions of the DNA descriptors used.**

| Symbol | Definition |
|---|---|
| **Primary descriptors: calculated directly from the sequence information** | |
| $IC_L$ | Information content for L-tuple (L= 1 – 10) |
| $IC_{Codon}$ | Information content for codon |
| $IC_{AA}$ | Information content for amino acid |
| $CSIC_L$ | Complementary sequence information content for L-tuple (L= 1 – 10) |
| $CSIC_{Codon}$ | Complementary sequence information content for codon |
| $CSIC_{AA}$ | Complementary sequence information content for amino acid |
| $SIC_L$ | Sequence information content for L-tuple (L = 0 – 10) |
| $SIC_{Codon}$ | Sequence information content for codon |
| $SIC_{AA}$ | Sequence information content for amino acid |
| $\chi^{pka}$ | Connectivity-based descriptors derived from DNA graph using $pK_a$ of the bases, $\chi^{pka}_i$, i = 0 to 10 |
| AT-GC | AT-GC ratio; Sum (A+T)/Sum(G+C) |
| **Secondary descriptors: calculated from graphical representations of DNA sequences** | |
| $G_{rN}$ | Graph radius based on Nandy representation |
| $G_{rY}$ | Graph radius on Yau-type representation |
| $\lambda_{N\text{-}D}$ | Leading eigenvalue for Nandy plot |
| $\lambda_{N\text{-}D/d}$ | Sum of off-diagonal elements for distance matrix for Nandy plot |
| $W_{N\text{-}D}$ | Wiener index of D matrix - Nandy plot |
| $W_{N\text{-}D/d}$ | Wiener index of D/d - Nandy plot |
| $\lambda_{Y\text{-}D}$ | Leading eigenvalue for Yau plot |
| $\lambda_{Y\text{-}D/d}$ | Sum of off-diagonal elements for distance matrix for Yau plot |
| $W_{Y\text{-}D}$ | Wiener index of D matrix – Yau plot |
| $W_{Y\text{-}D/d}$ | Wiener index for D/d matrix – Yau plot |
| O | Ovality |
| $\bar{O}$ | Ovality per nucleotide |

PCA is scale-dependent and the results are affected by the descriptors whose scales are several orders higher in magnitude than that of others. The input descriptors were scale transformed as $\log_e(Desc + x)$ where *Desc* is the numerical value of the descriptor and $x = 1$ when $Desc > -1$, which is true for most of the descriptors. However, higher-order information contents and ovality per nucleotide have values $\leq -1$ and for these descriptors, x is the smallest whole number which results in a positive sum for $(Desc + x)$. It is a general practice to retain principal components with eigenvalue $\geq 1.0$. Otherwise, principal components (PCs)

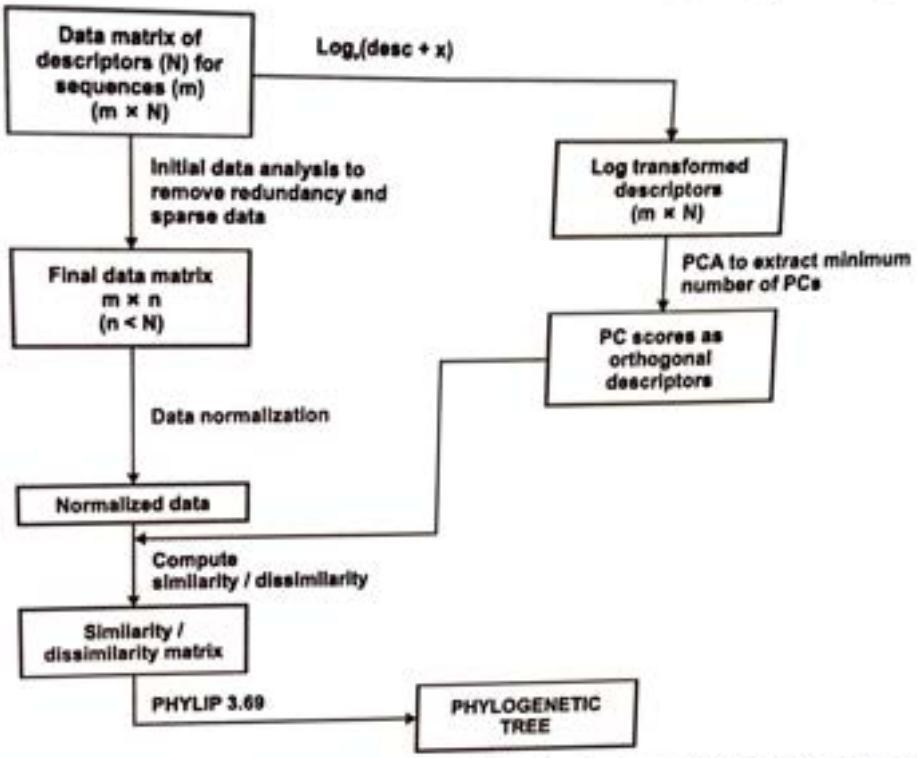



**Fig. (3).** Scheme followed for data reduction and extraction of orthogonal descriptors for alignment-free sequence comparison.

could be selected to account for maximum data variance. As the number of input descriptors used in the study is not very large, it was decided to extract PCs with eigenvalue ≥ 0.75 so that maximum data variance was obtained. The principal component scores (PCS) of the extracted factors were saved as new descriptors to carry out further statistical analyses. If needed, the PCs were subjected to *varimax* rotation to obtain rotated factor scores.

The extracted PCs were used to construct an *n*-dimensional similarity space in which similar sequences get clustered. The given set of sequences was grouped into number of clusters based upon the Euclidean distance from a centroid. Members (sequences) in the same cluster are similar, while those in different clusters are dissimilar. Thus, a similarity/dissimilarity clustering procedure was used to study similarity among sequences that are orthologous or non-orthologous.

### *3.5.1. Classification of Sequences from a Diverse Set of Genes*

Five hundred and sixty (560) coding region sequences (CDS) of different genes that are homogenous and heterogeneous were downloaded from GenBank (http://www.ncbi.nlm.nih/gov). An initial set of primary and secondary descriptors containing 70 descriptors were computed for the 560 DNA sequences using in-house computer programs developed by the group. Thus, a data matrix 560 × 70 was created. After initial data cleaning followed by PCA, the data matrix is reduced to 560 × 5 because five PCs accounted for 92% of data variance. The 5-dimensional similarity/dissimilarity space thus constructed clustered similar sequences in a cluster. The approach was extended to Coding Regions Sequences (CDS) of hemagglutinin gene of Influenza-A for subtypes H1 to H16 retrieved from GenBank (http://www.ncbi.nlm. nih. gov/). A total of 954 sequences were used and eight principal components were extracted and they accounted for 93% of data variance. The 8-dimensional similarity space clustered similar sequences together. The method is effective in classifying a diverse set of sequences in the first instance and sequences belonging to the same gene (HA) in the case of influenza virus A subtypes.

### *3.5.2. Phylogenetic Analysis using DNA Descriptors*

Though the new approach is useful in the classification of sequences, biologists are always more interested to see the effectiveness of the method in studying the phylogenetic relationship of homologous sets of sequences. An extension of the approach to carry out phylogenetic analysis to understand evolutionary relationships is presented in the following sections.

Principal component scores (PCS) extracted from the computed DNA descriptors used to construct the similarity/dissimilarity matrix were used as the input for PHYLIP 3.69 to construct a phylogenetic tree. The descriptors were also used as input for PHYLIP after normalization. Three methods of normalization, namely, z-score normalization, median absolute deviation normalization (MAD) and min-max normalization, were tested for each of the three different choices of descriptors *viz*, primary, secondary, and combined set of descriptors. Therefore, DNA-descriptors (primary and/or secondary) or principal component scores extracted after PCA were used to construct phylogenetic trees to choose the best alignment-free method for studying gene evolution. The results were compared with phylogenetic

trees generated by conventional sequence alignment. The procedure for phylogenetic analysis was standardized with respect to normalization, choice of descriptors, and combining descriptor sets, for fifteen β-globin genes' first exon sequences (15). The procedure was then extended to the following sets of sequences

i. Coding region sequences of β-globin (15)

ii. 16S rRNA genes (19)

iii. mtDNA (19)

iv. Mitochondrial genomic sequences of invertebrates and invertebrates (23)

The summary of important results of the phylogenetic analysis carried out using computed DNA descriptors is given below:

Best results were obtained when primary descriptors were used alone to construct the phylogenetic trees. Secondary descriptors alone or combination of primary and secondary descriptors did not give satisfactory phylogenetic trees. Secondary descriptors were derived from the DNA graphs constructed on the Euclidean plane. Any loss of information in generating the graph gets transferred to the descriptor and this might be the possible reason for the poor performance of secondary descriptors. Out of the three normalization procedures, MAD, z-score, and min-max normalization, z-score normalization was found to give trees that very well resembled species evolution. Data reduction using principal component and performing *varimax* rotation of the extracted principal component scores did not improve the results of phylogenetic analysis. Hence, it was concluded to carry out phylogenetic analysis using primary descriptors after normalization using z-scores. The classification of 16S rRNA sequences (19), mitochondrial DNA sequences (12) and genomic sequences mitochondrial DNA (23) gave highly acceptable phylogenetic trees. The phylogenetic trees constructed using the primary descriptors by alignment-free methods were better than those obtained by alignment algorithms. Certain anomalous groups found in phylogenetic classification by alignment method were eliminated in the classification by the alignment-free method.

Phylogenetic trees for the four sets of sequences are presented in Figs. (**4-7**).

## 4. SOME RECENT ALIGNMENT-FREE SEQUENCE COMPARISON METHODS AND TOOLS

A current approach [109] is BBO (Biogeography-based Optimization) which is based on the concept of emigration and immigration of species from one habitat to another. BBO is now improved, and it is named as IBBOMSA (An Improved Biogeography-based Approach for Multiple Sequence Alignment) [110]. The algorithm implements a mutation operator which calculates the probability of mutation in the given species according to their comparison [110]. Zhou *et al.*, [111] suggested an alignment-free comparison approach that is based on triplets of nucleotides in a gene that code for amino acids.

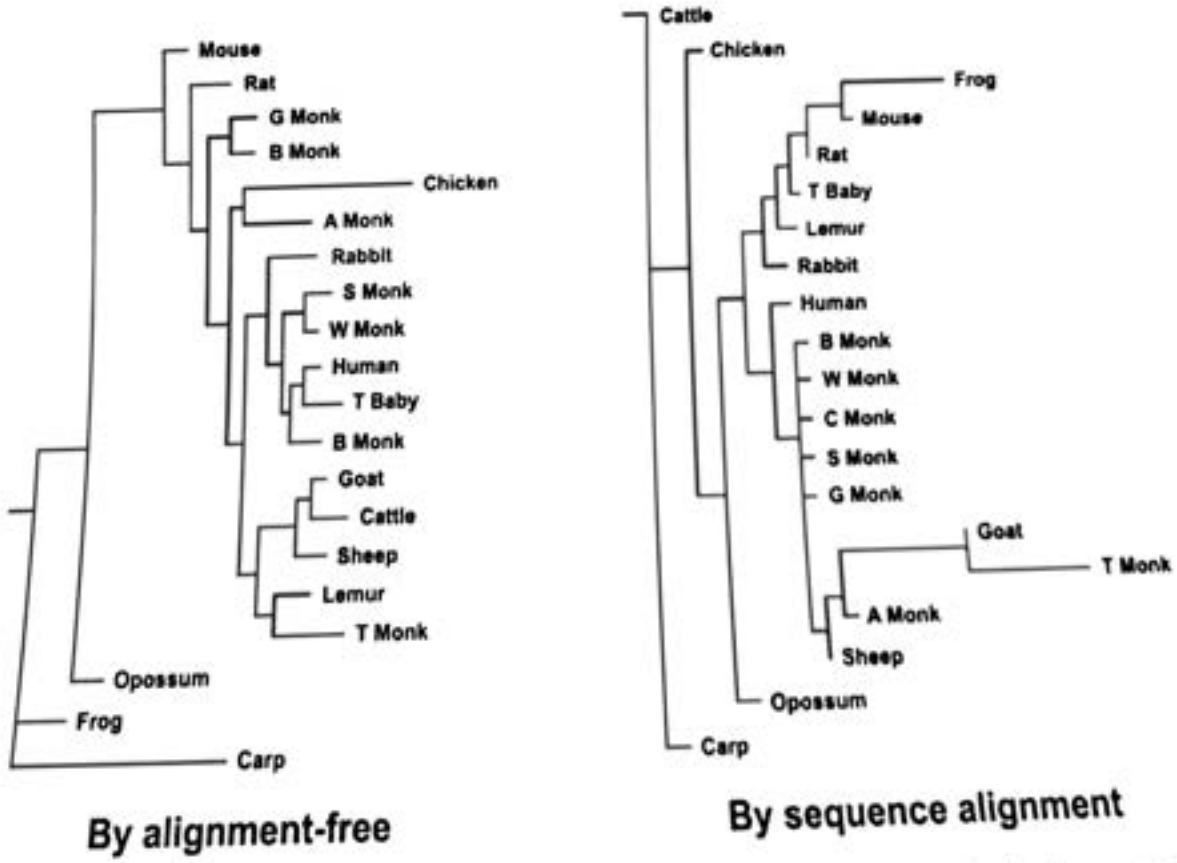



*Fig. (4).* Phylogenetic trees generated for CDS sequences of β globin gene using an alignment-free method and online sequence alignment method. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

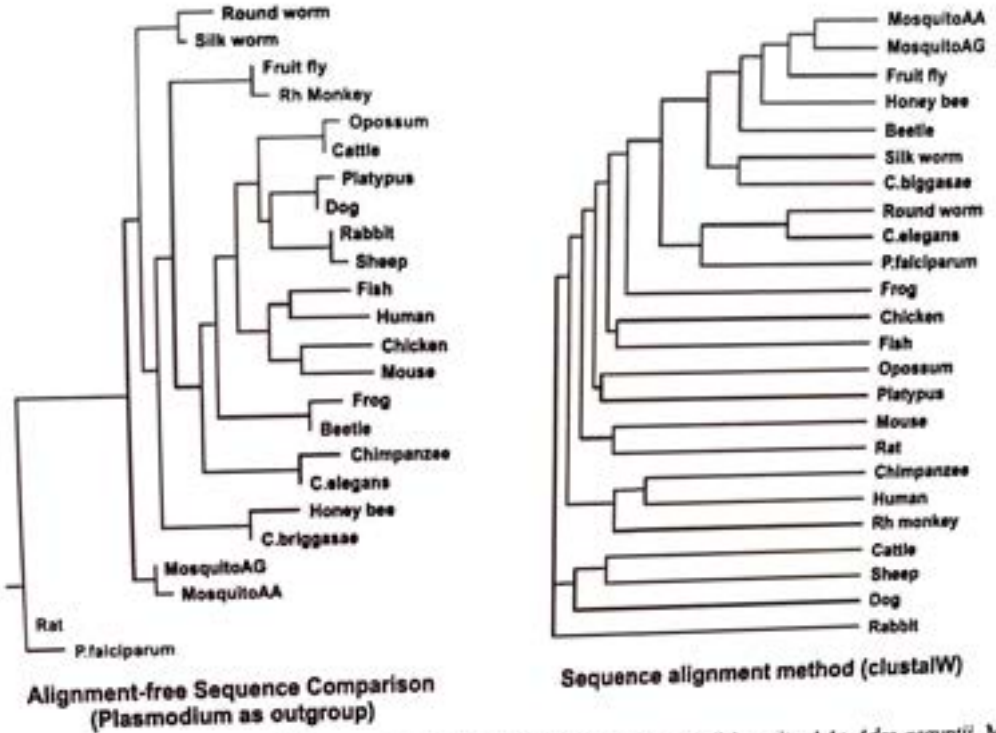



**Fig. (7).** Phylogenetic trees generated for whole genomic sequences of mitochondrial DNA (Mosquito AA: *Ades aegyptii*, Mosquito AG: *Anopheles gambiae*, Fruit: Fruit fly, Honey: Honey bee, Silk: Silk worm, Round: Round worm, Rh_monkey: Rhesus monkey). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

**Table 2. Alignment free sequence comparison tools based on Python.**

| Tools | Primary Feature | Reference | URL |
|---|---|---|---|
| Phy-Mer | Alignment-free and reference-independent mitochondrial haplogroup classifier. (Uses Python) | [112] | https://github.com/danielnavarrogomez/phymer |
| MICADo | MICADo is a tool to perform variant calling on targeted (third) next-generation sequencing data. Its algorithm is based on colored de Bruijn graphs. (Uses Python, R, Shell, Perl and TeX), | [113] | http://github.com/cbib/MICADo |
| AAF | Assembly and alignment-free method (*k*-mer based) of phylogeny reconstruction from next-generation sequencing data (Uses Python, C and R) | [114] | https://github.com/fanhuan/AAF |
| FOCUS | Find Organism by Composition Usage. Web-service software. An alignment-free model to identify organisms in metagenomes using non-negative least squares. | [115] | http://edwards.sdsu.edu/FOCUS/ |
| LMAT | Livermore Metagenomic Analysis Toolkit Livermore Metagenomic Analysis Toolkit. It assigns taxonomic labels to as many reads as possible in very large metagenomic datasets and report the taxonomic profile of the input sample. Reads by *k*-mer searches. Uses C++ and Python | [116] | https://sourceforge.net/projects/lmat/ |
| StringMLST | *k*-mer-based tool for detecting MLST directly from the genome sequencing reads. Uses Python. | [117] | http://jordan.biology.gatech.edu/page/software/stringMLST |
| d2tool | d2Tools are the toolbox for counting the frequency of *K*-tuple from sequencing datasets and then calculating the pairwise dissimilarity matrix between samples. Written by Python and R | [118, 119] | https://code.google.com/p/d2-tools/ |

With the advent of Python and Bio-python, sequence retrieval from databases such as PDB has become very easy. Especially Bio-python with its built-in capabilities has simplified string comparison and thus biological sequence comparison. Some of the tools/ software based on Python are listed in Table **2**. A more comprehensive list of alignment-free comparison tools is given in the review by Zielezinski *et al.* [19].

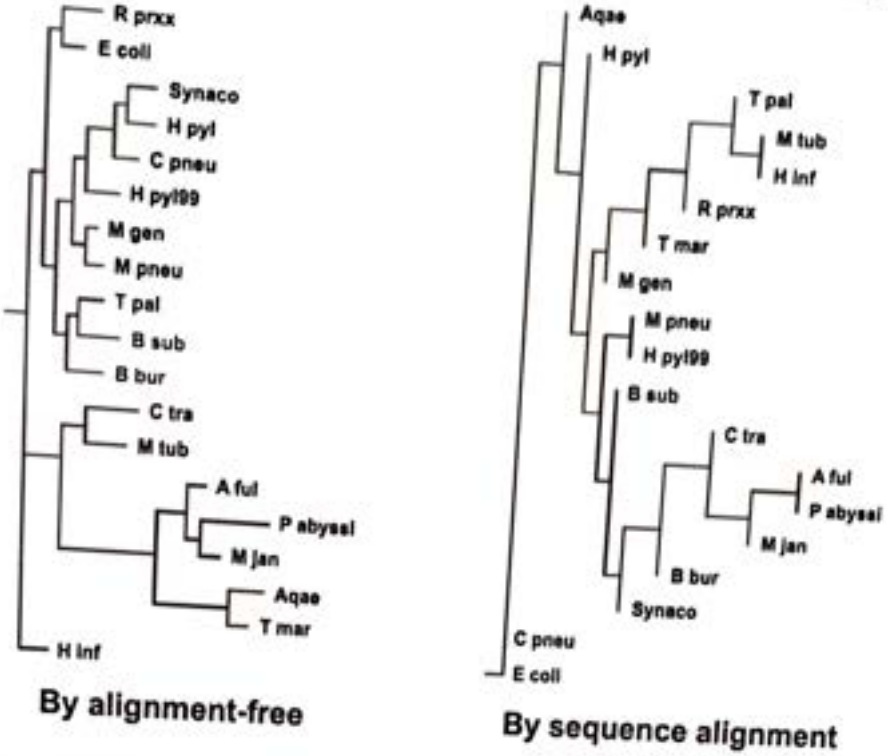



**Fig. (5).** Phylogenetic trees generated for 16S rRNA sequences using alignment-free methods and online sequence alignment tools. Hinf: *Haemophilus influenza*, Mgen: *Mycoplasma genetalium*, Synacho: *Synechocystis* sp, Mpneu: *Mycoplasma pneumonia*, *E. coli*: *Escherichia coli*, Bsub: *Bacillus subtilis*, Bbur: *Borrelia burgdorferi*, Aqea: *Aquifex aeolicus*, Mtub: *Mycobacterium tuberculosis*, Tpal: *Treponema pallidum*, Ctra: *Chlamydia trachomatis*, Rprxx: *Rickettisa prowazekii*, Cpneu: *Chlamydia pneumonia*, Tmar: *Thermotoga maritime*, Hpyl: *Helicobacter pylori*, Hpyl99: *Helicobacter pylori*, Mgan: *Methanococcus jannaschii*, Aful: *Archaeoglobus fulgidus*, Pabyssi: *Pyrococcus abyssi*. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

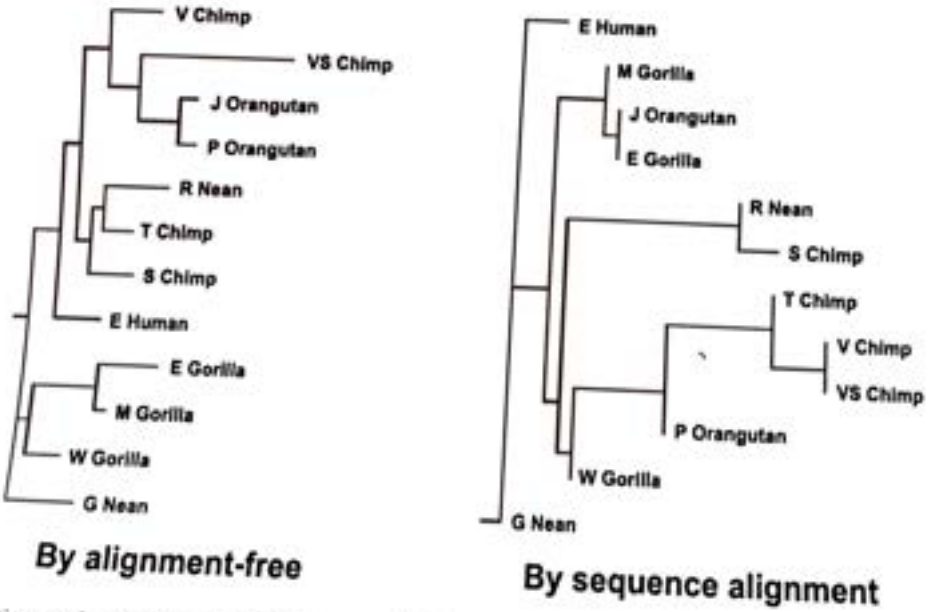



**Fig (6).** Phylogenetic trees for mitochondrial DNA sequences (G.Nean: German Neanderthals, R.Nean: Russian Neanderthals, E. Human: European Human, P Oranguta: Puti Orangutan, J Oranguta: Jari Orangutan, T Chimp: Chimp Troglodytes, V Chimp: Chimp verus, VS Chimp: Chimp Vellerosus, S Chimp: Chimp Schweinfurthii, M. Gorilla: Mountain gorilla, ELL Gorilla: Eastern Low Land Gorilla, WLL Gorilla: Western Low Land Gorilla). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

## CONCLUSION

There was a flow of ideas from different realms of science such as theoretical physics, mathematics, statistics, chemometrics, and computer science into the field of bioinformatics that emerged after the human genome project (GHP). This enabled well-tested and proven concepts from other fields to be adapted in bioinformatics to convert the large volume of data that started pouring into information and then knowledge. Though all approaches were not successful, some of them contributed considerably. In the case of alignment-free comparison of sequences by graphical representation, Chaos Game Representation (CGR) was more effective than others. However, the graphical representation approach introduced by Nandy [24] was extended by his team to study viral sequences of infectious diseases such as influenza [82, 83], and Zika virus fever [120]. They used their graphical method combined with numerical characterization to identify conserved regions [51] by non-alignment comparison of sequences and this has led to peptide-based vaccine design for zika virus [120, 53]. Bielińska-Wąż *et al.* [121] recently reported alignment-free comparison of sequences by a method called 4D-dynamic time evolution and origin of species. The method is an extension of 2D and 3D graphical representations of DNA/RNA sequences and could be used to analyze the complete genome of virus sequences. The authors applied the new approach to studying the complete genome sequences of the 2019 novel coronavirus. From the 3D classification, maps obtained, the authors could infer that the SARS-CoV-2 might have originated in bat and in pangolin. The preliminary results obtained on applying the method to zika virus sequence data show that it is possible to study the time evolution of genome sequences of viruses using this method. Alignment-free sequence comparison using numerical descriptors generated from sequences was shown to be very effective in phylogenetic analysis and evolutionary biology. The phylogenetic trees obtained by the new alignment-free method developed by Jayalakshmi *et al.* [100-103, 105] were better than those generated using an online alignment-based tool, namely PHYLIP. The L-tuple or the k-mer method based on information theory is a very effective alignment-free sequence comparison tool and is being used in several new bioinformatics tools. The alignment-free comparison methods and data mining have been given an impetus by Python programming language and R language - statistical computing freeware. With the increased use of deep learning and neural network models, an increase in algorithms for sequence comparison combining multiple techniques may be expected in the future.

## CONSENT FOR PUBLICATION

Not applicable.

## FUNDING

None.

## CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

## ACKNOWLEDGEMENTS

Declared none.

## REFERENCES

[1] Baxevanis, A.D.; Ouellette, B.F.F. *Bioinformatics: A practical guide to the analysis of genes and proteins*; John Wiley & Sons: New York, **2005**.

[2] Thompson, J.D.; Plewniak, F.; Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **1999**, *27*(13), 2682-2690. http://dx.doi.org/10.1093/nar/27.13.2682 PMID: 10373585

[3] Prjibelski, A.D.; Korobeynikov, A.I.; Lapidus, A.L. Sequence analysis. *Encyclopaedia of bioinformatics and computational biology*; Ranganathan, S.; Gribskov, M.; Nakai, K.; Schönbach, C., Eds.; Academic Press, **2019**, pp. 292-322. http://dx.doi.org/10.1016/B978-0-12-809633-8.20106-4

[4] Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **1970**, *48*(3), 443-453. http://dx.doi.org/10.1016/0022-2836(70)90057-4 PMID: 5420325

[5] Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.*, **1981**, *147*(1), 195-197. http://dx.doi.org/10.1016/0022-2836(81)90087-5 PMID: 7265238

[6] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25*(17), 3389-3402. http://dx.doi.org/10.1093/nar/25.17.3389 PMID: 9254694

[7] Pearson, W.R.; Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **1988**, *85*(8), 2444-2448. http://dx.doi.org/10.1073/pnas.85.8.2444 PMID: 3162770

[8] Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **1994**, *22*(22), 4673-4680. http://dx.doi.org/10.1093/nar/22.22.4673 PMID: 7984417

[9] Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **2004**, *32*(5), 1792-1797. http://dx.doi.org/10.1093/nar/gkh340 PMID: 15034147

[10] Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **2002**, *30*(14), 3059-3066. http://dx.doi.org/10.1093/nar/gkf436 PMID: 12136088

[11] Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E.L.; Tate, J.; Punta, M. Pfam: the protein families database. *Nucleic Acids Res.*, **2014**, *42*(Database issue), D222-D230. http://dx.doi.org/10.1093/nar/gkt1223 PMID: 24288371

[12] Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **2010**, *5*(6), e11147. http://dx.doi.org/10.1371/journal.pone.0011147 PMID: 20593022

[13] Schwartz, S.; Kent, W.J.; Smit, A.; Zhang, Z.; Baertsch, R.; Hardison, R.C.; Haussler, D.; Miller, W. Human-mouse alignments with BLASTZ. *Genome Res.*, **2003**, *13*(1), 103-107. http://dx.doi.org/10.1101/gr.809403 PMID: 12529312

[14] Blanchette, M.; Kent, W.J.; Riemer, C.; Elnitski, L.; Smit, A.F.; Roskin, K.M.; Baertsch, R.; Rosenbloom, K.; Clawson, H.; Green, E.D.; Haussler, D.; Miller, W. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **2004**, *14*(4), 708-715. http://dx.doi.org/10.1101/gr.1933104 PMID: 15060014

[15] Lynch, M. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA*, **2002**, *99*(9), 6118-6123. http://dx.doi.org/10.1073/pnas.092595699 PMID: 11983904

[16] Zhang, Y.X.; Perry, K.; Vinci, V.A.; Powell, K.; Stemmer, W.P.; del Cardayré, S.B. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, **2002**, *415*(6872), 644-646. http://dx.doi.org/10.1038/415644a PMID: 11832946

[17] Radomski, J.P.; Slominski, P.P. Primary sequences of proteins from complete genomes display a singular periodicity: Alignment-free *N*-gram analysis. *C.R. Biol.*, **2007**, *330*(1), 33-48. http://dx.doi.org/10.1016/j.crvi.2006.11.001 PMID: 17241946

[18] Wong, K.M.; Suchard, M.A.; Huelsenbeck, J.P. Alignment uncertainty and genomic analysis. *Science*, **2008**, *319*(5862), 473-476. http://dx.doi.org/10.1126/science.1151532 PMID: 18218900

[19] Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.*, **2017**, *18*(1), 186. http://dx.doi.org/10.1186/s13059-017-1319-7 PMID: 28974235

[20] Hamori, E. Visualization of biological information encoded in DNA. *Frontiers of scientific visualization*; Pickover, C.; Tewksbury, S.K., Eds.; John Wiley & Sons: New York, **1994**.

[21] Hamori, E. Novel DNA sequence repeesentations. *Nature*, **1985**, *314*(6012), 585-586. http://dx.doi.org/10.1038/314585a0 PMID: 3990794

[22] Hamori, E.; Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.*, **1983**, *258*(2), 1318-1327. http://dx.doi.org/10.1016/S0021-9258(18)33196-X PMID: 6822501

[23] Gates, M.A. Simpler DNA sequence repeesentations. *Nature*, **1985**, *316*(6025), 219. http://dx.doi.org/10.1038/316219a0 PMID: 3927167

[24] Nandy, A. New graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Curr. Sci.*, **1994**, *66*, 309-314.

[25] Leong, P.M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.*, **1995**, *11*(5), 503-507. http://dx.doi.org/10.1093/bioinformatics/11.5.503 PMID: 8590173

[26] Nandy, A. Investigation on evolutionary changes in base distribution in gene sequences. *Internet Electron. J. Mole. Des.*, **2002**, *1*, 545-558.

[27] Nandy, A.; Basak, S.C. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*(4), 915-919. http://dx.doi.org/10.1021/ci990117a PMID: 10955518

[28] Nandy, A. Graphical analysis of DNA Structure III. indications of evolutionary distictions and characteristics of introns and exons. *Curr. Sci.*, **1996**, *70*, 661-668.

[29] Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for macromolecular sequences. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*(2), 243-247. http://dx.doi.org/10.1021/ci980077v PMID: 10192941

[30] Yao, Y.; Wang, T.M. A class of New 2-D graphical representation of DNA sequences and their application. *Chem. Phys. Lett.*, **2004**, *398*, 318-323. http://dx.doi.org/10.1016/j.cplett.2004.09.087

[31] Yao, Y.; Nan, X.; Wang, T. A new 2D graphical representation-classification curve and the analysis of similarity/dissimilarity of DNA sequences. *J. Mol. Struct. Theochem.*, **2006**, *764*, 101-108. http://dx.doi.org/10.1016/j.theochem.2006.02.007

[32] Bai, F.L.; Liu, Y.Z.; Wang, T.M. A representation of DNA primary sequences by random walk. *Math. Biosci.*, **2007**, *209*(1), 282-291. http://dx.doi.org/10.1016/j.mbs.2006.06.004 PMID: 17306305

[33] Wang, J.; Zhang, Y. Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation. *Chem. Phys. Lett.*, **2006**, *423*, 50-53. http://dx.doi.org/10.1016/j.cplett.2006.03.030

[34] Wu, K.P.; Lin, H.N.; Sung, T.Y.; Hsu, W.L. A new similarity measure among protein sequences. *Proceedings of IEEE CSB 2003*, **2003**, pp. 347-352.

[35] Nandy, A.; Nandy, P. On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models. *Chem. Phys. Lett.*, **2003**, *368*, 102-107. http://dx.doi.org/10.1016/S0009-2614(02)01830-4

[36] Li, C.; Tang, N.; Wang, J. Directed graphs of DNA sequences and their numerical characterization. *J. Theor. Biol.*, **2006**, *241*(2), 173-177. http://dx.doi.org/10.1016/j.jtbi.2005.11.023 PMID: 16384585

[37] Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*(1), 50-56. http://dx.doi.org/10.1021/ci990084z PMID: 10661549

[38] Guo, X.; Randic, M.; Basak, S.C. A novel 2D graphical repeesentation of DNA sequences of low degenaracies. *Chem. Phys. Lett.*, **2001**, *350*, 106-112. http://dx.doi.org/10.1016/S0009-2614(01)01246-5

[39] Guo, X.; Nandy, A. Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy. *Chem. Phys. Lett.*, **2003**, *369*, 361-366. http://dx.doi.org/10.1016/S0009-2614(02)02029-8

[40] Liu, Y. The numerical characterization and similarity analysis of DNA primary sequences. *Internet Electron. J. Mole. Des.*, **2002**, *1*, 675-684.

[41] Bielinska-Waz, D.; Clark, T.; Nowak, W.; Nandy, A. 2D-dynamic representation of DNA sequences. *Chem. Phys. Lett.*, **2007**, *442*, 140-144. http://dx.doi.org/10.1016/j.cplett.2007.05.050

[42] Yau, S.S.T.; Wang, J.; Niknejad, A.; Lu, C.; Jin, N.; Ho, Y.K. DNA sequence representation without degeneracy. *Nucleic Acids Res.*, **2003**, *31*(12), 3078-3080. http://dx.doi.org/10.1093/nar/gkg432 PMID: 12799435

[43] Liao, B.; Zhang, Y.S.; Ding, K.Q.; Wang, T.M. Analysis of similarity/dissimilarity of DNA sequence based on a condensed curve representation. *J. Mol. Struct. Theochem.*, **2005**, *717*, 199-203. http://dx.doi.org/10.1016/j.theochem.2004.12.015

[44] Huang, G.; Liao, B.; Li, Y.; Liu, Z.H-L. Curve: A novel 2D graphical representation for DNA sequences. *Chem. Phys. Lett.*, **2008**, *462*, 129-132. http://dx.doi.org/10.1016/j.cplett.2008.07.046

[45] He, P.; Wang, J. Numerical characterization of DNA primary sequence. *Internet Electron. J. Mole. Des.*, **2002**, *1*, 668-674.

[46] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Analysis of similarity/dissimilarity of dna sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.*, **2003**, *371*, 202-207. http://dx.doi.org/10.1016/S0009-2614(03)00244-6

[47] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.*, **2002**, *368*, 1-6. http://dx.doi.org/10.1016/S0009-2614(02)01784-0

[48] Randic, M.; Zupan, J.; Vikic-Topic, D.; Plavsic, D. A novel unexpected use of a graphical representation of DNA: graphical alignment of DNA sequences. *Chem. Phys. Lett.*, **2006**, *431*, 375-379. http://dx.doi.org/10.1016/j.cplett.2006.09.044

[49] Li, C.; Wang, J. Numcerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences. *Comb. Chem. High Throuhput. Screen.*, **2003**, *6*, 795-799. PMID: 14683485

[50] Liao, B.; Wang, T.M. New 2D graphical representation of DNA sequences. *J. Comput. Chem.*, **2004**, *25*(11), 1364-1368. http://dx.doi.org/10.1002/jcc.20060 PMID: 15185330

[51] Ghosh, A.; Chattopadhyay, S.; Chawla-Sarkar, M.; Nandy, P.; Nandy, A. *In silico* study of rotavirus VP7 surface accessible conserved regions for antiviral drug/vaccine design. *PLoS One*, **2012**, *7*(7), e40749. http://dx.doi.org/10.1371/journal.pone.0040749 PMID: 22844409

[52] Nandy, A.; Basak, S.C. An emerging immunogenomics and computational appeoach for peptide vaccinology: Rational design of peptide vaccines. *Curr. Comput. Aided Drug Des.*, **2014**, *10*, 283-284. PMID: 25994637

[53] Nandy, A.; Basak, S.C. A brief review of computer-assisted approaches to rational design of peptide vaccines. *Int. J. Mol. Sci.*, **2016**, *17*, 666. 11p http://dx.doi.org/10.3390/ijms17050666

[54] Wang, S.; Tian, F.; Feng, W.; Liu, X. Applications of representation method for DNA sequences based on symbolic dynamics. *J. Mol. Struct. Theochem.*, **2009**, 999, 33-42. http://dx.doi.org/10.1016/j.theochem.2009.05.025

[55] Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.*, **1990**, *18*(8), 2163-2170. http://dx.doi.org/10.1093/nar/18.8.2163 PMID: 2336393

[56] Barnsley, M.F. *Fractals everywhere*. Academic Press: Boston **1993**.

[57] Hill, K.A.; Schisler, N.J.; Singh, S.M. Chaos game representatio of coding regions of human globin genes and alcoh

dehydrogenase genes of phylogenetically divergent species. *J. Mol. Evol.*, **1992**, *35*(3), 261-269. http://dx.doi.org/10.1007/BF00178602 PMID: 1518093

[58] Oliver, J.L.; Bernaola-Galván, P.; Guerrero-García, J.; Román-Roldán, R. Entropic profiles of DNA sequences through chaos-game-derived images. *J. Theor. Biol.*, **1993**, *160*(4), 457-470. http://dx.doi.org/10.1006/jtbi.1993.1030 PMID: 8501918

[59] Goldman, N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res.*, **1993**, *21*(10), 2487-2491. http://dx.doi.org/10.1093/nar/21.10.2487 PMID: 8506142

[60] Deschavanne, P.J.; Giron, A.; Vilain, J.; Fagot, G.; Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, **1999**, *16*(10), 1391-1399. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026048 PMID: 10563018

[61] Tino, P. Spatial representation of symbolic sequences through iterative function systems. *IEEE Trans. Syst. Man Cybern.*, **1999**, *29*, 386-393. http://dx.doi.org/10.1109/3468.769757

[62] Basu, S.; Pan, A.; Dutta, C.; Das, J. Chaos game representation of proteins. *J. Mol. Graph. Model.*, **1997**, *15*(5), 279-289. http://dx.doi.org/10.1016/S1093-3263(97)00106-X PMID: 9640559

[63] Almeida, J.S.; Carriço, J.A.; Maretzek, A.; Noble, P.A.; Fletcher, M. Analysis of genomic sequences by chaos game representation. *Bioinformatics*, **2001**, *17*(5), 429-437. http://dx.doi.org/10.1093/bioinformatics/17.5.429 PMID: 11331237

[64] Wang, Y.; Hill, K.; Singh, S.; Kari, L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*, **2005**, *346*, 173-185. http://dx.doi.org/10.1016/j.gene.2004.10.021 PMID: 15716010

[65] Almeida, J.S.; Vinga, S. Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, **2002**, *3*, 6. http://dx.doi.org/10.1186/1471-2105-3-6 PMID: 11895567

[66] Joseph, J.; Sasikumar, R. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, **2006**, *7*, 243. http://dx.doi.org/10.1186/1471-2105-7-243 PMID: 16677374

[67] Randic, M. Another look at the chaos-game representation of DNA. *Chem. Phys. Lett.*, **2008**, *456*, 84-88. http://dx.doi.org/10.1016/j.cplett.2008.03.011

[68] Hao, B.; Lee, H.C.; Zhang, S. Fractals related to long DNA sequences and complete genomes. *Chaos Solitons Fractals*, **2000**, *11*, 825-836. http://dx.doi.org/10.1016/S0960-0779(98)00182-9

[69] Liao, B.; Wang, T.m. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. *Chem. Phys. Lett.*, **2004**, *388*, 195-200. http://dx.doi.org/10.1016/j.cplett.2004.02.089

[70] Liao, B.; Wang, T.M. 3-D graphical representation of DNA sequences and their numerical characterization. *J. Mol. Struct. Theochem.*, **2004**, *681*, 209-212. http://dx.doi.org/10.1016/j.theochem.2004.05.020

[71] Li, C.; Wang, J. Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences. *Comb. Chem. High Throughput Screen.*, **2003**, *6*(8), 795-799. http://dx.doi.org/10.2174/138620703771826900 PMID: 14683485

[72] Song, J.; Tang, H. A new 2-D graphical representation of DNA sequences and their numerical characterization. *J. Biochem. Biophys. Methods*, **2005**, *63*(3), 228-239. http://dx.doi.org/10.1016/j.jbbm.2005.04.004 PMID: 15939477

[73] Yao, Y.H.; Nan, X.Y.; Wang, T.M. A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them. *J. Comput. Chem.*, **2005**, *26*(13), 1339-1346. http://dx.doi.org/10.1002/jcc.20271 PMID: 16021599

[74] Liao, B.; Tan, M.S.; Ding, K.Q. 4D representation of DNA sequences and its application. *Chem. Phys. Lett.*, **2005**, *402*, 380-383. http://dx.doi.org/10.1016/j.cplett.2004.12.062

[75] Liao, B.; Wang, T.M. A 3D graphical representation of RNA secondary structures. *J. Biomol. Struct. Dyn.*, **2004**, *21*(6), 827-832. http://dx.doi.org/10.1080/07391102.2004.10506972 PMID: 15107004

[76] Yao, Y.H.; Liao, B.; Wang, T.M. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. *J. Mol. Struct. Theochem.*, **2005**, *755*, 131-136. http://dx.doi.org/10.1016/j.theochem.2005.08.009

[77] Zhu, W.; Liao, B.; Ding, K.Q. A condensed 3D graphical representation of RNA secondary structures. *J. Mol. Struct. Theochem.*, **2005**, *757*, 193-198. http://dx.doi.org/10.1016/j.theochem.2005.04.042

[78] Bai, F.; Wang, T. On graphical and numerical representation of protein sequences. *J. Biomol. Struct. Dyn.*, **2006**, *23*(5), 537-546. http://dx.doi.org/10.1080/07391102.2006.10507078 PMID: 16494503

[79] Dai, Q.; Liu, X.Q.; Wang, T.M. C(i,j) matrix: a better numerical characterization for graphical representations of biological sequences. *J. Theor. Biol.*, **2007**, *247*(1), 103-109. http://dx.doi.org/10.1016/j.jtbi.2007.03.002 PMID: 17428502

[80] Randić, M.; Vračko, M.; Nandy, A.; Basak, S.C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*(5), 1235-1244. http://dx.doi.org/10.1021/ci000034q PMID: 11045819

[81] Bielinska-Waz, D.; Nowak, W.; Waz, P.; Nandy, A.; Clark, T. Distribution moments of 2D-graphs as descriptors of DNA sequences. *Chem. Phys. Lett.*, **2007**, *443*, 408-413. http://dx.doi.org/10.1016/j.cplett.2007.06.088

[82] Ghosh, A.; Nandy, A.; Nandy, P.; Gute, B.D.; Basak, S.C. Computational study of dispersion and extent of mutated and duplicated sequences of the H5N1 influenza neuraminidase over the period 1997-2008. *J. Chem. Inf. Model.*, **2009**, *49*(11), 2627-2638. http://dx.doi.org/10.1021/ci9001662 PMID: 19778054

[83] Ghosh, A.; Nandy, A.; Nandy, P. Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. *BMC Struct. Biol.*, **2010**, *10*, 6. http://dx.doi.org/10.1186/1472-6807-10-6 PMID: 20170556

[84] Zhang, Y.; Tan, M. Visualization of DNA sequences based on 3DD-curves. *J. Math. Chem.*, **2008**, *44*, 206-216. http://dx.doi.org/10.1007/s10910-007-9302-2

[85] Li, C.; Yu, X.; Helal, N. Similarity analysis of DNA sequences based on codon usage. *Chem. Phys. Lett.*, **2008**, *459*, 172-174. http://dx.doi.org/10.1016/j.cplett.2008.05.039

[86] Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A. A new similarity/diversity measure for the characterization of DNA sequences. *Croat. Chem. Acta*, **2008**, *81*, 657-664.

[87] Blaisdell, B.E. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. USA.*, **1986**, *83*(14), 5155-5159. http://dx.doi.org/10.1073/pnas.83.14.5155 PMID: 3460087

[88] Wu, T.J.; Hsieh, Y.C.; Li, L.A. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, **2001**, *57*(2), 441-448. http://dx.doi.org/10.1111/j.0006-341X.2001.00441.x PMID: 11414568

[89] Stuart, G.W.; Moffett, K.; Baker, S. Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **2002**, *18*(1), 100-108. http://dx.doi.org/10.1093/bioinformatics/18.1.100 PMID: 11836217

[90] Fichant, G.; Gautier, C. Statistical method for predicting protein coding regions in nucleic acid sequences. *Comput. Appl. Biosci.*, **1987**, *3*(4), 287-295. http://dx.doi.org/10.1093/bioinformatics/3.4.287 PMID: 3134115

[91] Wu, T.J.; Burke, J.P.; Davison, D.B. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, **1997**, *53*(4), 1431-1439. http://dx.doi.org/10.2307/2533509 PMID: 9423258

[92] van Helden, J. Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, **2004**, *20*(3), 399-406. http://dx.doi.org/10.1093/bioinformatics/btg425 PMID: 14764560

[93] Sims, G.E.; Jun, S.R.; Wu, G.A.; Kim, S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA*, **2009**, *106*(8), 2677-2682. http://dx.doi.org/10.1073/pnas.0813249106 PMID: 19188606

[94] Sims, G.E.; Jun, S.R.; Wu, G.A.; Kim, S.H. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc. Natl. Acad. Sci. USA*, **2009**, *106*(40), 17077-17082. http://dx.doi.org/10.1073/pnas.0909377106 PMID: 19805074

[95] Jing, J.; Burden, C.J.; Forêt, S.; Wilson, S.R. Statistical considerations underpinning an alignment-free sequence comparison method. *J. Korean Stat. Soc.*, **2010**, *39*(3), 325-335. http://dx.doi.org/10.1016/j.jkss.2010.02.009

[96] Qi, J.; Wang, B.; Hao, B.I. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.*, **2004**, *58*(1), 1-11. http://dx.doi.org/10.1007/s00239-003-2493-7 PMID: 14743310

[97] Qi, J.; Luo, H.; Hao, B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.*, **2004**, *32*(Web Server issue), W45-7. http://dx.doi.org/10.1093/nar/gkh362 PMID: 15215347

[98] Wu, X.; Cai, Z.; Wan, X.F.; Hoang, T.; Goebel, R.; Lin, G. Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics*, **2007**, *23*(14), 1744-1752. http://dx.doi.org/10.1093/bioinformatics/btm248 PMID: 17495995

[99] Lu, G.; Zhang, S.; Fang, X. An improved string composition method for sequence comparison. *BMC Bioinformatics*, **2008**, *9*(Suppl 6), 515. http://dx.doi.org/10.1186/1471-2105-9-S6-S15

[100] Jayalakshmi, R. *Numerical charaterization of DNA sequences and its application in alignment-free sequence ccomparsion*.. Ph.D. Theisis, Baharathidasan University: Tiruchirappali, Tamil Nadu, India, **2011**.

[101] Jayalakshmi, R.; Natarajan, R.; Vivekanandan, M. Extension of molecular similarity analysis approach to classification of DNA sequences using DNA descriptors. *SAR QSAR Environ. Res.*, **2011**, *22*(1-2), 21-34. http://dx.doi.org/10.1080/1062936X.2010.528255 PMID: 21391139

[102] Jayalakshmi, R.; Natarajan, R.; Vivekanandan, M.; Natarajan, G.S. Alignment-free sequence comparison using *N-dimensional* similarity space. *Curr. Computeraided Drug Des.*, **2010**, *6*(4), 290-296. http://dx.doi.org/10.2174/1573409911006040290 PMID: 20883198

[103] Jayalakshmi, R.; Natarajan, R.; Vivekanandan, M.; Ganapathy, N. Descriptors based on information theory for numerical characterization of DNA sequences. *Curr. Sci.*, **2010**, *99*, 370-375.

[104] Basak, S.C. *Topological indices and related descriptors in QSAR and QSPR*; Devillers, J.; Balaban, A.T., Eds.; Gordon and Breach Science: The Netherlands, **1999**, pp. 563-593.

[105] Natarajan, R.; Jayalakshmi, R.; Vivekanandhan, M. Numerical characterization of DNA sequences: Connectivity type indices derived from DNA line graphs. *J. Math. Chem.*, **2010**, *8*, 521-529. http://dx.doi.org/10.1007/s10910-010-9688-0

[106] Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.*, **1975**, *97*, 6609-6615. http://dx.doi.org/10.1021/ja00856a001

[107] Kier, L.B.; Murray, W.J.; Randić, M.; Hall, L.H. Molecular connectivity V: connectivity series concept applied to density. *J. Pharm. Sci.*, **1976**, *65*(8), 1226-1230. http://dx.doi.org/10.1002/jps.2600650824 PMID: 978443

[108] Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **1947**, *69*(1), 17-20. http://dx.doi.org/10.1021/ja01193a005 PMID: 20291038

[109] Simon, D. Biogeography-based optimization. *IEEE Trans. Evol. Comput.*, **2008**, *12*, 702-713. http://dx.doi.org/10.1109/TEVC.2008.919004

[110] Yadav, R.K.; Banka, H. IBBOMSA: An improved biogeography-based approach for multiple sequence alignment. *Evol. Bioinform. Online*, **2016**, *12*, 237-246. http://dx.doi.org/10.4137/EBO.S40457 PMID: 27812276

[111] Zhou, J.; Zhong, P.; Zhang, T. A novel method for alignment-free DNA sequence similarity analysis based on the characterization of complex networks. *Evol. Bioinform. Online*, **2016**, *12*, 229-235. http://dx.doi.org/10.4137/EBO.S40474 PMID: 27746676

[112] Navarro-Gomez, D.; Leipzig, J.; Shen, L.; Lott, M.; Stassen, A.P.M.; Wallace, D.C.; Wiggs, J.L.; Falk, M.J.; van Oven, M.; Gai, X. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. *Bioinformatics*, **2015**, *31*(8), 1310-1312. http://dx.doi.org/10.1093/bioinformatics/btu825 PMID: 25505086

[113] Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **2016**, *32*(14), 2103-2110. http://dx.doi.org/10.1093/bioinformatics/btw152 PMID: 27153593

[114] Fan, H.; Ives, A.R.; Surget-Groba, Y.; Cannon, C.H. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, **2015**, *16*, 522. http://dx.doi.org/10.1186/s12864-015-1647-5 PMID: 26169061

[115] Silva, G.G.Z.; Cuevas, D.A.; Dutilh, B.E.; Edwards, R.A. FOCUS: An alignment-free model to identify organisms in metagenomes using non-negative leastsquares. *PeerJ*, **2014**, *2*, e425. http://dx.doi.org/10.7717/peerj.425 PMID: 24949242

[116] Ames, S.K.; Hysom, D.A.; Gardner, S.N.; Lloyd, G.S.; Gokhale, M.B.; Allen, J.E. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, **2013**, *29*(18), 2253-2260. http://dx.doi.org/10.1093/bioinformatics/btt389 PMID: 23828782

[117] Gupta, A.; Jordan, I.K.; Rishishwar, L. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics*, **2017**, *33*(1), 119-121. http://dx.doi.org/10.1093/bioinformatics/btw586 PMID: 27605103

[118] Wang, Y.; Liu, L.; Chen, L.; Chen, T.; Sun, F. Comparison of metatranscriptomic samples based on k-tuple frequencies. *PLoS One*, **2014**, *9*(1), e84348. http://dx.doi.org/10.1371/journal.pone.0084348 PMID: 24392128

[119] Jiang, B.; Song, K.; Ren, J.; Deng, M.; Sun, F.; Zhang, X. Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, **2012**, *13*, 730. http://dx.doi.org/10.1186/1471-2164-13-730 PMID: 23268604

[120] Dey, S.; Nandy, A.; Basak, S.C.; Nandy, P.; Das, S. A bioinformatics approach to designing a Zika virus vaccine. *Comput. Biol. Chem.*, **2017**, *68*, 143-152. http://dx.doi.org/10.1016/j.compbiolchem.2017.03.002 PMID: 28342423

[121] Bielińska-Wąż, D.; Wąż, P. Non-standard bioinformatics characterization of SARS-CoV-2. *Comput. Biol. Med.*, **2021**, *131*, 104247. http://dx.doi.org/10.1016/j.compbiomed.2021.104247 PMID: 33611129

**Shodhsamhita** : Journal of Fundamental & Comparative Research
Vol. VII, No. 12 : 2021
ISSN: 2277-7067

# A STUDY ON CONSUMERS ATTITUDE TOWARDS ONLINE GROCERY SHOPPING IN KARUR TOWN

**Ms. S. Kavitha,** *Head, Department of Commerce,Sri Sarada Niketan College of Science for women, Karur-5*
**Mrs. K. Balambal,** *Assistant Professor of Commerce,Sri Sarada Niketan College of Science for women, Karur-5*

## Abstract

Internet is changing the way consumers shop and buy goods and services, and has rapidly evolved into a global phenomenon. Many companies have started using the Electronic Commerce with the aim of cutting marketing costs, thereby reducing the price of their products and services in order to stay ahead in highly competitive markets. Companies also use the Internet to convey communicates and disseminate information, to sell the product, to take feedback and also to conduct satisfaction surveys with customers. Customers use the Internet not only to buy the product online, but also to compare prices, product features and after sale service facilities the will receive if they purchase the product from a particular store. Many experts are optimistic about the prospect of online business. In addition to the tremendous potential of the E-commerce market, the Internet provides a unique opportunity for companies to more efficiently reach existing and potential customers. Although most of the revenue of online transactions comes from business-to-business commerce, the practitioners of business-to-consumer commerce should not lose confidence. It has been more than a decade since business-to-consumer E-commerce first evolved. Scholars and practitioners of electronic commerce constantly strive to gain an improved insight into consumer behavior in cyberspace. Along with the development of E-retailing, researchers continue to explain E-consumers behavior from different perspectives. Many of their studies have posited new emergent factors or assumptions which are based on the traditional models of consumer behavior, and then examine their validity in the Internet context. This study is an attempt to know the consumer perception towards purchasing grocery item through online.

**Keywords** : Global phenomenon, E-retailing, Business-to-Business

### INTRODUCTION

Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the internet using a web browser. Consumers find a product of interest by visiting the website of the retailer directly or by searching among best alternative vendors, which displays the same products availability and pricing at different e-retailers. As of 2018, customers can shop online using a range of different consumer and industrial products. An online shop evokes the physical analogy of buying products or services at a regular "bricks-and-mortar" retailer or shopping center; the process is called business-toconsumer (B2C) online shopping. When an online store is set up to enable businesses to buy from other businesses, the process is called business-to-business (B2B) online shopping. A typical online store enables the customer to browse the firm's range of products and services, view photos or images of the products, along with information about the product specifications, features and prices. Online stores typically enable shoppers to use "search" features to find specific models, brands or items. Online customers must have access to the Internet and a valid method of payment in order to complete a transaction, such as a credit card, an Interac-enabled debit card, or a service such as PayPal. For physical products (e.g., paperback books or clothes), the e-tailer ships the products to the customer; for digital products, such as digital audio files of songs or software, the e-tailer typically sends the file to the customer over the Internet.

### CONSUMER ATTITUDE

Consumer attitudes are a composite of a consumer's beliefs about, feelings about, and behavioral intentions toward some object--within the context of marketing, usually a brand or retail store. These components are viewed together since they are highly interdependent and together represent forces that influence how the consumer will react to the object. For example, we may have a very positive view of a particular sports car (for example, we believe it performs better than most), it makes us feel good, and we intend to buy it. Consumer attitude may be defined as a feeling of favorableness or unfavorableness that an individual has towards an object. As we, all know that an individual with a positive attitude is more likely to buy a product and

this result in the possibility of liking or disliking a product. Attitude can be a fixed way of thinking when it endures for a longer time. Attitude developed through experiences may change when new experiences are obtained. Attitude towards a general consumption behavior can also occur, For example how often the person should shop food. The more positive the attitude is regard to a behavior, the stronger is the individual's intention to perform the behavior under consideration.

## STATEMENT OF THE PROBLEM

Online grocers are confronted with numerous challenges. The general lack of practical experience in consumer's needs and demands renders the development of a profitable egrocery strategy even more difficult. These considerations raise the following questions: What are the prospectus for an e-grocery business to meet the consumer's needs and demands? What are the facilitating factors and barriers to the acceptance of online grocery shopping? What are the technological challenges are faced by consumers and what type of websites should be developed for the easy browsing? What kind of logistics and distribution channel consumers are expecting? What is the customer's expectation from this channel to fulfill their daily needs? In the light of above, role to understand the consumer's attitude towards online grocery shopping is very important. That"s why, when the online grocery market is spending their wings in India slow and steadily with the time, then this study will focus on to analyze the consumer's attitude towards OGS in karur where the internet usage density is high, for acceptance of this new innovative distributions and service channels for the fulfillment of consumer's daily needs. This study, however, will investigate key issues affecting their decision and find out the level of acceptance of OGS among Indian Consumers in near future.

## OBJECTIVES OF STUDY

- To study present status of online shopping
- To analyze the factors those affects consumers attitude towards online grocery shopping
- To study factors motivates consumers attitude towards online grocery shopping
- To study potential for development of online shopping

## SCOPE OF STUDY

E-grocery is happening in India in a big way. The E-shopping will replace traditional in-store shopping in the near future. Though the traditional shopping is still to stay but the online buying could change the way people experience shopping grocery. The study is to identify whether people are changing the way they shop their food and grocery from the next door store to online web stores. The study is conducted to reveal the consumer's attitude on e-grocery and their preference towards the same. This study will help to find out whether e-grocery will exceed the preference of using the traditional way of shopping and increase the response of online grocery shopping in future. There are numerous opportunities for innovative new services.

## LIMITATIONS OF THE STUDY

- This study is mainly confined to Karur town.
- This study is limited to 25 respondents.
- The limitation of time was another constraint in the study. Research period is not just much enough to know about the attitude of customers.

## THE ONLINE GROCERY COMPANIES IN INDIA ARE

### Bigbasket

It is one of the leading online grocery markets in India. The company operates from Bangalore, Mumbai. When you browse it, there are various categories of products. A customer places an order for preferred goods. The ordered goods are delivered at the customer's doorstep at the stipulated time.

### Star Bazaar

It has attracted many customers attention. The neatly laid out, brightly colored fresh vegetables, fruits and food products are what keeps the customer coming back to the store. For effective quality service for its consumers, they have introduced a Star Bazaar mobile app that makes you shop online.

### Zopnow

Once you place an order with Zopnow, within three hours you'll have your groceries at your doorstep. The company discourages use of polythene papers in its packaging. It uses plastic containers for large orders and environment-friendly green papers for small orders. Once the goods are dispatched, the customer gets an SMS and email. The two have a link which a customer can use to track the location of the order.

### Nature's Basket

Nature's Basket is another grocery shopping app in India. When the customer places an order, the product the quality and physical condition f a team of trained staff check the product before it's sent for delivery. The order is delivered at the customer's preferable time. Once the products leave the store for transportation, a notification email and SMS was sent to the client. With the app, you can change your order while the product is on the store.

### The Prime Pantry

The app has tons of grocery commodities. It's much easier to compare prices of products at your favorite store. In this store, Amazon Prime members can buy groceries and household products through in an easy way.

### Aaramshop

This hybrid retail platform that focuses on the marketing of FMCG and CPG brands. The platform connects its users with the nearby stores. It also provides prices of the commodities for the customer to compare. Users can shop for groceries under categories like rice, lentils, beverages, water, edible oil and ghee. The user selects the nearest retailer to confirm the order. Payment is only made by cash when the commodities are delivered.

### Bazaar Cart

Bazaar Cart is an online grocery store that does free doorstep delivery. The app has more than 15k+grocery products in its categories.

### Reliance Fresh

Reliance fresh industry is confident of becoming leading top India's online grocery store. Reliance Fresh store offers online services through mobile phones, websites and mobile app in all parts of Mumbai. Reliance Fresh is top leading consumer store synonymous with freshness and savings. From fresh fruits and vegetables to cereals we have an entire range of your grocery needs covered in one store.

### ADVANTAGES OF ONLINE GROCERY SHOPPING

#### 1. Ordering with Customer Service Assistance

You don't have to wait anymore for the next working day just to place your order or if you have inquiries. Normally, online shopping stores have 24/7 customer service assistance to accommodate all customers who wish to purchase or ask about their products and services.

#### 2. Goodbye Long Lines of Checkout Lanes

The traffic inside supermarkets, especially in peak hours, is heavy. So, spare yourself of the trouble of standing and waiting! Try shopping for your groceries with just the use of your computer or smartphone wherever you are.

#### 3. Time Saver

Most people in these modern times are time conscious. As much as possible, we want convenience and efficiency. Purchasing goods online wouldn't demand so much of your time. Unlike the actual store, you won't be taking up time roaming around trying to find the products you want. In online grocery shopping, you can find the items you want in just a few clicks! They are also grouped in categories for better and faster finding! Use your time by spending quality moments with your family and let these online stores do the shopping.

#### 4. No More Parking Hassles

Finding it hard to look for a space to park your car? No problem! Especially on holidays like Christmas, where all parking spaces are all filled in, you won't even need your car when you shop online! You can choose to have your orders ready for pick-up or have them delivered at your doorstep. Simple.

**5. Assured Product Quality and Freshness**

Online shopping companies are obliged to provide the freshest produce and quality products available in the market. Products in poor quality are subject to returns, refunds, or replacement.

**6. From Store to Door Delivery**

You'll never have think of leaving your house anymore just to buy groceries. All you have to do is sit back, relax and wait for your orders to arrive. Most online supermarkets would ask you to purchase your products 3-5 days ahead of desired delivery date to ensure availability and quality.

**5. Lessens Unnecessary Spending**

Most grocers would always have a list of items to purchase that fits their budget. When we go to the store, we often come across items that are so tempting to buy, cutting the budget we try to save. Oh, and don't forget the food carts and stalls after you pay for your groceries in the checkout counter!

## ANALYSIS AND INTERPRETATION

**USEAGE OF ONLINE SHOPPING**

| S. NO | Online shopping | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Yes | 13 | 52 |
| 2 | No | 12 | 48 |
| **Total** | | **25** | **100** |

The above table shows that 52 percent of the respondents have used online shopping and 48 percent of the respondents have not used online shopping.

**STATUS ABOUT ONLINE PURCHASE**

| S. NO | Status | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Very often | 10 | 40 |
| 2 | Sometimes | 8 | 32 |
| 3 | Rarely | 7 | 28 |
| **Total** | | **25** | **100** |

The above table shows that 40 percent of the respondents are very often to purchase from online, 32 percent of the respondents sometimes purchase from online and 28 percent of the respondents are rarely to purchase from online.

**DEVICE USED TO PURCHASE ONLINE**

| S. NO | Device | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Laptop | 8 | 32 |
| 2 | Smartphone | 10 | 40 |
| 3 | Tablet | 7 | 28 |
| **Total** | | **25** | **100** |

The above table shows that 40 percent of the respondents are purchase made through smartphone, 32 percent of the respondents are purchase made through laptop and 28 percent of the respondents are purchase through tablet.

**LEVEL OF SATISFACTION**

| S. NO | Level of satisfaction | No of Respondents | Percentages |
|---|---|---|---|
| 1 | Highly satisfied | 14 | 56 |
| 2 | Satisfied | 11 | 44 |
| 3 | Unsatisfied | --- | --- |
| **Total** | | **25** | **100** |

The above table shows that 56 percent of the respondents are highly satisfied and 11 percent of the respondents are satisfied for the eco friendly products.

**TYPES OF ONLINE STORE**

| S. NO | Types of online store | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Flipkart | 9 | 36 |
| 2 | Amazon | 6 | 24 |
| 3 | Snapdeal | 5 | 20 |
| 4 | Zomato | 3 | 12 |
| 5 | Jabong | 2 | 8 |
| Total | | 25 | 100 |

The above table shows that 36 percent of the respondents are select the flipkart online store, 24 percent of the respondents are select amazon online store, 20 percent of the respondents are select snapdeal online store , 12 percent of the respondents are select zomato online store and 8 percent of the respondents are select jabong online store.

**PAYMENT METHOD USE FOR ONLINE SHOPPING**

| S. NO | Payment method | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Credit card | 9 | 36 |
| 2 | Debit card | 3 | 12 |
| 3 | Net banking | 6 | 24 |
| 4 | Cash on delivery (COD) | 7 | 28 |
| Total | | 25 | 100 |

The above table shows that 36 percent of the respondents use credit card for online payment, 28 percent of the respondents are using cash on delivery(COD), 24 percent of the respondents use debit card for online payment and 12 percent of the respondents are using net banking.

**FACTORS INFLUENCING PURCHASE DECISION**

| S. NO | Factors | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Convenience | 8 | 32 |
| 2 | Quality of product | 6 | 24 |
| 3 | Better deal discount | 4 | 16 |
| 4 | Return policy | 2 | 8 |
| 5 | Recommendation from friends | 2 | 8 |
| 6 | Free samples | 3 | 12 |
| Total | | 25 | 100 |

The above table shows that 32 percent of the respondents says online grocery shopping is convenient, 24 percent of the respondents says quality of product, 16 percent of the respondents says they gave better deal discount, 12 percent of the respondents says they gave free samples and 8 percent of the respondents says recommendation from friends and return policy.

**SAFE OF ONLINE SHOPPING**

| S. NO | Online shopping | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Safe | 13 | 52 |
| 2 | Unsafe | 12 | 48 |
| **Total** | | **25** | **100** |

The above table shows that 52 percent of the respondents says online shopping is safe and 48 percent of the respondents says online shopping unsafe.

**SPEND AMOUNT ON ONLINE STORE**

| S. NO | Spend amount | No of Respondents | Percentage |
|---|---|---|---|
| 1 | Less than 1000 | 5 | 20 |
| 2 | 1000-3000 | 6 | 24 |
| 3 | 3000-5000 | 9 | 36 |

| | | | |
|---|---|---|---|
| 4 | 5000-10000 | 3 | 12 |
| 5 | More than 10000 | 2 | 8 |
| Total | | 25 | 100 |

The above table shows that 36 percent of the respondents spend 3000-5000 rupees, 24 percent of the respondents spend 1000-3000 rupees, 20 percent of the respondents spend less than thousand, 12 percent of the respondents are spend 5000-10000 rupees and 8 percent of the respondents spend more than 10000 rupees.

## SUGGESTIONS

- Companies should aware customers regarding how security regarding customers credit card no. is maintained by companies
- Companies should made aware customers regarding return policy and procedure if wrong or bad product arrived. Companies should make returning procedure simpler, like few companies are asking customers to resend products if any wrong or bad product arrived. Instead of these companies should collect product from customers and deliver write product to them in minimum time.
- Most of customers want to see product before purchase to make sure that same product arrived as per order. Most of companies are not having this facility. If companies want to increase no. of customers they should provide this facility because in manual purchase customers are getting chance to see and touch the product and this may be the important reason behind customers first preference for manual shopping on regular basis.
- in home and electronic appliances after sales service in very important aspect, companies should inform customers regarding how to install and use the product or send company representative for installation very soon after delivery. Companies should inform customers regarding nearest service station if any problem arrived in product. Majority of customers are preferring manual shopping for home and electronic appliances in fear of after sales service.

## CONCLUSION

The online grocery industry is one of the growing industries in India. Customers because they get access to the internet more often and tend to buy things online always prefer to shop from the brands in the online that give them the best customer service and satisfaction among other brands that the major reason for purchasing groceries online is saving of time and effort and there are no time restrictions in shopping. The expectation of a customer while buying groceries online and in physical market is totally different. Most of the respondents get to know about the brand from the Internet.

## BIBLIOGRAPHY

JOSEPH, S.J,P.T, - E-Commerce, An Indian Perspective, PHI Learning Private Limited, New Delhi,Ed.2009.

NATARAJAN, L, Dr. - Retail Marketing, Margham Publications, Chennai, Ed.2014.

RADHA,V.Dr. - Business Environment, Prasanna & Co., Chennai, Ed.2008.

TRIPATHI, P.C, - Research Methodology in Social Sciences, Sultan Chand & Sons , New Delhi,Ed.2003.

## WEBSITES

- Marketing eye.com
- Web marketingindia.com
- The marketing student.com
- Wikipedia.com
- Flipkart.com

Journal of Genetic Engineering and Biotechnology

RESEARCH Open Access

# Molecular docking and identification of G-protein-coupled receptor 120 (GPR120) agonists as SARS COVID-19 MPro inhibitors

Sellappan Mohan[1*], Jayagopal Dharani[2], Ramanathan Natarajan[3] and Arumugam Nagarajan[1]

**Abstract**

COVID-19 has become a pandemic, and any new drug for treating the disease could save millions of lives. Several drugs already in use for other diseases and medical conditions are repurposed for treating COVID-19 in an attempt to find treatment for the disease without spending research time on ADME TOX and other studies on side effects. In this exercise, the drugs repurposed are from antiviral, antibiotics, antiviral for HIV and HCV, anti-cancer, natural medicines, etc. Possible repurposing anti-diabetic GPR-120 agonists used as for SAR-CoV-2 is attempted in the study by carrying out docking of 68 GPR-120 agonists. Ten of these compounds were found to have docking scores −8.3 to −8.0, and the best docking score was observed for an arylsulfonamide and a biarylpropanoic acid belonging to GPR120 agonists previously evaluated for the treatment of type II diabetes. These GPR120 agonists could serve as start point for novel inhibitors for the discovery of drugs to treat COVID-19.

## Background

After suffering from a devastating spell of COVID-19, the world is slowly limping back to normalcy. This is one of the pandemics that has a better public awareness owing to the Internet and social media. As per the data reported to WHO Globally, as of 14 October 2021, there have been 239,007,759 confirmed cases of COVID-19, including 4,871,841 deaths. As of 13 October 2021, a total of 6,471,051,151 vaccine doses have been administered (https://covid19.who.int/ accessed on October 15, 2021). The world is undergoing the largest vaccination program to guard the people from any further spells of the deadly virus.

In addition to the vaccines, the pharmaceutical companies and the scientists in various organization are trying to develop drugs to combat the SARS CoV-2. There are several targets that could be explored to develop new drugs for COVID-19 [1]. The therapeutic targets include both structural and non-structural proteins [2]. Some of the targets considered to develop inhibitors are as follows:

- Spike protein (S-protein) [3–5]
- Angiotensin-converting enzyme-2 (ACE-2) [6, 7]
- Human proteases: Transmembrane protease, serine 2 (TMPRSS2) [7, 8], Furin [9], Papain like protease-2 (PLpro) [10–12] 3-chymotrypsin like protease (3-CLpro) or the main protease $M^{Pro}$ [13]
- Viral proteases (RNA-dependent RNA-polymerase (RdRp) [14]

One of the steps taken by the scientific community to combat the pandemic was to repurpose drugs already known and in use. This provides a shortcut and reduces the considerable amount of time spent on ADME Tox studies and the burden on assessing the new drug molecule's therapeutic efficacy, side effects, and risks. Several small molecules were considered [15]. The repurposed drugs are usually broad-spectrum antivirals that fall under the two therapeutic classes namely,

*Correspondence: selvaamohan@gmail.com
[1] Karpagam College of Pharmacy, Coimbatore, Tamil Nadu 641032, India
Full list of author information is available at the end of the article

protease inhibitor and nucleosides. Among the repurposed drugs favipiravir, remdesivir, molnupiravir, galidesirvir, sobosbivir, and azivudine are examples of nucleosides while boceprevir, narlaprevir, simeprevir, and calpain inhibitors belong to protease inhibitors.

The race in finding an antiviral for COVID-19 was given momentum by computer-aided drug design approach, especially with the aid of docking software such as Auto Dock and Schrödinger. Of the several targets mentioned above, the main protease ($M^{Pro}$) has been the most explored for the development of inhibitors. One of the main reasons for exploring the $M^{Pro}$ inhibitors is its important role played in the in the replication and transcription of SARS CoV-2 [16]. The main protease ($M^{Pro}$) is one of the proteins encoded in SARS-CoV-2 genome and is a dimer of cystine protease. This is called the 3-chymotrypsin-like protease (3-CLpro). $M^{Pro}$ presents a highly conserved active site in several coronaviruses, such as SARS-CoV and MERS-CoV. $M^{Pro}$ plays an important role in the cleavage of precursor polyproteins translated from viral RNA, and no other human protease does have a similar cleavage specificity. This makes $M^{Pro}$ an attractive target for developing inhibitors, and the inhibitor may thus be non-toxic.

The development of inhibitors targeting the main protease appears to have not left any stone unturned and these repurposed molecules may be grouped into (1) inhibitors of other CoV, (2) antiviral therapeutics of human immunodeficiency virus (HIV), (3) anti-viral that are being used in hepatitis C virus (HCV), (4) antimalarial and other antivirals for influenza, (5) anti-bacterial, (6) anti-cancer drugs, (7) Traditional Chinese medicines, and (8) chemicals in traditional spices and other natural compounds from marine origin.

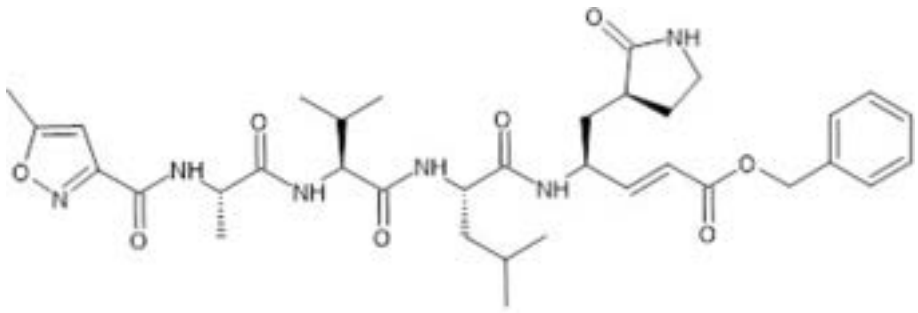
Based on the action of the main protease, Yang et al. has designed several inhibitors in 2005. The authors found molecule N3 (number assigned by Yang et al. [17]) (see Fig. 1) as the most potent inhibitor of CoV. This molecule was studied by Jin et al. [13] for SARS CoV-2 and from the results of docking study the authors proposed that N3 binds in $3CL^{Pro}$ binding pockets in an irreversible manner, and they thus exhibited good inhibitory potency. The $3CL^{Pro}$ complex withN3 molecule was used to identify new inhibitors and one such molecule is ebselen [13, 18]. Ebselen, a drug used for the treatment of stroke containing a selenium atom is repurposed for SARS-CoV-2. Crystal structure of $M^{Pro}$ without any ligand bound to the protein was reported by Zhang et al. [19] and Zhang et al. [20] studied the binding affinity of alpha-ketamide to $3CL^{Pro}$ and identified three binding pockets in the protein. By varying the four substructures (marked as A, B, C, & D in Fig. 1), they obtained the best fitting into the protein pockets from the inhibitory potency (Fig 1).

Calligari et al. [21] investigated thirteen proteinase inhibitors that are used as antiviral for human immunodeficiency virus (HIV) and hepatitis C virus (HCV). The ten anti-HIV drugs are saquinavir, indinavir, tipranavir, ritonavir, lopinavir, atazanavir, nelfinavir, amprenavir, darunavir, and fosamprenavir while the three anti-HCV aresimeprevir, faldaprevir, and asunaprevir. Among these simeprevir was found to have the highest docking score. Lopinavir/ritonavir, coformulation is sold under the brand name Kaletra as an antiretroviral medication for the treatment and prevention of HIV/AIDS. Repurposing of Kaletra for SARS CoV-2 was found to be effective. Nutho et al. [22] could explain the inhibitory efficacy of Kaletra based on the docking studies of lopinavir and ritonavir with $3CL^{Pro}$. Chang et al. [23] showed that indinavir binds with $3CL^{Pro}$ stronger than lopinavir and ritonavir and Calligari et al. [21] had also inferred this in their study. Nelfinavir was identified to be a potential inhibitor for CoV-2 from a docking that used 1903 candidates [24]. These authors went on to determine the inhibitory potency of nelfinavir [25]. Atazanavir the HIV antiviral was found to be a potential inhibitor of 3CLPro [26], and its ability to inhibit SAR CoV-2 Vero cells was studied by Fintelmen_Rodrigues et al. [27].

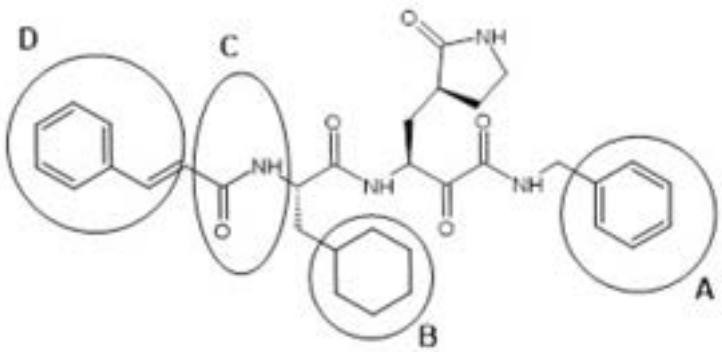
In addition to the three HCV drugs mentioned above, ledipasvir and velpatsavir were reported by Chen in 2020 [28]. Li et al. [29] ended up with four molecules namely, prulifloxacin, bictegravir, nelfinavir, and tegobuvir by high through put screening of 8000 clinical drug libraries based on the binding affinity with $M^{Pro}$. Khan et al. [30] screened 123 antiviral drugs to identify inhibitors of 3CLPro as well as 2′-O-MTase (2′-O-ribose methyltransferase). Paritaprevir and Raltegravir were found to have high binding affinity for $3CL^{Pro}$.

Talluri [31] carried out virtual screening of several clinically approved antiviral and the crystal structure of $M^{Pro}$ (PDB if 6LU7) and found saquinavir and beclabuvir as the best protease inhibitor candidate SARS CoV-2 among the compounds studied by them. Other anti-viral drugs that had been tested for repurposing by molecular docking and virtual screening include oseltamivir [32] and zanamivir [33].

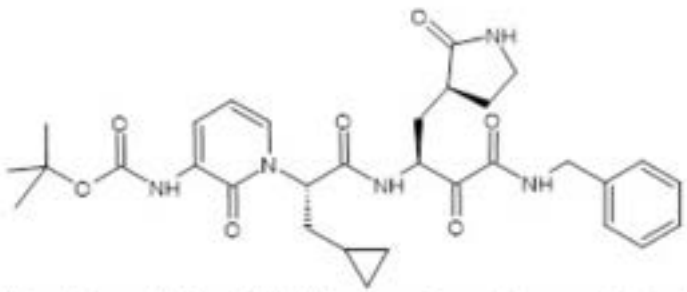
Some of the antibiotics that have been identified to be effective based on computer-aided virtual screening are the quinoline antibiotic prulifloxacin [29], tetracycline antibiotics eravacycline [34], and the polypeptide antibiotic colistin [35]. Non-steroidal anti-inflammatory drugs (NSAID) were also repurposed as potential inhibitors of MPro [36] by docking studies. In a similar study on NSAIDs, Gimeno et al. [37] identified Perampanel, Carprofen, Celecoxib, Alprazolam, Trovafloxacin,



N3 ($C_{38}H_{46}N_6O_8$)




α-ketamide (the four substructures subjected to change)



α-ketamide with best inhibitory potency (see substructures A, B, C and D for the best fitting molecule in the $M^{Pro}$ pockets)

**Fig. 1** Structure of N3, α-ketamide with sub-structures marked

Sarafloxacin, and ethyl biscoumacetate as possible inhibitors of $M^{Pro}$ by docking studies. The two compounds namely, Carprofen, a NSAID no longer use in human medicine but used for veterinary purpose, and Celecoxib another NSAID and a COX-2 inhibitor, were subjected to in vitro testing at 50 μM, and they showed 3.97% and 11.90% $M^{Pro}$ inhibition, respectively.

*Dipyridamole* (brand name Persantine) is a platelet inhibitor and is used to prevent blood clots after heart surgery was repurposed for CoV-2 by Liu et al. [38] and the inhibitory potency ($IC_{50}$) was studied targeting $3CL^{Pro}$. Odhar et al. [39] studied the drug molecules from US-FDA-approved drugs library from ZINC 5 database, and from their docking on to the $M^{Pro}$ (PDB id 6LU7), they identified ten hits that included drug that are used for cancer, epilepsy, and insomnia. The top ten hits based on the docking score are Perampanel (epilepsy), conivaptan (hyponatremia), sonidegib (basal-cell carcinoma), azelastine (allergy), idelalisib (leukemia and lymphoma), suvorexant (insomnia) olaparib (ovarian, breast, and pancreatic cancers), ponatinib (leukemia), loxapine (schizophrenia), and tolvaptan (hyponatremia). Wang et al. [34]

in the computational drug repurposing study identified carfilzomib (antineoplastic agent), valrubicin (chemotherapy drug), and elbasvir (antiviral for HCV) as inhibitors based on the docking with $M^{Pro}$, in addition to the antibiotic eravacycline.

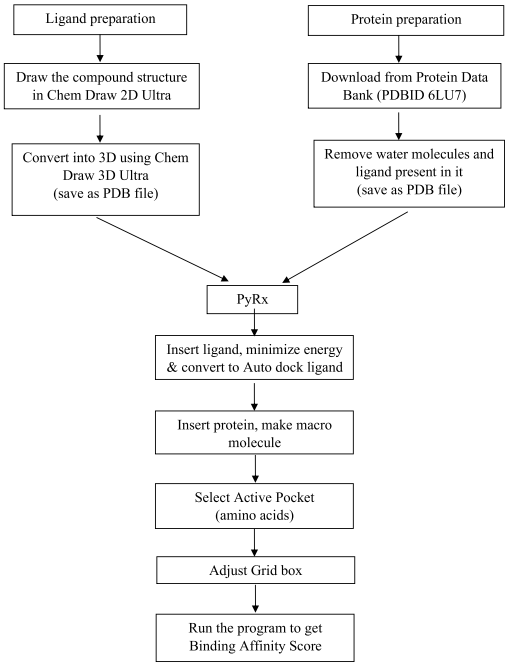
Traditional Chinese medicine (TCM) and traditional Indian medicines that fall under Ayurveda and Sidha were used as immune boosters to fight against COVID-19. Zhang et al. [19, 20] carried out docking studies of about 100 constituents of the Lung-toxin Dispelling Formula No. 1 (LDFN1) of TCM and found 22 of these chemicals are inhibitors of $3CL^{Pro}$. Of the several chemicals, baicalin and baicalein were found to have antiviral activities against $3CL^{Pro}$ [40] with $EC_{50}$ values of 10.27 μM and 1.69 μM, respectively. Liu et al. [41] studied the inhibitory activity of the ethanol extract of the herbal plant *Scutellaria baicalensis* and its major component, baicalein. They found that the plant extract and the constituent baicalein inhibited SARS-CoV-2 $3CL^{pro}$ activity in vitro with $IC_{50}$ of 8.52 mg/mL and 0.39 mM, respectively. The replication of SARS-CoV-2 in Vero cells were inhibited with EC50s of 0.74 mg/ml and 2.9 mM, respectively. In their study on screening several natural compounds that are constituents of TCM, Zhang et al. [19] and Zhang et al. [20] identified betulinic acid, coumaroyltyramine, cryptotanshinone, desmethoxyreserpine, dihomo-γ-linolenic acid, kaempferol, lignan, *N*-cis-feruloyltyramine, quercetin, sugiol, and tanshinoneiia to inhibit $3CL^{Pro}$. Cherrak et al. [42] studied several glycosylated flavonoids by docking them on the $M^{Pro}$ (6LU7) and identified quercetin-3-O-rhamnoside to have the highest binding affinity. Myricetin 3-rtinoside and rutin were also identified as potential inhibitors of $3CL^{Pro}$, and the binding affinities of these three compounds were greater than that of N3 with $3CL^{Pro}$. Shivanika et al. [43] carried out docking studies of several natural products that have been used as antiviral on to 6LU7 the 3CLPro protein structure and found theaflavin-3-3'-digallate, rutin, hypericin, robustaflavone, and (-)-solenolide as the compounds with highest binding energy. It might be noted that identification of rutin as a potential inhibitor is independently confirmed by two groups. Bhaliya and Shah [44] carried out docking studies of mono-carbonyl analogs of curcumin with 3CLPro and found one of the curcumin analogs was found to have potential to be used as an inhibitor. Joshi et al. [45] screened a library of ~7100 molecules that comprises of flavonoids, glucosinolates, anti-tussive, anti-influenza, anti-viral, terpenes, terpenoids, alkaloids, and other compounds predicted as potential therapeutic candidates against $M^{Pro}$. Molecules such as δ-viniferin, myricitrin, taiwanhomoflavone A, lactucopicrin 15-oxalate, nympholide A, afzelin, biorobin, hesperidin, and phyllaemblicin B were found to bind strongly with $M^{Pro}$ and hence suggested as potential inhibitors. Andrographolide a natural compound from *Andrographis paniculata* was studied [46] via docking on to MPro, and the in silico studies on ADME and toxicity prediction were also carried out. The molecule was predicted to have good solubility. Ramaiah et al. [47] studied the binding of natural molecules that are present in Indian spices and curry against MPro (6LU7). A similar study identified [48] carnosol a natural molecule as an inhibitor by docking studies using the protein structure PDBID: 6Y84, MPro. Bioactive compounds in medicinal plants were screened as potential MPro inhibitors [49] and natural compounds such as kaempferol, quercetin, luteolin-7-glucoside, demethoxycurcumin, naringenin, apigenin-7-glucoside, oleuropein, curcumin, catechin, and epicatechin-gallate as potential molecules for further exploration.

According to the latest report of the pharmaceutical company Merck, molnupiravir pills are able to reduce the hospitalization and deaths of people affected by COVID-19 [50]. They reported the results of Phase 2a trial (ClinicalTrials.govNCT04405570) in which safety, tolerability, and antiviral efficacy of molnupiravir in the treatment of COVID-19. Merck applied on October 11, 2021, for US-FDA emergency use authorization for the molnupiravir based-oral antiviral pill for COVID-19. This will not stop the hunt for new inhibitors, and the search for new molecules will continue.

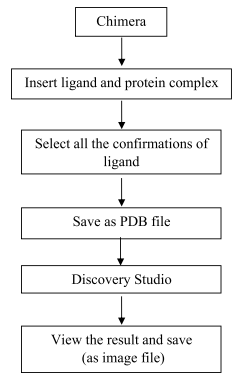
In one of the studies of repurposing drugs [51], virtually screened 1615 FDA approved drugs by docking each of them on to MPro and then refined the selection by employing molecular dynamics to identify nine compounds. The nine drugs selected as potential inhibitors vary from vasoconstrictor to microscopy dye. The potential inhibitors identified and their original use are:

1. Dihydroergotamine: vasoconstrictor
2. Midostaurin: treatment of acute myeloid leukemia
3. Ziprasidone: antipsychotic
4. Etoposide: antineoplastic
5. Apixaban: used to reduce the risk of stroke and blood clots
6. Fluorescein: a dye used in microscopy
7. Tadalafil: used to treat erectile dysfunction (ED), benign prostatic hyperplasia (BPH), and pulmonary arterial hypertension)
8. Rolapitant: used along with an antiemetic (anti-vomiting) agent in adults for the prevention of delayed nausea and vomiting associated with initial and repeat courses of emetogenic cancer chemotherapy
9. Palbociclib: used to treat HR-positive and HER2-negative breast cancer

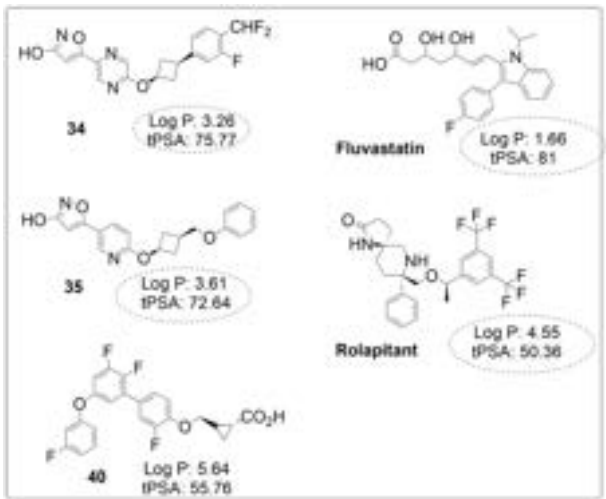
The above discussion indicates that repurposing of drugs belonging to different classes have been evaluated for CoV-2. Human GPR120 is a transmembrane protein, characterized by the interactions with the endogenous ligand linoleic acid and docosahexaenoic acid. Apart from the key role played by GPR 120 in diabetes, it is also involved in many other disease conditions, including cancer, inflammation, and central nervous system (CNS) disorders. GPR 120 presents itself in many metabolic pathways, and its pivotal role in controling obesity and diabetes is worth mentioning. Using gene knockdown studies, GPR120 has been shown to induce chemoresistance in breast cancer treatment with epirubicin and cisplatin-highlighting the relevance of GPR 120




a **Flowchart for docking procedure to obtains binding score for docking a ligand**




b **Flowchart for to view and save the protein-ligand complex**




c Structural similarity and drug like properties of docked GPR 120 agonists versus Rolapitant and Fluvastatin

| Classification | Target | Shorthand | Prediction | Probability |
|---|---|---|---|---|
| Organ toxicity | Hepatotoxicity | dili | Active | 0.64 |
| Toxicity end points | Carcinogenicity | carcino | Active | 0.51 |
| Toxicity end points | Immunotoxicity | immuno | Inactive | 0.64 |
| Toxicity end points | Mutagenicity | mutagen | Inactive | 0.63 |
| Toxicity end points | Cytotoxicity | cyto | Inactive | 0.71 |
| Tox21-Nuclear receptor signalling pathways | Aryl hydrocarbon Receptor (AhR) | nr_ahr | Inactive | 0.6 |
| Tox21-Nuclear receptor signalling pathways | Androgen Receptor (AR) | nr_ar | Inactive | 0.84 |
| Tox21-Nuclear receptor signalling pathways | Androgen Receptor Ligand Binding Domain (AR-LBD) | nr_ar_lbd | Inactive | 0.98 |
| Tox21-Nuclear receptor signalling pathways | Aromatase | nr_aromatase | Inactive | 0.83 |
| Tox21-Nuclear receptor signalling pathways | Estrogen Receptor Alpha (ER) | nr_er | Inactive | 0.74 |
| Tox21-Nuclear receptor signalling pathways | Estrogen Receptor Ligand Binding Domain (ER-LBD) | nr_er_lbd | Inactive | 0.94 |
| Tox21-Nuclear receptor signalling pathways | Peroxisome Proliferator Activated Receptor Gamma (PPAR-Gamma) | nr_ppar_gamma | Inactive | 0.91 |
| Tox21-Stress response pathways | Nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (nrf2/ARE) | sr_are | Inactive | 0.9 |
| Tox21-Stress response pathways | Heat shock factor response element (HSE) | sr_hse | Inactive | 0.9 |
| Tox21-Stress response pathways | Mitochondrial Membrane Potential (MMP) | sr_mmp | Inactive | 0.57 |
| Tox21-Stress response pathways | Phosphoprotein (Tumor Supressor) p53 | sr_p53 | Inactive | 0.86 |
| Tox21-Stress response pathways | ATPase family AAA domain-containing protein 5 (ATAD5) | sr_atad5 | Inactive | 0.59 |

d **Toxicity prediction for compound 34 (ProTox-II - Prediction of TOXicity of chemicals (charite.de))**

**Fig. 2** **a** Flowchart for docking procedure to obtain binding score for docking a ligand. **b** Flowchart for to view and save the protein-ligand complex. **c** Structural similarity and drug like properties of docked GPR 120 agonists versus Rolapitant and Fluvastatin.**d***Toxicity prediction for compound**34* (ProTox-II - Prediction of TOXicity of chemicals (charite.de))

antagonists for chemotherapy [52]. Further, Toelzer et al. recently identified a linoleic acid binding pocket in the SARS-CoV-2 spike protein which prompted us to look for alternate drugs for binding with *COVID-19 MPro.* The present study reports the results of docking studies carried out using G-protein-coupled receptor (GPR) agonists against the MPro to identify any potential inhibitor of SARS CoV-2.

## Methods

Auto Dock 4.2.6 was used to perform docking study. Chemical structures were drawn using Chemoffice 2002. Three-dimensional structures of proteins were downloaded from protein data bank (PDB id: 6LU7) (https://www.rcsb.org/).

### Protein preparation

The protein was prepared for docking process according to the standard protein preparation procedure integrated in Accelry's Discovery Studio 4 which is shown in the flow chart (Fig. 2a).

Ligand 2D structures were drawn using ChemDraw Ultra 8.0 (ChemOffice 2002) and converted into 3D structure using chem3D Ultra 8.0. The 68 molecules were used as ligands, and each one of them was docked on to the crystal structure of $M^{Pro}$ with PDB id 6LU7. The procedure for ligand preparation and docking is given as a flow chart in Fig.2a, b. Docking scores were obtained to understand any inhibitory potential of the 68 GPR120 agonists.

Zhang and Macielag [53] discussed the patented GPR 120 agonists for the treatment of diabetes. They reviewed the therapeutic patents of ten different classes of compounds that amounted to 68 therapeutic molecules. The 68 GPR-120 agonists collected by the authors from different patents and journals are grouped into ten classes. The ten classes and the number assigned in this paper along with abbreviation are given below:

1. Natural GPR 120 agonists and early synthetic GPR 120 agonists (1–9)
2. Bi-aryl-based phenyl propionic acid derivatives as GPR 120 agonists (BiAr-PPA 10–14)
3. Cycloalkenyl and heterocycloalkenyl-based phenyl propionic acid derivatives (CycA_Hcyc--PPA 15–23)
4. Dihydrobenzofuran derivatives (Metabolex) and benzo-fused heterocyclic derivatives (Metabolex 24 and Jansesen 25–27)
5. Chemcial scaffolds claimed by Merck (Merck 28–35)
6. Various carboxylic acid scaffolds claimed by Bristol-Meyers Squibb (BMS 36–45)
7. Patented structures by Piramal Enterprises Limited (PEL 46–52
8. Other carboxylic acid-based GPR 120 agonists (Calden (53–54; LG 55–59; Ajinamoto 60; DOMPE 61–62)
9. Non-acid-containing structures claimed by AXXAM (AXXAM 63-65)
10. GPR120 agonists/antagonists in the peer-reviewed journals (GSK 65; U of B 66; GSK 67; Ch Pharm U 68).

Different classes of these 68 molecules are patented as GPR120 agonists to treat type-2 diabetes by various pharmaceutical companies namely, Janssen Pharmaceutica NV, Merck Sharp & Dohme Corp., Bristol Myers Squibb, Piramal Enterprises Limited, Caldan Therapeutics Limited, LG Life Sciences Ltd., Ajinomoto Company, Accepted Manuscript Inc., Dompe' Farmaceutici S.P.A., and AXXAM S. P. A. Structures of these compounds with their abbreviated id number used in this paper are given in Table 1.

## Results and discussions

Binding scores for each of the 68 ligands are listed in Table 1 along with their molecular structures. Docking images each of the ligands in MPro (PDB ID 6LU7) are given in the Supplementary material while the ligands with best scores ($\geq -8.0$) are given in Fig. 3. The type of docking interaction for these ligands are presented in Table 2. Some of the ligands are having a docking score of $\geq -8.0$. The lowest value is $-8.3$ for the molecule with id GSK (65). This is a sulfonamide patented by GlaxoSmithKline as a selective antagonist against free fatty acid 4 (FFA4/GPR120) [54, 55] and to be used with the agonist GSK 137647A which is also a sulfonamide (id in this paper GSK137647A(8)). The compound with binding score $-8.2$ is a biaryl-based phenylpropanoic acid (13)) [56] patented by Janssen Pharmaceutica [57]. The compound 17 with binding score $-8.0$ is also phenylpropanoic acid derivative namely, cyclohexenyl-based phenyl propionic acid [58]. Three of the other compounds 47, 48, and 52, with binding score $-8.0$ are patented by Piramal Enterprises Limited as GPR120 agonists [47, 59–61] and phenylbutanoic acid with biarylsubstituent wherein one of the aryl groups is a heterocyclic or fused heterocyclic system. Cyclopropane carboxylic acid derivative with a phnoxybiphenyl substituent **40** is also found to have a binding score 8.0. This molecule is patented by Bristol-Meyers Squibb Company [62] as GPR120 modulators useful for treatment of diabetes and related diseases.

Based on the binding score of the top 10 compounds investigated here, their role in blocking the binding site through Glu 166 and Cys 145 could be considered

relevant for their potential role as novel ligands for Sars-COVID-19 virus protein. The observation that Remdesivir, Nelfinavir, and other antiviral compounds show similar interaction support our inference [63–65]. Additional support for such a claim has been found in the paper describing docking study of metocurine with M-Pro 6LUZ that indicate the drug occupies the binding site [66]. The important residues observed in the docking study of our GPR120 agonists as well as the above molecules studied by others including that of chlorquine [61] highlight the role of NH, COOH groups in manifesting pi bond formation with Glu 166 and aromatic pi

**Table 1** GPR120 agonists, ligand ids (used in this paper), docking scores, and molecular structures

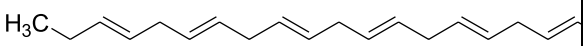
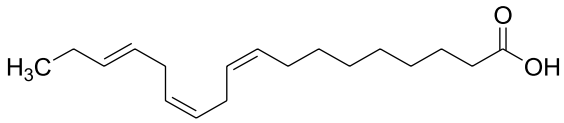
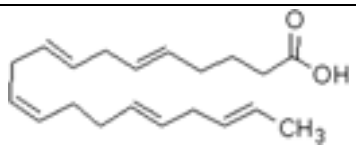
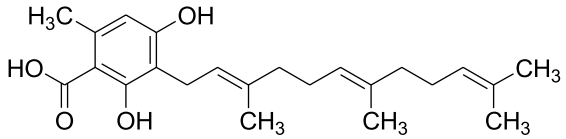
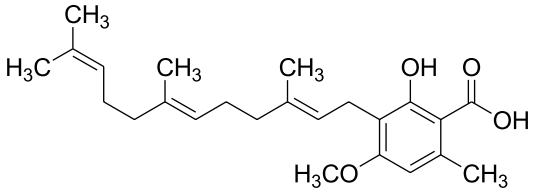
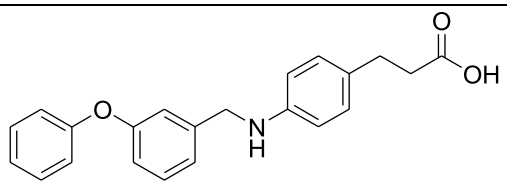
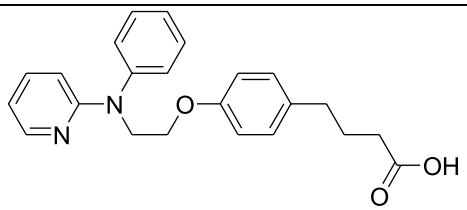
| Ligand id | Docking Score | Structure |
|---|---|---|
| DHA (1)<br>(3E,6E,9E,12E,15E,18E)-henicosa-3,6,9,12,15,18-hexaenoic acid | −5.6 |  |
| ALA (2)<br>Alpha linoleic acid | −4.6 |  |
| EPA (3)<br>Eicoasa pentanoic acid | −5.7 |  |
| Grifolic acid (4) | −6.4 |  |
| Grifolic acid methyl ether (5) | −5.6 |  |
| GW9508 (6)<br>4-(3-Phenoxybenzylamino)phenyl-propionic acid | −7.3 |  |
| NCG21 (7)<br>4-(4-(2-(phenyl(pyridin-2-yl)amino)ethoxy)phenyl)butanoic acid | −6.9 |  |

**Table 1** (continued)

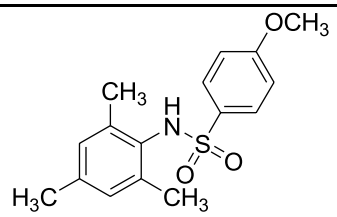
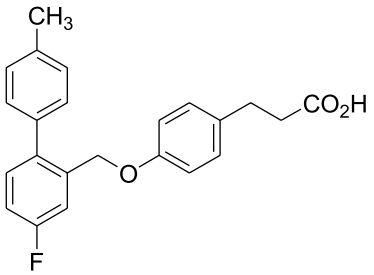
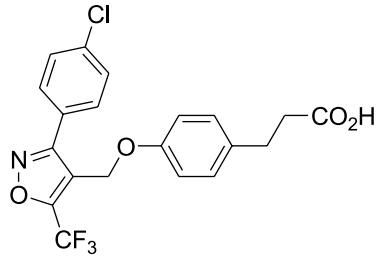
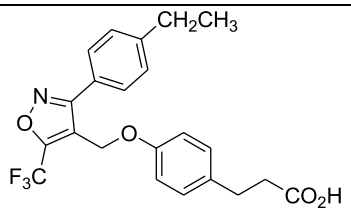
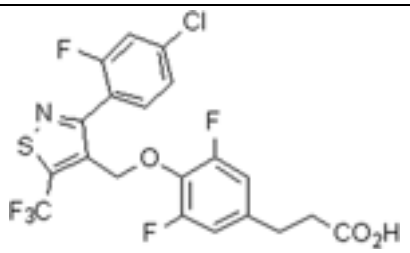
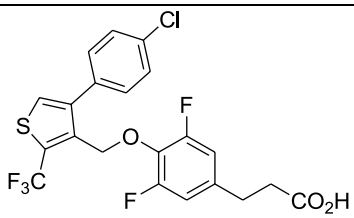
| | | |
|---|---|---|
| GSK137647A<br>(8)<br>N-mesityl-4-methoxybenzenesulfonamide | −6.4 |  |
| TUG891 (9)<br>3-(4-((4-fluoro-4'-methyl-[1,1'-biphenyl]-2-yl)methoxy)phenyl)propanoic acid | −6.8 |  |
| BiAr-PPA (10)<br>3-(4-((3-(4-chlorophenyl)-5-(trifluoromethyl)isoxazol-4-yl)methoxy)phenyl)propanoic acid | −7.9 |  |
| BiAr-PPA (11)<br>3-(4-((3-(4-ethylphenyl)-5-(trifluoromethyl)isoxazol-4-yl)methoxy)phenyl)propanoic acid | −7.4 |  |
| BiAr-PPA (12)<br>3-(4-((3-(5-chloro-2-fluorophenyl)-5-(trifluoromethyl)isothiazol-4-yl)methoxy)-3,5-difluorophenyl)propanoic acid | −7.5 |  |
| BiAr-PPA (13)<br>3-(4-((4-(4-chlorophenyl)-2-(trifluoromethyl)thiophen-3-yl)methoxy)-3,5-difluorophenyl)propanoic acid | −8.2 |  |

**Table 1** (continued)

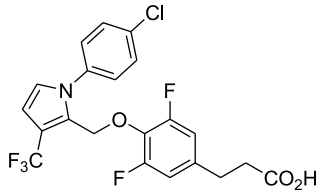
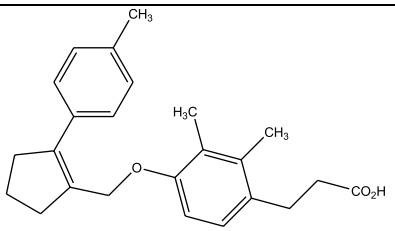
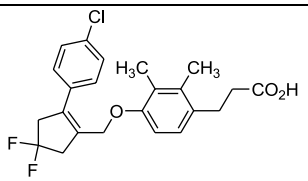
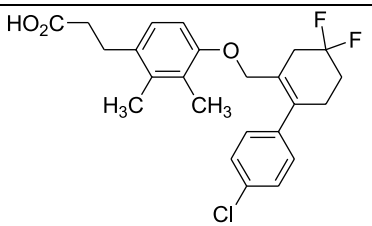
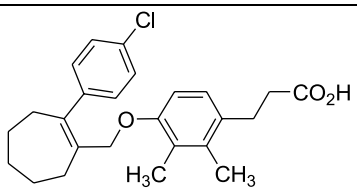
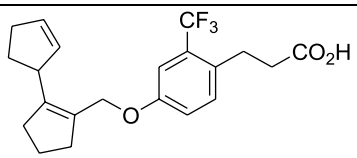
| | | |
|---|---|---|
| BiAr-PPA (14)<br>3-(4-((1-(4-chlorophenyl)-3-(trifluoromethyl)-1*H*-p | −7.1 |  |
| CycA_Hcyc-PPA (15)<br>3-(2,3-dimethyl-4-((2-(p-tolyl)cyclopent-1-en-1-yl)methoxy)phenyl)propanoic acid | −5.6 |  |
| CycA_Hcyc--PPA (16)<br>3-(4-((2-(4-chlorophenyl)-4,4-difluorocyclopent-1-en-1-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −7.9 |  |
| CycA_Hcyc--PPA (17)<br>3-(4-((4'-chloro-4,4-difluoro-3,4,5,6-tetrahydro-[1,1'-biphenyl]-2-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −8.0 |  |
| CycA_Hcyc--PPA (18)<br>3-(4-((2-(4-chlorophenyl)cyclohept-1-en-1-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −7.5 |  |
| CycA_Hcyc--PPA (19)<br>3-(4-([1,1'-bi(cyclopentane)]-1,2'-dien-2-ylmethoxy)-2-(trifluoromethyl)phenyl)propanoic acid | −7.3 |  |

**Table 1** (continued)

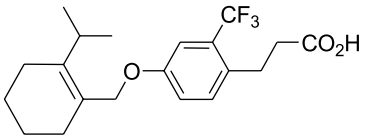
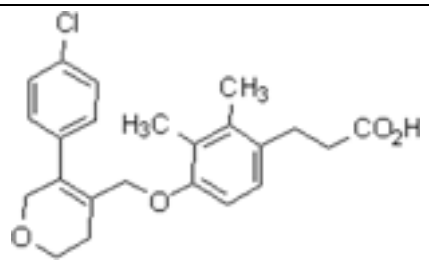
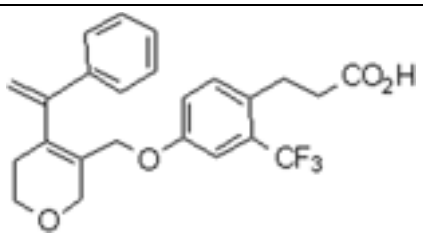
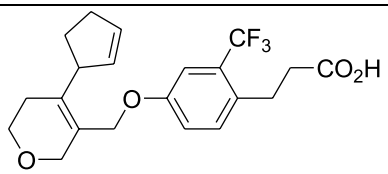
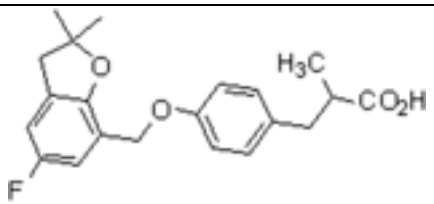
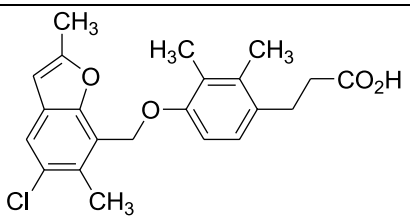
| | | |
|---|---|---|
| CycA_Hcyc--PPA (20 3-(4-((2-isopropylcyclohex-1-en-1-yl)methoxy)-2-(trifluoromethyl)phenyl)propanoic acid | −7.2 |  |
| CycA_Hcyc--PPA (21) 3-(4-((4-(4-chlorophenyl)-5,6-dihydro-2H-pyran-3-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −7.8 |  |
| CycA_Hcyc--PPA (22) 3-(4-((4-(1-phenylvinyl)-5,6-dihydro-2H-pyran-3-yl)methoxy)-2-(trifluoromethyl)phenyl)propanoic acid | −6.8 |  |
| CycA_Hcyc--PPA (23) 3-(4-((4-(cyclopent-2-en-1-yl)-5,6-dihydro-2H-pyran-3-yl)methoxy)-2-(trifluoromethyl)phenyl)propanoic acid | −7.7 |  |
| Metabolex (24) 3-(4-((5-fluoro-2,2-dimethyl-2,3-dihydrobenzofuran-7-yl)methoxy)phenyl)propanoic acid | −7.6 |  |
| Janssen (25) 3-(4-((5-chloro-2,6-dimethylbenzofuran-7-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −7.3 |  |

**Table 1** (continued)

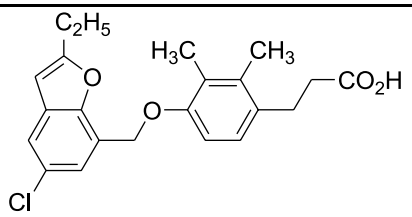
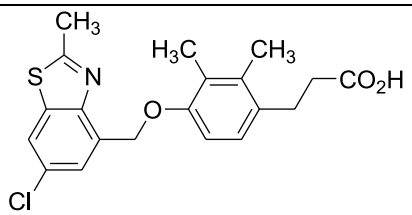
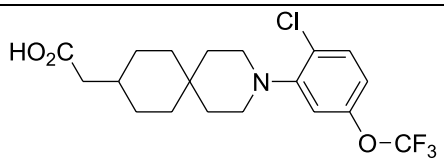
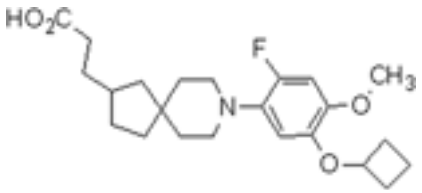
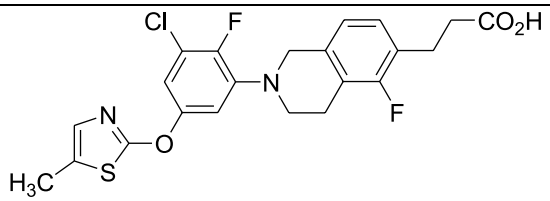
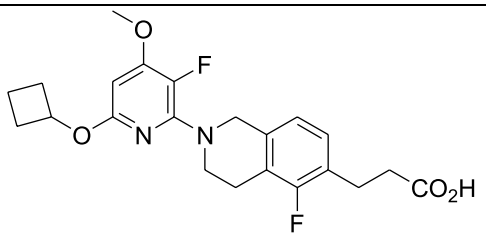
| | | |
|---|---|---|
| Janssen (26)<br>3-(4-((5-chloro-2-ethylbenzofuran-7-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −7.4 |  |
| Janssen (27)<br>3-(4-((6-chloro-2-methylbenzo[d]thiazol-4-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −7.0 |  |
| Merck (28)<br>2-(3-(2-chloro-5-(trifluoromethoxy)phenyl)-3-azaspiro[5.5]undecan-9-yl)acetic acid | −7.7 |  |
| Merck (29)<br>3-(8-(5-cyclobutoxy-2-fluoro-4-methoxyphenyl)-8-azaspiro[4.5]decan-2-yl)propanoic acid | −7.1 |  |
| Merck (30)<br>3-(2-(3-chloro-2-fluoro-5-((5-methylthiazol-2-yl)oxy)phenyl)-5-fluoro-1,2,3,4-tetrahydroisoquinolin-6-yl)propanoic acid | −7.3 |  |
| Merck (31)<br>3-(2-(6-cyclobutoxy-3-fluoro-4-methoxypyridin-2-yl)-5-fluoro-1,2,3,4-tetrahydroisoquinolin-6-yl)propanoic acid | −7.5 |  |

**Table 1** (continued)

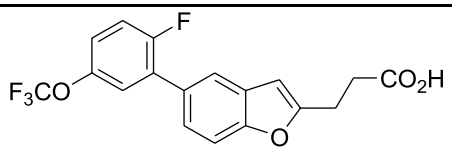
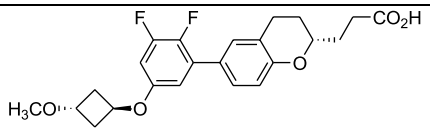
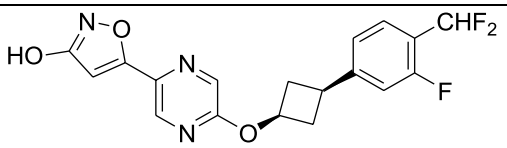
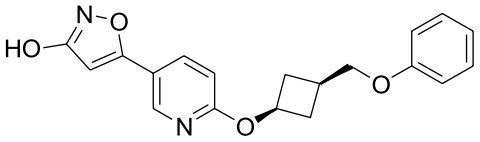
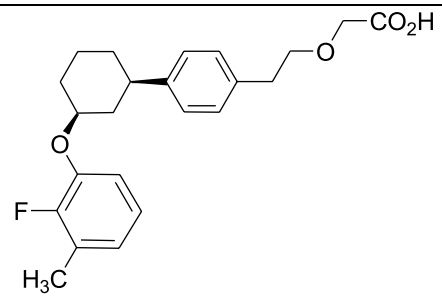
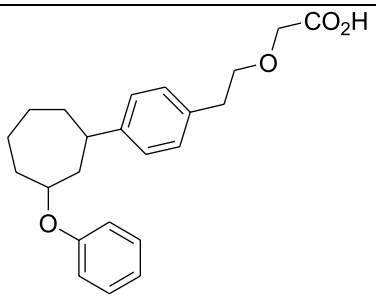
| | | |
|---|---|---|
| Merck (32)<br>3-(5-(2-fluoro-5-(trifluoromethoxy)phenyl)benzofuran-2-yl)propanoic acid | −7.3 |  |
| Merck (33)<br>3-((R)-6-(2,3-difluoro-5-((1r,3R)-3-methoxycyclobutoxy)phenyl)chroman-2-yl)propanoic acid | −7.3 |  |
| Merck (34)<br>5-(5-((1s,3s)-3-(4-(difluoromethyl)-3-fluorophenyl)cyclobutoxy)pyrazin-2-yl)isoxazol-3-ol | −7.9 |  |
| Merck (35)<br>5-(6-((1s,3s)-3-(phenoxymethyl)cyclobutoxy)pyridin-3-yl)isoxazol-3-ol | −7.6 |  |
| BMS (36)<br>2-(4-((1R,3S)-3-(2-fluoro-3-methylphenoxy)cyclohexyl)phenethoxy)acetic acid | −7.3 |  |
| BMS (37)<br>2-(4-(3-phenoxycycloheptyl)phenethoxy)acetic acid | −7.1 |  |

**Table 1** (continued)

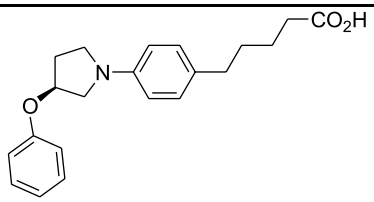
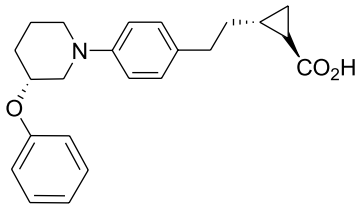
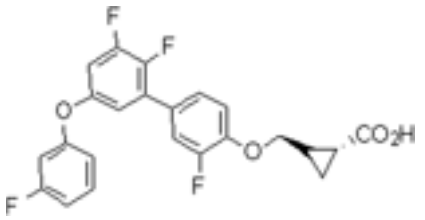
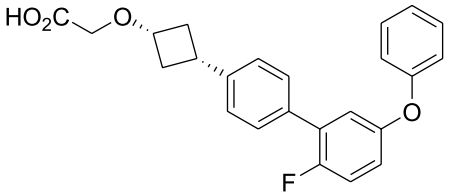
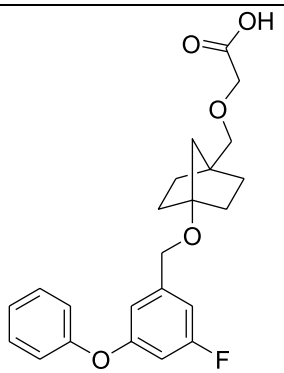
| | | |
|---|---|---|
| BMS (38)<br>(S)-5-(4-(3-phenoxypyrrolidin-1-yl)phenyl)pentanoic acid | −6.7 |  |
| BMS (39)<br>(1S,2R)-2-(4-((R)-3-phenoxypiperidin-1-yl)phenethyl)cyclopropane-1-carboxylic acid | −7.3 |  |
| BMS (40)<br>(1R,2R)-2-(((2',3,3'-trifluoro-5'-(3-fluorophenoxy)-[1,1'-biphenyl]-4-yl)oxy)methyl)cyclopropane-1-carboxylic acid | −8.0 |  |
| BMS (41)<br>2-((1s,3s)-3-(2'-fluoro-5'-phenoxy-[1,1'-biphenyl]-4-yl)cyclobutoxy)acetic acid | −7.5 |  |
| BMS (42)<br>2-(((1r,4r)-4-((3-fluoro-5-phenoxybenzyl)oxy)bicyclo[2.2.1]heptan-1-yl)methoxy)acetic acid | −7.3 |  |

**Table 1** (continued)

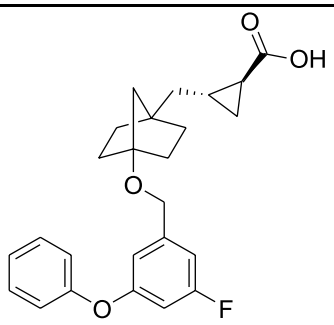
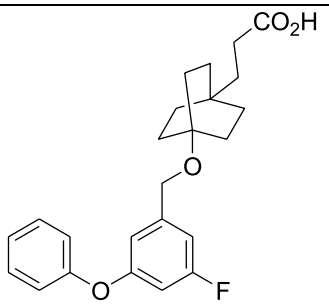
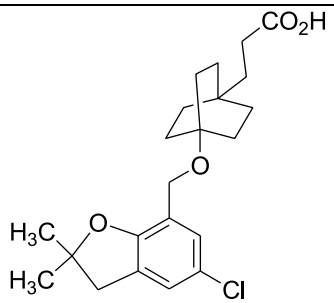
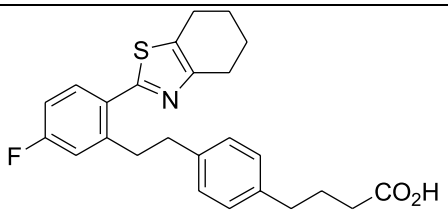
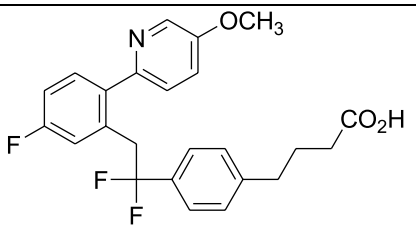
| | | |
|---|---|---|
| BMS (43)<br>(1S,2R)-2-(((1s,4R)-4-((3-fluoro-5-phenoxybenzyl)oxy)bicyclo[2.2.1]heptan-1-yl)methyl)cyclopropane-1-carboxylic acid | −7.8 |  |
| BMS (44)<br>3-(4-((3-fluoro-5-phenoxybenzyl)oxy)bicyclo[2.2.2]octan-1-yl)propanoic acid | −7.9 |  |
| BMS (45)<br>3-(4-((5-chloro-2,2-dimethyl-2,3-dihydrobenzofuran-7-yl)methoxy)bicyclo[2.2.2]octan-1-yl)propanoic acid | −7.3 |  |
| PEL (46)<br>4-(4-(5-fluoro-2-(4,5,6,7-tetrahydrobenzo[d]thiazol-2-yl)phenethyl)phenyl)butanoic acid | −7.6 |  |
| PEL (47)<br>4-(4-(1,1-difluoro-2-(5-fluoro-2-(5-methoxypyridin-2-yl)phenyl)ethyl)phenyl)butanoic acid | −8.0 |  |

**Table 1** (continued)

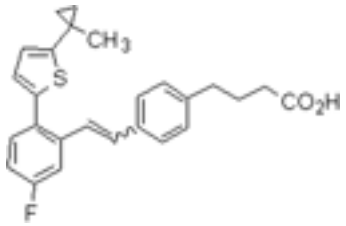
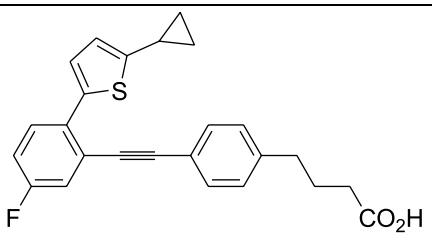
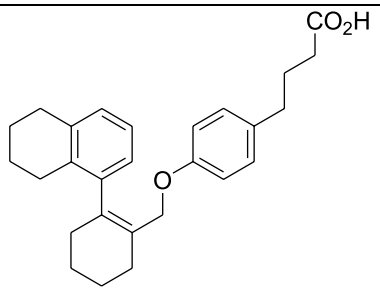
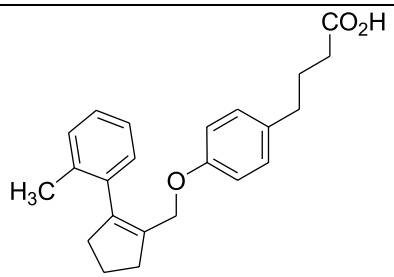
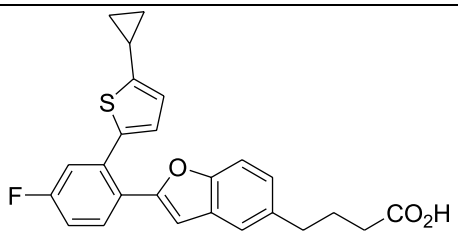
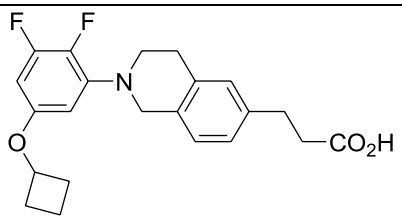
| | | |
|---|---|---|
| PEL (48)<br>4-(4-(5-fluoro-2-(5-(1-methylcyclopropyl)thiophen-2-yl)styryl)phenyl)butanoic acid | −8.0 |  |
| PEL (49)<br>4-(4-((2-(5-cyclopropylthiophen-2-yl)-5-fluorophenyl)ethynyl)phenyl)butanoic acid | −7.4 |  |
| PEL (50)<br>4-(4-((2-(5,6,7,8-tetrahydronaphthalen-1-yl)cyclohex-1-en-1-yl)methoxy)phenyl)butanoic acid | −7.5 |  |
| PEL (51)<br>4-(4-((2-(o-tolyl)cyclopent-1-en-1-yl)methoxy)phenyl)butanoic acid | −6.0 |  |
| PEL (52)<br>4-(2-(2-(5-cyclopropylthiophen-2-yl)-4-fluorophenyl)benzofuran-5-yl)butanoic acid | −8.0 |  |
| Calden (53)<br>3-(2-(5-cyclobutoxy-2,3-difluorophenyl)-1,2,3,4-tetrahydroisoquinolin-6-yl)propanoic acid | −7.6 |  |

**Table 1** (continued)

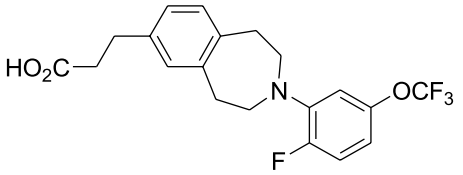
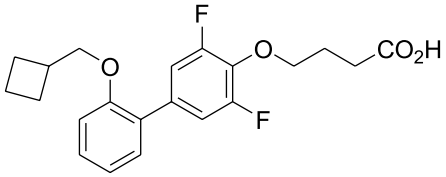
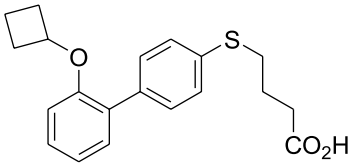
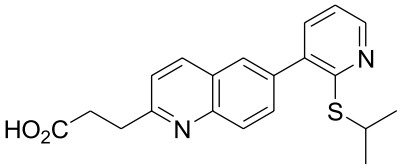
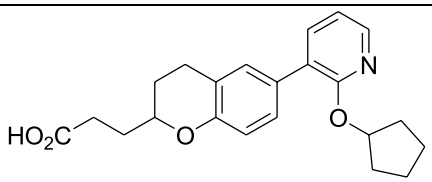
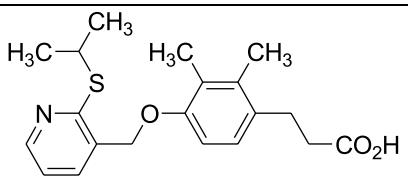
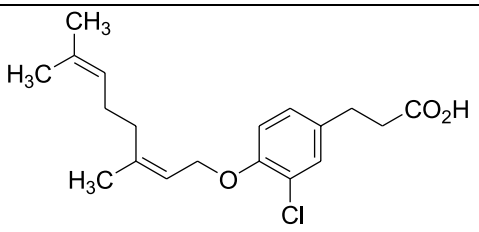
| | | |
|---|---|---|
| Calden(54)<br>3-(3-(2-fluoro-5-(trifluoromethoxy)phenyl)-2,3,4,5-tetrahydro-1H-benzo[d]azepin-7-yl)propanoic acid | −7.0 |  |
| LG (55)<br>4-((2'-(cyclobutylmethoxy)-3,5-difluoro-[1,1'-biphenyl]-4-yl)oxy)butanoic acid | −7.1 |  |
| LG (56)<br>4-((2'-cyclobutoxy-[1,1'-biphenyl]-4-yl)thio)butanoic acid | −6.6 |  |
| LG (57)<br>3-(6-(2-(isopropylthio)pyridin-3-yl)quinolin-2-yl)propanoic acid | −6.8 |  |
| LG (58)<br>3-(6-(2-(cyclopentyloxy)pyridin-3-yl)chroman-2-yl)propanoic acid | −7.5 |  |
| LG (59)<br>3-(4-((2-(isopropylthio)pyridin-3-yl)methoxy)-2,3-dimethylphenyl)propanoic acid | −6.8 |  |
| Ajinamoto (60)<br>(Z)-3-(3-chloro-4-((3,7-dimethylocta-2,6-dien-1-yl)oxy)phenyl)propanoic acid | −6.8 |  |

**Table 1** (continued)

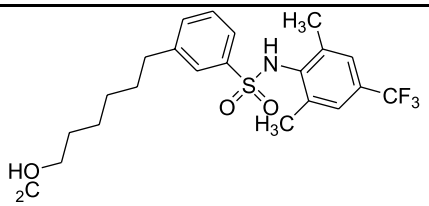
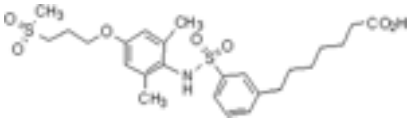
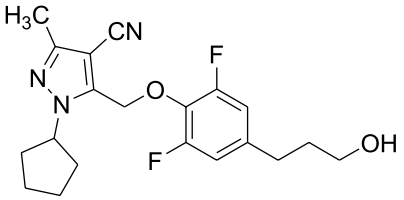
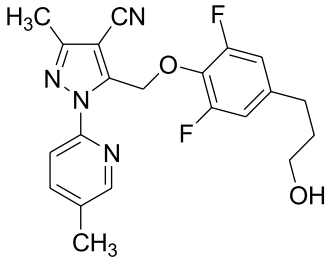
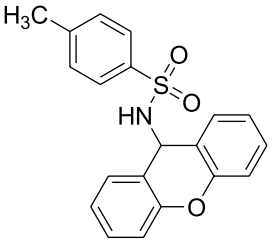
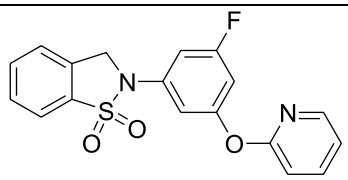
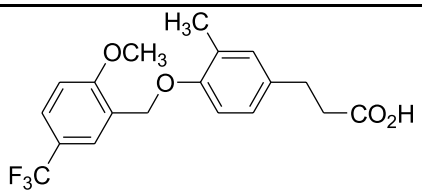
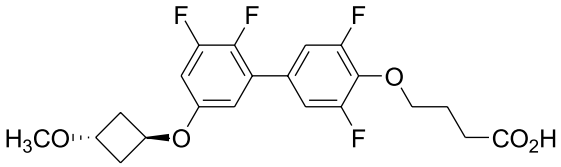
| | | |
|---|---|---|
| DOMPE (61)<br>7-(3-(N-(2,6-dimethyl-4-(trifluoromethyl)phenyl)sulfamoyl)phenyl) heptanoic acid | −6.8 |  |
| DOMPE (62)<br>7-(3-(N-(2,6-dimethyl-4-(2-(methylsulfonyl)ethoxy)phenyl)sulfamoyl) phenyl)heptanoic acid | −6.4 |  |
| AXXAM (63)<br>1-cyclopentyl-5-((2,6-difluoro-4-(3-hydroxypropyl)phenoxy)methyl)-3-methyl-1H-pyrazole-4-carbonitrile | −7.2 |  |
| AXXAM (64)<br>5-((2,6-difluoro-4-(3-hydroxypropyl)phenoxy)methyl)-3-methyl-1-(5-methylpyridin-2-yl)-1H-pyrazole-4-carbonitrile | −7.4 |  |
| GSK (65)<br>4-methyl-N-(9H-xanthen-9-yl)benzenesulfonamide | −8.3 |  |

**Table 1** (continued)

| | | |
|---|---|---|
| Uof B (66)<br>2-(3-fluoro-5-(pyridin-2-yloxy)phenyl)-2,3-dihydrobenzo[d]isothiazole 1,1-dioxide | −7.7 |  |
| GSK (67)<br>3-(4-((2-methoxy-5-(trifluoromethyl)benzyl)oxy)-3-methylphenyl)propanoic acid | −7.1 |  |
| ChPharmU (68)<br>4-((2',3,3',5-tetrafluoro-5'-((1r,3r)-3-methoxycyclobutoxy)-[1,1'-biphenyl]-4-yl)oxy)butanoic acid | −7.3 |  |

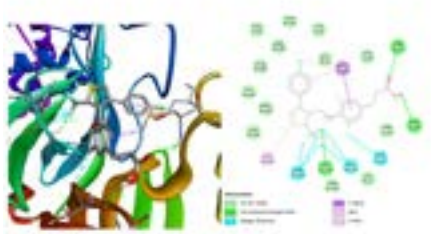
interaction with Cys 145, respectively. Dock score for the reference compounds (Fig. 3i–l) evaluated along with the GPR120 agonists ALPHA KETOMIDE (−7.4), LOPINAVIR (−7.9), SHIKONIN (−7.0), and TIDEGLUSIB (−8.0) indicate the interactions with Glu 166 and Cys 145 are present in these drugs also.

Compounds 13, 16, 17, 44, and 50 identified with high dock score are hydrophobic compounds having thiophenyl, cyclopentenyl, cyclohexenyl, norbornyl, and cyclopentenyl groups along with a phenylpropanoic acid function. Compound 40 is a phenoxyphenyl ether having a cyclopropropane carboxylic acid group while 34 has an isoxazolyl and pryrimidine compound with difluoromethane function. Compound 65 is a tricyclic compound having arylsulfonamide function. Compounds 13 and 52 are highly lipophilic having a log *P* value of 7.1 and 6.25, respectively, that might require vigorous optimization to make them orally available. On the other hand, linoleic acid has a log P value of 5.65 while that of LOPINAVIR is 4.56 and the value is 4.86 for TIDEGLUSIB. Hosseini et al. screened several classes of drugs and identified inhibitors for SARS-CoV-2 MPro and highlighted H bond interactions with Thr 26, Phe 140, Gly 143, Glu 166, and Gln 189 in addition to pi stacking interaction through His 140 as key contributors for receptor binding [67]. Our GPR 120 agonists, 40, 47, and 65 revealed H bonding interaction with Glu 189 in the docking against MPro while hydrophobic interactions with His 41, Met 165, and Glu 166 were shown by compounds 40, 48, 52, and 65. Fluvastatin on the other hand was found to interact with Thr 26 and Gly 143 by Maryam et al., and a similar interaction was observed in our compound 40, which also interacted with Glu 166 and Cys 145. Compounds 65 and 34 have a log P value of 4.45 and 3.26, respectively, indicating that they might have a good oral bioavailability, although they require further optimization. The comparison of topological polar surface area (TPSA) of the compounds evaluated show that only compound 34 has a TPSA value above 75, whereas the reference compounds have TPSA of 120 for lopinavir, a-ketoamide has a value of 113. This suggests that compounds 34 and 40 could be used as start points, and further optimizations could result in finding a drug

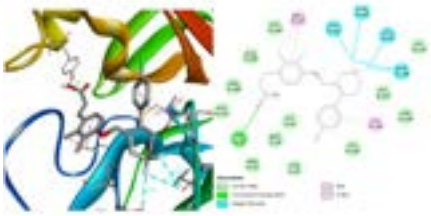
(See figure on next page.)

**Fig. 3** **a** GSK (65); Docking score −8.3. **b** BiAr-PPA (13); Docking score −8.2. **c** CycA_Hcyc--PPA (17); Docking score −8.0. **d** Merck (34); Docking score −7.9. **e** BMS (40); Docking score −8.0. **f.** PML (47); Docking score −8.0. **g** PML (48); Docking score −8.0. **h** PML (52); Docking score −8.0. Docking of the ligands onto MPro (6LU7). Docking images of ligands with docking score ≥ −7.9 are given (for other ligands please see Supplementary Information); **i** ALPHA KETOMIDE = -7.4; **j** LOPINAVIR = -7.9; **k** SHIKONIN = -7.0; **l** TIDEGLUSIB = -8.0

**a.** GSK (65); Docking score -8.3



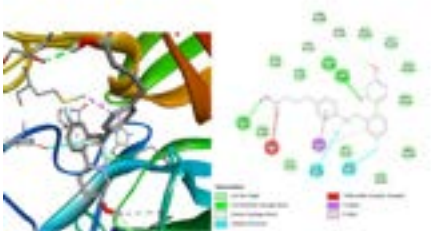
**b.** BiAr-PPA (13); Docking score -8.2



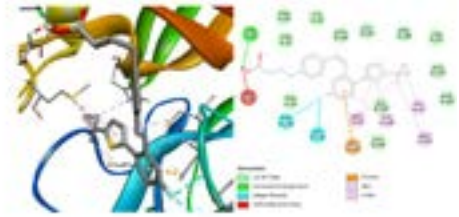
**c.** CycA_Hcye--PPA (17); Docking score -8.0


**d.** Merck (34); Docking score -7.9

**e.** BMS (40); Docking score -8.0

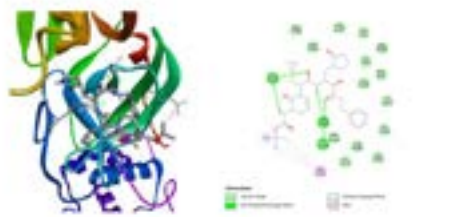


**f.** PML (47); Docking score -8.0



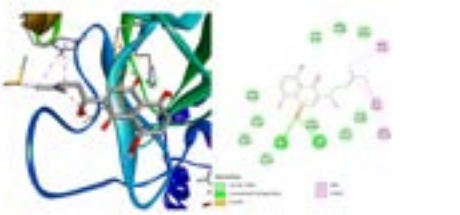
**g.** PML (48); Docking score -8.0

**h.** PML (52); Docking score -8.0

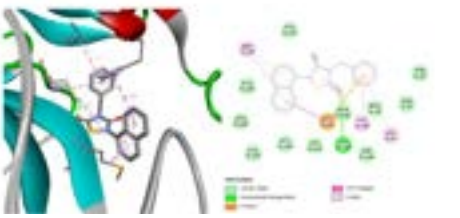


**i** ALPHA KETOMIDE = -7.4

**j** LOPINAVIR = -7.9



**k** SHIKONIN = -7.0



**l** TIDEGLUSIB = -8.0

**Fig. 3** (See legend on previous page.)

**Table 2** Docking interactions for the ligands with docking score ≥ −8.0

| Ligand id | Docking score | Docking details | | |
|---|---|---|---|---|
| | | Conventional H-bond | Alkyl and pi-alkyl | others |
| GSK (65) | −8.3 | GLN:189, GLU:166 | MET:49 | CYS:145 (Pi-donor hydrogen bond) |
| BiAr-PPA (13) | −8.2 | THR:24, SER:46, GLY:143 | MET:49, CYS:145 | MET:49 (Pi-sigma) THR:26, LEU:141, ASN:142 (Halogen) |
| CycA_Hcyc--PPA (17) | −8.0 | TYR:54 | CYS:145, MET:165 | PHE:140, LEU:141, ASN:142, GLU:166 (Halogen) |
| Merck (34) | −7.9 | Asp 187, Tyr 54, His 164, His 41 | Cys 145, Leu 27 | |
| BMS (40) | −8.0 | THR:25,26, GLY:143, SER:144, CYS:145, HIS:163 GLN:189, GLU:166 | CYS:145, MET:165 | GLN:189, CYS:145 (carbon hydrogen bond); GLY:143 (Pi-donor hydrogen bond); LEU:141, ARG:188 (Halogen) |
| PEL (47) | −8.0 | SER:46, CYS:145, HIS:164 | MET:49 | THR:24 (unfavorable acceptor-acceptor); MET:49 (Pi-sigma) LEU:141, PHE:140 (carbon hydrogen bond); GLN:189, ASP:187 (Halogen) |
| PEL (48) | −8.0 | THR:45 | HIS:41, MET:49, CYS:145, MET:165 | HIS:163 (Pi-cation); SER:46 (unfavorable donor-donor); PHE:140, LEU:141 (Halogen) |
| PEL (52) | −8.0 | HIS:41, CYS:145 | MET:49, MET:165 | MET:49 (Pi-sigma); CYS:145 (Pi-sulfur); ARG:188 (Halogen) |

for Sars-COVID-19. Compound 34 possesses a hydroxyl isoxazole group that would mimic COOH function and also the presence of polar heterocyclic ring providing it an ideal choice to improve its physicochemical characteristics. The compound **34** is in fact well anchored through H-bonded interactions, Pi interactions as shown in Fig. 3d. Similar to compound 40, having a cyclopropane carboxylic acid function could be optimized further to refine its log P to make this eligible as a lead compound. Obviously the GPR 120 agonists, designed as agonists for free fatty acid receptors, have functional groups and lipophilic characters designed for their receptor need to be tweaked to suit the binding interactions with MPro. These compounds identified through the present study are functionally similar to linoleic acid, a free fatty acid that has been found to occupy the binding pocket of spike protein in SARS-CoV-2 [68]. Comparing the free fatty acid, linoleic acid, the GPR 120 agonists identified herein possess several beneficial physicochemical properties in terms of favourable log *P* values and topological polar surface area, making them suitable for oral administration (Fig. 2c).

Toxicity prediction for compound 34, using online tool "protox_II," indicates that the molecule is safe for all the targets except showing carcinogenicity and hepatotoxicity of 0.51 and 0.64, respectively, further requiring structural modification. This compound also has an LD 50 value of 300 mg/kg and falls under predicted toxicity class 3, indicating it is only slightly toxic and slightly irritating.

## Conclusions

The present data supports the possibility of repurposing free fatty acid GPR 120 receptor agonists as potential inhibitors of Sars-COVID-19 M-Pro protein. Based on docking score and key interactions with the amino acid residues in the target protein, compound 34 could be used as a lead compound. The presence of COOH mimicking hydroxyl isoxazole group could provide necessary drug-like property in addition to maintaining H-bond and pi interactions with the receptor. Favorable logP and available physicochemical and toxicity data of compound 34 could shorten the drug development time to position the compound as an early lead candidate to overcome the hurdles in identifying therapeutic drugs in coronavirus infection.

#### Abbreviations

COVID-19 MPro: Severe acute respiratory syndrome coronavirus 2 main protease; GPR-120 agonists: G-protein-coupled receptor or free fatty acid receptor-4 agonists; Docking: Phenyl propanoic acid.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43141-022-00375-8.

**Additional file 1.** Docking images of the inhibitors (ligands) on $M^{Pro}$.

#### Acknowledgements

The authors wish to thank Karpagam Mnagement Trust for their support for the facilities.

#### Authors' contributions

Dr. Mohan, Sellappan conceived the idea and gave directions. Dr. Nagarajan Arumugam gave information about GPR 120 agonists and their bioactivities and relation to COVID-19 MPro. Dr. Natarajan Ramanathan collected GPR agonists and guided the workflow for docking study. Dr. Tharani Jayagopal did the docking study and collected the images of docking and scoring details. Dr. Nagarajan Arumugam and Dr. Natarajan R wrote the manuscript. Dr. Mohan S revised the manuscript. The author(s) read and approved the final manuscript.

#### Authors' information

Not applicable.

**Funding**
There is no funding received from any source.

**Availability of data and materials**
Yes. Attached as Supporting information

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
All authors consent for publication in the Journal of Genetic Engineering and Biotechnology.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Karpagam College of Pharmacy, Coimbatore, Tamil Nadu 641032, India. [2]Sri Sarada Niketan College for Women, Karur, Tamilnadu, India. [3]Saranathan College of Engineering, Panjappur, Tiruchirappalli, India.

Received: 28 February 2022 Accepted: 2 June 2022
Published online: 18 July 2022

**References**

1. Prajapat M, Sarma P, Shekhar N, Avti P, Sinha S, Kaur H, Kumar S, Bhattacharyya A, Kumar H, Bansal S, Medhi B (2020) Drug targets for corona virus: a systematic review. Indian J Pharm 52:56–65. https://doi.org/10.4103/ijp.IJP_115_20
2. Yadav R, Chaudhary JK, Jain N, Chaudhary PK, Khanra S, Dhamija P, Sharma A, Kumar A, Handu S (2021) Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19. Cells 10:821. https://doi.org/10.3390/cells10040821
3. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181:281–292. https://doi.org/10.1016/j.cell.2020.02.058
4. Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, Qi F, Bao L, Du L, Liu S, Qin C, Sun F, Shi Z, Zhu Y, Jiang S, Lu L (2020a) Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res 30:343–355. https://doi.org/10.1038/s41422-020-0305-x
5. Xia S, Zhu Y, Liu M, Lan Q, Xu W, Wu Y, Lu L (2020b) Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol 17:765–767. https://doi.org/10.1038/s41423-020-s0374-2
6. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nat Med 26:450–452. https://doi.org/10.1038/s41591-020-0820-9
7. Hoffmann M, KleineWeber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, Müller MA, Drosten C, Pöhlmann S (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cells 181:210–271. https://doi.org/10.1016/j.cell.2020.02.052
8. Bestle D, Heindl MR, Limburg H, Van TVL, Pilgram O, Moulton H, Böttcher-Friebertshäuser E (2020) TMPRSS2 and furin are both essential for proteolytic activation and spread of SARS-CoV-2 in human airway epithelial cells and provide promising drug targets. BioRxiv. https://doi.org/10.1101/2020.04.15.042085
9. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E (2020) The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176:104742. https://doi.org/10.1016/j.antiviral.2020.104742
10. Amin SA, Banerjee S, Ghosh K, Gayen S, Jha T (2021) Protease targeted COVID-19 drug discovery and its challenges: insight into viral main protease (Mpro) and papain-like protease (PLpro) inhibitors. Bioorg Med Chem 29:115860. https://doi.org/10.1016/j.bmc.2020.115860
11. McClain CB, Vabret N (2020) SARS-CoV-2: the many pros of targeting PLpro. Signal Transduct Target Ther 5:223. https://doi.org/10.1038/s41392-020-00335-z
12. Shen Z, Ratia K, Cooper L, Lee H, Kwon Y, Li Y, Algarni S, Huang F, Dubrovskyi O, Thatcher GRJ, Xiong R (2021) Design of SARS-CoV-PLpro inhibitors for COVID-19 antiviral therapy leveraging binding cooperativity. J Med Chem. https://doi.org/10.1021/acs.jmedchem.1c01307
13. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, Zhang B, Li X, Zhang L, Peng C, Duan Y, Yu J, Wang L, Yang K, Liu F, Jiang R, Yang X, You T, Liu X, Yang X, Bai F, Liu H, Liu X, Guddat LW, Xu W, Xiao G, Qin C, Shi Z, Jiang H, Rao Z, Yang H (2020) Structure of $M^{pro}$ from SARS-CoV-2 and discovery of its inhibitors. Nature 582:89–293. https://doi.org/10.1038/s41586-020-2223-y
14. Zhu W, Chen CZ, Gorshkov K, Xu M, Lo DC, Zheng W (2020) RNA-dependent RNA polymerase as a target for COVID-19 drug discovery. SLAS Discov: Adv Sci Drug Discov 25:1141–1151. https://doi.org/10.1177/2472555220942123
15. Namchuk MN (2021) Early returns on small molecule therapeutics for SARS-CoV-2. ACS Infect Dis 7:1298–1302. https://doi.org/10.1021/acsinfecdis.0c00874
16. Keretsu S, Bhujbal SP, Cho SJ (2020) Rational approach toward COVID-19 main protease inhibitors via molecular docking, molecular dynamics simulation and free energy calculation. Sci Rep 10:17716. https://doi.org/10.1038/s41598-020-74468-0
17. Yang H, Xie W, Xue X, Yang K, Ma J, Liang W, Zhao Q, Zhou Z, Pei D, Ziebuhr J, Hilgenfeld R, Yuen KY, Wong L, Gao G, Chen S, Chen Z, Ma D, Bartlam M, Rao Z (2005) Design of wide-spectrum inhibitors targeting coronavirus main proteases. PLoS Biol 3:e324. https://doi.org/10.1371/journal.pbio.0030324
18. Amporndanai K, Meng X, Shang W, Jin Z, Rogers M, Zhao Y, Rao Z, Liu ZJ, Yang H, Zhang L, Neill PM, Samar Hasnain S (2012) Inhibition mechanism of SARS-CoV-2 main protease by ebselen and its derivatives. Nat Commun 12:306. https://doi.org/10.1038/s41467-021-23313-7
19. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, Becker S, Rox K, Hilgenfeld R (2020a) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science (New York, NY) 368:409–412. https://doi.org/10.1126/science.abb3405
20. Zhang ZJ, Wu WY, Hou JJ, Zhang LL, Li FF, Gao L, Wu XD, Shi JY, Zhang R, Long HL, Lei M, Wu WY, Guo DA, Chen KX, Hofmann LA, Ci ZH (2020b) Active constituents and mechanisms of Respiratory Detox Shot, a traditional Chinese medicine prescription, for COVID-19 control and prevention: Network-molecular docking-LC-MS analysis. J integrat Med 18:229–241. https://doi.org/10.1016/j.joim.2020.03.004
21. Calligari P, Bobone S, Ricci G, Bocedi A (2020) Molecular investigation of SARS-CoV-2 proteins and their interactions with antiviral drugs. Viruses 12:445. https://doi.org/10.3390/v12040445
22. Nutho B, Mahalapbutr P, Hengphasatporn K, Pattaranggoon NC, Simanon N, Shigeta Y, Hannongbua S, Rungrotmongkol T (2020) Why are lopinavir and ritonavir effective against the newly emerged coronavirus 2019? Atomistic Insights into the Inhibitory Mechanisms. Biochemistry. 59:1769–1779. https://doi.org/10.1021/acs.biochem.0c00160
23. Chang Y, Tung Y, Lee K, Chen T, Hsiao Y, Chang H, Hsieh T, Su C, Wang S, Yu J, Shih S, Lin Y, Lin Y, Tu YE, Hsu C, Juan H, Tung C, Chen C (2020) Potential therapeutic agents for COVID-19 based on the analysis of protease and RNA polymerase docking. Preprints. https://doi.org/10.20944/preprints202002.0242.v2
24. Xu Z, Peng C, Shi Y, Zhu Z, Mu K, Wang X, Zhu W (2020a) Nelfinavir was predicted to be a potential inhibitor of 2019-nCov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. bioRxiv. https://doi.org/10.1101/2020.01.27.921627
25. Xu Z, Shen J, Wu N, Xu Y, Lu X, Zhu W, Li LJ (2020b) Nelfinavir is active against SARS-CoV-2 in Vero E6 cells. ChemRxiv. https://doi.org/10.26434/chemrxiv.12039888.v1
26. Beck BR, Shin B, Choi Y, Park S, Kang K (2020) Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model. Comput Struct Biotechnol J 18:784–790. https://doi.org/10.1016/j.csbj.2020.03.025
27. Fintelman-Rodrigues N, Sacramento CQ, Ribeiro Lima C, Souza da Silva F, Ferreira AC, Mattos M, de Freitas CS, Cardoso Soares V, da Silva Gomes

Dias S, Temerozo JR, Miranda MD, Matos AR, Bozza FA, Carels N, Alves CR, Siqueira MM, Bozza PT, Souza TML (2020) Atazanavir, alone or in combination with ritonavir, inhibits SARS-CoV-2 replication and proinflammatory cytokine production. Antimicrob Agents Chemother 64:e00825–e00820. https://doi.org/10.1128/AAC.00825-20
28. Chen YW, Yiu CB, Wong KY (2020) Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease ($3CL^{pro}$) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Research 9:129. https://doi.org/10.12688/f1000research.22457.2
29. Li Y, Zhang J, Wang N, Li H, Shi Y, Guo G, Liu K, Zeng H, Zou Q (2020) Therapeutic drugs targeting 2019-nCoV main protease by high-throughput screening. bioRxiv. https://doi.org/10.1101/2020.01.28.922922
30. Khan RJ, Jha RK, Amera GM, Jain M, Singh E, Pathak A, Singh RP, Muthukumaran J, Singh AK (2021) Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2′-O-ribose methyltransferase. J Biomol Struct Dyn 39:2679–2692. https://doi.org/10.1080/07391102.2020.1753577
31. Talluri S (2021) Molecular docking and virtual screening based prediction of drugs for COVID-19. Comb Chem High Throughput Screen 24:716–728. https://doi.org/10.2174/1386207323666200814132149
32. Muralidharan N, Sakthivel R, Velmurugan D, Gromiha MM (2021) Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19. J Biomol Struct Dyn 39:2673–2678. https://doi.org/10.1080/07391102.2020.1752802
33. Hall DC, Ji HF (2020) A search for medications to treat COVID-19 via in silico molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease. Travel Med Infect Dis 35:101646. https://doi.org/10.1016/j.tmaid.2020.101646
34. Wang C, Horby PW, Hayden FG, Gao GF (2020) A novel coronavirus outbreak of global health concern. Lancet 395:470–473. https://doi.org/10.1016/S0140-6736(20)30185-9
35. Liu X, Wang XJ (2020) Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. J Genet Genomics 47:119–121. https://doi.org/10.1016/j.jgg.2020.02.001
36. Sisakht M, Solhjoo A, Mahmoodzadeh A, Fathalipour M, Kabiri M, Sakhteman A (2021) Potential inhibitors of the main protease of SARS-CoV-2 and modulators of arachidonic acid pathway: non-steroidal anti-inflammatory drugs against COVID19. Comput Biol Med 136:104686
37. Gimeno A, Mestres-Truyol J, Ojeda-Montes MJ, Macip G, Saldivar-Espinoza B, Cereto-Massagué A, Pujadas G, Garcia-Vallvé S (2020) Prediction of novel inhibitors of the main protease (M-pro) of SARS-CoV-2 through consensus docking and drug reposition. Int J Mol Sci 21:3793. https://doi.org/10.3390/ijms21113793
38. Liu X, Li Z, Liu S, Sun J, Chen Z, Jiang M, Zhang Q, Wei Y, Wang X, Huang YY, Shi Y, Xu Y, Xian H, Bai F, Ou C, Xiong B, Lew AM, Cui J, Fang R, Huang H, Zhao J, Hong X, Zhang Y, Zhou F, Luo HB (2020) Potential therapeutic effects of dipyridamole in the severely ill patients with COVID-19. Acta Pharmaceut Sinica B 10:1205–1215. https://doi.org/10.1016/j.apsb.2020.04.008
39. Odhar HA, Ahjel SW, Albeer A, Hashim AF, Rayshan AM, Humadi SS (2020) Molecular docking and dynamics simulation of FDA approved drugs with the main protease from 2019 novel coronavirus. Bioinformation 16:236–244. https://doi.org/10.6026/97320630016236
40. Su H, Yao S, Zhao W, Li M, Liu J, Shang W, Xie H, Ke C, Hu H, Gao M, Yu K, Liu H, Shen J, Tang W, Zhang L, Xiao G, Ni L, Wang D, Zuo J, Jiang H, Bai F, Wu Y, Ye Y, Xu Y (2020) Anti-SARS-CoV-2 activities in vitro of Shuanghuanglian preparations and bioactive ingredients. Acta Pharmacol Sin 41:1167–1177. https://doi.org/10.1038/s41401-020-0483-6
41. Liu H, Ye F, Sun Q, Liang H, Li C, Lu R, Huang B, Tan W, Lai L (2021) *Scutellaria baicalensis* extract and baicalein inhibit replication of SARS-CoV-2 and its 3C-like protease in vitro. J Enzyme Inhib Med Chem 36:497–503. https://doi.org/10.1080/14756366.2021.187397
42. Cherrak SA, Merzouk H, Mokhtari-Soulimane N (2020) Potential bioactive glycosylated flavonoids as SARS-CoV-2 main protease inhibitors: a molecular docking and simulation studies. PLoS One 15. https://doi.org/10.1371/journal.pone.0240653
43. Shivanika C, Deepak Kumar S, Ragunathan V, Tiwari P, Sumitha A, Brindha Devi P (2020) Molecular docking, validation, dynamics simulations, and pharmacokinetic prediction of natural compounds against the SARS-CoV-2 main-protease. J Biomol Struct Dyn 8:1–27. https://doi.org/10.1080/07391102.2020.1815584
44. Bhaliya J, Shah V (2020) Identification of potent COVID-19 main protease (MPro) inhibitor from curcumin analogues by molecular docking analysis. Int J Adv Res, Ideas Innov Technol 6:664–672
45. Joshi RS, Jagdale SS, Bansode SB, Shankar SS, Tellis MB, Pandya VK, Chugh A, Giri AP, Kulkarni MJ (2021) Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. J Biomol Struct Dyn 39:3099–3114. https://doi.org/10.1080/07391102.2020.1760137
46. Enmozhi SK, Raja K, Sebastine I, Joseph J (2021) Andrographolide as a potential inhibitor of SARS-CoV-2 main protease: an in silico approach. J Biomol Struct Dyn 39:3092–3098. https://doi.org/10.1080/07391102.2020.1760136
47. Ramaiah MHS, Vishwakarma R, Umashaanker R (2020) Molecular docking analysis of selected natural products from plants for inhibition of SARS-CoV-2 main protease. Curr Sci 118:1087–1092
48. Umesh KD, Selvaraj C, Singh SK, Dubey VK (2021) Identification of new anti-CoV drug chemical compounds from Indian spices exploiting SARS-CoV-2 main protease as target. J Biomol Struct Dyn 39:3428–3434. https://doi.org/10.1080/07391102.2020.1763202
49. Khaerunnisa S, Kurniawan H, Awaluddin R, Suhartati S, Soetjipto S (2020) Potential inhibitor of COVID-19 main protease (Mpro) from several medicinal plant compounds by molecular docking study. Preprints 2020030226. https://doi.org/10.20944/preprints202003.0226.v1
50. Fischer W, Eron JJ, Holman W, Cohen MS, Fang L, Szewczyk LJ, Sheahan TP, Baric R, Mollan KR, Wolfe CR, Duke ER, Azizad MM, Borroto-Esoda K, Wohl DA, Loftis AJ, Alabanza P, Lipansky F, Painter WP. Molnupiravir, an Oral Antiviral Treatment for COVID-19. medRxiv [Preprint]. 2021. https://doi.org/10.1101/2021.06.17.21258639.
51. Gupta A, Zhou HX (2020) Profiling SARS-CoV-2 Main Protease ($M^{PRO}$) Binding to repurposed drugs using molecular dynamics simulations in classical and neural network -trained force fields. ACS Comb Sci 22:826–32. https://doi.org/10.1021/acscombsci.0c00140
52. Pal A, Curtin JF, Kinsella GK (2021) Structure based prediction of a novel GPR120 antagonist based on pharmacophore screening and molecular dynamics simulations. Comput Struct Biotechnol J 19:6050–6063
53. Zhang X, Macielag MJ (2020) GPR120 agonists for the treatment of diabetes: a patent review (2014-present). Expert Opin Ther Pat 30:729–742. https://doi.org/10.1080/13543776.2020.181185
54. Sparks SM, Chen G, Collins JL, Danger D, Dock ST, Jayawickreme C, Jenkinson S, Laudeman C, Leesnitzer MA, Liang X, Maloney P, McCoy DC, Moncol D, Rash V, Rimele T, Vulimiri P, Way JM, Ross S (2014) Identification of diarylsulfonamides as agonists of the free fatty acid receptor 4 (FFA4/GPR120). Bioorg Med Chem Lett 24:3100–3103
55. Sparks SM, Aquino C, Banker P, Collins JL, Cowan D, Diaz C, Dock ST, Hertzog DL, Liang X, Swiger ED, Yuen J, Chen G, Jayawickreme C, Moncol D, Nystrom C, Rash V, Rimele T, Roller S, Ross S (2017) Exploration of phenylpropanoic acids as agonists of the free fatty acid receptor 4 (FFA4): identification of an orally efficacious FFA4 agonist. Bioorg Med Chem Lett 27:1278–1283
56. Zhang X, Cai C, Winters M, Wells M, Wall M, Lanter J, Sui Z, Ma J, Novack A, Nashashibi I, Wang Y, Yan W, Suckow A, Hua H, Bell A, Haug P, Clapper W, Jenkinson C, Gunnet J, Leonard J, Murray WV (2017) Design, synthesis and SAR of a novel series of heterocyclic phenylpropanoic acids as GPR120 agonists. Bioorg Med Chem Lett 27:3272–3278
57. Janssen Pharmaceutica NV (2014) Preparation of GPR120 agonists for the treatment of type II diabetes. US 20140275179 A1
58. Janssen Pharmaceutica NV (2019) Cycloalkenyl derivatives useful as agonists of the GPR120 and /or GPR40 receptors and their preparation. WO 2019171277 A1
59. Piramal Enterprises Limited. Preparation of fused heterocyclic compounds as GPR120 agonists. WO 2016038540 A1; 2016b.
60. Piramal Enterprises Limited (2016c) Preparation of thiazole and thiophene derivatives containing carbon-carbon linker as GPR120 agonists for disease therapy. WO 2016125182 A1, p 74
61. da Silva Arouche T, Reis AF, Martins AY, Costa JFS, Junior RNC, Antonio Maia JC (2020) Neto interactions between Remdesivir, Ribavirin, Favipiravir, Galidesivir, Hydroxychloroquine and Chloroquine with fragment molecular of the COVID-19 main protease with inhibitor N3

complex (PDB ID:6LU7) using molecular docking. J Nanosci Nanotechnol 20(12):7311–23. https://doi.org/10.1166/jnn.2020.18955
62. Bristol-Meyers Squibb Company. Preparation of novel cyclopropane carboxylic acid compounds as GPR120 modulators useful for treatment of diabetes and related diseases. WO 2016040222 A1; 2016.
63. Selami E, Çınar E (2021) A molecular docking study of potential inhibitors and repurposed drugs against SARS-CoV-2 main protease enzyme. J Indian Chem Soc 98:100041
64. Smarajit M, Banerjee A, Nazmeen A, Kanwar M, Das S (2020) Active-site molecular docking of Nigellidine with nucleocapsid- NSP2-MPro of COVID-19 and to human IL1R-IL6R and strong antioxidant role of Nigella-sativa in experimental rats. J Drug Target. https://doi.org/10.1080/1061186X.2020.1817040
65. Rai H, Barik A, Singh YP, Suresh A, Singh L, Singh G, Nayak UY, Dubey VK, Modi G (2021) Molecular docking, binding mode analysis, molecular dynamics, and prediction of ADMET/toxicity properties of selective potential antiviral agents against SARS-CoV-2 main protease: an effort toward drug repurposing to combat COVID-19. Mol Divers 25(3):1905–1927. https://doi.org/10.1007/s11030-021-10188-5. Epub 2021 Feb 13. PMID: 33582935; PMCID: PMC7882058
66. Rashi J, Somdutt M (2020) Repurposing metocurine as main protease inhibitor to develop novel antiviral therapy for COVID-19. Struct Chem 31:2487–2499
67. Hosseini M, Chen W, Xiao D, Wang C (2021) Computational molecular docking and virtual screening revealed promising SARS-CoV-2 drugs. Precis Clin Med 4(1):1–16. https://doi.org/10.1093/pcmedi/pbab001 PMID: 33842834; PMCID: PMC7928605
68. Toelzer C, Gupta K, Yadav SKN, Borucu U, Davidson AD, Kavanagh Williamson M, Shoemark DK, Garzoni F, Staufer O, Milligan R, Capin J, Mulholland AJ, Spatz J, Fitzgerald D, Berger I, Schaffitzel C (2020) Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science 370(6517):725–730. https://doi.org/10.1126/science.abd3255 Epub 2020 Sep 21. PMID: 32958580; PMCID: PMC8050947
69. https://doi.org/10.1016/j.compbiomed.2021.104686
70. https://doi.org/10.1101/2021.06.17.21258639
71. Piramal Enterprises Limited (2015) Preparation of carboxylic acid derivatives for use as GPR120 agonists. WO 2015125085 A1, p 73
72. Piramal Enterprises Limited (2016a) Preparation of phenyl alkanoic acid derivatives for use as GPR120 agonists. WO 2016012965 A2, p 75
73. Usha YN, Dubey VK, Modi G Molecular docking, binding mode analysis, molecular dynamics, and prediction of ADMET/toxicity properties of selective potential antiviral agents against SARS-CoV-2 main protease: an effort toward drug repurposing to combat COVID-19. Mol Divers. https://doi.org/10.1007/s11030-021-10188-5
74. Wang J (2020) Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. J Chem Inf Model 60:3277–3286. https://doi.org/10.1021/acs.jcim.0c00179
75. Willyard C (2021) How antiviral pill molnupiravir shot ahead in the COVID drug hunt? Nature. https://doi.org/10.1038/d41586-021-02783-1

## Publisher's Note

*Jordan Journal of Pharmaceutical Sciences, Volume 15, No. 3, 2022*

# Chemical Constitution, *In-silico* Molecular Docking Studies and Antibacterial Activity of Flower Essential Oil of *Artabotrys hexapetalus*

***S. Sowndhariya[1], S. Ravi[1]*, J. Dharani[1], R. Sripathi[1]***

[1] Department of Chemistry, Karpagam Academy of Higher Education, India.

## ABSTRACT

The isolation of the volatile constituents from the flowers of *Artabotrys hexapetalus* was carried out using a simple headspace solvent-trapping technique and identified by GC-MS analysis. The major compounds are ethyl acetate 53.6%, isobutyl acetate (29.4%) and ethyl benzoate (14.2%). The odour of the solution obtained from this method was found to be similar to that of the fresh flowers. Further the essential oil from *A. hexapetalus* was obtained for the first time from India by hydro distillation using a Clevenger type apparatus and analysed by GC-MS. The plant yielded 1.26%, of the essential oils from the flower. The analysis lead to the identification of 28 compounds representing 96.17% of the total oil. The essential oil consists of predominantly oxygenated sesquiterpenes (51.91%) followed by sesquiterpenes (43.31%) and small quantities of monoterpenes (1.24%) and other compounds (1.34%). The main constituents of the essential oil obtained from the flowers of *A.hexapetalus* are β-caryophyllene (18.69%), caryophyllene oxide (14.54%), cubenol (12.53%) and ledol (11.5%). The essential oil showed antibacterial activity against bacterial strains *Streptococcus pneumonia*, *Staphylococcus aureus*, *Streptococcus pyogenes* and *Pseudomonas aeruginosa* exhibiting a zone of inhibition of 16.4, 15.7, 17.5 and 14.5 mm and MIC value of 2.5, 5.0, 2.5, 5.0 mg/ml respectively. Molecular docking analysis indicated that the essential oil constituents are nucleic acid and cell wall synthesis inhibitors. So it is worth to include this in cosmetics and fragrances.

**Keywords:** *Artabotrys hexapetalus*, essential oil, GC-MS, sesquiterpenes, β-caryophyllene, antibacterial, docking.

## 1. INTRODUCTION

*Artabotrys* species are traditionally used for a wide range of diseases like cholera, scrofula and malaria. The fruits and leaves of *Artabotrys* species are utilized as animal feeds, predominantly for goats, chimpanzees and cattle [1]. Due to the fragrance of the flowers of *Artabotrys* species, they are used as flavouring agents, in the manufacture of perfumes and for making stimulating tea-like beverages. Boiled juice of flowers is a stimulating beverage and used to treat vomiting, biliousness, blood diseases, heart and bladder disorders, itching and leucoderma[2]. They are used in the treatment of bad breath, headache, sweating, and thirst and also used as cardiotonic[3,4]. *A. hehapetalus* has numerous activities such as antispermatogenic, antiandrogenic, antioxidant, antimicrobial, and antidenaturation of protein, antiproteinase and anti-inflammatory.

The flowers from *A. hexapetalus* (Fig 1) have a sweet and fresh odour and however it was investigated only once from Thailand[5] to identify the volatile compounds responsible for its odour and from Vietnam to study the chemical composition of the essential oil[6]. To our knowledge we are investigating for the first time to identify these compounds from India. The sweet and fresh smell from this flower comes only between 5 to 8 a.m. in the morning and 6 to 8 p.m. in the evening[5]. It means that the compounds responsible for this odour from the flower are

***Corresponding author: S. Ravi**
ravisubban@rediffmail.com
Received: 9/10/2020 Accepted: 20/1/2022.
**DOI:** *https://doi.org/10.35516/jjps.v15i3.408*

released only within this period of time. Therefore, it is important that the onsite sampling and preconcentration steps are to be focussed in order to identify the volatile constituents of the flowers from *A. hexapetalus*. Thus the objective of the present work is to identify the chemical composition of the essential oil from flowers of *A. hexapetalus* after giving due importance to the onsite sampling and precondition step. Further another objective of the present work is to investigate the antimicrobial activity of the essential oil obtained from *A. hexapetalus* and to find a mechanism of the action of antimicrobial activity by molecular docking study. Several drugs that are currently available to the public for the treatment of different diseases have been developed based on in silico approaches. For example, Zanamivir, used to treat influenza, was developed using computer-assisted design[7] [A]. Nelfinavir and Saquinavar are used in the treatment of HIV and were also developed by computational methods[8]. [B].

**Fig.1. Flower from *A. hexapetalus***

## 2. Materials and Methods

### *2.1. Plant Material*

The flower of *A. hexapetalus* were collected (200 g) from the Coimbatore District (coordinates: 10.9880° N, 76.7740° E), Tamilnadu, India between 6 to 7 a.m. in the morning during the month of January 2018. The plant material was authenticated by Dr. R. Gopalan, Professor of Botany Department, KAHE, Coimbatore (Voucher No. KAHE/CHE/2018/102).

### *2.2. Extraction of Essential Oil*

After the onsite collection of the flowers between 6 to 7 a.m. in the morning the components were extracted immediately using a simple head space-solvent technique. In this method about 500 g of the flower were taken in an Erlenmeyer flask (500 ml capacity) fitted with a one holed rubber cork. Using an aquarium pump fresh air was blown in continuously through the inlet of the flask for nine hours. The vapour collected on the top surface of the flask was allowed to pass in to a round bottomed flask having 30 ml of methylene chloride solvent. This was repeated four times and the combined resulting solution was concentrated to 2 ml in a rotary evaporator and the concentrate was analysed by GC-MS.

Fresh flowers obtained (500 g) were chopped into small pieces and subjected to hydro distillation. A quantity of 60 g of the flowers *A. hexapetalus* was added to 300 ml of distilled water in a one litre flask fitted with a Clevenger apparatus and a condenser through which cold water was circulated to ensure condensation of essential oils for 2 h. This was repeated twice. At the end of the distillation, two phases were observed, an organic phase (essential oil) and an aqueous phase (aromatic water). The essential oil was collected, dried under anhydrous sodium sulphate. Until further analysis the resulted oil was stored at 4°C in a refrigerator.

### *2.3. Determination of Chemical Composition of Essential Oil*

GC-MS along with an ESI system with the ionization energy of 70 eV was utilized for essential oil composition

analysis. Agilent Technologies, 7890A, with a HP-5MS column (5 % phenyl methylpolysiloxane) 30 m × 0.25 mm ID × 0.25 μm film. The mass spectrometer with an ion-trap analyzer was set at 1508 for all analyses with an electron multiplier voltage of 1058V. Scanning was performed from m/z 39 to 400 in 70 eV EI (electronic impact) at 1 scan/ s-1 and the selected split ratio was 1:10. Helium (99.99%) with the flow rate of 1ml/min was used as the carrier gas. The injection part of the instrument was set at a temperature of 250ºC. The initial temperature of the column was maintained at 40°C for 1min, and then gradually increased to 240ºC at the rate of 30ºC/min. Essential oil constituents were tentatively identified by comparison of their GC retention indices (RI), determined with reference to a homologous series of C8-C20 n-alkanes and with those of available authentic standards and literature. Confirmation of such identification was done by comparing their mass spectral fragmentation patterns with those stored in the MS database (NIST 2005 and Wiley 7N libraries) and with mass spectra literature data. Components relative concentrations were obtained with the response factors to the FID.

***2.4. Antibacterial screening***

The antibacterial screening by zone of inhibition method and determination of Minimum inhibitory concentration (MIC) were determined using the bacterial strains like *Streptococcus pneumonia*, *Staphylococcus aureus*, *Streptococcus pyogenes* and *Pseudomonas aeruginosa* by the method as we reported earlier[9].

***2.5. Molecular Docking***

*2.5.1. Preparation of Proteins and ligands*

The three-dimensional structure of the proteins with PDB id: 3UDI, 3TYE, 3TTZ and 1JZQ were downloaded from the RCSB protein Data Bank and saved in PDB file format, for further studies in Auto dock vina under PyRx 0.8 Platform.

The compounds present in the essential oil obtained from *A. hexapetalus* were selected for docking studies. Molecular docking study has been carried out using the PyRx Version 0.8 docking program. Ligands 2D structures were drawn and converted into 3D using Chem Office 2002. After energy minimization of the ligands, it was docked with the protein's target sites (amino acids). Discovery studio was used to convert 2D in to 3D structure and the energy was minimized using AM1 method. To minimise the energy to minimum RMS gradient of 0.100 was set in each interaction. All structures were saved as PDB file format. All the ligand structures were then saved in SDF file format, to carry out docking in Autodock vina[10]. A grid box with dimension of 40 x 40 x 40A with 0.37A spacing and cantered on 29.47, 47.99, 8.86 was created around the binding site on proteins. The centre of the box was set a ligand centre, and grid energy calculations were carried out.

**3. Results and Discussion**

The isolation of the volatile constituents from the flowers of *A. hexapetalus* was carried out using a simple headspace solvent-trapping technique and the headspace vapour was flushed with air and collected in solvent methylene chloride. When analysed by GC-MS, Six compounds were identified from the resultant concentrated methylene chloride solution. The identified volatile compounds are ethyl acetate 53.6%, isobutyl acetate (29.4%) and ethyl benzoate (14.2%) as major compounds and ethyl propionate (1.6%), ethyl octonate (0.7%) and isobutyl valerate (0.43%) as minor compounds. The odour of the solution obtained from this method was identified to be similar to that of the fresh flowers. The presence of ethyl benzoate and ethyl propionate in the present investigation of the volatile constituents from the flowers of *A. hexapetalus* make the smell of the flowers of *A. hexapetalus* from India different from the flowers of *A. hexapetalus* from Thailand[5].

The chromatogram obtained from the GC-MS analyses was shown in Figure 2. It resulted in the identification of 28 compounds (Figure 3) representing 96.17% of the oil. The plant yielded 0.36%, (0.68 g) of the essential oils from the flower (the average yield of three

distillations). The essential oil consists of predominantly oxygenated sesquiterpenes (51.91%) followed by sesquiterpenes (43.31%), monoterpenes (1.24%) and small quantities of other compounds (1.82%). The GC-MS analysis results are summarized in the Table 1. Caryophyllene oxide (14.54%), β-caryophyllene (18.69%), cubenol (12.53%) and ledol (11.5%) were the main constituents of the essential oil of the flowers. It is having a strong green odour and differs a lot from the smell of the fresh flowers. This is due to the reason that in the high temperature prevailed during the hydrodistillation, the enzymatic processes that were responsible for the odour formation and release of the compounds would have denatured[11]. Hence the compounds which contribute to the significance odour could not be accumulated in the obtained essential oil.

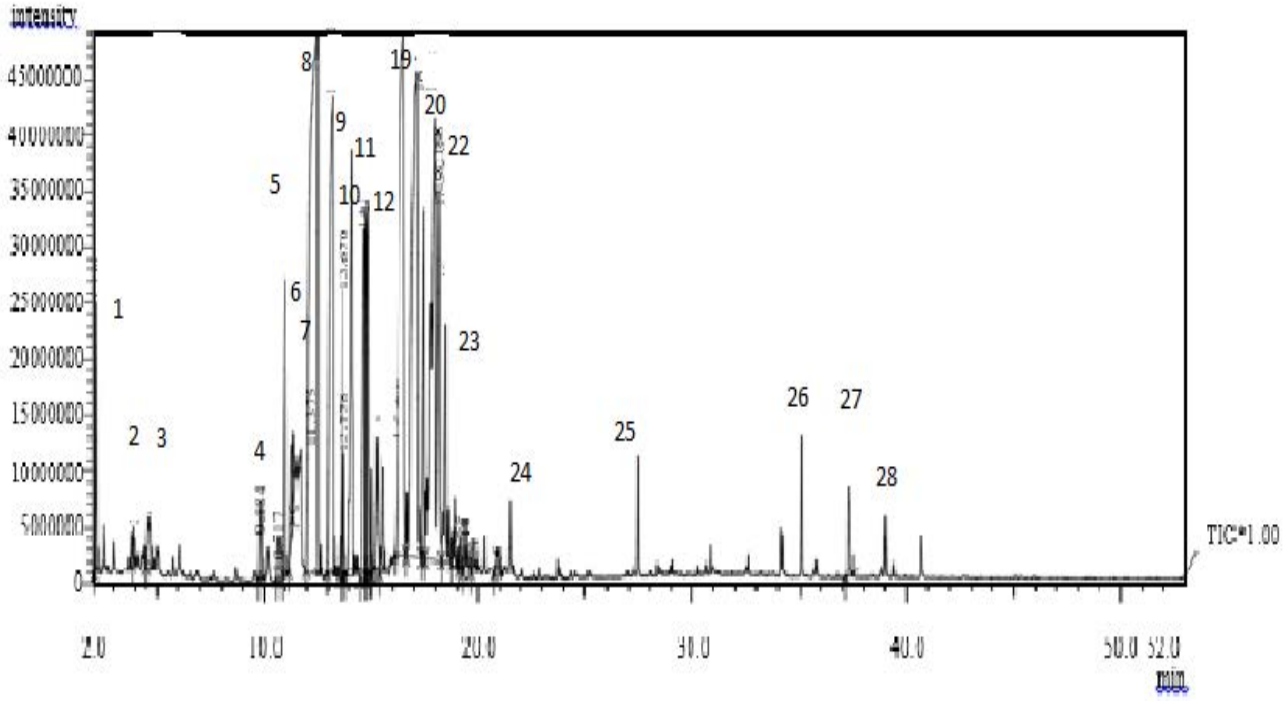



**Figure 2: GC-MS Chromatogram of essential oil obtained from *A. hexapetalus*.**

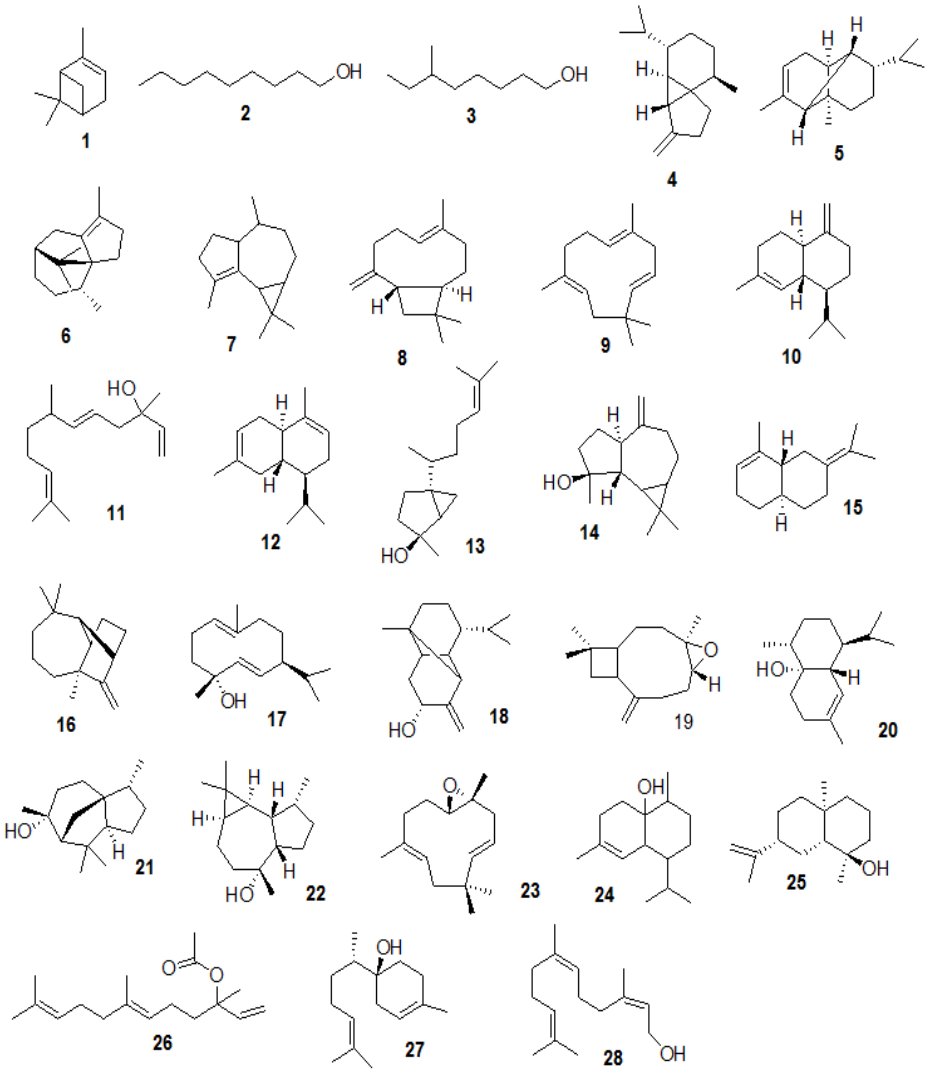



**Fig. 3 Structure of the compounds identified from the essential oil from the flowers of *A. hexapetalus***

In an earlier study twenty-eight components comprising of sesquiterpenes hydrocarbons (33% of the oil) and oxygenated sesquiterpenes (47.7%) were reported from the flower oil of *A. hexapetalus* collected from Vietnam. The major compounds are α-copaene (8.1%), β-elemene (1.0%), β-caryophyllene (11.4%), α-humulene (3.5%), γ-muurolene (3.5%), caryophyllene oxide (31.5%), and humulene epoxide (10.01%) [12].

**Table 1. Essential oil composition of *Artabotrys hexapetalus* as determined by GC-MS**

| Compound.No | Retention time[a] | Compound[b,c] | %[d] | Molecules formulae | Retention Index[e] |
|---|---|---|---|---|---|
| 1 | 2.1 | α-pinene | 1.24 | $C_{10}H_{16}$ | 934 |
| 2 | 3.94 | 1-Nonanol | 0.20 | $C_9H_{20}O$ | 1089 |
| 3 | 4.68 | 6-Methyloctan-1-ol | 0.64 | $C_9H_{20}O$ | 1109 |
| 4 | 10.71 | β-Cubeben | 0.41 | $C_{15}H_{24}$ | 1333 |
| 5 | 11.013 | Copaene | 3.91 | $C_{15}H_{26}$ | 1375 |
| 6 | 11.375 | cyperene | 1.00 | $C_{15}H_{24}$ | 1398 |
| 7 | 11.452 | α -Gurjunene | 0.59 | $C_{15}H_{24}$ | 1405 |
| 8 | 12.53 | β-Caryophyllene | 18.67 | $C_{15}H_{24}$ | 1420 |
| 9 | 13.263 | Humulene | 8.24 | $C_{15}H_{24}$ | 1449 |
| 10 | 13.670 | γ-cadinene | 1.97 | $C_{15}H_{24}$ | 1505 |
| 11 | 13.778 | Nerolidol | 0.60 | $C_{15}H_{26}O$ | 1520 |
| 12 | 14.19 | β-cadinene | 4.88 | $C_{15}H_{24}$ | 1530 |
| 13 | 14.70 | Sesquisabinene Hydrate | 3.86 | $C_{15}H_{26}O$ | 1534 |
| 14 | 14.79 | Spathulanol | 2.29 | $C_{15}H_{26}O$ | 1566 |
| 15 | 14.97 | Selina-3,7 (11)-diene | 3.18 | $C_{15}H_{24}$ | 1567 |
| 16 | 14.98 | Longifolene | 0.21 | $C_{15}H_{24}$ | 1568 |
| 17 | 15.067 | Germacrene D-4-ol | 0.46 | $C_{15}H_{26}O$ | 1569 |
| 18 | 15.29 | β-Copaene-4α-ol | 1.16 | $C_{15}H_{26}O$ | 1570 |
| 19 | 16.58 | Carryophyllene oxide | 13.46 | $C_{15}H_{26}O$ | 1573 |
| 20 | 17.18 | Cubenol | 12.53 | $C_{15}H_{26}O$ | 1590 |
| 21 | 17.52 | cedrol | 2.49 | $C_{15}H_{26}O$ | 1592 |
| 22 | 18.14 | Ledol | 11.57 | $C_{15}H_{26}O$ | 1594 |
| 23 | 18.51 | Humulene epoxide | 2.18 | $C_{15}H_{26}O$ | 1597 |
| 24 | 21.42 | 1-Cubenol, epi | 0.58 | $C_{15}H_{26}O$ | 1614 |
| 25 | 27.12 | Selin-11-en-4α-ol | 0.25 | $C_{15}H_{24}$ | 1641 |
| 26 | 35.04 | Nerolidol-Epoxyacetate | 0.50 | $C_{17}H_{28}O_2$ | 1687 |
| 27 | 37.21 | β-Bisabolol | 0.41 | $C_{15}H_{26}O$ | 1689 |
| 28 | 39.06 | Farnesol | 0.32 | $C_{15}H_{26}O$ | 1733 |

[a]Compounds are listed in order of their elution from a HP-5MS column.

[b]Identification: MS, based on comparison with NIST 14 MS databases;

[c]Retention index from NIST 14 and Wiley 275 mass spectral databases.

[d]Quantification was done by external standard method using calibration curves generated by running GC analysis of representative authentic components

[e]Retention index on the HP-5MS column, calculated using homologous series of $C_9$–$C_{18}$ alkanes.

From Thailand the essential oil was obtained by four different process like simple headspace solvent-trapping technique, solvent extraction, hydro distillation, and solid phase micro extraction (SPME) and the identified compounds were reported[5]. Oil from the hydro distillation method showed the presence of thirty one components, of which the major components were β-gurjunene (30.0%), Globulol (13.8%) and β-caryophyllene (10.1%). Essential oil obtained from the same source by solvent extraction led to the identification of thirty one components of which the major compounds were isopentyl acetate (12.6%), linalool (7.7%), 2-methylbutyl acetate (7.7%), limonene (5.7%) and 3-methylbutanol (5.7%). Alternatively when it was performed with solid-phase micro extraction (SPME) methods, thirty nine components were identified with ethyl acetate (12.8%) and isobutyl acetate (39.5%) as the major components[5]. Further Cadinol, spathulenol, β-caryophyllene oxide and cubenol (-) were reported from the essential oil obtained from Tanzania. The volatile constituents are dominated by sesquiterpene hydrocarbons and oxygenated sesquiterpenoids[14].

The present study showed that the chemical constituents from the essential obtained from Indian *A. hexapetalus* are found to be different with the essential oil obtained by the hydro distillation method from Tanzania[13] and Vietnam[12] and Thailand. In our study the quantity of β-caryophyllene is more than the caryophyllene oxide where as in the above studies caryophyllene oxide is more than that of β-caryophyllene. The study indicated that the sesquiterpenes β-caryophyllene oxide to be present in almost all essential oils obtained by hydro distillation of the *A. hexapetalus*. β-caryophyllene oxide was reported to exhibit mosquito repellent activity.

### *3.1. Antibacterial Screening*

The diameter of zone of inhibition was measured in mm and presented in the table 2. The essential oil from *A. hexapetalus* exhibited inhibitory activity against all the bacterial strains *Streptococcus pneumonia, Staphylococcus aureus, Streptococcus pyogenes* and *Pseudomonas aeruginosa* with MIC values of 2.5, 5.0, 2.5, 5.0 mg/ml and narrow inhibition zones of 16.4, 15.7, 17.5 and 14.5 mm respectively. Ampicillin was used as a positive control. Overall the results suggest that the essential oil of *A. hexapetalus* have a potential antibacterial activity. The activity is attributed to the various constituents present in the essential oil obtained from the flowers of *A. hexapetalus.*

### *3.2. Molecular Docking*

Antibiotics may either kill or inhibit the growth of bacteria by different mechanisms[15,16]. Now, in the current study, the knowledge on the target proteins of currently used antibiotics[17,18] is extended to the phytoconstituents which is identified from *A. hexapetalus* in order to examine their affinity with the bacterial proteins that are well known targets for some antibiotics with different mechanism of action such as cell wall synthesis, inhibitors of nucleic acid synthesis and antimetabolites. In the present study we carriedout the molecular docking studies with 3UDI (acinetobacter baumannii in complex with penicillin G), 3TYE (dihydropteroate synthase), 3TTZ (DNA gyrase) and 1JZQ (Isoleucyl-tRNA synthetase) proteins which represent the above three mechanisms. The docking score of the ligands with the protein 3TTZ and 1JZQ are not encourageable and hence not pursued further.

One of the target protein (PDB id: 3UDI) is from murD ligase which is involved in the cell wall synthesis and the other target is dihydropteroate synthase enzyme (DHPS; PDB id: 3TYE) a key component in the folate pathway of bacteria and primitive eukaryotes. The essential oil constituents were docked against these two targets and the compounds with a reasonable docking score (Kcal/mole) are presented in the table 3. Most of the ligands exhibited hydrophobic interactions (Figure 4a-4d) with the target proteins which are evidenced by their docking scores. This indicates that the essential oil components of *A. hexapetalus* behave as inhibitors of nucleic acids and cell wall synthesis inhibitors which involve in cell well synthesis. So we hypothesise that these essential oil constituents first interact with the cell wall to destruct the cell structure and then inhibits the normal synthesis of DNA that are

required for bacterial growth. β-Lactams act entirely outside the cell membrane, in the final phase of peptidoglycan biosynthesis. Sulfonamides inhibit the action of dihydropteroate synthetase (with p-aminobenzoic acid (PABA) as substrate), preventing the synthesis of dihydrofolic acid[17,20]. So, from the present study we can say that these compounds act on the multi targets and may serve as antibacterial agents.

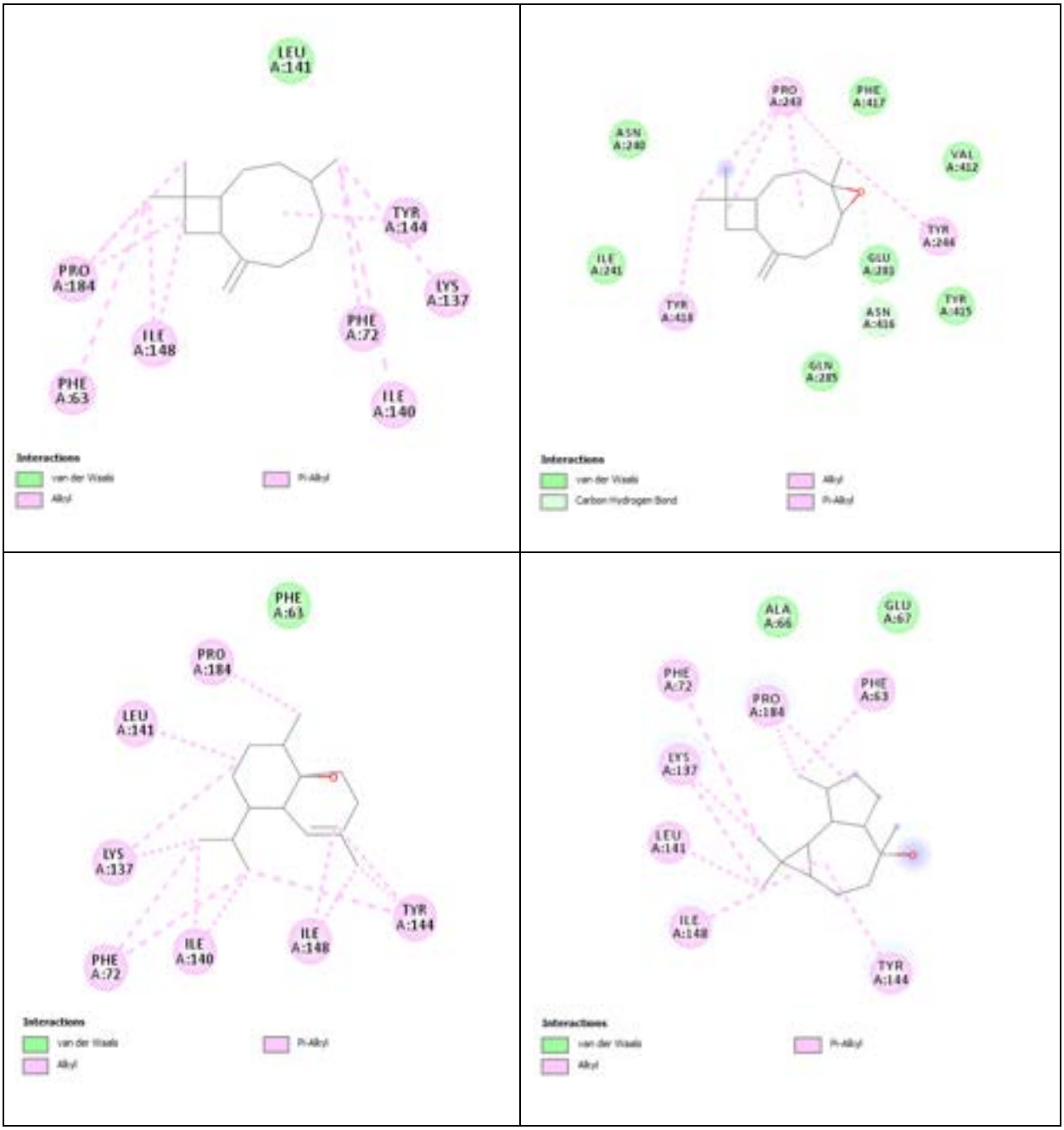



**Figure 4a: Molecular docking of Trans (beta)-caryophyllene, caryophyllene oxide, Cubenol and Ledol against the target protein with the PDB id: 3UDI**

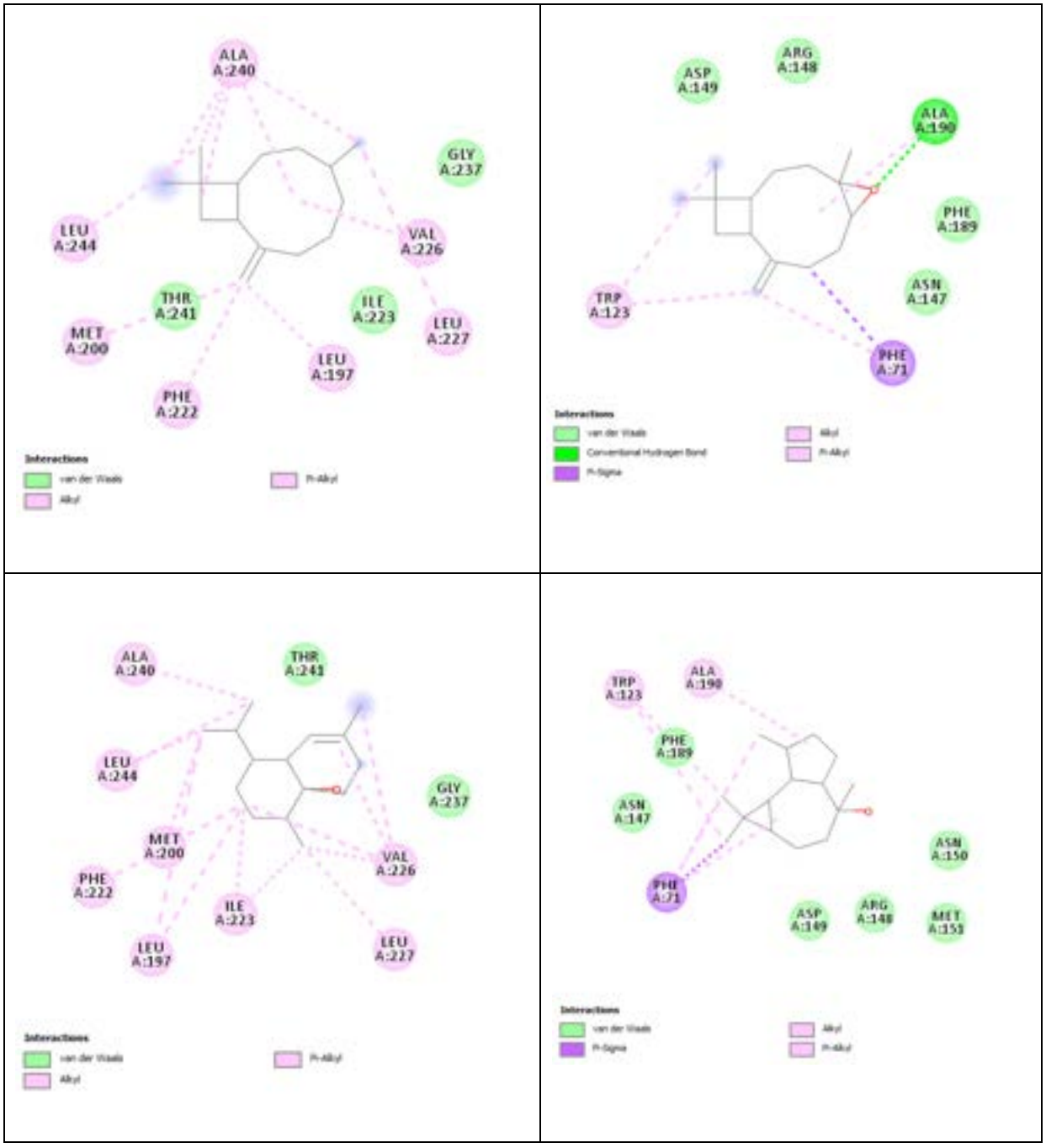



**Figure 4b: Molecular docking of Trans (beta)-caryophyllene, caryophyllene oxide, Cubenol and Ledol against the target protein with the PDB id: 3TYE**

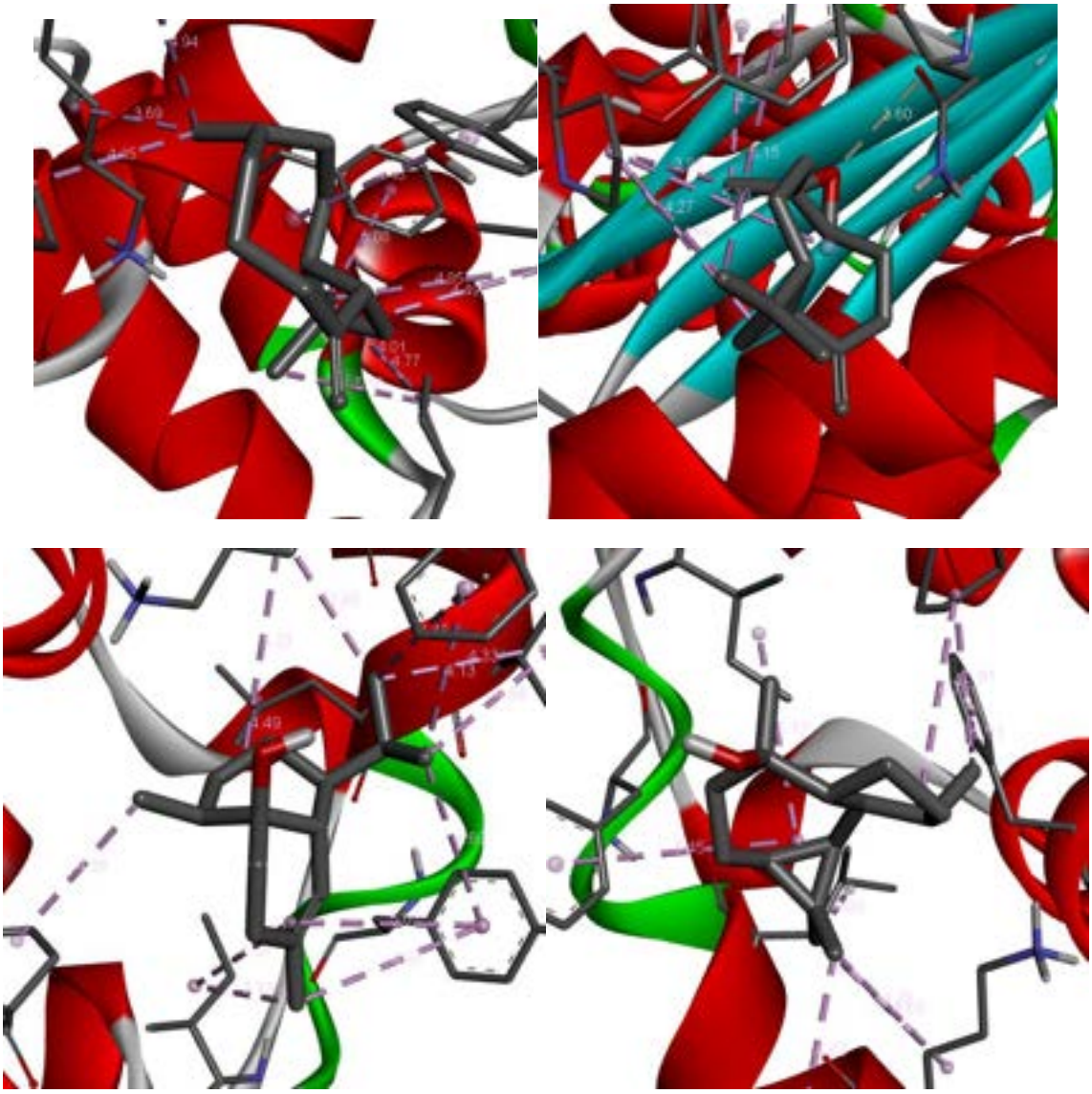


**Figure 4c: Molecular docking 3D images of Trans (beta)-caryophyllene, caryophyllene oxide, Cubenol and Ledol against the target protein with the PDB id: 3UDI**

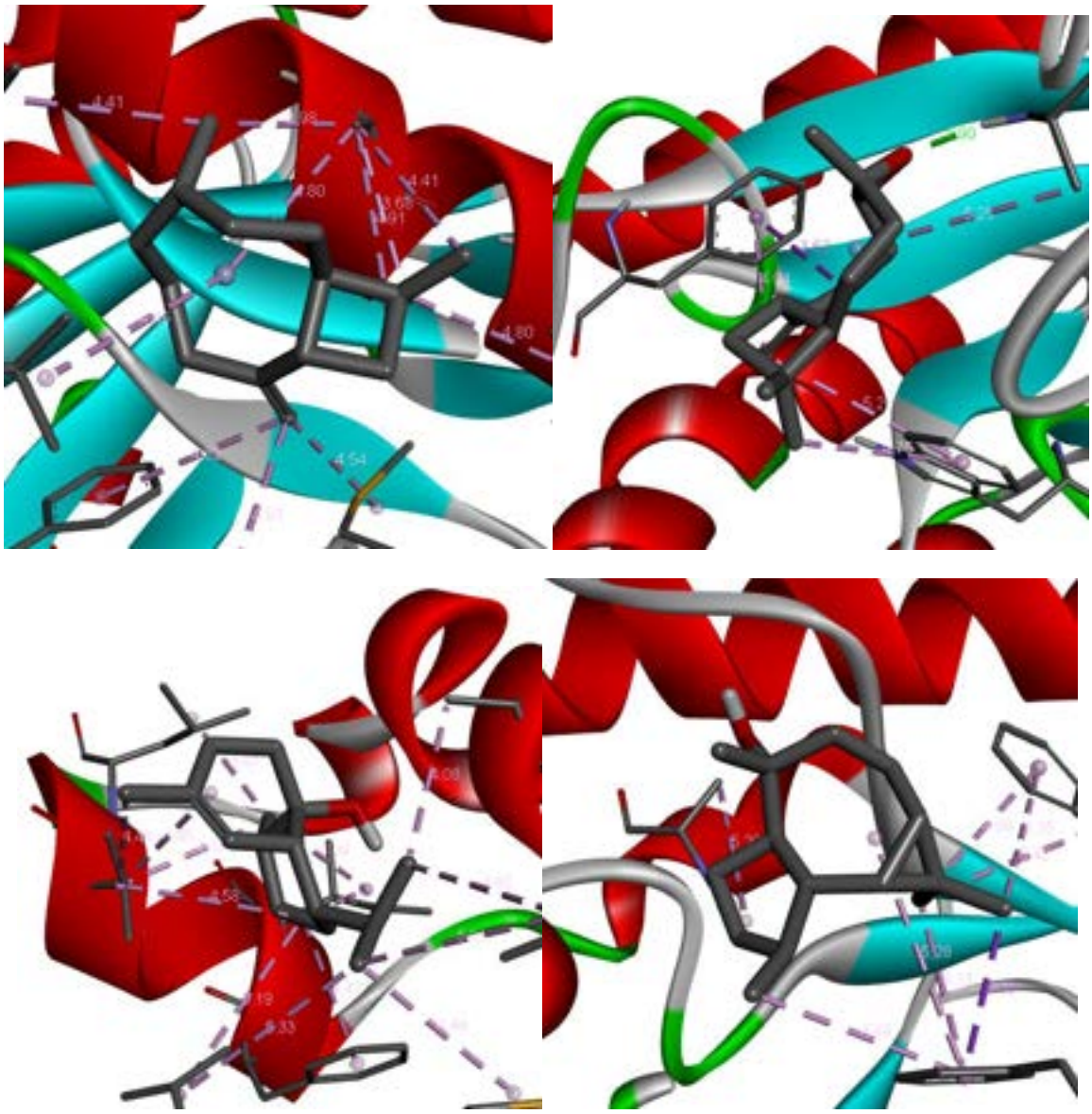


**Figure 4d: Molecular docking 3D images of Trans (beta)-caryophyllene, caryophyllene oxide, Cubenol and Ledol against the target protein with the PDB id: 3TYE**

**Table 2. Antibacterial activity of the essential oil obtained from the flowers of *A. hexapetalus***

| Bacterial strain | Zone of Inhibition(mm) | | Minimum Inhibitory Concentration(mg) | |
|---|---|---|---|---|
| | Essential oil | Ampicillin | Essential oil | Ampicillin |
| *Streptococcus pneumonia* | 16.4 | 19.5 | 2.5 | 2.5 |
| *Staphylococcus aureus* | 15.7 | 21.5 | 5.0 | 2.5 |
| *Streptococcus pyogenes* | 17.5 | 23.5 | 2.5 | 2.5 |
| *Pseudomonas aeruginosa* | 14.5 | 21.5 | 5.0 | 2.5 |

**Table 3. Molecular docking analysis of the essential oil constituents from *A. hexapetalus* against bacterial proteins**

| Ligands | Docking score 3TYE (Kcal/mole) | Docking score 3UDI (Kcal/mole) |
|---|---|---|
| Trans(beta)-caryophyllene | -6.8 | -6.9 |
| Carryophyllene oxide | -6.4 | -7.0 |
| Cubenol | -6.0 | -6.9 |
| Ledol | -6.6 | -6.8 |

**4. Conclusion**

Using the simple headspace solvent-trapping technique in association with GCMS the components responsible for the odour of the flowers of *A. hexapetalus* flowers were identified. The essential oil obtained from the flowers of *A. hexapetalus by* hydro distillation was analysed by GC-MS, and it lead to the identification of 28 compounds predominantly oxygenated sesquiterpenes (51.91%). Caryophyllene oxide (14.54%), β-caryophyllene (18.69%), cubenol (12.53%) and ledol (11.5%) were the main constituents of the essential oil. The essential oil showed antibacterial activity against bacterial strains *Streptococcus pneumonia*, *Staphylococcus aureus*, *Streptococcus pyogenes* and *Pseudomonas aeruginosa.* Molecular docking analysis indicated that the essential oil constituents act as inhibitors of cell well synthesis and nucleic acids synthesis. It can further be explored to use in the fragrances and cosmetics.

**Acknowledgements**

The authors acknowledge Central Instruments Facility, Karpagam Academy of Higher Education for providing all the equipments, instruments, chemicals and materials during this research work.

**Conflicts of Interest**

We declare that we have no conflict of interest.

## REFERENCES

1. Tan K. K. and Wiart C. (2014).Botanical descriptions, ethno medicinal and non-medicinal uses of the genus Artabotrys r.br. International Journal of Current Pharmaceutical Research, 6, 34-40.
2. Prakash A. O. (1980). Effect of *Artabotrys odoratissimus* extracts on rat uterine glycogen, protein and nonprotein nitrogen. Planta Medica, 38, 54-61. DOI: 10.1055/s-2008-1074837.
3. Van Valkenburg J. L. C. H. and Bunyapraphatsara N. (2001). Plant resources of South-East Asia, Medicinal and poisonous plants 85-89. Backhuys Publishers, Leiden Netherlands,
4. Savadi R. V. (2009).Phytochemical investigations and anti fertility properties of some medicinal plants. Ph.D thesis, Rajiv Gandhi University of Health Sciences.
5. Mahidol C., Chimnoi N., Chokchaichamnankit D. and Techasakul S. (20055). Identification of volatile constituents in *Artabotrys hexapetalus* flowers using simple headspace solvent-trapping technique in combination with gas

chromatography-mass spectrometry and retention indices. Acta Horticulturae, 677, 43-50.
DOI:10.17660/ActaHortic.2005.677.5

6. Phan G. M., Phan S.T. and König W.A. (2007) Chemical composition of the flower essential oil of *Artabotrys hexapetalus* (L. f.) Bhandare of Vietnam. Journal of Essential Oil Research, 19, 523-524.
DOI: 10.1080/10412905.2007.9699321
7. Prema L M., Gyanendra K., Stephen W W. and R.W. Thomas (2014). Recent advances in computer-aided drug design as applied to anti-influenza drug discovery, Curr Top Med Chem. 14(16):1875-89.
Doi: 10.2174/1568026614666140929153812.
8. Vinod P. R., and S. K. Shaju. (2020). Computational Evaluation of the Inhibition Efficacies of HIV Antivirals on SARS-CoV-2 (COVID-19) Protease and Identification of 3D Pharmacophore and Hit Compounds, Advances in Pharmacological and Pharmaceutical Sciences, 1-10 Doi.org/10.1155/2020/8818008
9. Sujina I. and Ravi S. (2012). In-vitro antimicrobial and cytotoxic activity of methanolic extract of *Osbeckia wynaadensis*. International Research Journal of Biological Science, 1, 33-38.
10. Dharani J., Sripathi R. and Ravi S. (2018). Chemical composition of *Cyanthillium cinereum (L.) H. Rob* essential oil and its molecular docking study against bacterial proteins. International Journal of Pharmaceutical Sciences and Research, 10, 2216-2220.
11. Tama A., Pecetti L., Povolo M. and G. Contarini, (2000). A Comparison between two systems of Volatile Sampling in Flowers of Alfalfa (*Medicago sativa* L.). Phytochemical Analysis, 11, 148-152.
DOI: 10.1002/(SICI)1099-1565(200005/06)11
12. Phan G. M., Phan S. T. and Konig W. A. (2007). Chemical composition of the flower essential oil of Artabotrys hexapetalus (L. f.) Bhandare of Vietnam. Journal of Essential Oil Research, 19, 523-524. DOI: 10.1080/10412905.2007.9699321
13. Suleiman R. A., Mgani Q. A. and Nandoro S. S. (2014). Chemical compositions and mosquito repellency of essential oils from *Artabotrys hexapetalus* and *Artabotrys rupestris.* International Journal of Biological and Chemical Sciences, 8, 2804-2812.
DOI: 10.4314/ijbcs.v8i6.37
14. Ravi S. and Sundaram K. (2020). The essential oil constituents of Artabotrys species – A review. Journal of Phytology, 12, 24-28.
15. Kohanski M.A., Dwyer D.J. and J. J. Collins. (2010). How antibiotics kill bacteria: from targets to networks. Nat Rev Microbiol. 8(6):423-35. Doi: 10.1038/nrmicro2333.
16. Dharani J., Sripathi R. and S. Ravi. (2018). Chemical Composition of Cyanthillium Cinereum (L.) H. Rob Essential Oil and its Molecular Docking Study against Bacterial Proteins, J. Pharm. Sci. & Res. 10(9), 2216-2220
17. Alves M. J., Froufe H. J., Costa A. F., Santos A. F., Oliveira L. G., Osório S. R., Abreu R. M., Pintado M. and Ferreira I. C. (2014). Docking Studies in Target Proteins Involved in Antibacterial Action Mechanisms: Extending the Knowledge on Standard Antibiotics to Antimicrobial Mushroom Compounds. Molecules, 19, 1672-1684. DOI: 10.3390/molecules19021672
18. Sripathi R. and Ravi S. (2017). Molecular Docking Studies of the Constituents Present in the Essential Oil of *Plectranthus hadiensis* against Bacterial Proteins. International Journal of Chemical Science, 15, 185.
19. Unni J., Mohammed Afzal A., Ashish Devidas W., Sameer K. V., Krishnan R., Susobhan M. (2020). Synthesis, Biological Evaluation and Molecular Modelling Studies of novel 2-[(2,4-dioxo-1,3-thiazolidin-3-yl) acetyl]-N-arylhydrazinecarbothioamides as Antibacterial Agents Targeting Alanine Racemase Enzyme. Jordan Journal of Pharmaceutical Sciences, 13, 337-361.
20. Kalirajan R., Gowramma B., Gomathi S., Vadivelan R. (2021) Activity of Isoxazole substituted 9-aminoacridines against SARS CoV-2 main protease for COVID19: A computational approach. Jordan Journal of Pharmaceutical Sciences, 14, 403-416.

# التركيب الكيميائي، في (سيلكو موليكيولار) دراسات الالتحام الجزيئي والنشاط المضاد للجراثيم في الزيوت الأساسية لزهور *Artabotrys hexapetalus*

***س. سونيهاريا[1]، س. رافي[1*]، ج.دارانى[1]، ر. سريباثي[1]***

[1] قسم الكيمياء، أكاديمية كارباجام للتعليم العالي، الهند.

## ملخص

عزل المكونات المتطايرة عن زهور ارتابورتيز هيكسابيتاولوس باستخدام عينة تقنية سولفنت–ترابينج والذي تم التعرف عليه من قبل تحاليل جي سي–ام اس. أكبر المكونات هي اثيل اكسيتات 53.6%، ايزوبوتيل اكسيتات (29.4) واثيل بنزوات (14.2). وقد اكتشف أن رائحة المحلول الذي تم الحصول عليه من هذه الطريقة تشبه رائحة الزهور الطازجة. بالإضافة للزيت الأساسي الناتج من أ. هيكسابيتاولوس والذي تم الحصول عليه أول مرة من الهند من قبل جهاز هيدرو ديستيلاشن يوسينج أكليفينجر تايب والذي تم تحليله من قبل جي سي–ام اس. وقد أخرجت النبتة 1.26% من الزيوت الأساسية للزهرة. وقد أسفر التحليل عن التعرف على 28 مكون تشكل 96.17% من الزيت. كما يتكون الزيت العطري في الغالب من اوكسجيناتد سيسكويتربنيس (51.91%) متبوعة ب سيسكويتربنيس (43.31%) وكميات قليلة من مونوتربينس (1.24%) ومكونات آخرى (1.34%). المكونات الرئيسية للزيت الأساسي الذي تم الحصول عليه من الزهور ل أ. هيكسابيتاولوس هي ب–كاريوفيلينى (18.69%)، كاريوفيلليني اوكسيد (14.54%)، كوبنول (12.53%) وليدول (11.5%). وقد أظهر الزيت العطري نشاطًا مضادًا للبكتريا ضد السلالات البكتيرية ستريبتوكوكوز بنومونيا، ستافيلوكوس أوريوس، ستربتوكوكوز بيوجينس أند بسويدوموناس أروجينوسا عارضة منطقة أوف انهيبيشن أوف 16.4، 15.7، 17.5 و14.5 مم وقيمة أم أي سي ب 2.5، 5.0، 2.5، 5.0 مج/مل على التوالي. وقد أشار تحليل موليكولار دوكينج أن مكونات الزيت العطري هي حمض نووي ومثبطات تخليق جدار الخلية. وعليه من المجدي إدخال ذلك في العطور ومستحضرات التجميل.

**الكلمات الدالة:** *Artabotrys hexapetalus*، الزيوت الأساسية، جي سي–ام اس، سيسكويتربنيس، أس كاريوفيلليني، انتيباكتيريال، الالتحام الجزيئي.

---

[*] ***المؤلف المراسل: اس. رافي***

ravisubban@rediffmail.com

تاريخ استلام البحث 2020/10/19 وتاريخ قبوله للنشر 2022/1/20.