Developing an Antituberculosis Compounds Database and Data Mining in the Search of a Motif Responsible for the Activity of a Diverse Class of Antituberculosis Agents

Om Prakash and Indira Ghosh*

Bioinformatics Centre, University of Pune, Ganeshkhind Road, Pune 411 007, India

Received April 9, 2005

A novel data mining procedure to look for new antitubercular agents and targets as well as to find a minimum common bioactive substructure (MCBS), has been reported here. The methodology extracts MCBS, both across the diverse chemical classes and within the particular chemical class, known to be present in the various marketed drugs alongside antimycobacterial compounds with known MICs. For this purpose a small in-house database of compounds has been created, for which MICs against *Mycobacterium* are known. The compounds have been collected from literature available on the synthetic compounds, having known MICs against *Mycobacterium tuberculosis*. An elaborate HQSAR (Hologram QSAR) study has been attempted to extract active fragment from a diverse class of compounds, in combination with the clustering technique to select a homogeneous group of compounds having good a profile toward the activity. The 2D pharmacophore (the 2D fragments extracted from HQSAR) has been validated searching the database. It has been found further that this validated 2D pharmacophore could be used for searching the orphan target in *Mycobacterium* effectively.

INTRODUCTION

Tuberculosis (TB) is primarily an illness of the respiratory system and is spread by coughing and sneezing. Each year about 2 million people die from this curable disease, and it is a case of genuine apprehension that there will be 10.2 million new cases by 2005, if the present trend continues.[2] In 1993, the World Health Organization (WHO) declared tuberculosis (TB) as a "global emergency" as the resurgence of TB over the last 15 years has been observed due to the pathogenic synergy with human immunodeficiency virus infection. TB and other atypical mycobacterioses are reported[2] as diseases frequently associated directly with AIDS; human immunodeficiency virus infection significantly increases the risk that new or latent TB infections will progress to active diseases.

Therefore, there is a more-than-pressing need to develop new and effective antituberculosis drugs. Focusing on the existing antituberculosis drug targets for drug development[3] may be of limited value because of potential cross-resistance on the part of *Mycobacterium tuberculosis.* Precisely because of this observed drug-resistance shown by the bacterium, it is imperative to develop smart new drugs that inhibit novel targets, structurally different from those currently known. Within this context, the major goal of this study has been identified as to arrive at some 'minimum common bioactive substructures' (MCBS), responsible for describing a diverse set of compounds, suitable to be labeled as potent candidates for previously unexplored antituberculotic compounds. These are irrespective of the targets they are known to be associated with. In this paper we have used clustering as a data mining technique for the purpose of exploring a large number of chemical structures. To extract the structural information derived from the subfragments across the various clusters, a new technique, namely the global and local information extraction technique (as has been elaborated in the Methods section below) has been developed. This helped us to identify the "hidden information" in the set of homogeneous or heterogeneous compounds studied.

METHODS

Database Creation. A database of 847 compounds with known minimum inhibitory concentration (MICs) against *Mycobacterium tuberculosis* was created, using ISIS Base. Structures of these compounds and respective MIC values have been taken from the published literature from 1996 onward, till March 2003.[4] For the purpose of analysis, the database had been divided into three different classes-actives (MIC<=4 μg/mL), moderately actives (4 μg/mL <MIC<=32 μg/mL), and in-actives (MIC>32 μg/mL) compounds set. With respect to this criterion, there are 232 active compounds, along with 334 moderately actives and 106 inactives; the rest were excluded because of the nonavailability of precise MIC values.

The data about the MIC for the drug sensitive strain of the *Mycobacterium* family, alongside that for the drug resistant (against most commonly used currently marketed drugs) strain of *Mycobacterium tuberculosis*, are included in the database; subjected to availability of data apart from these, cytotoxic activity in MTD50 in VERO cells (μg/mL), drug/prodrug classifications, IC_{50}, selectivity index, strain information, minimum bactericidal concentration, and minimum bacteriostatic concentration (MBC) data were also included.

The substructure searching operation could successfully retrieve most of the known antimycobacterial pharmacophores enlisted by C. E. Barry.[5] The database contains 28

* Corresponding author e-mail: indira@bioinfo.ernet.in.

10.1021/ci050115s CCC: $33.50
Published on Web 08/12/2005

Table 1. Summary of Different Classes of Chemical Compounds in the Database

class	total no.	average MIC (μg/mL)	max-min MIC (μg/mL)	% of active, moderate, inactive
imidazole	122	8.67	0.39−64	67.2, 32, 0.8
pyrazinoic acids	5	8.74	6.2−12.51	0, 100, 0
pyridines	304	29.51	0.05−367.6	30.9, 51.0, 18.1
isonicotinic acid hydrazides	73	13.83	0.05−118.6	46.6, 41.1, 12.3
thioamides	43	9.66	0.15−110.6	41.9, 53.5, 4.7
quinoxalines	22	3.86	0.39−6.25	54.5, 45.5, 0.0
quinolones	19	23.47	0.2−69	52.6, 21.1, 26.3
pyrroles	43	27.92	0.04−250	44.2, 46.5, 9.3

structurally different chemical classes of compounds in all; most of the well-known classes are enlisted in Table 1.

Limitation of the Database. Only 2D structures are included in the database. The exact stereochemical information (namely that for chirality) is included for limited one, subjected to their availability. Owing to the lack of consensus on the strains used by different research groups (with the most frequently occurring strains being H37Rv, CIP 103471, and CNCTC 331/81) MIC values used here are not standard with respect to one strain.

Clustering of the Database. The compounds were subjected to clustering using Distill (Tripos[6]). Distill is a hierarchical clustering tool, which classifies compounds according to their common substructures.[6] It (i) creates a structure based dendogram, with each node of it representing a substructure (grouping all the compounds containing it, sharing the same structural core) and (ii) colors each node of the dendogram according to an average property, typically biological activity (measured here with log [MIC]), which helps to relate the components of a structure (i.e., atoms, bonds, and connectivity) to the biological activity. The colored dendogram illustrates how incremental change in the structure adds or removes from the biological activity as well as indicating how combinations of these changes the biological activity.

The quality of the dendogram generated by Distill depends heavily upon the critical parameters,[7] namely 'time-out', 'target size', and 'minimum number of atom in the MCS'. These were optimized for the analysis of the database. The time-out parameter decides two characteristics of algorithm dynamics, namely (1) the incremental time-steps to find a larger MCS (the maximum common subgraphs program which Distill uses) and (2) agglomeration of more and more singletons under the highest scoring MCS, placing it as the parent node of the concerned cluster (scoring takes into account 'target size' and the 'minimum number of atom in the MCS', the MCS score is calculated as follows: # atoms + # bonds + (# ring bonds * ring bond weight) + (# heteroatoms * hetero weight) + (# of branch atoms * branch atom weight). 'Target grouping size' puts a lower threshold on the minimum number of targeted compounds having an MCS in it.

Three distill experiments using the hierarchical clustering mode (with Tanimoto similarity of 50%) have been used to arrive at the final clustering. A test run was performed on data sets of 672 compounds (the set of compounds exclusive of prodrugs and those not having accurate MIC data), and incremental 'time outs' of 40 s, 45 s, 55 s, and 75 s have been used. The starting 'time out' has been taken greater than (20 + (number of compounds/50)) as suggested in the Distill manual. For selecting natural clusters (the classification in which compounds naturally fall) no clear method yet has been reported in the literature.[7] To choose the best clustering (one that is nearly similar to natural cluster), the comparison of the dendogram generated at various 'time outs' was performed with one generated at no 'time out' (3600 s). The 'minimum number of atoms in MCS' has been increased to 5 atoms to increase the probability of having five- or six-membered rings in our common structures. After optimization, a dendogram generated with a 40-s 'time out' was found suitable for further study. The optimization of parameters had been performed with a small subset of compounds, as the Distill run is a time-consuming step. Following the suggestion by Barnard et al.,[7] the clustering at a 40-s 'time out' and a minimum of 5 atoms in MCS with 4 target groups was taken for further study, on the grounds of their being closest to the natural cluster specifications.

The dendogram thus generated is numbered from left to right as a cluster number and from top to bottom as the level-wise. The most active compounds are represented in blue dendograms, while the red color corresponds to the most inactive ones, Figure 1.

Search for the Active Fragments. To identify patterns in substructural fragments which will be relevant to biological activity, a QSAR technique, HQSAR,[8] was used. Hologram QSAR (HQSAR) relates biological activity to structural molecular composition where molecular composition is described in terms of patterns of substructural fragments. HQSAR eliminates the need for the generation of 3D structures, putative binding conformations, and molecular alignments. It uses the specialized fragment fingerprints (molecular holograms[8]) as predictive variables for biological activity. A molecular hologram is an array, containing counts of molecular fragments and is related to traditional binary 2D fingerprints[9] employed in database searching and molecular diversity techniques.[10]

HQSAR analysis involves three steps: (i) the generation of the substructural fragments for each of the molecules in a data set, (ii) encoding of these fragments in holograms, and (iii) the correlation of the latter with the available biological data.

In the first step (generation of molecular holograms) the input data set consists of the 2D chemical structures along with the respective biological data. The molecular structures are broken down into all possible linear and branched fragments of connected atoms with size varying between the minimum (M) and the maximum (N) number of atoms. Fingerprints were generated for all substructures for all molecules. Each unique fragment in the data set is assigned as a specific large positive integer by means of a cyclic redundancy check (CRC)[11] algorithm. Each of these integers corresponds to a bin in an integer array of fixed length L. Bin occupancies are incremented according to the fragments generated. Thus, all generated fragments are hashed into array bins in the range 1 to L. This array is called a 'molecular hologram', while the bin occupants are denoted as 'descriptor variables'.

In the second step (encoding of fragments) the HQSAR allows fragments to be distinguished based on the atoms, bonds, connections, hydrogens, and chirality parameters. This

Figure 1. Distill dendogram taken for the HQSAR study. The Distill parameters were 40-s time out; 4 target; and 5 MCS. The cluster marked in the green rectangle has been taken for the HQSAR study, and the cluster marked in red has not been taken for the HQSAR study. The cluster marked in red has either too low a population (less than 10 compounds as members) for the HQSAR study or singleton in nature. The color-coding has been done using an average of log [MIC], which has an indirect relationship with activity i.e., the lower the value of log [MIC], the higher the activity will be. The color-code is as follows: red, highest average log [MIC] or least active compounds and blue, lowest average log [MIC] or most active compounds.

step is sensitive to the data set in question. An atom distinction parameter {A} provides the ability to distinguish between fragments based on differences in their elemental types; for example, NH_3 (ammonia), PH_3 (phosphine), and CH_3 (methyl group) are distinguished upon their fragmentation. The bond distinction parameter {B} distinguishes between fragments based on differences in their bond types; for example C−C−H (in ethane) and C═C−H (in ethylene) are considered different. The connection parameter {Co} allows the holograms to retain information about the hybridization states of the atoms in the fragments; for example, in ethylene glycol ($OHCH_2CH_2OH$) the two carbons are sp^3 hybridized, while in acetic acid (CH_3-C(═O)OH) the first carbon is sp^3 hybridized and the second is sp^2 hybridized. The hydrogen parameter {H} provides the ability to distinguish between fragments based on whether hydrogen atoms are included; for example, C_6H_6 (benzene) and C_5H_5N (pyridine) would be considered identical if the distinction between the hydrogen atoms and the nitrogen atom is ignored. The chirality parameter {Ch} enables fragments to be distinguished based on their atomic and bond stereochemistry. Thus, stereochemistry allows cis double bonds to be distinguished from their trans counterparts, and R-enantiomers to be distinguished from S at all chiral centers. The donor and acceptor parameter {DA} initiates the search for donor and acceptor atoms; for example, CH_3C (═O), CH_2OH, CH_3C (═O), and CH_2NH_2 will look more similar when the {DA} flag is turned on.

In the third step (correlation with the biological data) the HQSAR generates a predictive model for each hologram length, with various combinations of the above-mentioned fragment distinction parameters. Two statistical parameters, q^2 ('cross-validation by leave-one-out' procedure) and r^2, were considered to study the quality of the models.

The quality of a model, generated in a HQSAR study, is very much dependent on the parameters of the HQSAR-like hologram length, fragment distinction parameters. But, it has been concluded by previous studies by Willet et al.,[12] that there is no direct correlation between the hologram length and the predictive quality of the HQSAR models. It has been observed too that for a highly diverse compound set the HQSAR is unable to generate good quality models. So, we have used compounds belonging to the selective clusters, to circumvent the limitation just mentioned. Upon being extracted from the clusters, the 'minimum common structures (MCS)' showing promising activity profile (referred to as 'active clusters' from here onward) were fused together in order to obtain a basic scaffold, Figure 2(D). These active clusters have been chosen excluding known drugs, to debar the possibility of biasing the outcome of HQSAR study. Cluster 101, cluster 60, cluster 57, cluster 26, cluster 10, and cluster 9 were considered for the HQSAR studies on the basis that these clusters contain sufficient population of active, moderately active, and inactive compounds besides satisfactory average range of MIC values, Figure 2.

Two types of HQSAR studies had been undertaken, namely (1) cluster-wise analysis, to extract only those 2D fragments from any given homogeneous set/class of compounds, which contribute toward activity. The 2D fragments, which are present in 'active compounds' only, have been considered for further analysis (kindly refer to Figure 2(D) for a summary of the local fragmental contribution). (2) Cluster-independent analysis, attempted to extract unique class independent motifs from the 2D fragments, was

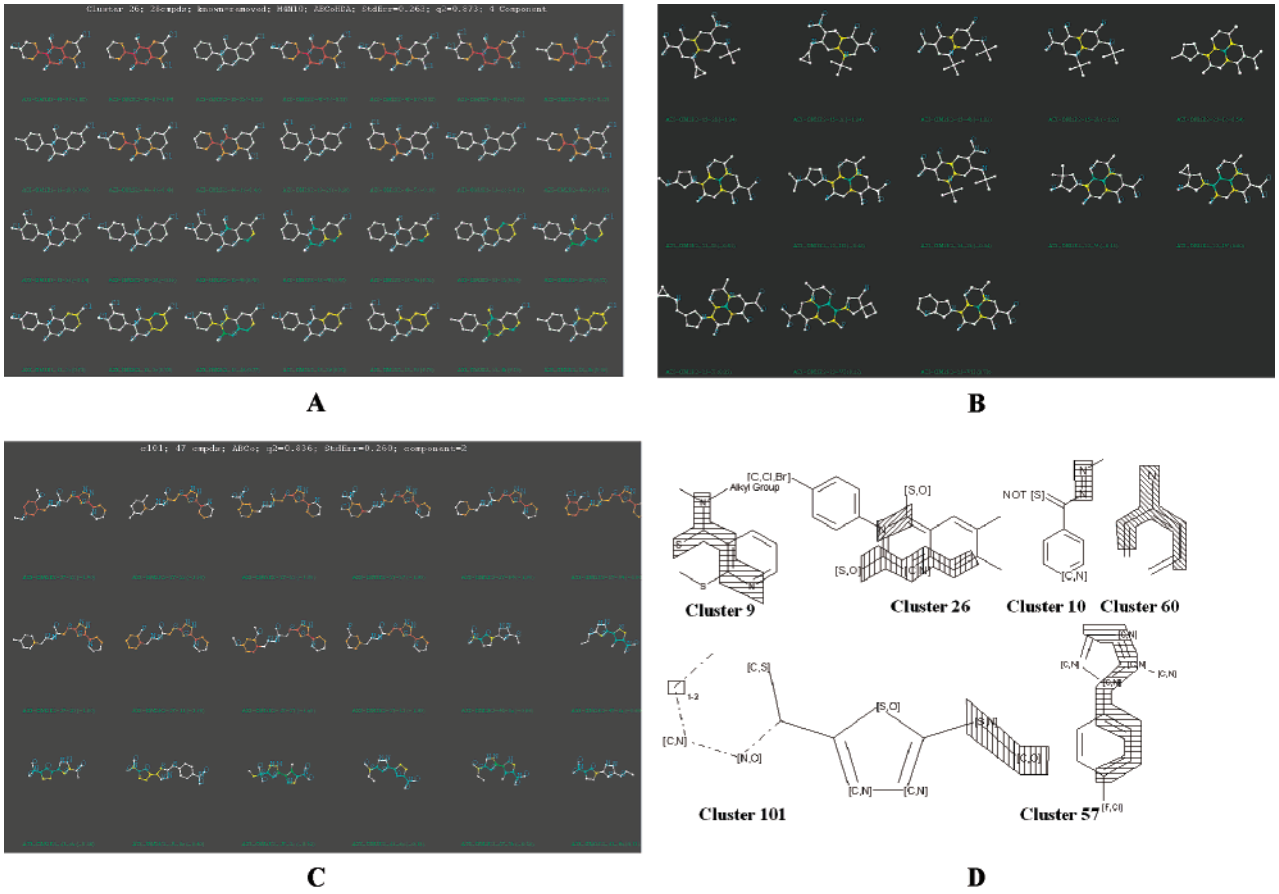

Figure 2. The HQSAR study of the clusters where parts A−C are HQSAR studies for the cluster 26, cluster 60, and cluster 101, respectively. Part D is a summary of the fragmental contribution. The colors at the red end of the spectrum (red, red orange, and orange) reflect poor (or negative) contributions, while colors at the green end (yellow, green blue, and green) reflect favorable (positive) contributions to the biological activity. Atoms with intermediate contributions are colored white. Atoms forming the MCS will be colored cyan.

considered across the different levels of structural similarity. Exploiting the strength of HQSAR (namely its ability to deal with structurally different compounds tolerating a certain degree of unrelatedness), we had tried to do a HQSAR study on a 'single subset of compounds' (102 compounds, called from here onward a "set all") accumulating from cluster 60, cluster 57, cluster 26, and cluster 10. The common motifs from these clusters could be used as combinatorial motifs. The motivation for combining these clusters was not only that the compound sets were from different classes of chemicals but also that presumably they inhibited different targets. This increases the probability of inhibition at different targets. It is well-known that the targets for the compounds such as aminohydrazones, isonicotinoylhydrazones,[13] and isonicotinohydrazides belong to cluster 10 are known as catalase-peroxidase (KatG) and 2-*trans*-enoyl-acyl carrier protein reductase (InhA). Regarding substitution, it has been found that the fluorine and trifluoromethyl substitutions on the benzene rings appeared to be critical for the activity of the compound classes mentioned just above. Similarly, the replacement of the arymethyl moiety (R═H) with the aryethyl one (R═CH_3) in these compounds was known to be not favorable for antituberculosis activity.[14] Cluster 57 containing pyrrole derivatives and azole antifungals were well-known to bind with a high affinity to CYP51 (a soluble protein from *M. tuberculosis* that is similar in sequence to CYP51, lanosterol-14alpha-demethylase) isozymes and are also known to impair the bacterial growth.[15,16] Cluster 26 containing oxazolidinones target bacterial protein synthesis by inhibiting DNA gyrase.[16,17] The other classes of compounds in the same cluster were 2-benzylthiopydidine-4-carbithioamides derivatives whose target proteins were not

Table 2. HQSAR Experiment for the Cluster 57 Has Been Summarized[a]

		known drugs included (18 compounds)			known drugs excluded (15 compounds)		
	M	4	4	4	4	4	4
	N	7	8	9	7	8	9
I	r^2	0.699	0.669	0.719	0.793	0.812	0.811
	StdErr	0.356	0.373	0.344	0.299	0.285	0.286
	q^2	0.538	0.453	0.53	0.68	0.714	0.674
	StdErrCV	0.441	0.48	0.45	0.374	0.352	0.376
	best length	7	37	17	7	7	29
	component	2	2	2	2	2	2
II	r^2	0.56	0.56	0.552	0.736	0.738	0.729
	StdErr	0.414	0.414	0.418	0.324	0.322	0.328
	q^2	0.414	0.421	0.415	0.639	0.647	0.64
	StdErrCV	0.477	0.475	0.477	0.379	0.374	0.378
	best length	101	113	311	101	131	457
	component	1	1	1	1	1	1

[a] I: Model 1 corresponds to ABCo; II: Model 2 corresponds to ABCoDA; q^2: cross-validation by LOO procedure, r^2: correlation, best length: molecular hologram length; component: number of components used to explain the variation in data. The ABCo should be read as the fragment distinction parameter related to atom, bond, and connection is used. Whereas ABCoDA corresponds to atom, bond, connection, and donor−acceptor fragment distinction parameter used.

Table 3. Summarizing the HQSAR Experiments

cluster	HQSAR experiments[a]	statistics	remarks
26	$M = 4$ $N = 10$; ABCoHDA	StdErr = 0.263; $q^2 = 0.873$	The fragment of a 2 to 3 aromatic bond spanning fused ring with nitrogen or carbon has a good contribution.
60	$M = 4$ $N = 8$; ABCo	StdErr = 0.507; $q^2 = 0.701$	The C-10 position of the pyridobenzoxazine nucleus shows that by increasing the lipophilicity of the compounds the contribution to enhancement of the activities. But the lipophilicity of the compounds is not the critical factor.
101	$M = 4$ $N = 7$; ABCo	StdErr = 0.836; $q^2 = 0.260$	The contribution of the fragment is conditional. When a five-membered hetroaromatic ring is attached to another five-membered hetroaromatic ring through a single bond the contribution is positive.

[a] See the legend of Table 2 for the definitions of the parameters written in the column.

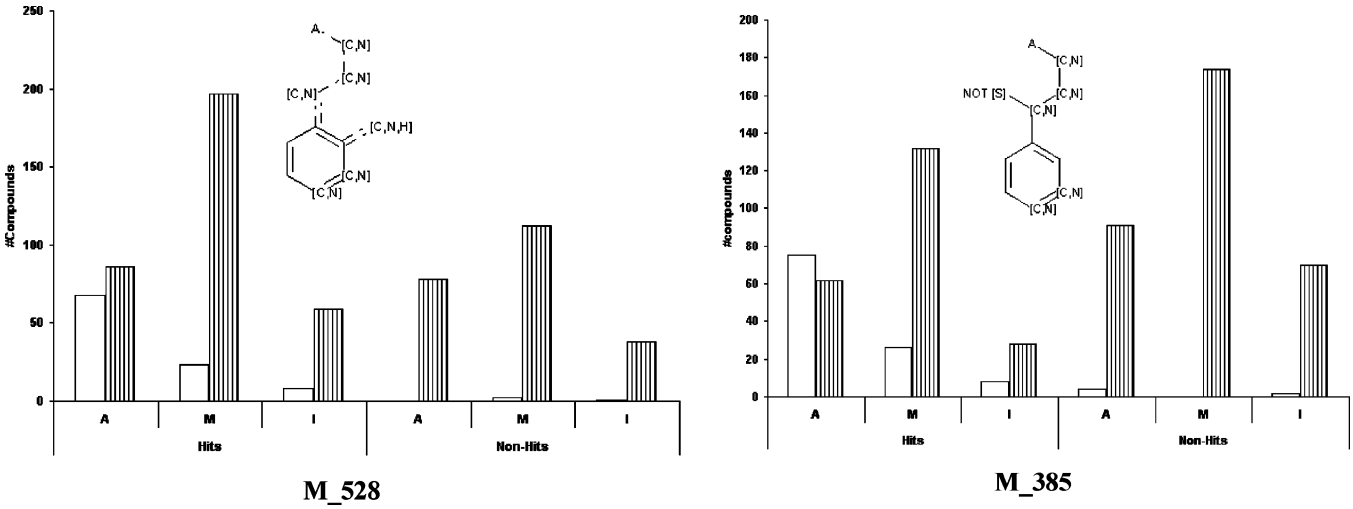

Figure 3. The total number of compounds taken for the HQSAR study = 102 (set all) taken from the clusters 10, 26, 57, and 60 after removal of known drugs, as our training set (called here set all—the nonshaded bar). The remaining compounds (672−102=570) comprise the non-HQSAR set (the shaded bar). Hits = compounds retrieved while database query with the particular motif (say M1), whereas nonhits = compounds not retrieved. Both the hits and nonhits have compounds spanning over active, moderately active, and inactive, from the HQSAR and non-HQSAR set. As can be observed the motifs M_528 and M_385 have a more discriminating nature for both the HQSAR and non-HQSAR sets. For information on other motifs refer to the Supporting Information (Tables S1−S3).

known. Cluster 60 contains the pyridobenzoxaine derivative of levofloxain, which targets the cell wall[14] synthesis, whereas the nitroquinolones target DNA gyrase.

Hence, accumulating this target based information and collectively using the cluster based study could increase the chances of designing inhibitors belonging to different target classes.

While carrying out the statistical analysis, PLS models were selected from cross-validation results (shown in Table 2) on the basis of the first $StdErr_{cv}$-minimum rather than the first q^2-maximum, coupled with an alternative rule-of-thumb, the "5%" rule, can be applied wherein an additional component is permitted only where q^2 is raised by 0.05 units or more.[12] With the assumption that the activity information is encoded in the fingerprints of the structure, the HQSAR study has been performed on the diverse class of antimycobacterial compounds using log [1/MIC] values in order to derive the 2D motif(s) responsible for biological activity.

Table 2 shows how the HQSAR model for cluster 57 was chosen for analysis, based on statistics. Two different runs (i) including and (ii) excluding known drugs were carried out. In the case of model I the best model had $M = 4$ and $N = 7$ with ABCo as it has more than 5% improvement in q^2 over the others and also has the lowest StdErr. While in model II, $M = 4$ and $N = 8$ with ABCo was found to be the best model as it had the lowest StdErr and more than 5% improvement in q^2 over the others, where the additional components were well justified.

RESULTS

Search for Globally Active Subfragments. Although the HQSAR study of clusters, formed by the subset called "set all", was performed in order to extract the common active subfragments from all these compounds, it was observed that most of the fragments contributed for intermediate activity only. The HQSAR study of the compounds derived from the combined subsets with the highest diversity showed that no distinctive active subfragment is present. This could well be explained as the effect of combining diverse classes along with a wide range of biological activity. Failing in this approach, we have tried to combine the MCS information and the subfragmental contribution from the individual clusters to derive a new hypothetical motif.

Search for Locally Active Subfragments. This approach was more like a 'hit and trial' method. At the same level of dendogram, the compound sets were of the same degree of similarity (as expected from the principle of clustering analysis) with important information on the locally active subfragments. These local active fragments proved to be useful while deriving the active hypothetical motif. Keeping

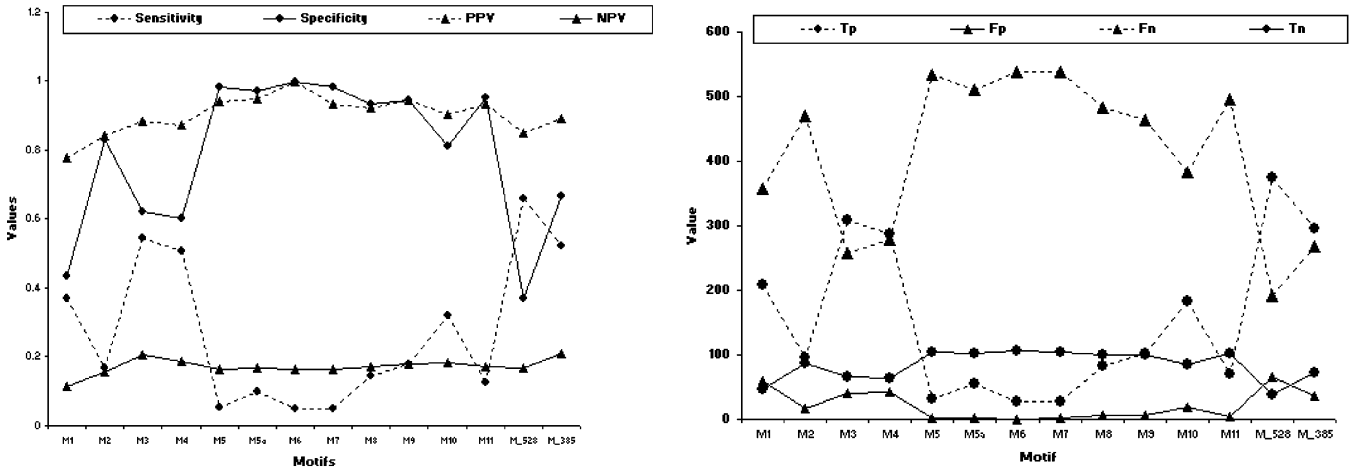

Figure 4. The plot showing the decision statistics. It can be seen that M_385 has quite high values of true positive and lowest false negative and quite low values of false positive and true negatives. All the motifs have the same predictiveness for FN, because we have used only the bioactive motifs and searched the database. Very little information is included for the inactivity of the motifs. Moreover, the M_385 has the highest value of the negative power of prediction and a quite high positive power of prediction. This motif has the highest value of sensitivity as well as, though not highest, a good value for specificity. The M_385 motif along with M_528 and M_6 are able to discriminate the actives from inactives more successfully. But, the number of hits in the case of M_6 is quite low. Also, it has been observed that M_385 is a substructure motif of M_6.

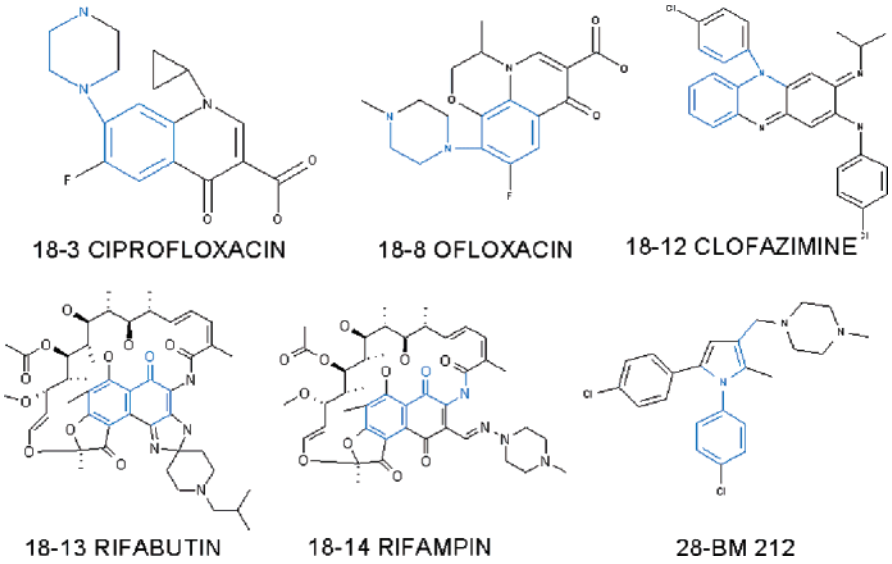

Figure 5. Some of the known drugs/compounds hit by the M_385 motif.

these two approaches in mind we had performed the HQSAR study on the various clusters, and the cluster specific results are presented in Table 3.

These observations showed that there exists a common generic scaffold that could be derived if we merge all the bioactive clusters meaningfully. The MCS information of the clusters 9 and 10 (from level 3); cluster 26 (from level 4); and clusters 57 and 60 (from level 6) had been combined in a chemical sense (Figure 3, also refer to Table S3 in the Supporting Information). The MCS from clusters 9 and 10 had been fused together, resulting in a new motif (say motif M). The possible combinations of individual fragments from the clusters 26, 57, and 60 were then incorporated in motif M to derive a set of new motifs such as M_528 and M_385, as shown in Figure 3. It was found that the limitation of the HQSAR study is in calculating the MCS for highly diverse subsets of compounds. However, for such subsets, inclusion of more similar compounds could improve the prediction.

Validation of the Novel Motifs. The best way to validate these active motifs is considered to be designing and synthesizing all the combinatorial chemical compounds. But this effort is highly time-consuming. The alternative method is to search already known databases using pertinent queries; for this particular case, the two motifs viz., M_528 and M_385 were used. To verify the probabilities for these motifs to be really active antimycobacterials, the substructure search was performed in the original database containing information for both of the known drugs as well as that for the inhibitors of *Mycobacterium tuberculosis*. All the other motifs too were validated on the basis of the quality and decision statistics. The hit rates found are elaborated in Figure 3, and the decision statistics are shown in Figure 4.

On the basis of the high hit rates we have narrowed down our choice of motifs, namely M_385 and M_528. The motifs have been shown in Figure 3. As described, they have been designed by fusing the active subfragments extracted both from level-wise and across the level in the cluster. The success hit rates of motif M_385 have been given in Figure

Table 4. Decision Statistics for Two the Motifs

	M-528 (in %)	M-385 (in %)
sensitivity	44.28	39.72
specificity	70.77	85.85
positive predictive power	81.30	82.78
negative predictive power	30.68	45.42

3. It can be seen from the histogram that the motif M_385 had been able to retrieve most the active compounds (60%) including some well-known drugs as shown in Figure 5.

The success rate of the method could be judged by using sensitivity = TP/(TP + FN) and specificity = TN/(TN + FP) along with positive predictive power or value = TP/(TP + FP) and negative predictive power or value = TN/(FN + TN), and TP = true positive; TN = true negative; FP = false positive; FN = false negative. Whereas TP = (A+M) from hits; FP = (I) from hits; FN = (A+M) from non-hits; TN = (I) from non-hits, whereas A = number of actives; M = number of moderately actives; I = number of inactives.

This study had tried to extract information, which may not be intuitively easy to visualize across the different classes of chemical compounds such as aminohydrazones, isonicotinoylhydrazones, pyrrole oxazolidinones, and pyridobenzoxaines. The hypothetical motifs suggested by this study were successful to retrieve some of the known drugs/active compounds, which were not included in our study. These include compounds referred to in reference 18, which were then used to develop 3D pharmacophores for searching in a database in the study proposed by Botta et al.[18]

CONCLUSIONS

A novel method of extracting the active subfragments, drawn from similar as well as dissimilar classes of chemical compounds, has been presented here using a combination of HQSAR and clustering. Two of the hypothetical motifs have shown better promise to discriminate actives from inactive. Searching the database (reported in this study) using one of the motifs as substructure query, some of the known antimycobacterial drugs, such as rifampin was successfully identified. The bioactive motif suggested by this study contains fuzzy atoms, which could be exploited in the combinatorial synthesis of novel antimycobacterial compounds. Further work is ongoing for expanding the motif and providing experimental evidences.

ACKNOWLEDGMENT

One of the authors, Om Prakash, would like to acknowledge the financial support provided from AstraZeneca Research Foundation, Bangalore, India while computing a part of the present work at their R&D center.

Supporting Information Available: Additional information for Figure 3 (Table S1), the clusters and MCS (Table S2), and the motif derivation, the combination of essential features from Table S2 (Table S3). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Stephen, H. Gillespie. MiniReview: Evolution of Drug Resistance in *Mycobacterium tuberculosis*: Clinical and Molecular Perspective. *Antimicrob Agents Chemother*. **2002**, *46* (2), 267−274.
(2) World Health Organization. The World Health Organization Global *Tuberculosis* Program. http://www.who.int/tb/en/
(3) Chopra, I.; Hesse, L.; O'Neill, A. J. Exploiting current understanding of antibiotic action for discovery of new drugs. *J. Appl. Microbiol*. **2002**, *92* (Suppl.), 4S-15S.
(4) A list of publications from which the database has been created can be made available upon request to the corresponding author.
(5) Barry, C. E.; Slayden, R. A.; Sampson, A. E.; Lee, R. E. Use of Genomics and Combinatorial Chemistry in the Development of New Antimycobacterial Drugs. *Biochem. Pharmacol*. **2000**, *59* (3), 221−231.
(6) Tripos Bookshelf and Technical notes. Tripos Inc. 1699 South Hanley, Road St. Louis, (http://www.tripos.com/sciTech/inSilicoDisc/media/LITCTR/DISTILL.PDF)
(7) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2003**, *18*, 1−40.
(8) Lowis, D. R. HQSAR: A new, Highly Predictive QSAR Technique; Tripos Technical Notes. *Tripos Technol. Notes* **1997**, *1* (5), 1−7.
(9) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representation of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886
(10) Schnur, D. Design and Diversity Analysis of large combinatorial library using cell based methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45
(11) Knuth, D. E. *Sorting and Searching*; Addison Wesley: 1973.
(12) Seel, M.; Turner, D. B.; Willet, P. Effect of parameter variation in HQSAR. *Quant. Struct.−Act. Relat*. **1999**, *18*, 245−252.
(13) Cocco, M. T.; Congiu, C.; Onnis, V.; Pusceddu, M. C.; Schivo, M. L.; De Logu, A. Synthesis and antimycobacterial activity of some isonicotinoylhydrazones. *Eur. J. Med. Chem*. **1999**, *34* (12), 1071−1076.
(14) Yew,W. W.; Piddock, L. J.; Li, M. S.; Lyon, D.; Chan, C. Y.; Cheng, A. F. Antimycobacterial Activities of Novel Levofloxacin Analogues. Antimicrob. *Agents Chemother*. **2000**, *44* (4), 2126−2129.
(15) Khasnobis, S.; Escuyer, V. E.; Chatterjee, D. Emerging therapeutic targets in *tuberculosis*: Post-genomic era. *Expert Opin. Ther. Targets* **2002**, *6* (1), 21−40.
(16) Biava, M.; Fioravanti, R.; Porretta, G. C.; Deidda, D.; Maullu, C.; Pompei, R. New Pyrrole derivatives as antimycobacterial agents analogues of BM212. *Bioorg. Med. Chem. Lett*. **1999**, *9*, 2983−2988.
(17) Zurenko, G. E.; Yagi, B. H.; Schaadt, R. D.; Allison, J. W.; Kilburn, J. O.; Glickman, S. E.; Hutchinson, D. K.; Barbachyn, M. R.; Brickner, S. J. In vitro activities of U-100592 and U-100766, novel oxazolidinone antibacterial agents. *Antimicrob. Agents Chemother*. **1996**, *40* (4), 839−845.
(18) Manetti, F.; Corelli, F.; Biava, M.; Fioravanti, R.; Porretta, G. C.; Botta, M. Building a pharmacophore model for a novel class of antitubercular compounds. *IL Farmco*. **2000**, *55* (6−7), 484−491.

CI050115S