Design and Analysis of a Combinatorial Library of HEPT Analogues: Comparison of Selection Methodologies and Inspection of the Actually Covered Chemical Space

Rosalia Pascual,[†] Marta Mateu,[†] Johann Gasteiger,[‡] José I. Borrell,[†] and Jordi Teixidó*[,†]

Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Via Augusta 390,
E-08017-Barcelona, Spain, and Computer-Chemie-Centrum, Institute of Organic Chemistry,
University of Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052-Erlangen, Germany

Received July 24, 2002

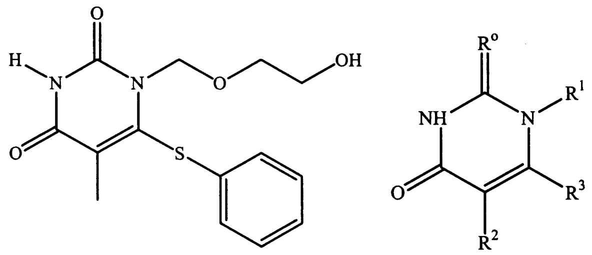
A large virtual library of 125 396 HEPT analogues, built by combining all fragments present in the published 180-compound HEPT family, has been studied in terms of diversity criteria and the goodness of the 11 available standard diversity selection methods analyzed. All the algorithms under study, except Cell-based Density, have rank above a random selection of compounds, with Optimum and Standard Deviation based Binning and Cell-based Fraction algorithms being the best choices. Furthermore, analysis of the actually tested compounds has been performed to compare the traditional drug discovery methodology versus a rational selection of combinatorial libraries approach.

INTRODUCTION

Once combinatorial chemistry[1] was overwhelmingly adopted by pharmaceutical and agrochemical industries both for lead discovery and optimization, a great need for suitable computational selection techniques to choose optimal substituents has arisen.[2−4] The combinatorial idea implies an explosion in the number of compounds to be considered for synthesis and thus, the goal of library design is to maximize the chances of discovering useful leads keeping the number of considered compounds at a reasonable size on behalf of the particular synthetic capabilities. This is achieved by maximizing the coverage of the chemical-property space, relying on the hypothesis that close compounds in the descriptor space should exhibit similar biological properties. Current literature is dominated by reports describing computational methods relevant for combinatorial library design in terms of the descriptors[5,6] and of the selection algorithms.[7−11] The reagent versus product-based selection controversy has been broadly discussed allowing recent development of new algorithms to select combinatorial subsets based on the properties of the resulting products.[12−16] Although these methods are frequently complemented by calculations on model libraries, there is less information concerning applications or case studies where opportunities and limitations of the different available approaches are considered in detail.

HEPT, 1-((2-hydroxyethoxy)methyl-6-phenylthio)thymine,[22] is a potent inhibitor of HIV-1 reverse transcriptase (RT) the enzyme responsible for catalyzing the synthesis of double strand viral DNA and therefore for the replication of HIV-1.[17,18] HEPT belongs to the nonnucleoside inhibitor class (NNRTIs) which unlike nucleoside-inhibitors (NRTIs) blocks the HIV−RT inducing conformational changes in the enzyme and exhibits fewer side effects and lower cytotoxicity.[19,20] However, the high rate of virus mutation, which leads to the emergence of drug-resistant viral strains,[21] has focused the research toward new molecules with improved activity and minor vulnerability to resistance.

Chart 1. HEPT and Its Combinatorial Library of Analogues



A large collection of HEPT analogues has been synthesized[22−29] mainly by Tanaka et al.[22−28] where new substituents were chosen by medicinal-chemistry intuition and synthetic accessibility. This also guarantees that the synthetically available space is unbiased due to exclusive rights. Due to both the attractiveness of being able to further clarify the structure−activity relationship for such an important disease and to the availability of large collection of homogeneous activity data, HEPT-analogues have been recently used in various QSAR studies.[30−35]

In the present work a library made up by combining all fragments present in the published HEPT analogues has been enumerated (virtually synthesized) and analyzed. The major purpose is to compare the outcome and performance of different selection methods applicable to that particular case. Analysis of the actually tested compound distribution within the whole synthetically accessible space of analogues will allow comparison between intuitive and experimentally based selection of fragments versus a rational selection of combinatorial libraries approach.

COMPUTATIONAL METHODS

The chemical fragments used to build the combinatorial library as well as the activity data for the synthesized

[*] Corresponding author phone: +34-93-267.20.00 fax: +34-93-205.62.66; e-mail: j.teixido@iqs.url.es.
[†] Grup d'Enginyeria Molecular.
[‡] Computer-Chemie-Centrum.

10.1021/ci0255681 CCC: $25.00 © 2003 American Chemical Society
Published on Web 11/06/2002

Table 1. Fragments Used To Build the 125 396 Compound Library of HEPT Analogs

F.I.N	R0	cells	test	R1	cells	test	R2	cells	test	R3	cells	test
0	**O**	217	157	Bu	35	3	Br	81	2	-(*R*)-CHOHPh	48	1
1	S	175	23	CH_2CH_2OMe	41	2	$C\equiv CMe$	77	1	-(*S*)-CHOHPh	50	1
2				$CH_2CH=CH$(2-benzofuranyl)	47	2	$C\equiv CH$	79	1	Br	53	1
3				$CH_2CH=CH$(2-furyl)	42	3	$C\equiv CPh$	75	1	$C\equiv CMe$	56	1
4				$CH_2CH=CH$(2-thienyl)	41	3	*c*-Pr	84	2	$C\equiv CH$	52	1
5				$CH_2CH=CH$(5-nitro-2-thienyl)	66	2	$CH_2CH=CH_2$	79	1	$C\equiv CPh$	59	1
6				$CH_2CH=CH$(3-Py)	38	1	CH_2Ph	93	1	CH_2CH_2Ph	76	1
7				$CH_2CH=CHC_6H_{11}$	46	1	**Me**	84	99	CH_2Ph	60	9
8				$CH_2CH=CHPh$	37	3	$CH=CH_2$	82	1	$CH_2Ph(3,5-Me_2)$	69	4
9				CH_2O-*c*-Hex	36	2	-*(Z)*-$CH=CHPh$	87	1	$CH=CH_2$	53	1
10				CH_2O-*i*-Pr	35	1	$CH=CPh_2$	81	1	-*(Z)*-$CH=CHMe$	63	1
11				CH_2OBu	33	1	Cl	81	2	-*(Z)*-$CH=CHPh$	53	1
12				CH_2OCH_2-*c*-Hex	37	2	$COCH(CH_3)_2$	80	1	Cl	71	1
13				$CH_2OCH_2C_6H_4$(4-Cl)	33	1	CONHPh	92	1	COPh	52	1
14				$CH_2OCH_2C_6H_4$(4-Me)	39	1	$COOCH_3$	77	1	I	53	5
15				$CH_2OCH_2CH_2Br$	30	1	COPh	72	1	Ph	45	1
16				$CH_2OCH_2CH_2Cl$	29	1	Et	83	40	SC_6H_{11}	72	1
17				$CH_2OCH_2CH_2F$	23	1	F	70	2	SBu	70	1
18				$CH_2OCH_2CH_2I$	31	1	H	89	3	SEt	67	1
19				$CH_2OCH_2CH_2N(CH_2CH_2)_2NH$	34	1	I	86	1	SMe	67	1
20				$CH_2OCH_2CH_2N(CH_2CH_2)_2O$	29	1	*i*-Pr	94	14	**SPh**	55	90
21				$CH_2OCH_2CH_2N(CH_2CN)_2$	37	1	Pr	87	2	SPh(2-Cl)	53	1
22				$CH_2OCH_2CH_2N(COPh)_2$	71	1	SPh	94	1	SPh(2-Me)	58	1
23				$CH_2OCH_2CH_2N(Ph)_2$	39	1				$SPh(2-NO_2)$	87	1
24				$CH_2OCH_2CH_2NH_2$	39	1				SPh(2-OMe)	53	1
25				$CH_2OCH_2CH_2NHCOCH_2CH_2Cl$	52	1				SPh(3-Br)	49	1
26				$CH_2OCH_2CH_2NHCOCH_2NHCH_2PO_3H$	47	1				$SPh(3-CF_3)$	56	1
27				$CH_2OCH_2CH_2NHCOPh$	62	1				$SPh(3-CH_2OH)$	63	1
28				$CH_2OCH_2CH_2NHPh$	41	2				SPh(3-Cl)	49	1
29				$CH_2OCH_2CH_2OC_5H_{11}$-*n*	33	1				SPh(3-CN)	58	1
30				$CH_2OCH_2CH_2OCH_2Ph$	33	1				SPh(3-COMe)	58	1
31				**$CH_2OCH_2CH_2OH$**	37	97				$SPh(3-CONH_2)$	59	1
32				$CH_2OCH_2CH_2OMe$	30	1				SPh(3-COOH)	69	1
33				$CH_2OCH_2CH_2Ph$	37	2				SPh(3-COOMe)	66	1
34				$CH_2OCH_2CH_2SiMe_3$	42	1				SPh(3-Et)	64	1
35				$CH_2OCH_2CH_2SPh$	44	2				SPh(3-F)	50	1
36				$CH_2OCH_2CH_2SPy$	47	2				SPh(3-I)	53	1
37				CH_2OCH_2Ph	38	7				SPh(3-Me)	58	1
38				CH_2OEt	34	15				$SPh(3-NO_2)$	77	1
39				CH_2OMe	38	1				SPh(3-OH)	51	1
40				CH_2OPr	33	1				SPh(3-OMe)	45	1
41				Me	37	1				SPh(3-t-Bu)	67	1
42				Et	41	1				$SPh(3,5-Cl_2)$	42	5
43				CH_2-*i*-Pr	35	1				$SPh(3,5-Me_2)$	56	13
44				H	42	1				SPh(4-Cl)	56	1
45				$CH_2OCH_2CH_2N_3$	63	1				SPh(4-CN)	57	1
46				$CH_2OCH_2CH_2NHCO(CH_2)_2PO_3H$	25	1				SPh(4-COMe)	58	1
47										SPh(4-F)	59	1
48										SPh(4-Me)	59	1
49										$SPh(4-NO_2)$	75	1
50										SPh(4-OH)	60	1
51										SPh(4-OMe)	58	1
52										SPy	57	3
53										NHC_6H_{11}	66	1
54										NHPh	52	1
55										OC_6H_{11}	90	1
56										OMe	68	1
57										OPh	72	1
	62698[a]			2668[a]			5452[a]			2162[a]		

[a] Last row: number of molecules in the sublibraries originated by fixing a single substituent; **F. I.N**: Fragment Index Number; HEPT = (0,31,7,20); **cells**: number of cells occupied by all molecules containing a particular fragment; **test**: number of tested molecules with that fragment.

analogues were taken from refs 22 to 28. In those studies, four different substitution positions were almost independently explored regarding their effect on the activity values. The fragments present at least in one tested compound are depicted in Table 1. The combinatorial combination of all of them results in a library of 2*47*23*58 = 125 396 molecules where the 180 actually synthesized and tested constitute just about a 0.15%.

The Pralins program,[36] a code developed at IQS to deal with combinatorial library selection topics, carried out the library enumeration, and Corina[37] converted the models to 3D. We decided to use two kinds of descriptors: on one hand 9 standard combinatorial chemistry descriptors calculated by Cerius2[38] (area, molecular volume, molecular weight, radius of gyration, density, Principal moment of inertia, number of rotable bonds, number of hydrogen-bond

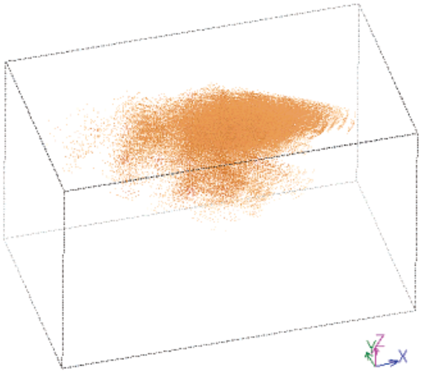


Figure 1. Visualization by the first three principal components that represent 61% of the variance.

acceptors and number of hydrogen-bond donors) and, on the other, spatial autocorrelation coefficients[39,40] and topological descriptors to account for atom-based physicochemical effects of the molecules as well as for their topology and 3D structure.

The autocorrelation function is defined by eq 1, where *A(d)* is the correlation coefficient referring to atom pairs *i* and *j* separated by *d* bonds in case of 2D-topological or by a *d* Å distance for 3D-spatial autocorrelation, and $p_{i,}$ p_j refer to a given property of atoms *i* or *j* respectively.

$$A(d) = \sum_{i,j} p_i \cdot p_j \qquad (1)$$

To be able to deal with such a large data set we used the rapid empirical methods for the calculation of atom-based physicochemical properties collected in the PETRA[41] program package. We selected 8 properties (atomic identity function, atom polarizability,[42] σ charge,[43] σ electronegativity,[44] π charge, π electronegativiy,[45] lone pair electronegativity and total charge) and computed for each property the autocorrelation values for 6 topological (number of intervening bonds) and 6 spatial distances (6 equidistant intervals from 1 to 7 Å). Thus, 8*(6+6) = 96 autocorrelation functions together with the 9 aforementioned combichem variables were used to characterize the virtual library within a 105-dimensional chemical space. The dimensionality was reduced by principal component analysis (PCA) down to 12 components that account for 90% of the variance (Figure 1). Statistical analysis of the whole library was only possible by means of the binary data file system provided by Cerius2, which is able to work without holding libraries and descriptors in memory.

Once the chemical space was set, we explored 11 different diverse selection criteria that can be grouped into four categories:

1. Binning algorithms, which divide the space in bins or cells by a specific criterion method and select a compound from each bin, are as follows:

•Optimum Binning: divides the property ranges in such a way that the number of occupied cells is always less than or equal to the number of molecules to select. Properties are binned with a bias toward the ones that exhibit the largest variation. Thus, properties with large ranges tend to have more divisions than the rest, searching for bins as much equal faced as possible while remaining in the desirable number of total occupied cells.

•Standard Deviation based Binning: every property range is divided into three bins. i) from the minimum value to the mean minus two times the standard deviation, ii) from there up to the mean plus two standard deviations, and iii) from the mean plus two standard deviations to the maximum value.

2. D-Optimal design: a popular computer aided experimental design technique based on maximizing the determinant of the information matrix. This maximization is an NP-complete problem managed by Monte Carlo Optimization. The conditions for Monte Carlo optimizations were kept constant to 1 000 000 steps, a temperature of 300 and a termination condition of 100 000 idle steps, and for all the cases stochastic optimization was used (vide infra), as they gave stabilized results for all runs.

3. Distance-based functions: measure intermolecular distances (Euclidean was selected), optimized as well, using a Monte Carlo algorithm to achieve quick convergence. The four different selection criteria considered in the present work are as follows:

•Max MinSpanTree (based on calculating the minimum spanning tree for each subset of selected points in the stochastic optimization process, and computing an error function based on the length of each edge in the tree[46]),

•MaxMin (2),

•Product (3)

•PowerSum (4)

$$\textit{MaxMin}: \qquad \max\,\{\min\,\{D_{ij}^{\;2}\}\} \qquad (2)$$

$$\textit{Product}: \qquad \max\,\{[\prod D_{ij}^{\;2}\,]^{1/0.5\cdot n\cdot(n-1)}\} \qquad (3)$$

$$\textit{PowerSum}: \qquad \max\left\{\frac{0.5\cdot n\cdot(n-1)}{\sum 1/D_{ij}^{\;2}}\right\} \qquad (4)$$

where D_{ij} is the intermolecular distance and *n* is the number of selected compounds.

4. Cell-based functions for R-group subsetting purposes. Actually, these methods are derived from an optimum binning of space into cells, in the sense described above, followed by a reagent or R-group substituent selection where the cell-coverage of the resulting products is evaluated by four different criteria:

•Cell-based Fraction: Cells occupied by subset/Number of occupied cells

•Cell-based Chi^2: Sum $(n_\mathrm{i} - n_\mathrm{ave})^2$

•Cell-based Entropy: $-$Sum (n_i * Log(n_i))

•Cell-based Density: $-$Sum ($n_\mathrm{i}/n_\mathrm{ave}$ * Log($n_\mathrm{i}/n_\mathrm{ave}$)) where n_i is the number of compounds in cell i and n_ave is the average number of compounds per cell.

We decided to fix the size of the selections at $\sqrt{N} \approx 355$ molecules ($N = 125\,396$ compounds); however, as our aim was the intrinsic comparison of the methods, we let initially this number adapt to their specific particularities. In the case of the R-group subsetting selections the number of fragments to select for each position had to be fixed. Therefore the next integer boundary to the square root of the size of each group of fragments was taken to give up combinatorial sublibraries of 2*7*5*8 = 560 compounds. In the Binning algorithms, the number of molecules selected corresponds to the number of filled cells, which in the Optimum approach was found to be 262 occupied bins (from 2048 total cells) and in the Standard Deviation based method 1091 occupied bins.

Once the selections were made by the eleven aforementioned methods, we intended to analyze and compare the selected subsets. As there is no universal accepted criterion to judge the quality of a subset in terms of diversity, several different evaluation criterion have been used in theoretical studies,[10−13,15,16,47] which are mostly based on the ability of the subset to represent the entire library. We chose space coverage as an objective evaluation measure, considering thereby important, the capacity of the selection to spread as much as possible in the range of properties offered by the entire library. Therefore we chose the following two techniques for diversity evaluation: the Cell-based method[48] and the Diversity Integral method. Cell-based techniques have been introduced above as selection strategies; however, they are also very useful for evaluation and comparison purposes. In this case the entire property space was divided into 1000 total bins and cell occupancy of each subset was determined. Higher values indicate better coverage and therefore greater diversity. The Diversity Integral method is specially designed to evaluate diversity[38] because of the importance of judging compound diversity in library acquisition strategies. It is based on comparing the sum of the distances between random points in the common property space of two libraries and the closest molecule to those points from each of these libraries. The library with the lowest value of the two samples better the property space and is therefore considered more diverse than the other.

To be able to compare the different selection methods, overcoming the possible size effect of the resulting subsets, we set the number of selected compounds to match the different initial values in those adjustable methods: D-Optimal Design and the four Distance-based functions. We have also considered for each case a random selection run intended to prove the effectiveness of a rational proceeding. The new equal-sized subsets were newly evaluated and compared.

A further analysis was a comparison of the similarity of the selected subsets to draw conclusions about the equivalence of the different selection criteria. This was accomplished by using Cell-based comparisons. Now, only the common property space of the two selected subsets to be compared is binned, and several metrics are defined as functions of the cell occupancy of each subset: Carbó index (5), Hodgkin index (6), Tanimoto coefficient (7) and Hamming distance (8). The binning based comparison was conducted at two levels: 1000 total cells and 355 cells occupied with at least one molecule, where 355 stands for the desired number of molecules to select in the study.

$$\text{Carbó Index: } \frac{Cab}{(Ca\cdot Cb)^{1/2}} \tag{5}$$

$$\text{Hodgkin Index: } \frac{2\cdot Cab}{Ca + Cb} \tag{6}$$

$$\text{Tanimoto coefficient: } \frac{Cab}{Cao + Cbo + Cab} \tag{7}$$

$$\text{Hamming distance: } Coc - Cab \tag{8}$$

where Ca = cells occupied by library A, Cb = cells occupied by library B, Cao = cells occupied by library A only, Cbo = cells occupied by library B only, Cab = cells occupied by libraries A and B, and Coc = cells occupied by libraries A or B.

Table 2. Cell-based Evaluation of the Selected Subsets

selection method	% coverage	selection method	% coverage
Optimum Binning	100.00	Max Min Span Tree	25.00
Std. dev. Based Binnig	73.53	Cell-based Fraction	58.82
MaxMin	41.91	Cell-based Chi^2	55.15
Product	28.68	Cell-based Entropy	53.68
Power Sum	28.68	Cell-based Density	10.29
D-Optimal Design	28.68	Random	19.85

The last point of our study was the analysis of the actually tested compounds within the whole synthetically available chemical space of the virtual library in order to compare the distribution of an intuitive choice versus a diversity-based rational one. First, we calculated the number of cells occupied by the sublibraries originated by fixing a substituent in a single position (2+47+23+58 = 130 sublibraries, i.e., fix at R0 an oxygen atom and count the number of cells occupied by the 64 032 elements of the respective sublibrary). Afterward, we divided the synthesized and tested molecules into five categories: A) those with EC_{50} not reached at the highest concentration tested because of low solubility; B) those with $EC_{50} < 7\ \mu M$ better than HEPT = (0,31,7,20); C) $7\ \mu M < EC_{50} < 10\ \mu M$ similar activity than HEPT; D) $10\ \mu M < EC_{50} < 100\ \mu M$; and E) finally those compounds with $EC_{50} > 100\ \mu M$. Then, we analyzed both the cell coverage and the compound distribution (Table 6). These calculations were done through programmed Visual Basic Applications (VBA) in Microsoft Excel 2000.

RESULTS AND DISCUSSION

The results of the Cell-based method are shown in Table 2. Certainly, the Optimum Binning algorithm manages a complete representativity of the space in terms of cells, as both the selection and evaluation criteria coincide and there are no further restrictions as in the case of the R-group subsetting selections. Standard Deviation Based Binning ranks next, where the higher coverage may be due both to high diversity managed by this method and to the greater number of compounds inherent to this strategy. Surprisingly three out of four R-group subsetting techniques rank better than general selection algorithms despite the combinatorial restrictions, having Cell-based Fraction the best result. Maxmin stands above all other distance-based method and also above D-Optimal design. Cell-based Density shows lesser coverage than even a random selection. Going on to the Diversity Integral comparison, the results are shown in Table 3. The values in each cell $\boldsymbol{c_{rc}}$ correspond to the result of Formula 9, where Dmi_{ci} and Dmi_{ri} are the minimum distance between a molecule, corresponding to the subset obtained by the method indicated in column c or row r respectively, and a random point in the common property space of the two subsets.

$$c_{rc} = \sum_{i=1}^{1000} Dmi_{ci} - \sum_{i=1}^{1000} Dmi_{ri} \tag{9}$$

Consequently, according to the diversity integral criterion, negative values indicate a better sampling of the methods

Table 3. Comparison of the Selected Subsets by the Diversity Integral Criterion

	Optimum Binning	Std. dev. Based Binnig	MaxMin	Product	Power Sum	D-Optimal design	Max Min SpanTree	Cell-based Fraction	Cell-based Chi^2	Cell-based Entropy
Std. Dev. Based Binnig	0.289									
MaxMin	−0.423	−0.873								
Product	−1.152	−1.501	−0.669							
Power Sum	−0.925	−1.535	−0.602	0.076						
D-Optimal design	−1.490	−2.022	−1.000	−0.219	−0.254					
Max MinSpanTree	−1.341	−2.076	−0.936	−0.293	−0.375	−0.255				
Cell-based Fraction	0.057	−0.167	0.658	1.399	1.416	1.693	1.904			
Cell-based Chi^2	−0.055	−0.264	0.461	1.239	1.392	1.627	1.756	−0.081		
Cell-based Entropy	−0.247	−0.545	0.446	1.002	1.081	1.419	1.512	−0.411	−0.216	
Cell-based Density	−3.700	−4.852	−3.227	−2.315	−2.383	−2.555	−2.042	−4.459	−4.266	−4.062

[a] c_{rc} values calculated by formula (9). 1000 random points have been used for this calculation. Negative values indicate better sampling of the methods pointed by columns, while positive better sampling of those pointed by rows.

Table 4. Comparison of the Selections by the Diversity Integral Criterion Avoiding the Size Effect of the Subsets (1000 Random Points)

	MaxMin	Product	Power Sum	D-Optimal Design	Max Min Design	Random	# Products
MaxMin						1.539	262
Product	−0.688					0.688	
Power Sum	−0.855	0.018				0.753	
D-Optimal design	−0.709	0.083	0.042			1.019	
Max MinSpanTree	−1.069	−0.172	−0.157	−0.293		0.446	
Optimum Binning	0.696	1.235	1.246	1.471	1.423	2.076	
MaxMin						1.041	1091
Product	−0.370					0.599	
Power Sum	−0.385	−0.018				0.688	
D-Optimal design	−0.744	−0.367	−0.383			0.206	
Max MinSpanTree	−1.064	−0.563	−0.671	−0.212		0.162	
Std. dev. Based Binnig	0.504	0.738	0.701	1.136	1.471	1.523	
MaxMin						1.414	560
Product	−0.654					0.783	
Power Sum	−0.580	0.041				0.833	
D-Optimal design	−0.967	−0.412	−0.349			0.307	
Max MinSpanTree	−1.131	−0.381	−0.471	−0.034		0.159	
Cell-based Fraction	0.453	1.174	1.137	1.452	1.397	1.793	
Cell-based Chi^2	0.347	1.074	1.030	1.385	1.414	1.824	
Cell-based Entropy	0.237	0.948	0.723	1.129	1.265	1.534	
Cell-based Density	−4.167	−2.804	−2.933	−2.810	−1.919	−1.688	

pointed by columns, while positive values show a better sampling of those pointed by rows. Ranking in this frame of reference the different methodologies from best to worst, we obtained the following order: Standard Deviation Based Binnig > Cell-based Fraction ≈ Optimum Binning ≈ Cell-based Chi^2 > Cell-based Entropy > MaxMin > Power Sum > Product > D-Optimal design > Max MinSpanTree > Cell-based Density. It has to be noted that Cell-based Fraction, Optimum binning and Cell-based Chi^2 gave identical good results. It should be emphasized that the values in the tables are qualitative and may exhibit deviations depending on the random points used in the evaluation; however, the tendency and the order of magnitude is always preserved, even when the number of random points is increased by a 10-fold factor. The ranking fully coincides with the one obtained by the Cell-based method, showing good agreement between the criteria behind those methods.

It strikes again that some R-group subsetting techniques rank better than general selection methods. A first explanation for these observations could be that the number of molecules in each selected subset influences the results of the comparison. Bigger sizes would lead to better diversity scores. Therefore, we repeated the selections to match the different initial values in those adjustable methods: D-Optimal Design and the four Distance-based functions. Thereafter, different size-consistent comparisons to match the different initial selection values were performed. The new results are shown in Table 4. There is no significant change concerning the observations above. Surprisingly, for this particular library and the Monte Carlo conditions used in the optimization, the R-group subsetting techniques still rank better than nonrestricted selection algorithms even with the number of selected compounds being equal for both kinds of methods. Std. dev. Based Binnig, Cell-based Fraction and Optimum Binning are each the best behaved within their size-category. As the size is an inherent property in this methodology there is no way to evaluate their performance without including the effect of the different sizes. It is also remarkable that the Cell-based Density method positions itself at the bottom of the list and its c_{rc} values are significantly bigger in magnitude than all others. It must also be noticed that all methods except Cell-based Density give better results than a random selection of compounds. In fact, the Cell-based Density criteria is equal to the Cell-based Entropy method but for the occupancy variable in the equation. The number of compounds in each cell is substituted by a density rate giving thus more importance to the coverage of the most populated cells than to a homogeneous spreading of the property space offered by the entire library. The results suggest that the cell-density concept, which is sustained by this criterion, can be counterproductive when

Table 5. Distance Values of the Selected Subsets Considering 355 Occupied Cells[a]

	Carbó index										Hodgkin index									
	OptBin	StdBin	MxMn	Prod	PowS	Dopd	MMST	CbF	CChi^2	CbEnt	OptBin	StdBin	MxMn	Prod	PowS	DOpd	MMST	CbF	CChi^2	CbEnt
StdBin	0.627										0.616									
MxMn	0.372	0.576									0.372	0.553								
Prod	0.282	0.483	0.416								0.282	0.430	0.413							
PowS	0.242	0.513	0.402	0.377							0.242	0.467	0.398	0.377						
DOpd	0.230	0.444	0.370	0.415	0.457						0.228	0.388	0.367	0.415	0.457					
MMST	0.179	0.401	0.386	0.403	0.422	0.369					0.178	0.353	0.380	0.403	0.421	0.369				
CbF	0.556	0.652	0.463	0.372	0.390	0.392	0.367				0.556	0.649	0.460	0.359	0.378	0.377	0.350			
CChi^2	0.574	0.644	0.455	0.420	0.538	0.398	0.445	0.680			0.574	0.640	0.450	0.398	0.510	0.381	0.418	0.680		
CbEnt	0.387	0.603	0.498	0.363	0.489	0.425	0.429	0.661	0.775		0.382	0.597	0.494	0.351	0.475	0.405	0.407	0.659	0.775	
CbDen	0.117	0.227	0.331	0.303	0.302	0.349	0.398	0.293	0.324	0.318	0.113	0.163	0.321	0.296	0.298	0.345	0.396	0.260	0.266	0.277

	Tanimoto coefficient										Hamming distance									
	OptBin	StdBin	MxMn	Prod	PowS	Dopd	MMST	CbF	CChi^2	CbEnt	OptBin	StdBin	MxMn	Prod	PowS	DOpd	MMST	CbF	CChi^2	CbEnt
StdBin	0.445										162									
MxMn	0.228	0.382									274	189								
Prod	0.164	0.274	0.261								305	199	298							
PowS	0.138	0.304	0.249	0.232							326	199	290	291						
DOpd	0.129	0.241	0.225	0.262	0.296						325	296	324	285	257					
MMST	0.098	0.214	0.235	0.252	0.267	0.226					341	338	274	282	283	308				
CbF	0.385	0.480	0.299	0.219	0.233	0.232	0.212				179	172	296	318	309	291	301			
CChi^2	0.403	0.471	0.291	0.249	0.342	0.235	0.264	0.515			169	172	310	287	200	293	259	179		
CbEnt	0.236	0.425	0.328	0.213	0.311	0.254	0.255	0.492	0.632		324	227	234	315	199	256	216	186	128	
CbDen	0.060	0.089	0.191	0.174	0.175	0.208	0.247	0.150	0.153	0.161	314	380	309	304	306	285	269	324	293	240

[a] OptBin: Optimum Binning; StdBin: Standard Deviation Based Binning; MxMn: MaxMin; Prod: Product; PowS: Power Sum; Dopd: D-Optimal Design; MMST: Max MinSpanTree; CbF: Cell-based Fraction; CChi^2: Cell-based Chi^2; CbEnt: Cell-based Entropy; CbDen: Cell-based Density.

aiming only for diversity in the sense defined in the present work.

In the previous study there were subsets or sublibraries which showed similar diversity marks. Now, the question arises if those subsets not only rank similar but consist, in fact, of similar molecules. Similarity comparison was therefore performed for all subsets obtained from the different selection methods. The resulting values are shown in Table 5 for 355 occupied cells. The Cell-based Chi^2 and Cell-based Entropy subset pair exhibit the higher similarity scores followed by the pair Cell-based Chi^2 and Cell-based Fraction, which agrees with those three having very similar dissimilarity scores in all the previous comparisons. This lets us conclude that the methods not only achieve the same quality of subsets regarding the Diversity Integral criterion but also their selections are almost identical. The same reasoning applies to Optimum Binning, Standard Deviation and Cell-based Fraction which reproduce similar results by the Cell-based comparison. On the other hand, the Cell-based Density criterion, as expected, stands out again for being the most dissimilar of all methods. Equivalent observations are derived when examining the result of the Cell-based comparison with 1000 total cells.

Finally, the number of cells occupied by all molecules containing a particular fragment together with the number of tested molecules with that fragment are shown in Table 1. The cells are defined as those obtained from an Optimum Binning partition of space (2048 total cells), as the Optimum Binning division is adequate for comparison and therefore for coverage evaluation purposes. The last row in the table shows the size of the subsets obtained when fixing in the library one of the available fragments to the position indicated by the column. This table allows us to draw some conclusions about the influence of varying groups for every position in the scaffold. Values obtained for fragments in position R3 are especially remarkable. The subsets characterized by fixing those fragments are the smallest, this means, fixing a fragment in position R3 reduces the size of a combinatorial library from 125 396 to 2162. However, the average number of cells covered by all the compounds having a particular fragment in that position is greater than the collections obtained when fixing a fragment to R1 despite the bigger size of these collections. For instance, the highest coverage achieved when fixing a fragment in the library corresponds to OC6H11, a substituent in R3, whose subset occupies 90 out of the 262 cells that build the total property space of the 125 396 compound library. These statements bring us to the conclusion that R3 has lesser influence on coverage than R1. On the other hand, the biggest coverage achieved by R2, 94, is not much bigger than the one achieved by R3, where the size of its subsets is twice as large. Dividing the size of the subsets R1 and R2 by the average number of occupied cells in each case, it can be said that both substituent positions do not present significant differences. As for R0 just to say that the oxygen atom allows a greater coverage than the sulfur atom.

Contrasting the intuitive medicinal-chemistry process with the observations above, it can be seen that more molecules with oxygen in R0 were synthesized and tested, which agrees with achieving higher space coverage; however, the contrary happens with the procedure for position R3. On one hand, its fragment group has the biggest size, and, on the other hand, the proportion of fragments used only once (87.9%) is bigger than for the other positions (60% for both R2 and R3). Consequently, the coverage of the 180 synthesized compounds is less (23 cells) than the coverage obtained by randomly picking 180 compounds from the complete library (38 cells in average). The distribution of compounds can also be regarded from another frame of reference when taking into account the population density in each cell. Dividing the number of compounds in each cell by the size of the library, it can be seen that the tested cells stand for 81.5%

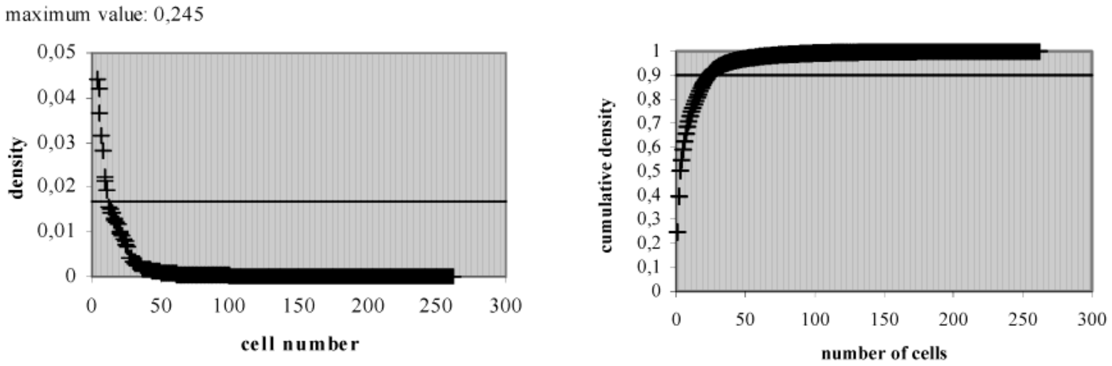


Figure 2. Density and cumulative density in each Optimum Binning cell versus cell number, where cells have been reordered according to size. (Density = number compounds in every cell/125 396).

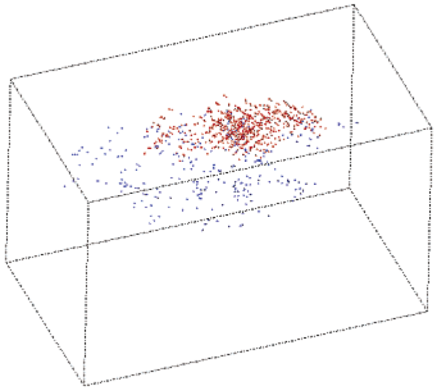


Figure 3. Blue points show an Optimum Binning based selection while red points correspond to compounds selected by the Cell-based Density algorithm.

of the molecules. This means that 9% of the cells are able to represent 81.5% of the population. Furthermore it has been observed that the 24 most populated cells achieve a representativity of 90% (Figure 2). The high number of singletons (19% of the cells) is remarkable as well, and also the fact that no cell with a density inferior to 0.0018 has been experimentally explored. These observations lead to the following assessment: the tested compounds are very representative from a density point of view, a characteristic that may be induced by the fact that the library was built by combining all the fragments used in the synthesized analogues and that most of these analogues have little distance between them. However, as stated above, the tested compounds are able to cover only a small part of the property space offered by the whole library. This study of the distribution of compounds enables also a better explanation of the poor diversity results obtained by the Cell-based Density selection. As just mentioned, the amount of compounds is very high in certain regions or cells, while other regions are covered by just one molecule. This causes a method that cares for representativity in terms of density to ignore all this low-density regions, thus weakening in terms of diversity (Figure 3).

Table 6 shows the spreading of the tested molecules among the occupied cells. The compounds are represented in form of their Fragment Index Numbers (F.I.N) and given as (R0, R1, R2, R3), i.e., HEPT = (0,31,7,20). The cell division does not generally separate active compounds from less active ones or from those with solubility problems. However, this does not imply that close compounds in the descriptor space do not show similar biological properties. One has to bear in mind two important factors. First the cells under consideration were obtained by a partitioning technique, as the computational demands of conventional clustering algorithms were unworkable, and though it constitutes a good pattern for coverage evaluation, the resulting cells are not internally as homogeneous as clusters are. The second thing to point out is that in diversity selections the descriptors are not adjusted and properly weighted, as in the case of QSAR studies, where activity data allows a descriptor validation.

The analysis of Table 6 offers an important evidence for the need of making combinatorial design for focused libraries. When representing every tested molecule in function of their constituent fragments, one realizes that fragments present in HEPT analogues derivatized in only one position and with low activity values (columns E, D in Table 6) are not tried anymore (R2 = 24,25; R3 = 17, 22, 23, 24, 26, 28, 38, 39, 40, 41, 48, 57). This procedure arises from an intuitive Free-Wilson approach,[49,50] which assumes an additive and independent effect of substituents on activity. However, the following example shows the importance of not extracting conclusions by only a single combination. The EC_{50} value of the analogue (0,31,7,7) is 23 μM, poorer than HEPT, although the other eight tested analogues having fragment 7 in R3 [(0,31,20,7) (0,38,16,7) (0,31,16,7) (0,-38,20,7) (0,1,16,7) (0,0,16,7) (0,0,20,7) and (0,1,20,7)] show all of them better activity than HEPT (column B in Table 6) such as (0,38,20,7) analogue with an EC_{50} of 0.0042 μM.

CONCLUSIONS

We have built a virtual combinatorial library of 125 396 compounds starting from reported fragments and have thereby increased the space of HEPT analogues by a factor of 670. The programs Corina and PETRA have shown their worth in combinatorial chemistry for being able to deal with the commonly required huge quantities of molecules in this field. Also the binary data file system has been essential to work with this library within Cerius2.

We have tested the performance of different selection methods for diversity based combinatorial design. The best-classified methods from a coverage point of view are Optimum and Standard Deviation Based Binning algorithms and surprisingly also Cell-based Fraction despite the combinatorial restraints imposed by this criterion. Within the distance based methods MaxMin ranks best. All the algorithms under study have rank above a random selection of compounds except for the Cell-based Density criterion. This

Table 6. Distribution of the Tested Compounds[a]

cell	total	test	A			B			C	D	E
207	220	1								(0,31,7,38)	
239	1802	4	(0,5,16,20)	(0,31,7,49)	(0,5,7,20)						(0,31,7,23)
334	1039	1	(0,23,7,20)								
335	1804	1				(0,3,16,43)					
366	1910	2	(0,22,7,20)			(0,31,10,20)					
367	3556	8	(0,2,16,20)			(0,3,16,20)	(0,4,16,20)	(0,4,16,43)		(0,31,7,57)	
						(0,8,16,43)	(0,37,7,20)	(0,37,16,20)			
368	553	2				(0,38,7,20)	(0,39,7,20)				
379	1474	1				(0,45,7,20)					
455	999	3	(0,31,7,32)	(0,26,7,20)						(0,31,7,26)	
463	18696	57	(0,19,7,20)	(0,20,7,20)	(0,31,7,21)	(0,31,7,25)	(0,31,7,34)	(0,31,7,37)	(0,31,7,30)	(0,31,7,24)	(0,31,7,48)
			(0,31,7,50)	(0,31,7,51)	(0,31,7,47)	(0,31,7,42)	(0,31,7,43)	(0,31,21,20)	(0,31,7,33)	(0,31,7,41)	
			(0,21,7,20)	(0,31,7,27)	(0,31,7,45)	(1,31,16,41)	(0,31,5,20)	(0,31,16,8)	(0,31,7,36)	(0,31,7,22)	
			(0,31,7,46)	(0,31,1,20)	(0,31,7,6)	(0,31,20,20)	(0,31,16,43)	(0,31,20,7)	(0,31,7,29)	(0,31,7,28)	
			(0,29,7,20)	(0,30,7,20)	(0,31,7,44)	(0,31,20,8)	(0,31,16,20)	(0,31,20,43)	(0,31,7,16)	(0,31,7,39)	
			(0,31,22,20)	(0,31,3,20)	(0,27,7,20)	(0,38,16,7)	(0,38,16,8)	(0,38,20,8)		(0,31,7,40)	
			(0,31,6,20)	(0,31,7,13)		(0,9,16,20)	(0,31,7,35)	(0,31,16,7)		(0,25,7,20)	
						(0,28,7,20)				(0,24,7,20)	
										(0,31,7,7)	
464	1951	13	(0,31,7,18)	(0,31,7,19)	(0,31,11,20)	(0,31,19,20)			**(0,31,7,20)**		(0,31,7,17)
			(0,31,17,20)	(0,31,0,14)	(0,31,0,20)						
			(0,31,7,2)	(0,31,7,14)	(0,31,11,14)						
			(0,31,17,14)								
487	1541	1	(0,31,14,20)								
495	30829	40	(0,31,13,20)	(0,35,7,52)	(0,31,12,20)	(0,28,7,52)	(0,31,16,42)	(0,38,16,42)	(0,36,7,52)		
			(0,34,7,20)	(0,31,7,55)	(0,36,7,20)	(1,14,16,19)	(1,31,7,42)	(1,31,16,42)	(0,32,7,20)		
			(0,35,7,20)	(0,31,15,20)		(1,31,20,42)	(1,33,16,19)	(1,37,16,42)			
						(1,37,20,19)	(1,38,16,42)	(0,31,8,20)			
						(0,38,4,20)	(0,38,20,7)	(0,38,16,43)			
						(1,9,16,19)	(1,12,16,19)	(1,13,16,19)			
						(1,37,16,19)	(1,38,16,41)	(0,12,16,20)			
						(0,33,16,20)	(0,31,9,20)	(0,10,16,20)			
						(0,11,7,20)	(0,37,16,43)	(0,37,20,20)			
						(0,38,16,20)	(0,38,20,20)	(0,40,7,20)			
496	5539	11	(0,31,18,20)	(0,31,2,20)	(0,31,7,12)	(1,31,20,19)	(0,17,7,20)	(1,38,4,19)			
			(0,31,7,56)	(0,31,18,14)	(0,18,7,20)	(0,15,7,20)	(0,16,7,20)				
1359	1619	3	(0,2,7,20)	(0,7,7,20)		(0,1,16,7)					
1360	516	1				(0,0,7,20)					
1391	2424	5	(0,3,7,20)	(0,4,7,20)		(0,6,7,20)	(0,8,7,20)	(0,8,16,20)			
1392	1221	3	(0,44,7,20)	(0,41,7,20)		(0,42,7,20)					
1479	1601	1	(0,46,7,20)								
1487	3943	8	(0,31,7,31)	(0,31,7,53)	(0,31,7,11)	(0,0,16,7)	(0,0,20,7)	(0,1,20,7)			
			(0,31,7,0)	(0,31,7,1)							
1488	897	9	(0,31,7,3)	(0,31,7,4)	(0,31,7,9)	(1,31,7,19)	(1,43,16,19)		(1,31,21,19)		
			(0,31,7,10)	(0,31,7,5)	(0,31,7,15)						
1519	13532	1	(0,31,7,54)								
1520	4561	4	(1,31,18,19)			(1,31,16,19)	(1,38,20,19)	(1,38,16,19)			

[a] A: EC_{50} not reached; B: $EC_{50} < 7\ \mu M$; C: $7\ \mu M < EC_{50} < 10\ \mu M$; D: $10\ \mu M < EC_{50} < 100\ \mu M$; E: $EC_{50} > 100\ \mu M$.

is due to the high grouping tendency of compounds of this library in certain regions of space together with the representativity or density optimization implicit in this criterion. Furthermore the Cell-based evaluation method and the Diversity Integral method have shown to give equivalent results.

In addition we have analyzed the distribution of the tested compounds in the combinatorial accessible space of analogues and have shown that indeed the conventional and intuitive procedure of one position variation starting from the lead compound, which would correspond to a qualitative Free-Wilson approach, restrains the coverage giving smaller values than a random selection. It is arguable whether this lower coverage, focusing toward the cells near to the lead, is desirable in a library of analogues, or whether it is valuable to cover the whole combinatorial descriptor space allowed by the core structure.

Finally, it is important to remark the great need of efficient tools and algorithms to computationally manage the huge amount of information generated in combinatorial library design.

ACKNOWLEDGMENT

One of us (R.P.) would like to thank the Generalitat de Catalunya − DURSI for a grant within the Formació de Personal Investigador (2000FI) program. Support of this work by a grant from the Dirección General de Enseñanza Superior e Investigación Científica within the Programa Sectorial de Promoción General del Conocimiento (PM98-0012-C02-02) is gratefully acknowledged.

REFERENCES AND NOTES

(1) Kassel, D. B. Combinatorial Chemistry and Mass Spectrometry in the 21st Century Drug Discovery Laboratory. *Chem. Rev.* **2001**, *101*, 255−267.

(2) Leach, A. R.; Hann, M. M. The in silico world of virtual libraries. *Drug Discovery Today* **2000**, *5*, 326−336.

(3) Van Drie, J. H.; Lajiness, M. S. Approaches to virtual library design. *Drug Discovery Today* **1998**, *3*, 274−283.

(4) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.

(5) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363−372.

(6) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.

(7) Hassan, M.; Bielawski. J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Mol. Diversity* **1996**, *2*, 64−74.

(8) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.

(9) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(10) Agrafiotis, D. K. A constant time algorithm for estimating the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159−167.

(11) Lobanov, V. S.; Agrafiotis, D. K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 460−470.

(12) Reynolds, C. H.; Tropsha, A.; Pfahler, L. B.; Druker, R.; Chakravorty, S.; Ethiraj, G.; Zheng, W. Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1470−1477.

(13) Graham, E. T.; Jacober, S. P.; Cardozo, M. G. A Novel Frequency Distribution Selection Method for Efficient Plate Layout of a Diverse Combinatorial Library. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1508−1516.

(14) Stanton, R. V.; Mount, J.; Miller, J. L. Combinatorial Library Design: Maximizing Model-Fitting Compounds within Matrix Synthesis Constraints. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 701−705.

(15) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

(16) Jamois, E. J.; Hassan, M.; Waldman. M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63−70.

(17) Jacobo-Molina, A.; Arnold, E. HIV Reverse transcriptase structure−function relationships. *Biochemistry* **1991**, *30*, 6351−6361.

(18) Garg, R.; Gupta, S. P.; Gao, H.; Babu, M. S.; Debnath, A. K.; Hansch, C. Comparative Quantitative Structure−Activity Relationship Studies on Anti-HIV Drugs. *Chem. Rev.* **1999**, *99*, 3525−3601.

(19) Spence, R. A.; Kati, W. M.; Anderson, K. S.; Johnson, K. A. Mechanism of Inhibition of HIV-1 Reverse Transcriptase by Nonnucleoside Inhibitors. *Science* **1995**, *267*, 988−993.

(20) De Clercq, E. The role of nonnucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Antiviral Res.* **1998**, *38*, 153−179.

(21) De Clercq, E. HIV Resistance to Reverse Transcriptase Inhibitors. *Biochem. Pharmacol.* **1994**, *47*, 155−169.

(22) Miyasaka, T.; Tanaka, H.; Baba, M.; Hayakawa, H.; Walker, R. T.; Balzarini, J.; De Clercq, E. A Novel Lead for Specific Anti-HIV-1 Agents: 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1989**, *32*, 2507−2509.

(23) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Balzarini, J.; De Clercq, E. A New Class of HIV-1-Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV-1 Activity of 5- or 6-Substituted Analogues of 1-[(2-Hydroxyethoxy)-methyl]- 6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349−357.

(24) Tanaka, H.; Baba, M.; Saito, S.; Miyasaka, T.; Takashima, H.; Sekiya, K.; Ubasawa, M.; Nitta, I.; Walker, R. T.; Nakashima, H.; De Clercq, E. Specific Anti-HIV-1 "Acyclonucleosides" Which Cannot Be Phosphorylated: Synthesis of Some Deoxy Analogues of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1991**, *34*, 1508−1511.

(25) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Anti-HIV Activity of 2-, 3-, and 4-Substituted Analogues of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 1394−1399.

(26) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Structure−Activity Relationships of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)-thymine Analogues: Effect of Substitutions at the C-6 Phenyl Ring and the C-5 Position on Anti-HIV-1 Activity. *J. Med. Chem.* **1992**, *35*, 337−345.

(27) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of Deoxy Analogues of 1-[(2-Hydroxyethoxy)-methyl]- 6-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1992**, *35*, 4713−4719.

(28) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Inouye, N.; Baba, Masanori; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of 6-Benzyl Analogues of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1995**, *38*, 2860−2865.

(29) Pontikis, R.; Benhida, R.; Aubertin, A.; Grierson, D.; Monneret, C. Synthesis and Anti-HIV Activity of Novel N-1 Side Chain-Modified Analogues of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1997**, *40*, 1845−1854.

(30) Luco, J. M.; Ferretti, F. H. QSAR Based on Multiple Linear Regression and PLS Methods for the Anti-HIV Activity of a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392−401.

(31) Gupta, S.; Singh, M.; Madan, A. K. Predicting anti-HIV activity: computational approach using a novel topological descriptor. *J. Comput-Aided Mol. Des.* **2001**, *15*, 671−678.

(32) Kireev, D. B.; Chretien, J. R.; Grierson, D. S.; Monneret, C. A 3D QSAR Study of a Series of HEPT Analogues: The Influence of Conformational Mobility on HIV-1 Reverse Transcriptase Inhibition. *J. Med. Chem.* **1997**, *40*, 4257−4264.

(33) Hannongbua, S.; Nivesanond, K.; Lawtrakul, L.; Pungpo, P.; Wolschann, P. 3D-Quantitative Structure−Activity Relationships of HEPT Derivatives as HIV-1 Reverse Transcriptase Inhibitors, Based on Ab Initio Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 848−855.

(34) Jalali-Heravi, M.; Parastar, F. Use of Artificial Neural Networks in a QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147−154.

(35) Garg, R.; Kurup, A.; Gupta, S. P. Quantitative structure−activity relationship studies on some acyclouridine derivatives acting as anti-HIV-1 drugs. *Quant. Struct.-Act. Relat.* **1997**, *16*, 20−24.

(36) Pascual, R.; Borrell, J. I.; Teixidó, J. Effective Methods for Combinatorial Sublibrary Selection. MGMS Meeting on "Structure-Based Drug Design"; Oxford, December, 2000.

(37) Sadowski, J.; Schwab, C. H.; Gasteiger, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(38) Cerius2, version 4.6; Molecular Simulations Inc.: 9685 Scranton Rd., San Diego, CA 92121.

(39) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.

(40) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205−1213.

(41) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, 1988; pp 119−138.

(42) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Application to Studies of X-ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559−564.

(43) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity-A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(44) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity − An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.

(45) Gasteiger, J.; Saller, H. Berechnung der Ladungsverteilung in konjungierten Systemen durch eine Quantifizierung des Mesomeriekonzepts. *Angew. Chem.* **1985**, *97*, 699−701. Calculation of Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem., Int.. Ed. Engl.* **1985**, *24*, 687−689.

(46) Sedgewick, R. *Algorithms*; Addison-Wesley: 1983. Baase, S. *Computer Algorithms*, *Introduction to Design and Analysis*; Addison-Wesley: 1988.

(47) Zheng, W. Z.; Waller, C. L.; Cho, S. J.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem. Inf. Comput. Sci.* **1963**, *39*, 738−746.

(48) Pearlman, R. S.; Wang, X. C.; Xu, Y.; Green, M. Novel methods for assessing and comparing the diversity of chemical libraries. 218th ACS meeting, New Orleans, August 22−26 1999.

(49) Free. S. M.; Wilson, J. W. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395.

(50) Kubinyi, *QSAR: Hansch Analysis and Related Approaches, Vol. 1 of Methods and Principles in Medicinal Chemistry*; Mannhold, R., et al., Eds.; VCH: Weinheim, 1993.

CI0255681