Journal of Computational Methods in Molecular Design, 2014, 4 (4):15-25

Scholars Research Library
(http://scholarsresearchlibrary.com/archive.html)

ISSN : 2231- 3176
CODEN (USA): JCMMDA

QSAR modeling of HEPT compounds: An attempt to anti HIV drug design

Shubhra Sarkar and Sisir Nandi*

Division of Pharmaceutical Chemistry, Global Institute of Pharmaceutical Education and Research (GIPER), Affiliated to Uttarakhand Technical University, Kashipur, India

---

ABSTRACT

A number of quantitative structure-activity relationship (QSAR) models has been developed utilizing theoretical molecular descriptors such as topological, geometrical and functional group indices calculated solely from the structures of 80 set of synthesized 1-(2-hydroxyethoxy-methyl)-6-(phenylthio) thymine (HEPT) derivatives which are non nucleoside reverse transcriptase inhibitors shown to have both potent anti-HIV activity and inhibit HIV-1 at nanomolar concentration. The QSAR models are generated by using multiple linear regression method. Impacts of such computed structural descriptors towards reverse transcriptase inhibitory activities of these compounds were analysed by stepwise forward backward variable selections.The developed QSAR training models are statistically validated. Topological indices can contribute the maximum impact on biological activity obtained in terms of model quality such as R^2 *= 0.903,* Q_{Loo}^2 *= 0.850,* R_{pred}^2 *= 0.620 respectively. Whereas geometrical and functional group descriptors have produced almost similar influences on the activity although external predictability using functional group indices is higher than the geometrical descriptors. Most significant descriptors having crucial influences on the reverse transcriptase inhibition include path/walk 4 - Randic shape index (PW4), topological charge indices and Eigen value from edge adjacency matrix weighted by resonance integrals, information content and total information index on atomic composition, quadrupole x component value / weighted by polarizability (QXXp), Folding degree index (FDI) and sum of geometrical distances between S..Cl at 6-phenyl thio moiety of the HEPT derivatives, number of acceptor atoms for H-bonds (N,O,F), number of un substituted benzene C(sp2) and CHR3 (C-003). The derived significant models in such chemo metric descriptors may be used to design and synthesize new potential compounds in this series and the studies in this direction would focus designing of potent anti HIV-1 HEPT compounds considering the predicted essential structural features predicted by our developed models.*

Key words: HEPT compounds, non nucleoside reverse transcriptase inhibitor, computed structural descriptors, stepwise-MLR, QSAR and antiHIV-1 drug design, prediction of important molecular features.

INTRODUCTION

Human immunodeficiency virus (HIV) is a lenti virus or slowly replicating retrovirus that causes the acquired immunodeficiency syndrome (AIDS) which leads to progressive failure of the human immune system and allows life-threatening opportunistic infections and cancers to thrive. HIV infection occurs due to unprotected sex with the multiple sex partners or infected needles by the transfer of blood, semen, vaginal fluid, pre-ejaculate, or breast milk. Within these bodily fluids, HIV is present as both free virus particles and virus within infected immune cells. HIV infects vital cells in the human immune system such as helper T cells (specifically $CD4^+$ T cells), macrophages and dendritic cells, thus decrease in levels of $CD4^+$ T cells and cell-mediated immunity is lost and the body becomes progressively more susceptible to opportunistic infections [1-3].

HIV causes death of millions of people throughout the world. Hence researchers are concentrating to discover potent anti HIV leads. A lot of experimental and theoretical research has been carried out in this context. 1-(2-hydroxyethoxy-methyl)-6-(phenylthio) thymine derivatives (HEPT) are the first non nucleoside reverse transcriptase inhibitors (NNRTI) shown to have both potent anti-HIV activity and inhibit HIV-1 at nanomolar concentration. These compounds are non nucleoside analogue and thus not necessary for undergoing intracellular phosphorylation. These compounds can directly inhibit the viral reverse transcriptase to stop copying of double-stranded DNA from the template of single-stranded RNA genome of HIV virus. Viral resistance through point mutation and cross resistance is common among different NNRTI. Therefore a comprehensive theoretical study is necessary to design potent active non nucleoside reverse transcriptase inhibitors including HEPT derivatives having least resistance. A number of 6-benzyl analogs of 1-[(2-hydroxyethoxy) methyl]-6 phenylthio)thymine **(**HEPT) compounds were synthesized and evaluated for their anti-HIV-1 activity. Structure-activity relationships were studied wherein a ring structure at the C-6 position of the pyrimidine moiety was predicted an important determinant for the anti-HIV-1 activity [4].

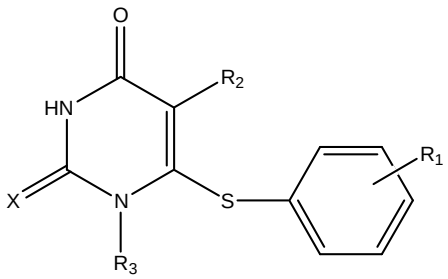
For the designing of potent congeners of HEPT compounds, biochemical mechanisms should be focused. These can be achieved by quantitative structure-activity relationship studies considering theoretical molecular descriptors calculated solely from the structure of HEPT compounds. Although a number of QSAR modeling has been performed using multiple linear regression (MLR), partial least squares (PLS) and Principal Component Analysis (PCA) methods and also non linear methodology by artificial neural network (ANN), but the above mentioned studies were made by considering few number of topological descriptors [5-11]. Thus, an attempt has been made in the present study to predict structural requirements involving biochemical mechanisms of these HEPT compounds utilizing QSAR modeling under the framework of various sets of computed molecular descriptors including topological, geometrical and functional group indices respectively. This may focus to design potent active leads in this congeneric series.

MATERIALS AND METHODS

Biological Activity Data

A number of 80 1-(2-hydroxyethoxy-methyl)-6-(phenylthio) thymine (HEPT) derivatives were taken into consideration in the present study. These compounds act as non-nucleoside reverse transcriptase inhibitors (NNRTI) and produce anti HIV activity by achieving 50% protection of MT-4 cells against the cytopathic effect of HIV-1. The biological activities of these 80 compounds were reported by Luco and Fereti [7]. Biological activity data is given in Table 1.

Table 1: Biological activity data of HEPT compounds



Compound No.	R^1	R^2	R^3	X	Experimental biological activity value
1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15
2*	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85
3*	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72
4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59
5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57
6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92
7*	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35
8*	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48
9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89

10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24
11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00
12*	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47
13*	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09
14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66
15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59
16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89
17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66
18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10
19*	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14
20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00
21	H	$CH_2CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.60
22*	H	Et	$CH_2OCH_2CH_2OH$	S	6.96
23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00
24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.23
25*	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11
26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.30
27	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37
28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92
29*	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47
30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.20
31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89
32*	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.57
33*	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85
34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66
35	H	Me	$CH_2OCH_2CH_2OH$	O	5.15
36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01
37*	H	I	$CH_2OCH_2CH_2OH$	O	5.44
38*	H	$CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.69
39	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.22
40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37
41*	H	CH=CPh	$CH_2OCH_2CH_2OH$	O	6.07
42	H	Me	$CH_2OCH_2CH_2Me$	O	5.06
43*	H	Me	$CH_2OCH_2CH_2Ac$	O	5.17
44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12
45	H	Me	CH_2OCH_2 Me	O	6.48
46*	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82
47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24
48	H	Me	$CH_2OCH_2CH_2F$	O	5.96
49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48
50	H	Me	$CH_2OCH_2CH_2Ph$	O	7.06
51	H	Et	CH_2OCH_2Me	O	7.72
52	H	Et	CH_2OCH_2Me	S	7.58
53*	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24
54*	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30
55	H	Et	CH_2OCH_2Ph	O	8.23
56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55
57*	H	Et	CH_2OCH_2Ph	S	8.09
58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14
59*	H	i-Pr	CH_2OCH_2Me	O	7.99
60	H	i-Pr	CH_2OCH_2Ph	O	8.51
61*	H	i-Pr	CH_2OCH_2Me	S	7.89
62	H	i-Pr	CH_2OCH_2Ph	S	8.14
63	H	Me	CH_2OMe	O	5.68
64	H	Me	CH_2OBu	O	5.33
65	H	Me	Et	O	5.66
66	H	Me	Bu	O	5.92
67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89
68	H	Et	CH_2O-i-Pr	S	6.66
69	H	Et	CH_2O-c-Hex	S	5.79
70	H	Et	CH_2OCH_2 -c-Hex	S	6.45
71*	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11
72*	H	Et	CH_2OCH_2 C_6H_4(4-Cl)	S	7.92
73	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04
74	3,5-Cl_2	Et	CH_2OCH_2Me	S	8.13
75	H	Et	CH_2O-i-Pr	O	6.47

76	H	Et	CH_2O-c-Hex	O	5.40
77	H	Et	CH_2OCH_2-c-Hex	O	7.02
78	H	Et	$CH_2OCH_2CH_2Ph$	O	6.35
79	H	c-Pr	CH_2OCH_2 Me	S	7.266
80	H	c-Pr	CH_2OCH_2Me	O	7.00

** test set molecules*

Computational Details

Structure computation and energy minimization: The structures of 80 HEPT compounds were drawn using 2D Chemdraw. The drawn structures were then converted into 3D modules and the geometries of all compounds were fully optimized using MM2 force field considering the default conversion procedure implemented in Chem3D Ultra [12]. The total energy of a conformation can be calculated using MM2 by the following relation

$E_{total} = E_B + E_A + E_{BA} + E_{OOP} + E_T + E_{vdw} + E_{elec}$,

Where,
E_B = energy of bond stretching
E_A = energy of angle bending
E_{BA} = energy of bond stretching and angle bending
E_{OOP} = out-of-plane bending energy
E_T = torsion energy term
E_{vdw} = van der Waals energy
E_{elec} = electrostatic energy

Structure optimization is performed to obtain most energetically stable geometry of conformations which is being taken into consideration for the calculation of theoretical structural descriptors. [13].

Calculation of structural descriptors: Theoretical molecular descriptors are the numerical representation of molecule, achieved by applying the principles of graph theory to molecular structure. It encodes molecular architecture and quantifies such aspects of molecular structure as size, shape, symmetry, complexity, branching, cyclicity, stereoelectronic character, etc. These significant parameters of chemical structure have been used for the development of specific quantitative structure-activity relationship models for HEPT derivatives considered in the present investigation optimized structures of HEPT compounds were browsed into DRAGON software for the computation of different class of descriptors such as topological, geometrical and functional group respectively [14]. This program can calculate a number of 1664 molecular descriptors including electrostatic, topological, constitutional, geometrical,functional group and physicochemical descriptors. The input file may be created by MDL mol file of the compound concerned. The topological descriptors are the largest set of molecular descriptors and may be subdivided into two classes: topostructural and topochemical descriptors. Topostructural descriptors encode information strictly on the neighborhood and connectivity of atoms within the molecule, while the topochemical descriptors encode information related to both the topology of the molecule and the chemical nature of atoms and bonds within it. The three-dimensional or shape descriptors (3D) are more complex, encoding information about the 3D geometrical aspects of molecular structure. Functional group descriptors represent the contribution of different functional groups upon biological activity of the compounds. Electrostatic descriptors constitute charged polarization, polarity parameter, local dipole index, maximum positive charged, maximum negative charged, total absolute atomic charge, total negative charge, total positive charge. The constitutional descriptors consist of molecular descriptors such as molecular mass, molecular formula, formal charges, fraction of rotatable bonds, and number of rigid bonds, rings, charged groups, and so forth. The physicochemical descriptors include AlogP98 value, AMR value, buffer solubility, polarizability, vapor density, water solubility, solvation free energy, and so forth.

A total number of 1284 molecular descriptors, useful for our purpose, were calculated via DRAGON, and before model development, these were reduced to 898. The reduction in the descriptors was due to keeping a constant value for, or nearly all, of the compounds, and for those that perfectly correlated (r = 1.0) with other descriptors. Table 2 represents the symbols of the calculated molecular descriptors used in our present study along with their corresponding groups. The reduced sets of descriptors were then treated by multiple linear regressions (MLR) algorithm for developing QSAR models.

Table 2: Calculated molecular descriptors

Descriptor classes	Descriptor names
Topological Descriptors	First Zagreb index (ZM1),first Zagreb index by valence vertex degrees(ZM1V), second Zagreb index(ZM2), second Zagreb index by valence vertex degrees(ZM2V), quadratic index(Qindex), Narumi simple topological index (log function)(SNar), Narumi harmonic topological index(HNar), Narumi geometric topological index (GNar), total structure connectivity index(Xt), Pogliani index (Dz), Pogliani index (Ram), polarity number(Pol), log of product of row sums (PRS), log of product of row sums (LPRS),(VDA),mean square distance index (MSD), Schultz Molecular Topological Index (SMTI), Schultz Molecular Topological Index by valence vertex degrees(SMTIV),Gutman Molecular Topological Index (GMTI),Gutman Molecular Topological Index by valence vertex degrees (GMTIV),Xu index(Xu),superpendentic index (SPI),W,WA,Har,Har2, quasi-Wiener index (Kirchhoff number) from Laplace matrix (QW), first Mohar index from Laplace matrix(TI1), second Mohar index from Laplace matrix(TI2), spanning tree number (log function) from Laplace matrix (STN),HyDp,RHyDp, Wiener-like index from topological distance matrix (w),ww,Rww, Wiener-like index from distance/detour matrix (D/D), all-path Wiener index (Wap),WhetZ,Whetv, Whete,Whetp,J,JhetZ,Jhetv,Jhete,Jhetp, maximal electrotopological negative variation (MAXDN), maximal electrotopological positive variation (MAXDP), molecular electrotopological variation (DELS), E-state topological parameter (TIE), Kier symmetry index (S0K), 1-path Kier alpha-modified shape index (S1K), 2-path Kier alpha-modified shape index (S2K), 3-path Kier alpha-modified shape index (S3K), Kier flexibility index (PHI), Kier benzene-likeliness index (BLI), path/walk 2 - Randic shape index(PW2), path/walk 3 - Randic shape index(PW3), path/walk 4 - Randic shape index(PW4), path/walk 5 - Randic shape index(PW5), 2D Petitjean shape index(PJI2), eccentric connectivity index (CSI), eccentricity (ECC), average eccentricity (AECC), eccentric (DECC), mean distance degree deviation (MDDD),unipolarity (UNIP), centralization(CENT), variation (VAR),Balaban centric index (BAC), lopping (Lop), radial centric information index (ICR),D/Dr06, sum of topological distances between N..O(T(N..O)),sum of topological distances between N..S(T(N..S)), sum of topological distances between O..O(T(O..O)), sum of topological distances between O..S(T(O..S)), molecular walk count of order 2 (MWC02), molecular walk count of order 3(MWC03), molecular walk count of order 4(MWC04), molecular walk count of order 5(MWC05), molecular walk count of order 6(MWC06), molecular walk count of order 7(MWC07), molecular walk count of order 8(MWC08), molecular walk count of order 10(MWC10), total walk coun(TWC), SRW01, self-returning walk count of order 2(SRW02), self-returning walk count of order 4(SRW04), self-returning walk count of order 6(SRW06), self-returning walk count of order 7(SRW07), self-returning walk count of order 8(SRW08), self-returning walk count of order 10(SRW10), molecular path count of order 2 (Gordon-Scantlebury index)(MPC02), molecular path count of order 3(MPC03), molecular path count of order 4(MPC04), molecular path count of order 5(MPC05), molecular path count of order 6(MPC06), molecular path count of order 7(MPC07), molecular path count of order 8(MPC08), molecular path count of order 9(MPC09), molecular path count of order 10(MPC10), molecular multiple path count of order 1(piPC01), molecular multiple path count of order 2(piPC02), molecular multiple path count of order3 (piPC03), x molecular multiple path count of order 4(piPC04 molecular multiple path count of order 4), molecular multiple path count of order 5(piPC05), molecular multiple path count of order 6(piPC06), molecular multiple path count of order 7(piPC07), molecular multiple path count of order 8(piPC08), molecular multiple path count of order 9(piPC09), molecular multiple path count of order 10(piPC10), total path count(TPC), conventional bond order ID number(piID), ratio of multiple path count over path coun(PCR), difference between multiple path count and path count(PCD),Randic ID number(CID), Balaban ID number(BID), connectivity index of order 0(X0), connectivity index of order 1 (Randic connectivity index)(X1), connectivity index of order 2(X2), connectivity index of order 3(X3), connectivity index of order 4(X4), connectivity index of order 5(X5), average connectivity index of order 0(X0A), average connectivity index of order 1(X1A), average connectivity index of order 2(X2A), average connectivity index of order 3(X3A), valence connectivity index of order 0(X0v), valence connectivity index of order 1(X1v), valence connectivity index of order 2(X2v), valence connectivity index of order 3(X3v), valence connectivity index of order 4(X4v), valence connectivity index of order 5(X5v), average valence connectivity index of order 0(X0Av), average valence connectivity index of order 1(X1Av), average valence connectivity index of order 2(X2Av), average valence connectivity index of order 3(X3Av), average valence connectivity index of order 4(X4Av), solvation connectivity index of order 0(X0sol), solvation connectivity index of order 1(X1sol), solvation connectivity index of order 2(X2sol), solvation connectivity index of order 3(X3sol), solvation connectivity index of order 4(X4sol), solvation connectivity index of order 5(X5sol), modified Randic index(XMOD), reciprocal distance sum Randic-like index(RDCHI), reciprocal distance sum inverse Randic-like index(RDSQ), information index on molecular size(ISIZ), x total information index on atomic composition(IAC), mean information content on the distance equalit(IDE), x mean information content on the distance magnitude(IDM mean information content on the distance magnitude), mean information content on the distance degree equality(iDDE), mean information content on the distance degree magnitude(IDDM), total information content on the distance equality(IDET), total information content on the distance magnitude(IDMT), mean information content on the vertex degree equality(IVDE), mean information content on the vertex degree magnitude(IVDM), graph vertex complexity index(HVcpx), graph distance complexity index (log function)(HDcpx), Balaban U index(Uindex), Balaban V index(Vindex), Balaban X index(Xindex), Balaban Y index(Yindex), Information Content index (neighborhood symmetry of 0-order)(IC0), Total Information Content index (neighborhood symmetry of 0-order)(TIC0), Structural Information Content index (neighborhood symmetry of 0-order)(SIC0), Complementary Information Content index (neighborhood symmetry of 0-order)(CIC0), Bond Information Content index (neighborhood symmetry of 0-order)(BIC0), Information Content index (neighborhood symmetry of 1-order)(IC1), Total Information Content index (neighborhood symmetry of 1-order)(TIC1), Structural Information Content index (neighborhood symmetry of 1-order)(SIC1), Complementary Information Content index (neighborhood symmetry of 1-order)(CIC1), Bond Information Content index (neighborhood symmetry of 1-order)(BIC1), Information Content index (neighborhood symmetry of 2-order)(IC2), Total Information Content index (neighborhood symmetry of 2-order)(TIC2), Structural Information Content index (neighborhood symmetry of 2-order)(SIC2), Complementary Information Content index (neighborhood symmetry of 1-order)(CIC1), Bond Information Content index (neighborhood symmetry of 2-order)(BIC2), Information Content index (neighborhood symmetry of 3-order)(IC3), Total Information Content index (neighborhood symmetry of 4-order)(TIC4), Structural Information Content index (neighborhood symmetry of 3-order)(SIC3), Complementary Information Content index (neighborhood symmetry of 3-order)(CIC3), Bond

	Information Content index (neighborhood symmetry of 3-order)(BIC3), Information Content index (neighborhood symmetry of 4-order)(IC4), Total Information Content index (neighborhood symmetry of 4-order)(TIC4), Structural Information Content index (neighborhood symmetry of 4-order)(SIC4), Complementary Information Content index (neighborhood symmetry of 4-order)(CIC4), Bond Information Content index (neighborhood symmetry of 4-order)(BIC4), Information Content index (neighborhood symmetry of 5-order)(IC5), Total Information Content index (neighborhood symmetry of 5-order)(TIC5), Structural Information Content index (neighborhood symmetry of 5-order)(SIC5), Complementary Information Content index (neighborhood symmetry of 5-order)(CIC5), Bond Information Content index (neighborhood symmetry of 5-order)(BIC5),
Geometrical descriptors	Gravitational index G1(G1), gravitational index G2 (bond-restricted)(G2), radius of gyration (mass weighted)(RGyr), span R(SPAN), average span R(SPAM), molecular eccentricity(MEcc), spherosity(SPH), asphericity(ASP), 3D Petitjean shape index(PJI3), length-to-breadth ratio by WHIM(L/Bw), Folding degree index (FDI), Harmonic Oscillator Model of Aromaticity index(HOMA), ring complexity index(RCI), aromaticity index(AROM), HOMA total(HOMT), displacement value / weighted by mass(DISPm), quadrupole x-component value / weighted by mass(QXXm), quadrupole y-component value / weighted by mass(QYYm), quadrupole z-component value / weighted by mass(QZZm), displacement value / weighted by van der Waals volume(DISPv), quadrupole x-component value / weighted by van der Waals volume(QXXv), quadrupole y-component value / weighted by van der Waals volume(QYYv), quadrupole z-component value / weighted by van der Waals volume(QZZv), displacement value / weighted by Sanderson electronegativity(DISPe), quadrupole x-component value / weighted by Sanderson electronegativity(QXXe), quadrupole y-component value / weighted by Sanderson electronegativity(QYYe), quadrupole z-component value / weighted by Sanderson electronegativity(QZZe), displacement value / weighted by polarizability(DISPp), quadrupole x-component value / weighted by polarizability(QXXp), quadrupole y-component value / weighted by polarizability(QYYp), quadrupole z-component value / weighted by polarizability(QZZp), sum of geometrical distances between N..N(G(N..N)), sum of geometrical distances between N..O(G(N..O)), sum of geometrical distances between N..S(G(N..S)), sum of geometrical distances between N..F(G(N..F)), sum of geometrical distances between N..Cl(G(N..Cl)), sum of geometrical distances between N..I((N..I)), sum of geometrical distances between O..O(G(O..O)), sum of geometrical distances between O..S(G(O..S)), sum of geometrical distances between O..F(G(O..F)), sum of geometrical distances between O..Cl(G(O..Cl)),sum of geometrical distances between O..I(G(O..I)), sum of geometrical distances between S..S(G(S..S)),sum of geometrical distances between S..F(G(S..F)), sum of geometrical distances between S..Cl(G(S..Cl)), sum of geometrical distances between S..I(G(S..I)),x sum of geometrical distances between Cl..Cl(G(Cl..Cl)
Functional Group descriptors	number of terminal primary C(sp3)(nCp), number of total secondary C(sp3)(nCs), number of total tertiary C(sp3)(nCt), number of ring secondary C(sp3)(nCrs), number of ring tertiary C(sp3)(nCrt), number of aromatic C(sp2)(nCar), number of unsubstituted benzene C(sp2)(nCbH), - number of substituted benzene C(sp2)(nCb), number of non-aromatic conjugated C(sp2)(nCconj), number of terminal primary C(sp2)(nR=Cp), number of aliphatic secondary C(sp2)(nR=Cs), number of aliphatic tertiary C(sp2)(nR=Ct), number of esters (aromatic)(nArCOOR), number of positively charged N(nN+), number of nitro groups (aromatic)(nArNO2), number of hydroxyl group(nROH), number of ethers (aliphatic)(nROR), number of ethers (aromatic)(nArOR), number of CH2RX(nCH2RX), number of X on aromatic ring(nArX), number of donor atoms for H-bonds (N and O)(nHDon), number of acceptor atoms for H-bonds (N,O,F)(nHAcc), CH3R / CH(C-001), CH2R2(C-002), CHR3(C-003), CH3X(C-005), CH2RX(v)(C-006), CHR2X(C-008), =CH2(C-015), =CHR(C-016), =CR2(C-017), R--CH—R(C-024), R--CR—R(C-025), R--CX—R(C-026), R-C(=X)-X / R-C#X / X=C=X(C-040), H attached to C0(sp3) no X attached to next C(H-046), H attached to C1(sp3)/C0(sp2)(H-047), H attached to C2(sp3)/C1(sp2)/C0(sp)(H-048), H attached to heteroatom(H-050), H attached to alpha-C(H-051), H attached to C0(sp3) with 1X attached to next C(H-052), alcohol(O-056), #NOME?(O-058), Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X(O-060), O—(O-061),Cl attached to C1(sp2)(Cl-089), I attached to C1(sp2)(I-099), R=S(S-108).

Linear statistical modelling by Stepwise-MLR: For QSAR modeling of chemical compounds, it is necessary to consider a large number of physicochemical parameters as well as other calculated theoretical molecular descriptors such as constitutional, geometrical, electrostatic, and topological descriptors. In the present study, we have used a very large number of different types of topological as well as various geometrical and functional group physicochemical descriptors to develop QSAR of HEPT compounds. Multivariate regression analysis (MRA), one of the oldest data reduction methodologies, continues to be widely used in QSAR [15], as it does not impose any restriction on the type and number of graphical invariants used in structure–property–activity studies. For a valid statistical significance of the MRA, it is necessary to restrict the maximal number of descriptors, which depends on the number of compounds investigated [16-17] To avoid ambiguities in the interpretation of regression, only a few parameters, or ideally a single parameter, may be used. Consideration of theoretical molecular descriptors such as constitutional, geometric, functional group and electrostatic, and topological descriptors have found wide applications in quantitative structure–activity relationship modeling [18-20]. To establish such a relationship between activity and structural descriptors of the HEPT compounds under consideration, it is essential to develop a regression or an input–output model. Conventional regression (OLS) cannot be used when the number of molecular descriptors exceeds the number of observations [21]. When the number of theoretical structural descriptors greatly exceeds the number of compounds feature selection is essential to screen the significant descriptors supposed to produce considerable impact on the biological activity. In the present study stepwise forward-backward based feature selection method incorporated in Minitab software [22] has been used for the QSAR model generation. One has to choose the values of the F statistic for the partial F tests that will determine if a variable is to enter or be removed from the model; F = 4.0 has been chosen as the threshold for inclusion and exclusion of variables in the present statistical calculation.

The stepwise forward-backward based feature selection method begins with no candidate variables in the model. Predictor variables are then checked one at a time using the partial correlation coefficient (equivalently F to enter) as a measure of importance in predicting the dependent variable. At each stage the variable with the highest significant partial correlation coefficient (F to enter) is added to the model. Once this has been done the partial F statistic (F to remove) is computed for all variables present in the model to check if any of the variables previously added can now be deleted. This procedure is continued until no further variables can be added or deleted from the model. The partial correlation coefficient for a given variable is the correlation between the given variable and the response when the present independent variables in the equation are held fixed. It is also the correlation between the given variable and the residuals computed from fitting an equation with the present independent variables in the equation. After variable selection, MLR is used to derive a number of training QSAR models using different types of descriptor including topological, geometrical and functional group indices [23]. The developed models should undergo for statistical validation prior to applications.

RESULTS AND DISCUSSION

Analysis of QSARs

For proper validation of these training models, the total HEPT compound data set is divided into training and test sets. 70% of the 80 molecules are considered as training set to build QSAR models while remaining 30% is taken as test set. The division is done by random selection. Test set molecules are indicated by asterisk given in Table 1. The quality of each model is denoted by R^2 (R is the square root of multiple R-square for regression), Q_{Loo}^2 (Leave one out cross-validated r^2) values for the training set and an external validation was performed by calculating predictive R^2 (R_{pred}^2). R^2 and Q^2 of a model can be obtained from:

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum (Y_{obs} - \bar{Y})^2}$$

R^2 is a measure of explained variance. Each additional X variable added to a model increases R^2.

Calculation of Q_{Loo}^2 (Leave one out cross-validated r^2) is called as internal validation.

$$Q^2_{Loo} = 1 - \frac{\sum (Y_{obs} - Y_{pred})^2}{\sum (Y_{obs} - \bar{Y})^2}$$

where, Y_{obs} and Y_{pred} indicate observed and predicted activity values respectively and $\overline{Y}$ indicates mean activity value. A model is considered acceptable when the value of Q^2 exceeds 0.5.

External validation or predictability of the models are performed by calculating predictive R^2 (R_{pred}^2).

$$R^2_{Pred} = 1 - \frac{\sum (Y_{pred(Test)} - Y_{(Test)})^2}{\sum (Y_{(Test)} - \bar{Y}_{training})^2}$$

where, $Y_{pred(test)}$ and $Y_{(test)}$ indicate predicted and observed activity values respectively of the test set compounds and $\overline{Y}_{training}$ indicates mean of observed activity values of the training set. For a predictive QSAR model, the value of R^2_{pred} should be more than 0.5 [24].

The validated training QSAR models framed by different types of descriptors alongwith statistical parameters related to describe the quality of models are listed in Table 3.SE represents standard error and PRESS indicate predicted sum of square deviation.

Table 3: Different QSARs alongwith model quality parameters

Model number	Descriptor type	Model Equation	Statistical parameters related to quality of the model				
			R^2	SE	PRESS	Q_{Loo}^2	R^2_{pred}
1	Topological	pIC_{50} = 25.42 + (88) × path/walk 4 - Randic shape index(PW4) + (4.70) × Eigen value 06 from edge adjacency matrix weighted by resonance integrals (EEig06r) + (-1.62) × Information content index neighborhood symmetry of 3 order (IC3) +(-4.0) × Eigenvector coefficient sum from adjacency matrix (VEA1) + (6.6) × topological charge index of order 7(GGI7) + (-285) × mean topological charge index of order 4 (JGI4) + (9.2) × topological charge index of order 9(GGI9) + (-0.193) × total information index on atomic composition (IAC) + (0.150)× shape profile no. 18(SP18) + (-2.76) × Eigen value 15 from edge adjacency matrix weighted by resonance integrals (EEig15r)	0.903	0.426	0.127	0.850	0.620
2	Geometrical	pIC_{50} = 4.308 + (0.0191) × quadrupole x component value / weighted by polarizability (QXXp) + (20.8) × Folding degree index (FDI) + (0.072) × sum of geometrical distances between S..Cl(G(S..Cl)) + (-0.090) × displacement value / weighted by van der Waals volume (DISPv) + (-19.8) × molecular eccentricity (MEcc)	0.650	0.764	0.360	0.569	0.567
3	Functional Group	pIC_{50} = 8.22 + (- 0.69) × number of acceptor atoms for H-bonds (N,O,F) (nHAcc) + (0.61) × number of un substituted benzene C(sp2) (nCb-) + (1.14) × CHR3 (C-003)	0.588	0.860	0.430	0.583	0.621

The above models are then used to predict the biological activities of the test molecules, as indicated in Table 4.

Table 4: Predicted activities of the test set compounds using three developed training QSAR models

Test set	Experimental activity	Predicted activity (using model 1)	Predicted activity (using model 2)	Predicted activity (using model 3)
2	3.85	6.103	5.342	3.920
3	4.72	5.005	5.475	4.610
7	4.35	2.343	6.092	3.230
9	4.89	5.138	5.562	5.310
12	4.47	4.299	4.295	3.922
13	4.09	5.482	5.607	4.611
19	5.14	4.156	6.051	4.612
22	6.96	6.176	6.745	5.383
25	8.11	7.381	7.698	6.611
29	5.47	5.800	6.883	4.692
32	8.57	8.549	7.496	7.052
33	7.85	6.618	6.226	5.912
37	5.44	5.650	5.210	4.694
38	5.69	6.936	4.764	4.692
41	6.07	4.882	7.410	5.911
43	5.17	5.211	6.363	4.691
46	5.82	5.407	5.565	5.384
53	8.24	8.908	8.479	6.621
54	8.3	8.679	7.676	7.291
57	8.09	7.550	7.591	6.685
59	7.99	7.236	6.098	6.520
61	7.89	7.052	6.971	7.210
71	7.11	6.982	6.547	7.291
72	7.92	6.538	8.219	7.292

The plot of observed versus predicted activities for the test compounds is represented in figures 1-3. From the Table 4 it is evident that the predicted activities of all the compounds in the test set are in good agreement with their corresponding experimental activities and optimal fit is obtained generated by the QSARs utilizing different set of topological, geometrical and functional group descriptors. The square correlation coefficients between experimental vs predicted activities of the test set molecules calculated using QSAR models 1, 2 and 3 are 0.656, 0.573 and 0.831 respectively.

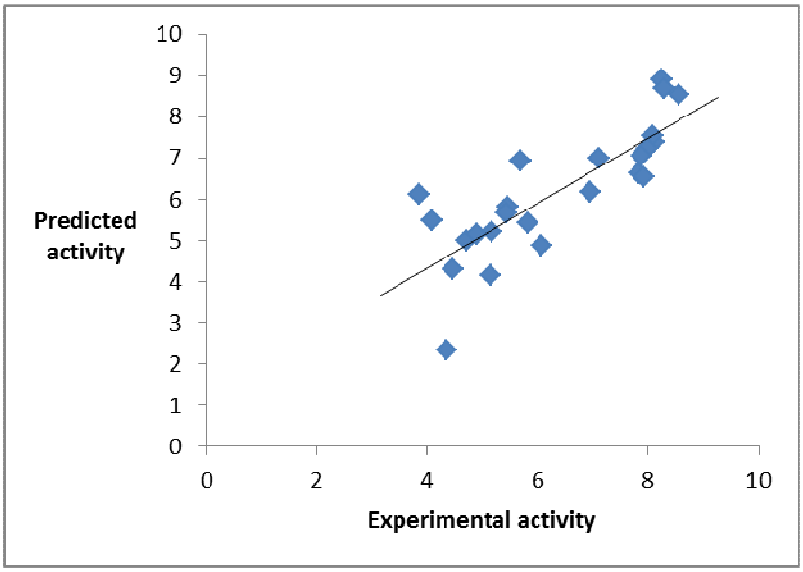


Figure 1: Experimental vs predicted activity of the test molecules (using model 1)

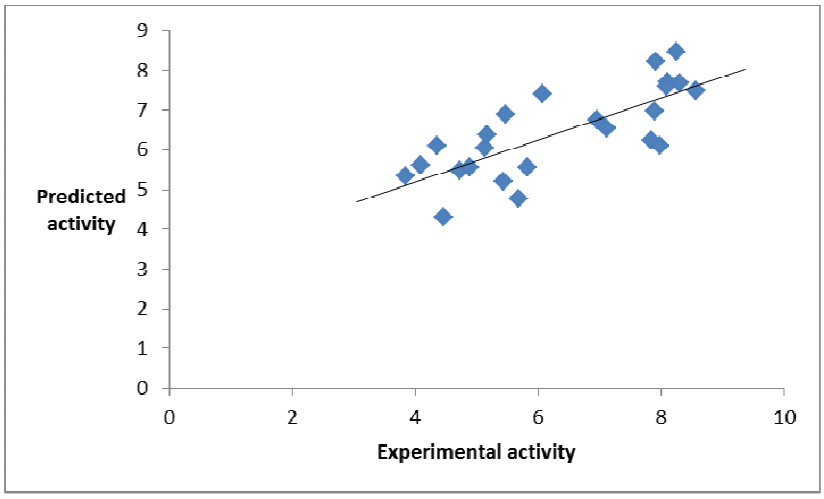


Figure 2: Experimental vs predicted activity of the test molecules (using model 2)

Role of topological indices to predict anti HIV-1 activity

The topological indices contribute maximum influence followed by geometrical and functional group descriptors on the reverse transcriptase inhibitory activities. Topological descriptors can explain and predict 90% and 85% of variances of the reverse transcriptase inhibitory activities of the studied compounds. This model can also produce 62% external predictability.

The above QSAR modeling states that the topological indices play a significant role on the reverse transcriptase inhibitory activities of HEPT compounds. Significant influence is produced by the presence of path/walk 4 - Randic shape index (PW4) descriptor which denotes shape and branching of the compounds which may further indicate steric hindrance while increasing the branching of the molecule. Branching of these ligand may reduce the biological activity. Shape characteristics of these compounds are the most important variable which has also predicted according to Luco et al. [7].

Topological charge indices and Eigen value from edge adjacency matrix weighted by resonance integrals are responsible for contributing charge and resonating effects of these molecules. Charge distribution and resonating effect may produce different electronic structures with similar internal energy of the concerned molecule. The

specific resonating hybrid structure with higher electron dencity interacts with the reverse transcriptase protein and thus enhances the binding affinity of these ligands to the reverse transcriptase protein to stop DNA replication. Information content and total information index on atomic composition are also predicted as significant modeled parameters in this class of descriptors.

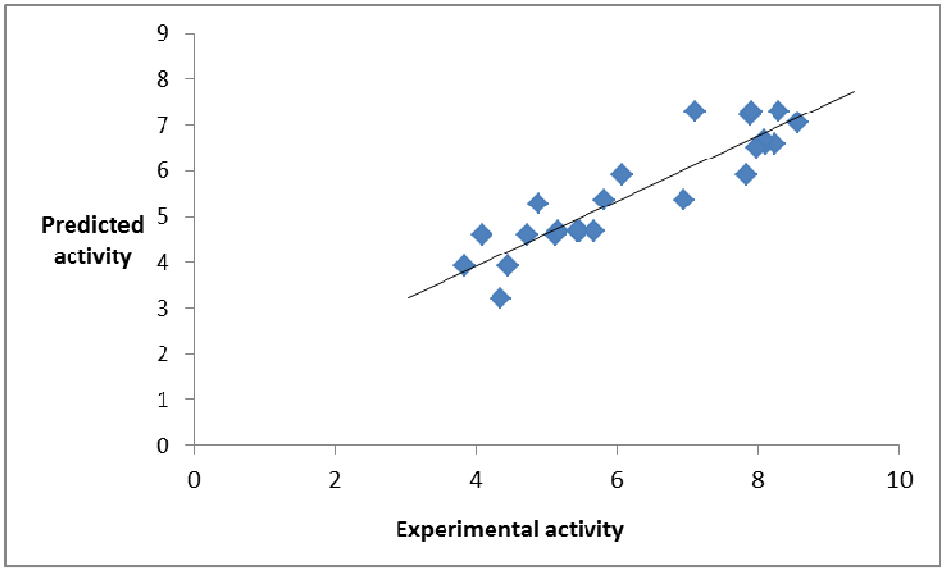


Figure 3: Experimental vs predicted activity of the test molecules (using model 3)

Role of geometrical and functional group descriptors to predict anti HIV-1 activity

Geometrical descriptors can explain and predict 65% and 57% of variances of the reverse transcriptase inhibitory activities of the studied compounds. This model can also produce 57% of external predictability whereas functional group descriptors can explain 59% of the variances and can produce 58% of the internal and 62% of the external predictability, respectively.

Geometrical descriptor based model clears that the descriptors with positive coefficients including QXXp, FDI and G(S..Cl) contribute positive impact on reverse transcriptase inhibition. Quadrupole x component value / weighted by polarizability (QXXp) can produce electric polarizability towards x – direction which is the relative tendency of a charge distribution, like the electron cloud of an atom or molecule. Folding degree index (FDI) is based on graph spectral moments of a matrix representing the dihedral angles of a protein backbone and thus increase the affinity of HEPT ligand for binding with the protein. Increasing the sum of geometrical distances between S..Cl at 6-phenyl thio moiety of the HEPT derivatives may increase in the number of loan pairs in the molecule and thus enhance the ligand inhibitory activities. Decrease in the values of displacement value / weighted by van der Waals volume (DISPv) and molecular eccentricity (MEcc) may enhance anti HIV-1 activities of these compounds.

The QSAR model developed by using functional group descriptors contains significant parameters such as number of acceptor atoms for H-bonds (N,O,F), number of un substituted benzene C(sp2) and CHR3 (C-003) respectively. Presence of more hetero atoms such as N, O, F and S in the ligand may produce more hydrogen bond interaction with the receptor, thus it may increase the compound's duration of action which is very crucial for producing pharmacological activities of these congeners. Number of un substituted benzene C (sp2) may produce aromaticity and polarizability of the compound. Increase in aromaticity & polarizability may enhance the biological activity of the compound. CHR3 (C-003) represents atom centered aromatic –CH fragments which is an essential feature for the inhibition of reverse transcriptase. Polarizability can create van dar waal force between closely approached atoms.

CONCLUSION

The QSAR models derived by using computed molecular descriptors including topological, geometrical and functional groups have provided rationales to explain the reverse transcriptase inhibitory activity of 1-(2-

hydroxyethoxy-methyl)-6-(phenylthio) thymine (HEPT) derivatives. The stepwise-MLR training models state that path/walk 4 - Randic shape index ,topological charge indices and Eigen value from edge adjacency matrix weighted by resonance integrals, information content and total information index on atomic composition belong to topological indices are crucial for producing anti HIV-1 inhibition. Increasing branching of the molecule may increase size and steric hindrance which may further decrease resonance & thus decrease in activity. Quadrupole x component value / weighted by polarizability (QXXp), Folding degree index (FDI) and sum of geometrical distances between S..Cl at 6-phenyl thio moiety of the HEPT derivatives is significant amongst geometrical indices. Functional group descriptors such as number of acceptor atoms for H-bonds (N,O,F), number of un substituted benzene C(sp2) and CHR3 (C-003) are predicted as significant parameters responsible for increasing ligand affinity towards reverse transcriptase inhibition, because these features may produce more electrostatic interactions between ligand- receptor complexes.

Acknowledgement
Authors are sincerely thankful to GIPER India for providing necessary reseearch facilities. Shubhra shows deep sense of gratitude to her supervisor Dr. Nandi. SN is thankful to NIC Slovenia for availing DRAGON Professional version 5.4-2006 software for calculation of theoretical molecular descriptors used in the present work.

REFERENCES

[1] RA Weiss, *Science* **1993,** 260, 1273–1279.
[2] DC Douek; M Roederer; R A Koup, *Annu. Rev. Med.* **2009,** 60**,** 471–484.
[3] A Cunningham; H Donaghy; A Harman; M Kim; S Turville, *Curr. Opin. Microbial.,* **2010**, 13, 524–529.
[4] H Tanaka; H Takashima; M Ubasawa.; K Sekiya, I Nitta; M Baba; Sh Shigeta; RT De Clercq; E Walker; T Miyasaka; *J. Med. Chem.* **1995**, 38**,** 2860–2865.
[5] C Hansch; L Zhang, Bioorg.. *Med. Chem. Lett*., **1992**, 2, 1165–1169.
[6] S Hannongbua; L Lawtrakul; *J* Limtrakul, *J. Comput. Aided Mol. Des*. **1996**, 10, 145–152.
[7] JM Luco, FH Ferreti , *J. Chem. Inf. Comput. Sci*., **1997**, 37**,** 392–401.
[8] M Jalali-Heravi; FJ Parastar, *Chem. Inf. Sci*. **2000**, 40, 147–154.
[9] CN Alves; JC Pinheiro; AJ Camargo; MMC Ferreira; AB Silva, *J Mol Struc. (Theochem).* **2000**, 530**,** 39–47.
[10] H Bazoui; M Zahouily; S Boulajaaj; S Sebti; D Zakarya, *SAR QSAR Environ. Res*. **2002**, 13, 567–577.
[11] L Duali; D Villemin; D Cherqaoui,, *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1200–1207.
[12] N Mills, Chem Draw Ultra 10.0. *J. Am. Chem. Soc*., **2006**, 128, 13649–13650.
[13] S Nandi, MC Bagchi, *Mol Divers*. **2010**, 14, 27-38.
[14] R Todeschini, V Consonni, Dragon software. Milano, Italy. (version 5.4-**2006**)
[15] ARKatritzky; R Petrukhin; D Tatham; S Basak; EBenfenati; M Karelson; U. Maran,*J. Chem. Inf.Comput. Sci*.,**2001**, 41, 679–685.
[16] CR. Rao, *Linear statistical inference and its applications*, 2nd ed. John Wiley & Sons, New York**. 1973.**
[17] M Randic, *J. Chem. Inf. Comput. Sci*., **2001**, 41**,** 607–613.
[18] SC Basak, *Med. Sci. Res*.,**1987,** 15, 605–609.
[19] S.C Basak; GD Grunwald; GJ Niemi, In Plenum Press, New York, **1997**, pp, 73–116.
[20] E Estrada; J Devillers; AT Balaban (eds), Gordon and Breach, Amsterdam, **1999**, pp 403–453.
[21] AJ Miller, Chapman and Hall, New York, **1990**.
[22] Minitab® Statistical Software: Minitab, **2010**. www.minitab.com
[23] S Nandi; M C Bagchi, *Mol. Simul.,* **2011**, 37, 196 — 209.
[24] PP Roy; K. Roy, *QSAR Comb. Sci.,* **2008**, 27, 302 – 313.