Send Orders for Reprints to reprints@benthamscience.ae

Current Computer-Aided Drug Design, 2018, *14*, 000-000

RESEARCH ARTICLE

Combinatorial Design of Molecule using Activity-Linked Substructural Topological Information as Applied to Antitubercular Compounds

Chandan Raychaudhury, Md. Imbesat Hassan Rizvi, Debnath Pal*

Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

Abstract: ***Background*:** Generating a large number of compounds using combinatorial methods increases the possibility of finding novel bioactive compounds. Although some combinatorial structure generation algorithms are available, any method for generating structures from activity-linked substructural topological information is not yet reported.

***Objective*:** To develop a method using graph-theoretical techniques for generating structures of antitubercular compounds combinatorially from activity-linked substructural topological information, predict activity and prioritize and screen potential drug candidates.

***Methods*:** Activity related vertices are identified from datasets composed of both active and inactive or, differently active compounds and structures are generated combinatorially using the topological distance distribution associated with those vertices. Biological activities are predicted using topological distance based vertex indices and a rule based method. Generated structures are prioritized using a newly defined Molecular Priority Score (MPS).

***Results*:** Studies considering a series of Acid Alkyl Ester (AAE) compounds and three known anti-tubercular drugs show that active compounds can be generated from substructural information of other active compounds for all these classes of compounds. Activity predictions show high level of success rate and a number of highly active AAE compounds produced high MPS score indicating that MPS score may help prioritize and screen potential drug molecules. A possible relation of this work with scaffold hopping and inverse Quantitative Structure-Activity Relationship (iQSAR) problem has also been discussed.

***Conclusion*:** The proposed method seems to hold promise for discovering novel therapeutic candidates for combating Tuberculosis and may be useful for discovering novel drug molecules for the treatment of other diseases as well.

ARTICLE HISTORY

Received: November 10, 2017
Revised: April 10, 2018
Accepted: April 30, 2018

DOI:
10.2174/1573409914666180509152711

Keywords: Combinatorial drug design, activity-linked substructure, graph theory, topological vertex index, rule based method, activity prediction, prioritization, screening.

1. INTRODUCTION

The need for new drugs to combat diseases and to cater to increasing plethora of resistant infections is clinching towards an urgent situation. This demand can possibly be satisfied through the discovery of novel bioactive compounds that remain underexplored in the traditional drug discovery pipeline. Several efforts have already been made in this direction. For example, Ruddigkeit *et al*. [1] have built a data base of all possible compounds of 17 atom size made from C, N, O, S and halogens. While such an effort of finding a suitable drug candidate out of several billions of compounds is certainly a useful endeavour, it is felt to use serendipity intuitively, instead, to search a relatively smaller set of molecules that is exhaustive with respect to the defined limits, activity linked and is rationally guided as well may have higher chances of success and may help accelerate the drug discovery process.

Current drug discovery pipelines seek to search for new bioactive compounds mainly using data modelling and activity prediction [2-4], through 3D virtual High Throughput Screening (vHTS) using molecular docking and scoring studies [5], and by carrying out 3D Quantitative Structure-Activity Relationship (QSAR) studies [6]. To enhance the chances of discovery, chemical diversity of combinatorial compounds is enforced and approaches that can generate structures having different scaffolds in the sense of scaffold hopping [7] may be preferable which offers drug designers wider options for looking at diverse

*Address correspondence to this author at the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India; Tel: +91-80-2293-2901; Fax: +91-80-2360-6332; E-mail: dpal@iisc.ac.in

1573-4099/18 $58.00+.00

molecular structures. This may be particularly useful in designing novel bioactive compounds for emerging challenges such as for discovering novel antitubercular agents [8] and for handling drug resistance problems [9].

Molecular topology-based approaches are most suitable for generating and guiding the design of novel molecular structures [10, 11] and graph theoretical methods [12] have been found to be most useful for serving the purpose. However, existing methods are primarily used as engines for generating structures with no connection to their biological activities unless subjected to separate activity prediction studies. Therefore, a method that *de novo* generates compounds combinatorially in such a way that they get linked to their activities automatically may better help the drug discovery effort and may bring down the number of structures to be generated, helping accelerate the process. Presumably, use of topological molecular descriptors [2] and the associated connectivity information can be helpful in this regard. Moreover, doing that using a single molecular / substructural descriptor that is used for activity prediction too may be even a simpler method and linked to inverse QSAR (iQSAR) [13] approach which is another interesting area of research in terms of getting structures back from quantitative molecular descriptors. Therefore, an integrated method that can be used for generating structures of diverse topological architecture / scaffold from a single molecular (structural or, substructural) descriptor *de novo* coupled with activity prediction and prioritization and screening of potentially active compounds may be an attractive approach for designing / discovering novel drug candidates.

The present work is aimed at developing an integrated method and the corresponding algorithms (computer programs with different modules) that can be used for activity-linked combinatorial drug design based primarily on graph-theoretical techniques, predict activities and prioritize and screen potential drug candidates. Activity–linked combinatorial structure generation is the most important part of this approach and we have leveraged a non-isomorphic rooted tree generation algorithm [14] and a cycle enumeration method [15] to design novel compounds in the form of reconstructed molecular graphs as outlined earlier [16, 17]. Activity prediction is done using a rule-based method [16, 17] and compounds are prioritized using a newly defined "Molecular Priority Score (MPS)".

In order to investigate the usefulness of the proposed integrated method with special emphasis on antitubercular drug discovery, we have carried out studies with a series of 41 antitubercular Acid Alkyl Ester (AAE) derivatives [18] and three known antitubercular drugs - Isoniazid, Pyrazinamide and Ethionamide - in a drug resistance scenario. The method has been able to predict the activities of AAE derivatives with a high percentage of success rate. Regarding structure generation / reconstruction, we have been able to reconstruct structures of some active AAE compounds from the substructural information associated with activity related vertices of other active AAE compounds using the corresponding algorithm. Moreover, several highly active compounds have been prioritized by MPS showing the usefulness of MPS in screening potential antitubercular drug molecules. For the AAE series, we have also reported some validation results related to activity prediction in terms of 'accuracy', 'sensitivity' and 'specificity' that shows the usefulness of the vertex index and the method used. In studies with three known antitubercular drugs, mentioned above, the structure generation has been done by relaxing distance constraint in getting the structure of one of the compounds which is not drug resistant from the distance distribution information associated with the activity related vertex of another compound which is resistant to tuberculosis. It appears from the results obtained from this study that the proposed method may find useful applications in discovering novel antitubercular drugs and to overcome drug resistance problem such as that for the treatment of tuberculosis [19]. The method seems to have the potentiality to emerge as a useful drug discovery tool for other diseases as well.

2. MATERIALS AND METHODS

The proposed integrated method has three components in a broader sense – (1) prediction of activity from substructural information; (2) generation of chemical structures from substructural information and conversion to SMILES notation; and (3) prediction of activity of the compounds obtained from the generated structures for a biological end-point of interest followed by compound prioritization and screening. Topological substructural information is the basis for carrying out these studies and a vertex index (D^x), the distance exponent index [20], has been used for that purpose. The index D^x helps to consider the connectivity properties of the atoms situated in the closer neighbourhood of a given atom more with a negative exponent value. Since D^x index having $x = -4$ i.e., D^{-4} index has been found to be useful in predicting activities for different series of bioactive compounds [18, 20, 21], we have chosen to use this index for the present study.

2.1. Computation of Vertex Index and Activity Prediction

The computation of D^{-4} index has been illustrated by considering the molecular graph G of the carbon skeleton of pentane and the corresponding distance matrix $D(G)$ shown in Fig. (**1**).

G: $\bullet^1$-$\bullet^2$-$\bullet^3$-$\bullet^4$-$\bullet^5$

$D(G)$:

	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

Fig. (1). Graph G representing carbon skeleton of the straight chain isomer of pentane and the corresponding vertex labelled distance matrix $D(G)$.

Therefore, D^{-4} index for the five vertices v_i, $i = 1, 2, \ldots, 5$ of G may be computed as:

$$D^{-4}(v_1) = 1^{-4} + 2^{-4} + 3^{-4} + 4^{-4} = 1.0787$$

$$D^{-4}(v_2) = 1^{-4} + 1^{-4} + 2^{-4} + 3^{-4} = 2.0748$$
$$D^{-4}(v_3) = 1^{-4} + 1^{-4} + 2^{-4} + 2^{-4} = 2.1250$$
$$D^{-4}(v_4) = 1^{-4} + 1^{-4} + 2^{-4} + 3^{-4} = 2.0748$$
$$D^{-4}(v_5) = 1^{-4} + 2^{-4} + 3^{-4} + 4^{-4} = 1.0787$$

By following the same procedure, the index D^{-4} can be computed for all the atoms (vertices) of all the compounds (molecular graphs) in the data set considering the hydrogen-suppressed graphs of the compounds for carrying out further studies. It may be noted that consideration of hydrogen-suppressed graph is particularly important here since we intend to generate structures *de novo* and doing that using hydrogen-filled graphs may pose computational bottlenecks as a very large number of structures are expected to be generated during this process. Moreover, the hydrogen-filled graphs can always be created from hydrogen-suppressed graphs if required.

Therefore, a data set containing both active and inactive compounds for a biological end-point of interest has to be gathered first from the literature (or, experimental laboratories). The dataset is then divided suitably into a training set and a test set and the training set is used for the system to learn about the structural requirement that makes a compound active and thus it is standardized for activity prediction for a biological end-point of interest. Subsequently, the vertex index (D^{-4}) values are computed for the vertices of the molecular graphs representing training set compounds to identify ranges and the activity related vertices falling in the identified ranges which is required for using the rule based method [17, 18]. Therefore, once the indices are computed, they are arranged in an ascending order and ranges of values coming from both active and inactive compounds are found in the ordering. The ranges are termed "Active" or, "Inactive" based on a set of rules applied on the number of D^{-4} index values, coming from active and inactive compounds, falling in the ranges. The rules [16, 17] used for the present study are given below:

1. Three or, more vertex index values coming exclusively from active compounds and exclusively from inactive compounds are said to form an "active range" and an "inactive range", respectively. However, at least three index values in a range should be distinct if they come from the same compound and at least two index values in a range have to be distinct if they come from different compounds.
2. Some single vertex index value coming from both active and inactive compounds is not considered to form an 'active range' or, 'an inactive range' by itself or, along with other vertex index values unless two-thirds of that single vertex index comes from active compounds or, inactive compounds respectively.

The vertex index values forming active ranges may be regarded as a set of features constituting "Topological Biophore" which are responsible for a compound to exhibit certain biological activities [22]. Therefore, if the index values for some of the vertices of the molecular graph of a compound fall in active ranges then those vertices may be regarded as representing the same or, similar features to form a set of features representing certain topological biophore which may make the compound active. However, some of the vertex indices for a compound may fall in inactive ranges too. In order to investigate that and predict activities of chemical compounds, another set of rules are applied on the number vertices of a compound falling in active and inactive ranges [16, 17]:

A compound is predicted ACTIVE if all or, some of its vertices (atoms) fall: -

1. Only in active ranges; or,
2. In both active and inactive ranges and the number of index values falling in active ranges is greater than those falling in inactive ranges.

Otherwise the compound is predicted 'INACTIVE'.

If the activity prediction for both the training set and the test set compounds are of very high percentage with no or, very few (acceptable) wrong predictions, the system is considered to be standardised for the prediction of activity for the given biological end-point.

2.2. Theory of Structure Generation and Implementation

For the purpose of designing novel bioactive compounds, structures are generated using the topological distance distribution associated with an identified activity related vertex (root vertex). Topological distance distribution gives the topological distances of all the vertices in a molecular graph from the root vertex and this is the key distinguishing idea from the generic structure generation for a given number of vertices. It also stems from the reasonably correct activity predictions obtained using topological distance-based vertex indices [16-18, 20, 21]. The structure generation exercise is composed of generating rooted trees [14], generating cyclic compounds [15] and imposing topological distance restriction [16, 17]. Some relaxation has also been allowed in the distance criteria for generating structures having increased or decreased number of vertices representing non-hydrogen atoms and a matching criterion for distance distribution has been suitably changed to accommodate the addition, deletion and migration of the vertices over the tree structures with exact distance restriction.

2.2.1. Tree-Based Structure Generation for a Given Number of Vertices

Beyer *et al*. [14] have proposed an iterative algorithm to reverse-lexicographically generate non-isomorphic canonical trees for a given number of nodes. The trees generated by the algorithm can in general have any number of children for any parent node. In context of chemical structures of carbon atoms, only those trees are filtered and kept where the root has at the most four children and the rest of the nodes have at most three children. This restriction can later be further refined for hetero-atoms in accordance with their valency. The nodes in the tree can be chosen in a combinatorial manner and joined by edges to create cycles. We have considered the algorithm by Gibbs [15] to enumerate all the cycles present. From the entire cycle set, the fundamental cycles are then obtained in accordance with the IUPAC

convention of the number of rings in polycyclic systems [23] where the number of rings is equal to the minimum number of scissions required to convert the system into an open chain compound or structure. During this process of cycle introduction, duplicate structures are expected to be generated owing to combinatorial nature of vertex selection for cycle completion. Such duplicate structures are identified and eliminated through canonicalization in conjunction with unique SMILES generation [24].

2.2.2. Structure Generation with Distance Distribution Constraint

Since, the purpose is to design novel drug molecules possessing better desirable activities (for the given end-point) than the existing compounds, picking an activity related vertex (substructure) from the most or a highly active compound seems reasonable as such a substructure may be believed to contain the topological structural requirement for making the compound exhibit better activity. Hence to start this exercise, a vertex of the most or a highly active compound is chosen from an active range in the ordering of the D^x values of the training set compounds. However, while picking a vertex from an active range, the composition of the active range such as the length of the range *i.e.* the number of index (D^x) values in the range, the number of compounds contributing to forming the range *etc*. may be taken into consideration. In the present work, the length of the active range has been considered to pick up the vertex. From an intuitive point of view, if the length of an active range is large and / or is formed by contribution from a large number of active compounds (even if not by as many compounds) then the range may be regarded as a "STRONG" range. Therefore, a vertex picked-up from a strong range as well as coming from the most or a highly active compound may be considered to be a reasonable starting point for generating structures *de novo*. The topological distance distribution associated with the vertex is considered to generate structures.

2.2.2.1. Non-Isomorphic Canonical Tree Generation with Given Distance Distribution

The basic idea remains the same as in case of general tree distribution. However, as the present work requires generation of trees having same distance distribution as that of the starting vertex, the starting level sequence that can be used to initiate the tree generation algorithm is the largest lexicographic representation corresponding to the distance distribution which is obtained as explained below:

For a given distance distribution, let's say there are a_i vertices at distance i where $a_i \geq 1 \; \forall \; i$ such that $1 \leq i \leq e$, e being the eccentricity of the vertex from where the values of distance distribution is obtained. Then the lexicographically largest representation will be given by the level array $[1,2 \ldots e, \underbrace{e,e \ldots e}_{a_e - 1 \; times}, \ldots \underbrace{i,i \ldots i}_{a_i - 1 \; times}, \ldots \underbrace{2,2 \ldots 2}_{a_2 - 1 \; times}]$ *i.e.* first values are strictly increased up to e starting from 1 and are then monotonically decreased from e to 2.

Additionally, as the algorithm builds trees successively, it suffices to use only those level sequences *i.e.* trees, which have the same distance distribution as starting vertex, to further generate the compound structures.

2.2.2.2. Cycle Introduction by Adding Edges While Maintaining Distance Distribution

Once again, the generic procedure outlined previously for cycle introduction remains valid except that in order to preserve the distance distribution, the edges to be introduced between any two vertices, say i and j which do not have a parent-child relationship between them and the levels they occupy with respect to the root vertex denoted by l_i and l_j respectively must satisfy the criterion $|l_i - l_j| \leq 1$.

We call it *Exact matching*.

Rest of the procedure including the SMILES notation generation remains the same.

2.2.3. Structure Generation with Slightly Relaxed Distance Distribution

The approach taken so far *(exact matching)* suffers from the drawback that only those compound structures will be generated that have the same number of non-hydrogen atoms as the starting molecule from which the distance distribution was obtained. This subsection tries to tackle this drawback by slightly relaxing the distance distribution matching criteria for the trees with number of nodes deviating from the source or starting distribution. This deviation can either lead to increased or decreased number of nodes.

2.2.3.1. Non-Isomorphic Canonical Tree Generation with Relaxed Distance Distribution

The first step involves specifying the number of vertices (after factoring in the deviation) and then generating the trees. Positive deviation means required number of vertices is greater than that in the current tree while negative deviation means the required number of vertices is lesser. However, since exact distance distribution matching is not possible in this case, two variants of relaxed distribution matching are considered as explained below:

Strong matching – This matching corresponds a situation when the distance distribution of the generated tree can be thought of as obtained from the source distance distribution by either addition or deletion of vertices at any level. However, simultaneous insertion or deletion of vertices is not allowed for a given deviation. The obtained distance distribution corresponds to a pruned tree of the source distance distribution or vice-versa depending on whether the deviation is negative (by deletion of vertices) or positive (by addition of vertices) respectively.

Now, let n be the number of vertex deviations allowed either by increasing or decreasing the number of vertices. Further, let c_i^s denote the count of vertices at level i in the source distance distribution, c_i^p denote the count of vertices at level i in the present distance distribution and e denote the maximum of the eccentricity of the source and present distance distribution. Then, the criteria for *strong matching* is given by:

$$\left| \sum_{i=1}^{e} \left(c_i^s - c_i^p \right) \right| = n$$

Weak matching – In this situation the distance distribution matching criteria is even further relaxed in the sense that simultaneous addition and deletion of vertices at any level is allowed resulting in migration of vertices from one level to another. If this is allowed without a cap on the number of vertex migrations, then all the possible structure generation will be considered a match including the linear chain. Thus, for *weak matching*, number of such vertex migrations allowed should also be provided and in general should be low in order to match the source distance distribution as closely as possible.

Now, let n, c_i^s, c_i^p and e carry the same meanings as defined in the case of strong matching. Further, let m denote the number of vertex migrations allowed, m_p denote the sum of vertex surplus and m_n denote the sum of vertex deficit in the source distance distribution over the present distance distribution. Then, the criteria for *weak matching* is given by:

$$\left|\sum_{i=1}^{e}\left(c_i^s - c_i^p\right)\right| = n$$

and

$$\min\left(m_p, m_n\right) = m$$

where

$$m_p = \sum_{i=1}^{e} max\left(\left(c_i^s - c_i^p\right), 0\right)$$

$$m_n = \left|\sum_{i=1}^{e} min\left(\left(c_i^s - c_i^p\right), 0\right)\right|$$

Rest of the procedure of cycle introduction, canonicalization and unique SMILES notation generation is the same as before.

In addition to the generic implementation of the algorithm explained above, some user-defined parameters are provided in the program which may be used to restrict the number and size of the cycles to be created in the 2D structures. Similarly, other user-defined parameters, incorporated in the program, may be used to add multiplicity of bonds (double and triple bonds) between pairs of vertices as well as other hetero-atoms (*e.g.*, nitrogen, oxygen, halogens *etc*.) to get complete 2D structures of the compounds. For investigating structural details of the generated compounds one can use any standard molecular modelling software (commercial or, those available in public domains) that can interpret SMILES line notation. Once the compounds are generated, their activities can be predicted using the earlier mentioned rule-based activity prediction method [16, 17] standardized for a biological end-point. It may be noted that being a molecular topology-based approach; the activity prediction can be done using the molecular graphs of the compounds where bond multiplicity and hetero-atom factors are not required. For other approaches, explicit 2D and 3D structures may be generated. In this way one can identify those compounds which are predicted active by this method from the large number of structures generated *de novo* from sub-structural information.

2.3. Compound Prioritization

Presumably, one would get a large number of *de novo* generated compounds classified active. Therefore, a scheme may be devised to further screen these classified active compounds and rank them so that one can pick a reasonable number of top-ranking candidates. To do that it is important to consider the details of the ranges where the vertex index values are falling since the activity of a compound is predicted based on the occurrences of vertex index values in different ranges. Two factors may be given special attention - one is the number of vertex index values in an active range (Active Range Length: ARL) and the other one is the number of compounds contributing to form the range (Active Range Weight: ARW). Intuitively too, consideration of a joint effect of these two factors may help prioritize predicted active compounds from the large number of *de novo* generated structures. To do that, we first propose a measure, Active Range Value (ARV), as the algebraic sum of ARL and ARW values and is given by:

$$ARV = (ARL + ARW) \quad (1)$$

Thus, a range larger in length and contributed by more number of compounds contributed in forming the range would have higher ARV value and such a range of higher ARV value may be regarded as a "STRONGER" range compared to those which have lower ARV values. Now, let's assume that M out of N vertices of a molecular graph G (representing a chemical compound) have fallen in different active ranges. If the vertices are denoted by $v_1, v_2, \ldots, v_M$ one would get M number of ARV measures as $ARV(v_1)$, $ARV(v_2), \ldots, ARV(v_M)$. To get a measure of the contribution of the vertices falling in different active ranges (*i.e.*, contribution of activity related vertices) we further propose a Molecular Activity Index (MAI) as:

$$MAI(G) = \sum_{i=1}^{M} ARV(v_i) \quad (2)$$

It may also be noted that while considering the length of an active range and the number of compounds contributing to form the range, some single-value that comes from both active and inactive compounds are considered since they are part of the active range according to the second rule of range selection mentioned earlier.

At the same time, there is a possibility that some of the vertex index values of molecular graph G may fall in inactive ranges too (the second rule for activity prediction) and that may be considered to pose a negative effect on the activity of the compound. For the prediction purpose, therefore, vertices falling in inactive ranges should be considered. Thus in-line with what we have defined for active ranges, let us define IRL (Inactive Range Length) as the number of vertex index values in an inactive range and IRW (Inactive Range Weight) as the number of compounds contributing to form the range. IRV (Inactive Range Value) is then defined as:

$$IRV = (IRL + IRW) \quad (3)$$

Now, let's assume that M' vertices of G viz. $u_1, u_2, \ldots, u_{M'}$ fall in inactive ranges. We, thus, propose a measure, Molecular De-Activity Index (MDI) for G and it is defined as:

$$MDI(G) = \sum_{j=1}^{M'} IRV(u_j) \quad (4)$$

Therefore, by considering a combined effect of *MAI* and *MDI*, one can prioritize the newly generated active compounds and curate some high-ranking compounds for further studies. Thus, to get a measure of combined effect of the vertices falling in active ranges and inactive ranges (if any) and prioritizing (ranking) the compounds according to their activities, we propose a measure, Molecular Priority Score (*MPS*), for *G* and it may be computed using equation (5):

$$MPS(G) = MAI(G) - MDI(G) \quad (5)$$

Understandably, a compound with higher *MPS* value will occupy a higher position in the ranking and may be prioritized for screening purposes. It may be noted that if a molecular graph doesn't have its vertex indices falling in any of the active or inactive ranges, then both *MAI(G)* and *MDI(G)* will become zero resulting in the *MPS(G)* value being zero. Also, clearly, if both *MAI(G)* and *MDI(G)* measures get the same value then *MPS(G)* value will be zero. However, prioritization using *MPS* is not mandatory and one may wish to consider all the predicted active compounds for further studies.

3. RESULTS AND DISCUSSION

In order to investigate the performance of the proposed integrated drug discovery method, we have considered a series of 41 antitubercular Acid Alkyl Ester (AAE) compounds along with their experimentally determined Minimum Inhibitory Concentration (MIC) values [18a, 18b] and three known antitubercular compounds – Isoniazid, Pyrazinamide and Ethionamide. We report here the results obtained for activity prediction using vertex index D^{-4} and the rule based method [16, 17], prioritization of active compounds using MPS and combinatorial generation of structures from substructural information of activity related vertices for the AAE series of compounds. The number of active and inactive compounds in the training set (and test set) for studies with AAE series have been kept balanced for getting unbiased estimates. We also report here the structure generation studies with the three above mentioned antitubercular drugs using relaxed distance distribution algorithm in drug resistance scenario.

3.1. Studies with Acid Alkyl Ester (AAE) Series of Compounds

The 41 compounds of the AAE series have been divided into two groups - active compounds and inactive compounds - using a cut-off MIC value of 3.9 μM. A compound having MIC value ≤ 3.9 is considered active leading to almost equal number of active and inactive molecules (15 and 14, respectively) in the training set of 29 molecules. The remaining 12 compounds, 6 active and 6 inactive, have been kept as a test set. The activity prediction results along with their MPS values for these training set and the test set compounds are given in Table **1**.

At first, the D^{-4} index values have been computed for all the atoms (vertices) from the hydrogen-suppressed molecular graphs of all the 41 AAE compounds (molecular graphs). However, D^{-4} index values computed for the training set compounds have only been arranged in an ascending order to identify active and inactive ranges in the ordering. The ordering of these D^{-4} index values is given as a supplementary material in Supplementary File **1**, along with the details of Compound No. (Atom No.), the Atom Symbol and the Activity Type (+/-) of the molecule from which the D^{-4} indices were obtained. A few sample ranges identified in the ordering have been shown in Table **2**. It may be noted that there are certain regions in the ordering where active or, inactive ranges are not found since those portions don't satisfy the rules for forming a range (method section). It may also be noted that D^{-4} is not a unique substructural descriptor *i.e.*, more than one substructure may have the same D^{-4} index value. However, the same value (*e.g.*, 2.233137 in Table **2**) mostly represent the same or, very similar substructures and it helps bring index values of these substructures coming from the same or, different compounds within an identified range. Both these cases are illustrated in Figures **2** and **3**. The distance distribution in all these cases is (1, 2, 3, 3, 1, 1, 2, 2, 1, 2, 1, 1) and hence yields the same vertex index value of 2.233137 (Serial No. 10, 11, 12 in Table **2**).

Table 1. Experimentally determined MIC values, assigned and predicted activities and Molecular Priority Score (MPS) of 41 acid alkyl ester derivatives divided into 29 training set and 12 test set compounds.

Sr. No.	Compound No. Ref. 18a (18b)	Experimental MIC Value (μM) As in Ref. 18a, 18b	Activity*		MPS** Value
			Assigned	Predicted	
Training Set					
1	23(30)	0.59	+	+	190
2	4(9)	0.73	+	+	178
3	9(14)	0.98	+	+	210
4	22(29)	1.4	+	+	92
5	1(6)	1.6	+	+	174

(Table 1) Contd…

Sr. No.	Compound No. Ref. 18a (18b)	Experimental MIC Value (μM) As in Ref. 18a, 18b	Activity*		MPS** Value
			Assigned	Predicted	
6	24(31)	1.6	+	#	0
7	18(25)	1.7	+	+	200
8	26(33)	2.0	+	+	85
9	33(41)	2.0	+	+	165
10	27(35)	2.5	+	+	190
11	10(15)	3.1	+	+	88
12	8(13)	3.5	+	+	173
13	19(26)	3.7	+	+	51
14	28(36)	3.8	+	+	183
15	6(11)	3.9	+	+	174
16	7(12)	5.3	-	-	-43
17	16(22)	6.3	-	-	-169
18	12(17)	6.5	-	-	-132
19	20(27)	7.5	-	-	-132
20	21(28)	7.8	-	-	-132
21	25(32)	7.9	-	-	0
22	36(44)	8.3	-	-	-108
23	3(8)	11.7	-	-	-59
24	29(37)	15.2	-	-	-97
25	17(23)	17.3	-	-	-44
26	30(38)	>128	-	-	-84
27	31(39)	>128	-	-	-64
28	39(47)	>128	-	-	-56
29	41(49)	>128	-	-	-65
Test Set					
1	14(20)	0.96	+	+	210
2	38(46)	1.1	+	+	27
3	32(40)	1.8	+	+	112
4	11(16)	2.0	+	+	164
5	13(18)	2.3	+	- #	-132
6	5(10)	3.9	+	+	32
7	15(21)	5.4	-	-	-65
8	37(45)	7.3	-	-	-29
9	34(42)	7.9	-	-	-10
10	35(43)	10.9	-	-	-36
11	40(48)	14.2	-	+ #	178
12	2(7)	15.5	-	-	-59

***** (+) means active, (-) means inactive and (#) means incorrect prediction.
Compounds with MIC ≤ 3.9 μM have been considered as active compounds.

Table 2. Examples of active and inactive range formation in the ordering of D^{-4} values taken from Supplementary File 1.

Serial No.	D^{-4} Index Value	Compound No. (Atom No.)	Activity* and Range Type Active Range
1	2.232499	33 (3)	+
2	2.232590	4 (3)	+
3	2.232605	8 (14)	+
4	2.232605	10 (15)	+
5	2.232610	28 (3)	+
6	2.232630	9 (3)	+
7	2.232639	18 (3)	+
8	2.232985	23 (3)	+
9	2.232985	27 (3)	+
10	2.233137	9 (16)	+
11	2.233137	22 (13)	+
12	2.233137	22 (15)	+
13	2.233220	33 (11)	+
14	2.233220	33 (13)	+
15	2.233290	1 (3)	+
			Inactive Range
1	1.158398	29 (24)	-
2	1.161405	7 (19)	-
3	1.161490	31 (9)	-
4	1.161490	31 (16)	-
5	1.161606	30 (9)	-
6	1.161606	30 (10)	-
7	1.161666	12 (18)	-
8	1.161666	20 (18)	-
9	1.161666	21 (18)	-
			Not a Range
1	2.163005	1(11)	+
2	2.163797	9(14)	+
3	2.166911	17(20)	-
4	2.166997	6(14)	+
5	2.167055	30(18)	-
6	2.167172	1(10)	+
7	2.168772	17(19)	-

* (+) sign corresponds to index value coming from active compound.
(-) sign corresponds to index value coming from inactive compound.

Once the active and inactive ranges have been identified in the ordering of the AAE training set compounds, the anti-tubercular activities (active or, inactive) of the training set and the test set compounds are predicted using the prescribed activity prediction rules (method section). For the prediction of activity of a compound, we have considered those vertices of the corresponding molecular graph which have fallen in active and inactive ranges.

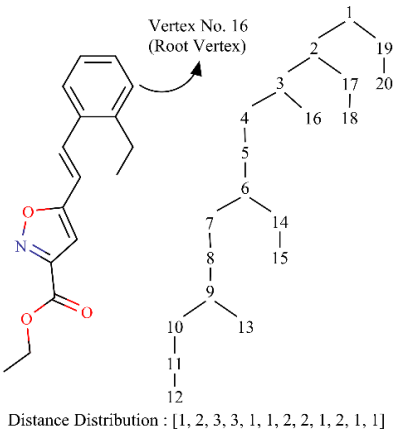

Fig. (2). Compound No. 9 along with the rooted vertex (Vertex No. 16) and the corresponding rooted tree.

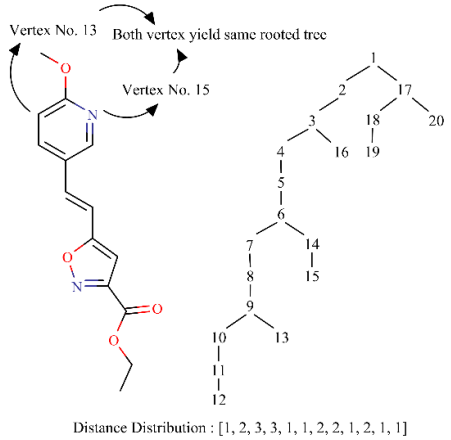

Fig. (3). Compound No. 22 along with the root vertices (Vertex No. 13 and 15) and the corresponding rooted tree.

For illustration purpose, the information about the vertices of compound no. 11 falling in active and inactive ranges (and those not falling in any range) and activity prediction is given in Table **3**. It is seen that 13 out of 22 vertices of compound no. 11 have fallen in the active ranges, 8 of them are not falling in any of the ranges while only one of the vertices is falling in an inactive range. Hence according to the rules formulated for activity prediction

Table 3. Details of vertex along with their range information for activity prediction of Compound No. 11.

Serial No.	D^{-4} value	Falling in the Range	No. of Values in the Range:	
			Active	Inactive
1	2.302931	2.302758 - 2.303630	10	0
2	3.236536	3.236531 - 3.237360	3	0
3	2.232763	2.232499 - 2.233290	15	0
4	2.238963	2.238831 - 2.239319	9	0
5	3.287517	3.287385 - 3.287873	9	0
6	3.232793	3.232685 - 3.232990	9	0
7	2.259333	Not falling in a range	-	-
8	2.263418	Not falling in a range	-	-
9	1.177774	1.177742 - 1.177889	7	0
10	2.224022	2.223990 - 2.224137	7	0
11	2.100442	2.100422 - 2.100514	5	0
12	3.257709	3.249896 - 3.271529	0	7
13	2.302350	Not falling in a range	-	-
14	3.237043	3.236531 - 3.237360	3	0
15	2.295105	Not falling in a range	-	-
16	3.237043	3.236531 - 3.237360	3	0
17	2.302350	Not falling in a range	-	-
18	2.167172	Not falling in a range	-	-
19	2.167172	Not falling in a range	-	-
20	1.102877	1.098767 - 1.104672	5	0
21	1.102877	1.098767 - 1.104672	5	0
22	1.088766	Not falling in a range	-	-

Prediction: Compound No. 11 is ACTIVE.

(method section), this compound is predicted active and this prediction agrees with the observed MIC value of the compound (Table **1**).

It can be seen that successful prediction has been obtained for 28 out of 29 compounds of the training set (96.55%) and 10 out of 12 compounds of the test set (83.33%). Clearly the number (and percentage) of correct predictions for both training set and the test set are high. Thus, the activity prediction system may be considered to be standardized for prediction of the given anti-tubercular activity of unknown compounds. Hence, the prediction of anti-tubercular activities of the combinatorially generated compounds may be done using this standardised system and some predicted active compounds may be curated for further studies related to drug discovery.

To further assess the performance of the activity prediction method [16, 17] for the 41 anti-tubercular AAE compounds, we randomly selected 10,000 combinations of training and test sets, with the number of compounds in training set varying between 26 to 32 molecules from the total compound set. The training set lengths were not kept too high or too low to prevent test set from becoming too small or the training set itself losing significant structural information required for training. Moreover, as the number of iterations was quite large, an increment of 2 was considered over previous number of molecules in the training set. The threshold MIC values are suitably chosen to keep the number of active and inactive compounds in the training sets almost equal to maintain the class balance as much as possible. The validation parameters - Accuracy, Sensitivity and Specificity - were evaluated for the test set of these combinations. The mean of the validation parameters (with their standard error of mean error bars) are shown in Fig. (**4**).

It is seen that the validation parameters show an increasing trend as the training set length is increased and more and more instances are included to represent the data to learn from. It is understandable that the validation metrics will vary greatly based on the training set in consideration and the method will perform poorly when these are not a proper representation of the total set or the unseen data. The molecule details of these 41 compounds have been obtained from the '.MOL' file given in Supplementary File **2** and the Compound No. given in Table **1** and Table **2** and Atom No. in Table **2** are derived and used as depicted in the corresponding MOL files. The third line in the MOL file for each molecule i.e. the comment line has been populated by the MIC value of the corresponding compound. The structures of these 41 AAE compounds are given as a supplementary material in Supplementary File **3** with the same compound numbering as used in Supplementary File **1**.

3.1.1. Identification of Vertex and Structure Generation

Once the system is standardized for activity prediction, an activity related vertex is identified from a strong range for generating structures combinatorially. One of the purposes of the present structure generation work from sub-structural information is to get newly designed bioactive compounds having a diverse topological architecture / scaffold. We report below the results of the structure generation studies

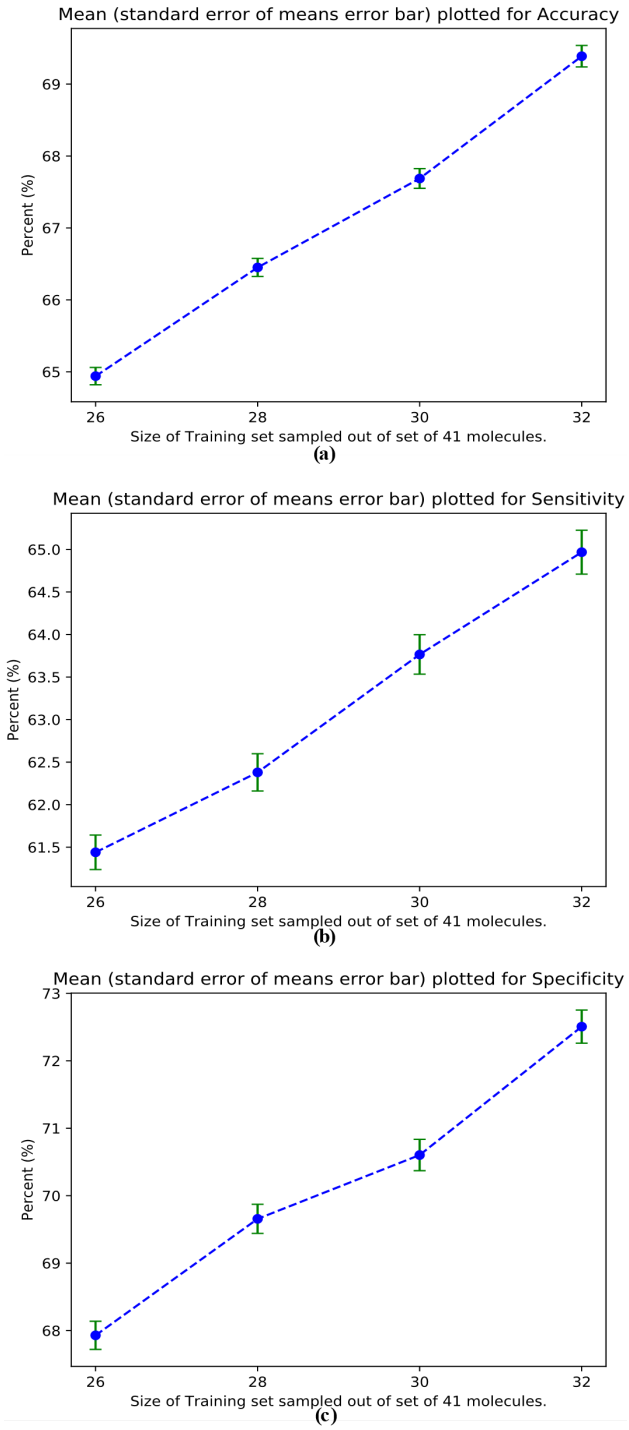

Fig. (4). Plots of (**a**) Accuracy (**b**) Sensitivity (**c**) Specificity against the number of compounds in training set.

for the AAE series of compounds using distance distribution information associated with an activity related vertex identified from a strong active range.

3.1.1.1. Structure Generation Using Distance Distribution Information

As discussed in the method section, the vertex which lies in a "strong" range and belongs to a comparatively highly active compound in the available set of molecules is believed to contain structural information that may be required to

Table 4. Details of the range in which vertex 15, in the molecular graph of compound no. 22, lies in.

Serial No.	D^{-4} index value	Compound No. (Atom No.)	Activity
1	2.232499	33 (3)	+
2	2.232590	4 (3)	+
3	2.232605	8 (14)	+
4	2.232605	10 (15)	+
5	2.232610	28 (3)	+
6	2.232630	9 (3)	+
7	2.232639	18 (3)	+
8	2.232985	23 (3)	+
9	2.232985	27 (3)	+
10	2.233137	9 (16)	+
11	2.233137	22 (13)	+
12	**2.233137**	**22 (15)**	**+**
13	2.233220	33 (11)	+
14	2.233220	33 (13)	+
15	2.233290	1 (3)	+

(+) means active, (-) means inactive.

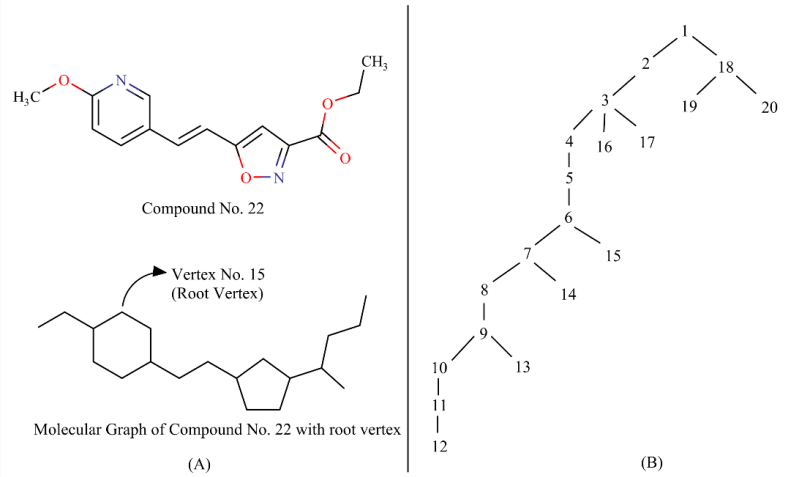

Fig. (5). (A) Compound No. 22, its hydrogen-suppressed molecular graph and the root vertex (vertex no. 15). **(B)** Sample rooted tree structure generated. In the tree, the root vertex is labelled as vertex 1.

make a compound highly active. Pursuant to this, to investigate the possibility of designing a novel bioactive compound through the proposed structure generation techniques we have considered two vertices from "strong" ranges. The compound numbers used here correspond to that given in Table **1**.

For the first structure generation exercise, we have selected vertex No. 15 in the molecular graph representing compound No. 22 for generating structures using the distance distribution associated with the vertex since it belongs to a fairly active compound (*MIC* = 1.4 μM) and falls in a strong active range. The details of this strong active range is given in Table **4**.

The compound No. 22 along with its molecular graph and the chosen structure generation vertex (root vertex) is given in Fig. (**5A**). The distance distribution associated with this vertex (Vertex No. 15) starting with distance 0 is (1, 2, 3, 3, 1, 1, 2, 2, 1, 2, 1, 1).

As described in the method section, a graph theoretical algorithm [14] has been used to generate all possible non-isomorphic rooted trees using the distance distribution. A sample rooted tree is shown in Fig. (**5B**) with the

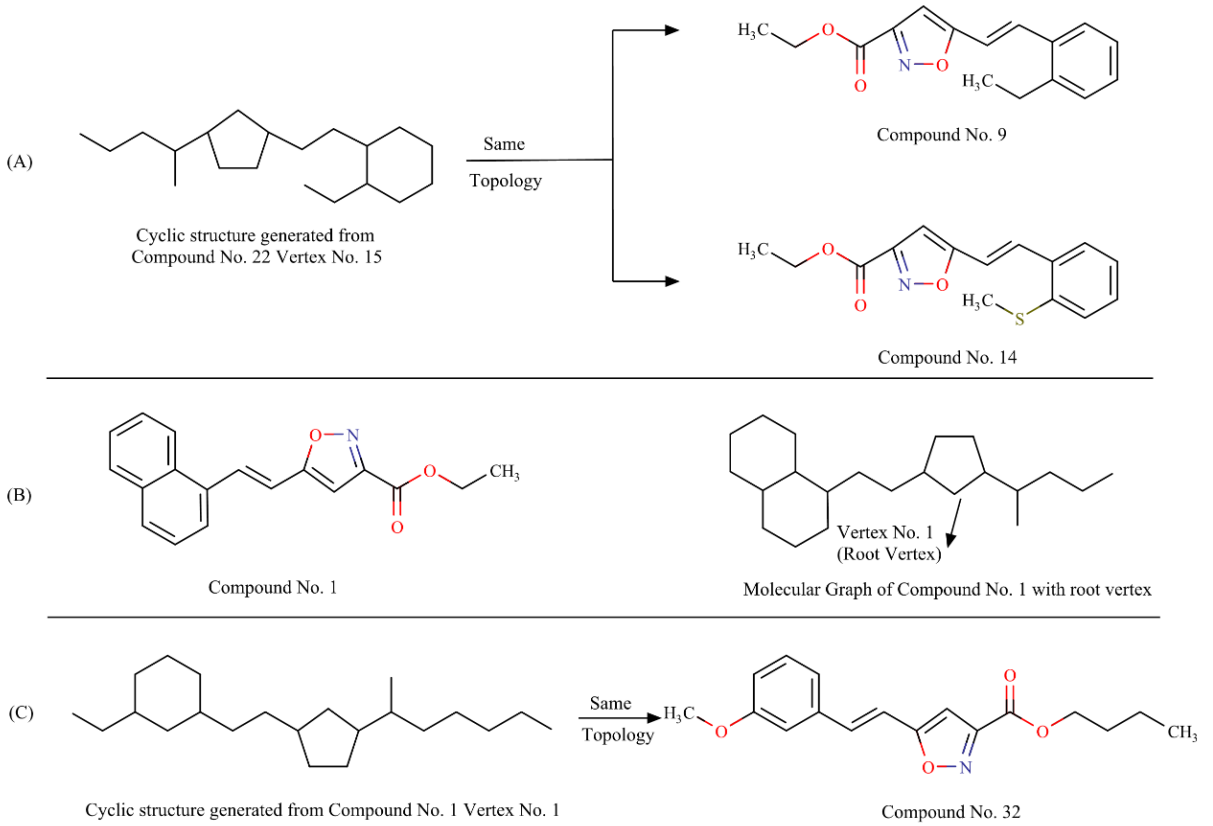

Fig. (6). (A) One of the cyclic structures generated considering vertex 15 of Compound 22 (Table **1**) as root vertex and with same topology as the structures of existing compounds 9 and 14 (Table **1**). **(B)** Compound No. 1, its hydrogen-suppressed molecular graph and the root vertex. **(C)** Cyclic structure generated from Compound 1 Vertex 1 and its topological equivalence to existing molecular structures.

corresponding distance distribution which gets generated. The SMILES notations of the generated rooted trees are given as a supplementary material in Supplementary File **4** where the one corresponding to Fig. (**5B**) is the 3rd SMILES notation. The root vertex about which the distance distribution gets matched is also mentioned alongside.

Once the rooted trees are generated, the computer program generates structures which contain cycles to generate the topology of the structural formula of a variety of chemical compound while still maintaining the distance distribution. At this step, one can use user-defined parameters to decide the number of cycles to be created and the size of the cycles too. In the present study, we have chosen to generate structures containing two cycles, having a number of sides either 5 or 6, to investigate whether we are able to generate any other active compound present in the studied dataset.

The SMILES notations of the structures generated with the criteria are given in Supplementary File **5**. As in the previous case, the root vertex about which the distance distribution gets matched is also mentioned alongside. It has been found that the generated structures contain one such structure (Fig. **6A**) corresponding to the SMILES notation no. 71 in the Supplementary File **5** and the topology of this structure matches with that of compound No. 9 (*MIC* = 0.98) of the training set and compound No. 14 (*MIC* = 0.96) of the test set. Clearly, the method has been able to generate structures of those compounds which are more active from the sub-structural information of a less active compound.

To further substantiate the capability of this method to generate structure of active compounds, we have carried out another structure generation exercise using distance distribution associated with a vertex of a different active compound. This time we have considered vertex No. 1 of compound No. 1 (*MIC* = 1.6) as depicted in Fig. (**6B**) to start structure generation, the range details of which are provided in Table **5**.

In this case too, the topology of one of the generated structures (Fig. **6C**) matches with that of compound No. 32 (*MIC* = 1.8) belonging to the test set. Here also the method has helped generate structure of an almost equally active compound. The ability of the method to generate structure of a test set compounds starting from that of a training set compound seems to indicate that the proposed method may be able to generate many more novel structures of highly active compounds. Additionally, the current structure generation exemplifies the case where an existing ring got disassembled and some of the vertices got moved resulting in a different scaffold.

Subsequently, we have investigated structure generation for several other acid alkyl ester compounds that we have taken for the present study. The details of starting structures and generated structures are given in Table **6**. The terms 'Top position Molecule' refers to the activity position of the

Table 5. Details of the range in which vertex 1, in the molecular graph of compound no. 1, lies in.

Serial No.	D^{-4} Index Value	Compound No. (Atom No.)	Activity
1	2.302758	4 (1)	+
2	2.302777	28 (1)	+
3	2.302797	9 (1)	+
4	2.302807	18 (1)	+
5	2.303152	23 (1)	+
6	2.303152	27 (1)	+
7	2.303194	33 (1)	+
8	**2.303458**	**1 (1)**	**+**
9	2.303569	8 (1)	+
10	2.303630	6 (1)	+

(+) means active, (-) means inactive.

Table 6. The starting and generated structures from the Acid Alkyl Ester (AAE) data set.

Top Position	Starting Structure No.	Generated Structures (Molecule No.)	
Molecule / Atom	(Molecule, Atom)	In training set	In test set
1 / 1	(23, 3)	27	
3 / 1	(9, 3)		14
	(9, 16)	22	14
3 / 7	(9, 2)		14
	(9, 15), (9, 18)	19	14
4 / 1	(22, 13), (22, 15)	9	14
5 / 2	(1, 1)		32

molecule with respect to their sorted activity values *e.g*. Top position Molecule being 3 means the corresponding molecule is the 3rd most active molecule. Similarly, 'Top position Atom' refers to the position of the atom with respect to the length of the active range in which the atom lies *e.g*. Top position Atom being 7 means once a molecule has been considered (of any given position, in this case 3) the given atom in this molecule has the 7th largest length of active range among all the atoms of this molecule. Multiple values on more than one line corresponding to the same Top Position Molecule / Top position Atom represent that these starting structures representing tuples *i.e*. (Molecule No., Atom No.) tie for the same position and either of them or all of them can be used for structure generation as desired. When multiple (Molecule No., Atom No.) pairs are mentioned on the same line then it means that either of these starting structures representing tuples are leading to the same set of generated structures.

It is interesting to note that all the active structures generated by the proposed structure generation technique, given in Table **6**, have also been predicted active by the rule based method used for activity prediction in the present study (Table **1**).

3.2. Studies with Existing Antitubercular Drugs – Drug Resistance Problem

For the present study, we have considered three known antitubercular drugs Isoniazid, Pyrazinamide and Ethionamide to demonstrate the application potential of the relaxed distance distribution algorithm incorporated in the proposed integrated method. Some of these three drugs can be resistant to some TB strains while others may not. This study demonstrates how the structure of a different compound which is not resistant may be obtained from that of a drug which is resistant.

3.2.1. Structure Generation with Relaxed Distance Distribution Constraint

As proposed in the method section, slight relaxation on the distance distribution method, termed as 'strong matching' and 'weak matching', can be used to generate structures with either increased or decreased number of

vertices while trying to maintain the topological properties of the starting structure. To investigate the usefulness of this "relaxed distance" based structure generation algorithm, we have applied this algorithm to address drug discovery problem in drug resistance scenario considering three known TB drugs to find out whether this algorithm can help generate structure of a non-resistant antitubercular drug from a known Multidrug Resistant first line TB drug. Thus, we have generated structures by picking an atom from "Isoniazid" (containing 10 non-hydrogen atoms) a known first line anti-tubercular drug [25] which is also considered as one of the drugs in categorizing Multiple Drug Resistant (MDR) / Extensively Drug Resistant (XDR) tuberculosis (Center for Disease Control and prevention (CDC). https://www.cdc.gov/tb). The vertex picked up as the root / starting vertex for structure generation is known to take part in making "Isoniazid" active [26] and therefore may be regarded as an activity related vertex (substructure) to be considered for structure generation. Important steps of this structure generation exercise are described in Fig. (**7A-C**).

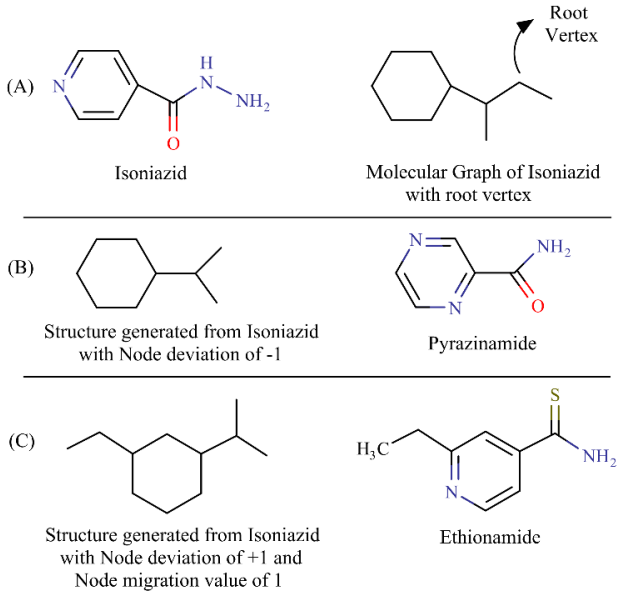

Fig. (7). (A) Isoniazid and its molecular graph with starting (root) vertex for structure generation. **(B)** Structure generated from Isoniazid with decreased node count of 9 and the resembling compound Pyrazinamide. **(C)** Structure generated from Isoniazid with increased node count of 11 and the resembling compound Ethionamide.

In the current study, the deviations considered in the number of vertices in the structures to be generated are -1 and +1 corresponding to 9 and 11 as the vertex count in the structures generated. Moreover, the structure generation was performed with both strong matching and the weak matching approaches with the number of vertex migration allowed in case of weak matching being 1.

Now, starting with the activity related (root) vertex as shown in Fig. (**7A**), one of the structures generated with 9 vertices corresponds to another anti-TB drug, Pyrazinamide [27] and the structure (Fig. **7B**) was obtained with both strong and weak matching approaches.

Similarly, when the structures are generated with 11 vertices with weak matching criteria and the number of vertex migrations allowed is 1, one of the structures thus generated resembles that of another antibiotic, Ethionamide, a potent antitubercular drug [26]. Although Ethionamide has become a resistant TB drug, some research on this compound seems to indicate that this drug's resistance to TB may be reversed [26]. In that case it is possible that Ethionamide will resurface as a resistance free potent antitubercular compound. It is to be noted, however, that this structure cannot be generated with strong matching criteria. The topological equivalence in the structure of Ethionamide and the generated structure are shown in Fig. **7C**.

Therefore, it is apparent from the results obtained for structure generation using "relaxed distance" criteria that this method can add value to *de novo* structure generation exercise in search of potent drug molecules. Although the study performed here has been done for molecules having small and relatively simpler structures, the relaxed matching both in the strong and weak sense shows a promising way of extending the distance distribution based molecule design approach to search for wider range of drug candidates, particularly in drug resistant scenario such as that for MDR / XDR tuberculosis. It may also me noted that we have used hydrogen suppressed graphs for the molecular structures in our activity prediction. Using hydrogen filled graph allows us to explore more topologies and potentially improves the power of the method.

CONCLUSION

The objective of the present study is to develop an integrated method for drug discovery using combinatorial generation of compounds coupled with activity prediction using substructural topological information followed by screening of potential drug candidates. To evaluate the performances of different modules / functionalities of the computer program / proposed method, a series of 41 Acid Alkyl Ester (AAE) derivatives and three known antitubercular drugs – Isoniazid, Pyrazinamide and Ethionamide - have been considered since discovery of potent antitubercular drug is of special interest globally. The proposed method has been found to predict activities of AAE series of compounds with high percentage of success rate and the newly developed MPS values have been found useful for prioritization and screening of active AAE derivatives. The method has also been successfully used to generate structures of active AAE compounds from topological distance information of activity related vertices (substructures) coming from other active compounds of this series. Activity-linked combinatorial structure generation holds a key part of the proposed method since this method is meant for designing / discovering novel drug candidates having diverse scaffold / topological architecture using single substructural (vertex) index. In the process, it also links the method with two key areas of drug discovery research - scaffold hopping and iQSAR approaches. Also, successful standardization of the activity prediction module indicates that it may help screen potential active compound from the combinatorially generated structures.

Therefore, having different algorithms for discovering novel drug molecules, the proposed topological substructural information based activity linked drug discovery method may find useful application in discovering novel antitubercular drug candidates including handling drug resistance problem. As a part of future work, it would be interesting to investigate whether inclusion of some quantitative activity prediction method, ADME/TOX filters and more user defined parameters can help improve the system's capability of discovering novel drug candidates. Also, additional rooted tree generation algorithms may be developed for combinatorial structure generation a. The rooted trees can also be used for searching databases to find new lead compounds. Since the goal is to build a useful tool for the discovery of novel and effective therapeutic candidates to combat tuberculosis, the results obtained so far seems to be quite encouraging for going forward in that direction and may even be helpful for discovering novel therapeutic agents for the treatment of various other diseases as well.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

We would like to thank the Department of Biotechnology (DBT), Government of India, New Delhi, for financial support. All the authors have contributed equally

REFERENCES

[1] Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. model.* **2012**, *52*, 2864-2875.

[2] Hansch, C.; Sammes, P. G.; Taylor, J. B.; Ramsden, C., Eds., *Comprehensive Medicinal Chemistry: Quantitative Drug Design*, Vol. 4; Pergamon Press: New York, **1990**.

[3] Kier, L. B.; Hall, L. H., *Molecular Connectivity in Structure-Activity Analysis*; Research Studies: Chichester, **1986**.

[4] Basak, S.C.; Restrepo, G.; Villaveces, J. L. Eds. *Advances in Mathematical Chemistry and Applications*, 1st Ed.; Vol 1-2 (Revised Edition); Elsevier **2015**.

[5] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.,* **2004**, *3*, 935-949.

[6] Cramer, R. D., Topomer CoMFA: A design methodology for rapid lead optimization, *J. Med. Chem.,* **2003**, *46*, 374-389.

[7] Sun, H.; Tawa, G.; Wallqvist, A., Classification of scaffold-hopping approaches. *Drug Discov. Today,* **2012**, *17*, 310-324.

[8] Prathipati, P.; Ma, N. L.; Keller, T. H., Global bayesian models for the prioritization of antitubercular agents. J. Chem. Inf. Model., 2008, 48, 2362-2370.

[9] Tanwar, J.; Das, S.; Fatima, Z.; Hameed, S., Multidrug resistance: An emerging crisis. *Interdisciplinary Perspectives on Infectious Diseases* **2014**, http://dx.doi.org/10.1155/2014/541340.

[10] Gálvez, J.; García-Domenech, R., On the contribution of molecular topology to drug design and discovery. *Curr. Comput.-Aided Drug Des.*, **2010,** *6*, 252-268.

[11] Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A., MOLGEN 5.0, A molecular structure generator. *Advances in mathematical chemistry and applications* **2014**, 1, 113-138.

[12] Faulon, J.-L.; Bender, A., *Handbook of chemoinformatics algorithms*. CRC press: Boca Raton, **2010**.

[13] Wong, W. W.; Burkowski, F. J., A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem. *J. Cheminf.,* **2009**, *1*, 4.

[14] Beyer, T.; Hedetniemi, S. M., Constant time generation of rooted trees. *SIAM Journal on Computing* **1980**, *9*, 706-712.

[15] Gibbs, N. E., A cycle generation algorithm for finite undirected linear graphs. *Journal of the ACM (JACM)* **1969**, 16, 564-568.

[16] Klopman, G.; Raychaudhury, C., Vertex indexes of molecular graphs in structure-activity relationships: a study of the convulsant-anticonvulsant activity of barbiturates and the carcinogenicity of unsubstituted polycyclic aromatic hydrocarbons. *J. Chem. Inf. Comput. Sci.,* **1990**, *30*, 12-19.

[17] Raychaudhury, C.; Pal, D., Use of vertex index in structure-activity analysis and design of molecules. *Curr. Comput.-Aided Drug Des.,* **2012**, *8*, 128-134.

[18] Raychaudhury, C.; Kandel, D. D.; Pal, D., Role of vertex index in substructure identification and activity prediction: a study on antitubercular activity of a series of acid alkyl ester derivatives. *Croat. Chem. Acta,* **2014**, *87*, 39-47; (b) Pieroni, M.: Lilienkampf, A.; Wan, B.; Wang, Y.; Franzblau, S. G.; Kozikowski, A. P., Synthesis, biological evaluation, and structure-activity relationships for 5-[(E)-2-arylethenyl]-3-isoxazolecarboxylic acid alkyl ester derivatives as valuable antitubercular chemotypes. *J. Med. Chem.*, **2009**, *52*, 6287-6296.

[19] Günther, G., Multidrug-resistant and extensively drug-resistant tuberculosis: A review of current concepts and future challenges. *Clin. Med.,* **2014**, *14*, 279-285.

[20] Raychaudhury, C.; Klopman, G., New Vertex Indices and their Applications in Evaluating Antileukemic Activity of 9-Anilinoacridines and the Activity of 2′, 3′-Dideoxy-Nuclosides Against HIV. *Bull. Soc. Chim. Belg.,* **1990**, *99*, 255-264.

[21] Raychaudhury, C.; Dey, I.; Bag, P.; Biswas, G.; Das, B.; Roy, P.; Banerjee, A., Use of a rule based graph-theoretical system in evaluating the activity of a class of nucleoside analogues against human immunodeficiency virus. *Arzneim.-Forsch. / Drug Res.,* **1993**, *43*, 1122-1125.

[22] Kandel, D. D.; Raychaudhury, C.; Pal, D., Two new atom centered fragment descriptors and scoring function enhance classification of antibacterial activity. *J. Mol. Model.,* **2014**, *20*, 2164.

[23] Moss, G., Extension and revision of the von Baeyer system for naming polycyclic compounds (including bicyclic compounds). *Pure Appl. Chem.,* **1999**, *71*, 513-529.

[24] Weininger, D.; Weininger, A.; Weininger, J. L., SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.,* **1989**, *29*, 97-101.

[25] Zumla, A.; Nahid, P.; Cole, S. T., Advances in the development of new tuberculosis drugs and treatment regimens. *Nat. Rev. Drug Discov.,* **2013**, *12*, 388-404.

[26] Timmins, G. S.; Deretic, V., Mechanisms of action of isoniazid. *Mol. Microbiol.,* **2006**, *62*, 1220-1227.

[27] DrugBank. https://www.drugbank.ca/drugs/DB00339 (Accessed on October 12, 2017).