13rd International Electronic Conference on Synthetic Organic Chemistry (ECSOC-13), 1-30 November 2009.
http://www.mdpi.org/ecsoc-13/ & http://www.usc.es/congresos/ecsoc/13/

[G008]

Molecular Classification of Thiocarbamates with Cytoprotection Activity against Anti-human Immunodeficiency Virus

Francisco Torrens*,[1] and Gloria Castellano[2]

[1] Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, E-46071 València, Spain

[2] Instituto Universitario de Medio Ambiente y Ciencias Marinas, Universidad Católica de Valencia *San Vicente Mártir*, Guillem de Castro-94, E-46003 València, Spain

Classification algorithms are proposed based on *information entropy*. It is studied the molecular classification of anti-human immunodeficiency virus thiocarbamates. The 62 thiocarbamates (TCs) are classified by their structural chemical properties. Many classification algorithms are based on information entropy. An excessive number of results appear compatible with the data and suffer combinatorial explosion. However, after the *equipartition conjecture* one has a selection criterion. According to this conjecture, the best configuration of a flowsheet is that in which entropy production is most uniformly distributed. The *structural elements* of an inhibitor can be *ranked* according to their inhibitory activity in the order: $B_{1/2} > R > R_1 > R_2$ substitution. In TC 17, $B_{1/2} = B_1$, R = 4-CH_3 and $R_1 = R_2 = H$; its associated vector is unary. The TC 17 is selected as a *reference*. In some TCs $B_{1/2} = B_1$, in some others $B_{1/2} = B_2$. The analysis is in qualitative agreement with other classification taken as *good* based on *k*-means clustering. Program MolClas is a simple, reliable, efficient and fast procedure for molecular classification, based on the equipartition conjecture of entropy production. The structural elements allow the periodic classification of the TCs. A validation is performed with an external property, cytoprotection activity, not used in the development of the table.

Keywords: Periodic property; Periodic table; Periodic law; Classification; Information entropy; Equipartition conjecture; Cytoprotection; Thiocarbamate

1. Introduction

Nucleoside (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs) targeting the human immunodeficiency virus type 1 (HIV-1) encoded reverse transcriptase (RT)[1] must be proved effective in treating the HIV infection and acquired immunodeficiency syndrome (AIDS).[2] The NNRTIs bind to an *allosteric site* (non-nucleoside binding site, NNBS) largely contained within the RT p66 subunit, some 10Å from the polymerase *active site*.[3–14] Despite their chemical diversity, NNRTIs interact with the NNBS showing a similar three-dimensional arrangement, the so-called *butterfly-like conformation* typical of first-generation NNRTIs,[15] as demonstrated by X-ray crystallography of HIV-1 RT–NNRTI complexes.[16–24] However, the relatively unconserved amino-acid sequence of the NNBS favours the rapid selection of NNRTI-resistant viruses, both *in vitro* and *in vivo*.[25] As a result of single-point mutations in the NNBS,[26] first-generation NNRTIs, *e.g.*, nevirapine and delavirdine, show a loss of potency of several orders of magnitude. In contrast, second-generation NNRTIs, *e.g.*, efavirenz[27] and some thiocarboxanilide[28] and quinoxaline[29] derivatives, result in minor losses of activity against variants carrying either single or double NNRTI resistance mutations. Nevertheless, the fact that cross-resistance extends to the whole NNRTI class calls for development of new agents capable of inhibiting clinically relevant NNRTI-resistant mutants.

Ranise *et al*. described a novel class of NNRTIs, *i.e.*, O-substituted *N*-acyl-*N*-arylthiocarbamates (ATCs)[30] structurally related to *N*-phenethyl-*N'*-thiazolylthiourea (PETT) derivatives.[31,32] Among the ATCs, the phthalimidoethyl-ATCs proved to be potent inhibitors of the multiplication of wild-type (WT) HIV-1, significantly active against Y181C mutants but ineffective against K103R mutants. The thiocarbamate (TC) UC-38 was selected as an anti-HIV-1 agent in the early 1990s for pre-clinical development.[33] Ranise *et al*. described structure-based ligand design, synthetic strategy and structure–activity relationship (SAR) studies that led to the identification of TCs, a novel class of NNRTIs, isosteres of phenethylthiazolylthiourea (PETT) derivatives.[34] Assuming as a lead compound *O*-[2-(phthalimido)ethyl]-phenylthiocarbamate, one of the precursors of the previously described ATCs, they prepared two targeted solution-phase TC libraries by parallel synthesis. The lead optimization strategy led to nine *para*-substituted TCs, which were active against WT HIV-1 in MT-4-based assays at

nanomolar concentrations (50% effective concentration, EC_{50}, range: 0.04–0.01μM). The most potent congener (EC_{50} = 0.01μM) bears a methyl group at position 4 of the phthalimide moiety and a nitro group at the *para* position of the *N*-phenyl ring. Most of the TCs showed good selectivity indices, since no cytotoxic effect was detected at concentrations as high as 100μM. Five TCs significantly reduced the multiplication of the Y181C mutant, but they were inactive against K103R and K103N + Y181C mutants. Nevertheless, the fold increase in resistance of a TC was not greater than that of efavirenz against the K103R mutant in enzyme assays. Their docking model predictions were consistent with *in vitro* biological assays of the anti-HIV-1 activity of the TCs and related synthesized compounds. The *k*-means clustering of compounds using standardized descriptor matrix was taken as reference classification. The TCs are classified in three classes: class 1 (33–39,41–51,53,54), class 2 (1–3,5–9,11,13,15–19,22–28,30–32,56,58–61) and class 3 (4,10,12,14,20,21,29,40,52,55,57,62), *cf*. Figs. 1 and 2.

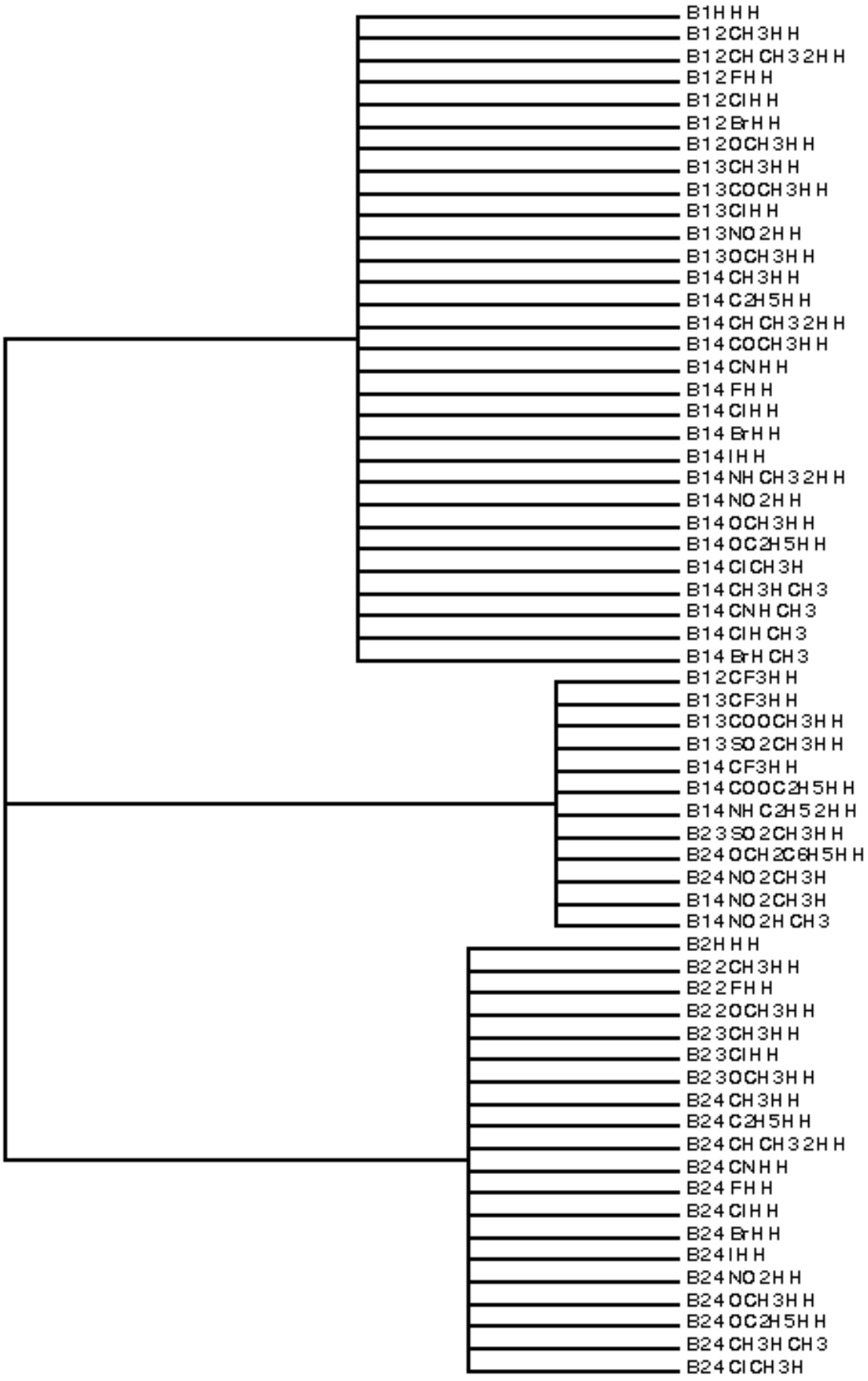


Fig. 1. Reference dendrogram of thiocarbamates with anti-HIV cycloprotection activity at level b_1.

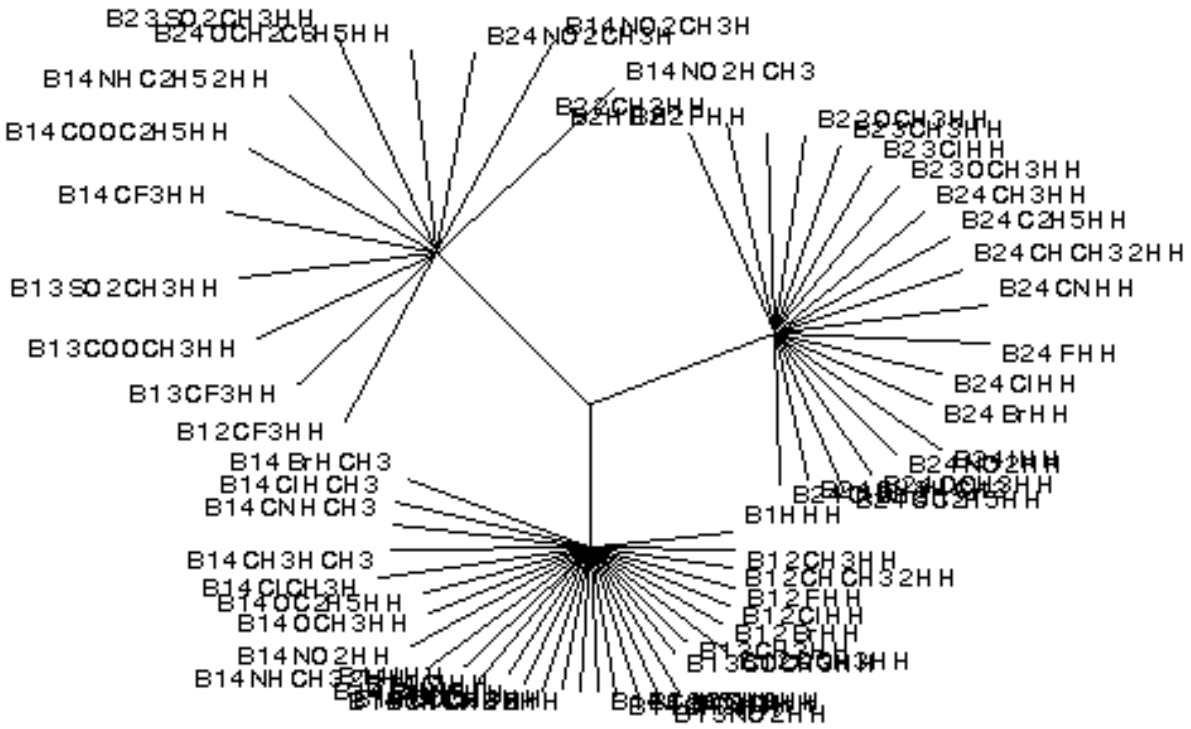


Fig. 2. Reference radial tree of thiocarbamates with anti-HIV cytoprotection activity.

A simple computerized algorithm, useful for establishing a relationship between chemical structures and their biological activities or significance, is proposed and exemplified.[35,36] The starting point is to use an informational or configurational entropy for pattern recognition purposes. The entropy is formulated on the basis of a *matrix of similarity* between two biochemical species. As entropy is weakly discriminating for classification purposes, the more powerful concepts of *entropy production* and its *equipartition conjecture* are introduced.[37] In earlier publications, the periodic classifications of local anaesthetics[38] and HIV inhibitors[39–41] were analyzed. The aim of the present report is to develop the learning potentialities of the code and, since molecules are more naturally described *via* a varying size structured representation, to study general approaches to the processing of structured information. A second goal is to present a periodic classification of the TCs. A further objective is to carry out a validation of the periodic table with an external property, cytoprotection activity, not used in the development of the table.

2. Classification Algorithm

The *grouping algorithm* uses the *stabilized* matrix of similarity, obtained by applying the *max–min composition rule o* defined by:

$$(\mathbf{R}\,\mathrm{o}\,\mathbf{S})_{ij} = \max_k \left[\min_k \left(r_{ik}, s_{kj}\right)\right] \tag{2}$$

where $\mathbf{R} = [r_{ij}]$ and $\mathbf{S} = [s_{ij}]$ are matrices of the same type, and $(\mathbf{R}\mathrm{o}\mathbf{S})_{ij}$ the (i,j)-*th* element of the matrix $\mathbf{R}\mathrm{o}\mathbf{S}$.[42–45]

It can be shown that when applying the max–min composition rule iteratively, so that $\mathbf{R}(n+1) = \mathbf{R}(n)$ o $\mathbf{R}$, there exists an integer n such that: $\mathbf{R}(n) = \mathbf{R}(n+1) = \ldots$ The resulting matrix $\mathbf{R}(n)$ is called the *stabilized similarity matrix*. The importance of stabilization lies in the fact that in the classification process, it will generate a partition into disjoint classes. From now on it is understood that the stabilized matrix is used and designated by $\mathbf{R}(n) = [r_{ij}(n)]$. The *grouping rule* is the following: i and

j are assigned to the same class if $r_{ij}(n) \geq b$. The class of i noted i' is the set of species j that satisfies the rule $r_{ij}(n) \geq b$. The matrix of classes is:

$$\mathbf{R}'(n) = \left[r'_{ij}\right] = \max_{s,t}(r_{st}) \quad (s \in i',\ t \in j') \tag{3}$$

where s stands for any index of a species belonging to the class i' (similarly for t and j'). Rule (3) means finding the largest similarity index between species of two different classes.

3. Information Entropy

In information theory, the *information entropy* h measures the surprise that the source emitting the sequences can give.[46,47] Consider the use of a qualitative spot test to determine the presence of iron in a water sample. Without any sample history the testing analyst must begin by assuming that the two outcomes, *viz*. 0 (Fe absent), and 1 (Fe present), are equiprobable with probabilities 1/2. When up to two metals may be present in the sample solution (*e.g*., Fe or Ni or both), there are four possible outcomes, ranging from neither (0, 0) to both being present (1, 1) with equal probabilities 1/4. Which of these four possibilities turns up can be determined using two tests, each having two observable states. Similarly with three elements there are eight possibilities each with a probability of $1/8 = 1/2^3$. Three tests are needed to resolve the question. The following pattern clearly relates the uncertainty and the information needed to resolve it. The number of possibilities is expressed to the power of 2. The power to which 2 must be raised to give the number of possibilities N is defined as the logarithm to base 2 of that number. Information and uncertainty can be defined, quantitatively, in terms of the logarithm to base 2 of the number of possible analytical outcomes: $I = H = \log_2 N$, where I indicates the amount of information, and H the amount of uncertainty. The initial uncertainty can also be defined in terms of the probability of the occurrence of each outcome; *e.g*., by referring to the probabilities above the following definition can be written: $I = H = \log_2 N = \log_2 1/p = -\log_2 p$, where I is the information contained in the answer given that there were N possibilities, H the initial uncertainty resulting from the need to consider the N possibilities, and p the probability of each outcome if all N possibilities are equally likely to occur. The

expression can be generalized to the situation in which the probability of each outcome is not the same. If one knows from past experience that some elements are more likely to be present than others, the equation is adjusted so that the logarithms of the individual probabilities, suitable weighted, are summed: $H = -\Sigma\, p_i \log_2 p_i$, where: $\Sigma\, p_i = 1$. Consider the original example, except that now past experience showed that 90% of the samples contained no iron. The degree of uncertainty is calculated using the equation as: $H = -(0.9 \log_2 0.9 + 0.1 \log_2 0.1)$ bits = 0.469 bits. In summary for a single event occurring with probability p the degree of surprise is proportional to $-\ln p$. Generalizing the result to a random variable X (which can take N possible values $x_1, \ldots, x_N$ with probabilities $p_1, \ldots, p_N$), the average surprise received on learning the value of X is $-\Sigma\, p_i \ln p_i$.

The information entropy associated with the matrix of similarity **R** is:

$$h(\mathbf{R}) = -\sum_{i,j} r_{ij} \ln r_{ij} - \sum_{i,j} \left(1 - r_{ij}\right) \ln\left(1 - r_{ij}\right) \tag{4}$$

Denote also by C_b the set of classes and by $\mathbf{R}'_b$ the matrix of similarity at the grouping level b. The information entropy satisfies the following properties.

1. $h(\mathbf{R}) = 0$ if $r_{ij} = 0$ or $r_{ij} = 1$.
2. $h(\mathbf{R})$ is maximum if $r_{ij} = 0.5$, *i.e.*, when the imprecision is maximum.
3. $h\left(\mathbf{R}'_b\right) \leq h(\mathbf{R})$ for any b, *i.e.*, classification leads to a loss of entropy.
4. $h\left(\mathbf{R}'_{b_1}\right) \leq h\left(\mathbf{R}'_{b_2}\right)$ if $b_1 < b_2$, *i.e.*, the entropy is a monotone function of the grouping level b.

4. The Equipartition Conjecture of Entropy Production

In the classification algorithm, each *hierarchical tree* corresponds to a dependence of entropy on the grouping level, and thus an h–b diagram can be obtained. The Tondeur and Kvaalen *equipartition conjecture of entropy production* is proposed as a selection criterion among different variants resulting from classification among hierarchical trees. According to the conjecture for a given charge or duty, the best configuration of a flowsheet is the one in which entropy production is most uniformly distributed, *i.e.*, closest to a kind of equipartition. One proceeds here by analogy using *information entropy* instead

of thermodynamic entropy. Equipartition implies a linear dependence, *i.e.*, a constant production of entropy along the b scale, so that the *equipartition line* is described by:

$$h_{\mathrm{eqp}} = h_{\max} b \tag{5}$$

Since the classification is discrete, a way of expressing equipartition would be a regular staircase function. The best variant is chosen to be that minimizing the sum of squares of the deviations:

$$SS = \sum_{b_i} \left(h - h_{\mathrm{eqp}}\right)^2 \tag{6}$$

5. Learning Procedure

Learning procedures similar to those encountered in *stochastic methods* are implemented as follows.[48] Consider a given partition into classes as *good* or ideal from practical or empirical observations, which corresponds to a *reference* similarity matrix $\mathbf{S} = [s_{ij}]$ obtained for equal weights $a_1 = a_2 = \ldots = a$ and for an arbitrary number of fictious properties. Next consider the same set of species as in the good classification and the actual properties. The similarity degree r_{ij} is then computed with Equation (1) giving the matrix **R**. The number of properties for **R** and **S** may differ. The learning procedure consists in trying to find classification results for **R**, as close as possible to the *good* classification. The first weight a_1 is taken constant and only the following weights a_2, a_3,… are subjected to random variations. A new similarity matrix is obtained using Equation (1) and the new weights.

The distance between the partitions into classes characterized by **R** and **S** is given by:

$$D = -\sum_{ij} \left(1 - r_{ij}\right) \ln \frac{1 - r_{ij}}{1 - s_{ij}} - \sum_{ij} r_{ij} \ln \frac{r_{ij}}{s_{ij}} \qquad \forall\, 0 \le r_{ij}, s_{ij} \le 1 \tag{7}$$

The definition was suggested by that introduced in information theory by Kullback to measure the distance between two probability distributions.[49] In the present case it is a measure of the distance between matrices **R** and **S**. Since for every matrix there is a corresponding classification, the two classifications will be compared by the distance. The D is a nonnegative quantity that approaches zero as the resemblance between **R** and **S** increases.

The result of the algorithm is a set of weights allowing adequate classification. The procedure was applied to the synthesis of complex flowsheets using information entropy.[50]

Our program MolClas is a simple, reliable, efficient and fast procedure for molecular classification, based on the equipartition conjecture of entropy production according to Equations (1) to (7). It reads the number of properties and the molecular properties. MolClas allows the optimization of the coefficients. It optionally reads the starting coefficients and the number of iteration cycles. The correlation matrix can be either calculated by the program or read from the input file. MolClas allows the transformation of the correlation matrix from the range [–1, 1] to [0, 1]. It calculates the similarity matrix of the properties in symmetric storage mode, calculates the classifications, tests if the classifications are different, calculates the distances between classifications, calculates the similarity matrices of the classifications, calculates the information entropy of classifications, optimizes the coefficients, performs both single- and complete-linkage hierarchical cluster analyses, and plots the cluster diagrams. Molclas was written not only to analyze the equipartition conjecture of entropy production, but also to explore the world of molecular classification.

6. Calculation Results and Discussion

The cytoprotection data of anti-human immunodeficiency virus type 1 (HIV-1) TCs reported by Ranise *et al*. were used as the model dataset: the cytoprotection data [EC_{50} (μM)] of substituted TCs were converted to the logarithmic scale [pEC_{50}, (EC_{50} in mM)] and then used for subsequent classification analyses based on molecular structure. The *k*-means clustering of compounds using standardized descriptor matrix, by Mitra *et al*., was taken as reference classification. They classify the TCs in three classes: class 1 (33–39,41–51,53,54), class 2 (1–3,5–9,11,13,15–19,22–28,30–32,56,58–61) and class 3 (4,10,12,14,20,21,29,40,52,55,57,62).

The Pearson correlation coefficient matrix has been calculated between the pairs of vector properties $<i_1,i_2,i_3,i_4>$ of the 62 TCs. The Pearson intercorrelations are illustrated in the partial correlation diagram, which contains high ($r \geq 0.75$), medium ($0.50 \leq r < 0.75$), low ($0.25 \leq r < 0.50$) and

zero ($r < 0.25$) partial correlations. Pairs of inhibitors with high partial correlations show a similar vector property. However, the results should be taken with care, because the 16 TCs with constant <1111> vector (Entries 17–32) show null standard deviation, causing high partial correlations ($r = 1$) with any inhibitor, which is an artifact. With the equipartition conjecture the intercorrelations are illustrated in the partial correlation diagram, which contains 506 high, 488 medium (*orange*), 473 low (*yellow*) and 424 *zero* (*black*) partial correlations. Notice that 624 out of 976 (16×39/61) high partial correlations of Entries 17–30 were corrected; *e.g.*, for Entry 17 the correlations with Entries 1–16 are medium, its correlations with Entries 41–55 are low and its correlations with Entries 33–40 are *zero* partial correlations.

The grouping rule in the case with equal weights $a_k = 0.5$ for $0.88 \le b_1 \le 0.93$ allows the classes:

C–b_1 = (1–16)(17–32)(33–40)(41–52)(53)(54,55)(56,57)(58–62)

The eight classes are obtained with the associated entropy h–**R**–b_1 = 32.66. The dendrogram (binary tree)[51–53] matching to <i_1,i_2,i_3,i_4> and C–b_1 is calculated;[54] it provides a binary taxonomy of Table 1, which separates the same eight classes: the data bifurcates into classes 5, 1–4, 6–8 with 1, 16, 16, 8, 12, 2, 2 and 5 TCs, respectively. In particular TC 17, 27, *etc*. with the greatest cytoprotection activity are grouped into the same class. The TCs belonging to the same class appear highly correlated in the partial correlation diagram, in qualitative agreement with the reference clustering.

Table 1. Vector properties of anti-HIV thiocarbamates for molecular substitutions ($B_{1/2}$, R, R_1, R_2).

1. –B_1 –H –H –H <1011>	32. –B_1 4-OC_2H_5 –H –H <1111>
2. –B_1 2-CH_3 –H –H <1011>	33. –B_2 –H –H –H <0011>
3. –B_1 2-$CH(CH_3)_2$ –H –H <1011>	34. –B_2 2-CH_3 –H –H <0011>
4. –B_1 2-CF_3 –H –H <1011>	35. –B_2 2-F –H –H <0011>
5. –B_1 2-F –H –H <1011>	36. –B_2 2-OCH_3 –H –H <0011>
6. –B_1 2-Cl –H –H <1011>	37. –B_2 3-CH_3 –H –H <0011>
7. –B_1 2-Br –H –H <1011>	38. –B_2 3-Cl –H –H <0011>
8. –B_1 2-OCH_3 –H –H <1011>	39. –B_2 3-OCH_3 –H –H <0011>
9. –B_1 3-CH_3 –H –H <1011>	40. –B_2 3-SO_2-CH_3 –H –H <0011>
10. –B_1 3-CF_3 –H –H <1011>	41. –B_2 4-CH_3 –H –H <0111>
11. –B_1 3-$COCH_3$ –H –H <1011>	42. –B_2 4-C_2H_5 –H –H <0111>
12. –B_1 3-$COOCH_3$ –H –H <1011>	43. –B_2 4-$CH(CH_3)_2$ –H –H <0111>
13. –B_1 3-Cl –H –H <1011>	44. –B_2 4-CN –H –H <0111>
14. –B_1 3-SO_2-CH_3 –H –H <1011>	45. –B_2 4-F –H –H <0111>

15. $-B_1$ 3-NO_2 –H –H <1011>	46. $-B_2$ 4-Cl –H –H <0111>
16. $-B_1$ 3-OCH_3 –H –H <1011>	47. $-B_2$ 4-Br –H –H <0111>
17. $-B_1$ 4-CH_3 –H –H <1111>	48. $-B_2$ 4-I –H –H <0111>
18. $-B_1$ 4-C_2H_5 –H –H <1111>	49. $-B_2$ 4-NO_2 –H –H <0111>
19. $-B_1$ 4-$CH(CH_3)_2$ –H –H <1111>	50. $-B_2$ 4-OCH_3 –H –H <0111>
20. $-B_1$ 4-CF_3 –H –H <1111>	51. $-B_2$ 4-OC_2H_5 –H –H <0111>
21. $-B_1$ 4-$COOC_2H_5$ –H –H <1111>	52. $-B_2$ 4-$OCH_2C_6H_5$ –H –H <0111>
22. $-B_1$ 4-$COCH_3$ –H –H <1111>	53. $-B_2$ 4-CH_3 –H –CH_3 <0110>
23. $-B_1$ 4-CN –H –H <1111>	54. $-B_2$ 4-Cl –CH_3 –H <0101>
24. $-B_1$ 4-F –H –H <1111>	55. $-B_2$ 4-NO_2 –CH_3 –H <0101>
25. $-B_1$ 4-Cl –H –H <1111>	56. $-B_1$ 4-Cl –CH_3 –H <1101>
26. $-B_1$ 4-Br –H –H <1111>	57. $-B_1$ 4-NO_2 –CH_3 –H <1101>
27. $-B_1$ 4-I –H –H <1111>	58. $-B_1$ 4-CH_3 –H –CH_3 <1110>
28. $-B_1$ 4-$NH(CH_3)_2$ –H –H <1111>	59. $-B_1$ 4-CN –H –CH_3 <1110>
29. $-B_1$ 4-$NH(C_2H_5)_2$ –H –H <1111>	60. $-B_1$ 4-Cl –H –CH_3 <1110>
30. $-B_1$ 4-NO_2 –H –H <1111>	61. $-B_1$ 4-Br –H –CH_3 <1110>
31. $-B_1$ 4-OCH_3 –H –H <1111>	62. $-B_1$ 4-NO_2 –H –CH_3 <1110>

At level b_2 with $0.82 \leq b_2 \leq 0.87$ the set of classes turns out to be:

C–b_2 = (1–16)(17–32,58–62)(33–40)(41–53)(54,55)(56,57)

Six classes result in this case and the entropy decreases to h–**R**–b_2 = 18.02. The dendrogram matching to $<i_1,i_2,i_3,i_4>$ and C–b_2 divides the same six classes: 1–6 with 16, 21, 8, 13, 2 and 2 TCs, respectively. Again TC 17, 27, *etc*. with greater cytoprotection activity are grouped into the same class. The TCs belonging to the same class appear highly correlated in the partial correlation diagram, in qualitative agreement with the reference clustering and previous results.

At level b_3 with $0.69 \leq b_3 \leq 0.81$ the set of classes results:

C–b_3 = (1–16)(17–32,56–62)(33–40)(41–55)

Four classes result and the entropy decreases to h–**R**–b_3 = 8.09. The dendrogram matching to $<i_1,i_2,i_3,i_4>$ and C–b_3 is computed; it provides a binary taxonomy of Table 1, which splits the same four classes: 1–4 with 16, 23, 8 and 15 TCs, respectively. Once more TC 17, 27, *etc*. with the greatest cytoprotection activity are grouped into the same class. The TCs belonging to the same class appear highly correlated in the partial correlation diagram, in qualitative agreement with the reference clustering and previous results.

At level b_4 with $0.44 \leq b_4 \leq 0.56$ the set of classes is:

C–b_4 = (1–32,56–62)(33–55)

Two classes result and the entropy decreases to h–**R**–b_4 = 1.84. The dendrogram matching to <i_1,i_2,i_3,i_4> and C–b_4 separates the same two classes: 1–2 with 39 and 23 TCs, respectively. One more time TC 17, 27, *etc*. with the greatest cytoprotection activity are grouped into the same class. The TCs belonging to the same class appear highly correlated in the partial correlation diagram, in qualitative agreement with the reference clustering and previous results.

A comparative analysis of the set containing 1–62 classes, in agreement with previous results.

In view of the previous partial correlation diagram and dendrograms we suggest to split the data into three classes: class 1 (1–16), class 2 (17–32,56–62) and class 3 (33–55). The dendrogram shows, again, that TC 17, 27, *etc.* are grouped into the same class. The results are in qualitative agreement with the reference clustering, corresponding class 1 with cluster 2, class 2 with cluster 1 and class 3 with cluster 3.

The illustration of the classification above in a radial tree shows the same classes, in qualitative agreement with partial correlation diagram, dendrogram and previous results. Once more TC17, 27, *etc*. are grouped into the same class.

SplitsTree is a program for analyzing cluster analysis (CA) data.[55] Based on the method of *split decomposition*, it takes as input a *distance matrix* or a set of CA data and produces as output a graph, which represents the relationships between the taxa. For ideal data this graph is a tree whereas less ideal data will give rise to a tree-like network, which can be interpreted as possible evidence for different and conflicting data. Furthermore as split decomposition does not attempt to force data onto a tree, it can provide a good indication of how *tree*-like given data are. The splits graph for the 64 TCs in Table 1 reveals no conflicting relationship between the inhibitors. Most groups of TCs appear superimposed, *viz*. 1–16, 17–32, 33–40, 41–52, 54–55, 56–57, and 58–62. The splits graph is in qualitative agreement with partial correlation diagram, dendrograms, radial tree and previous results.

Usually, in quantitative structure–property relationship (QSPR) studies, the data file contains less than one hundred objects and several thousands of *X*-variables. In fact, there are so many *X*-variables

that no one can discover by *inspection* patterns, trends, clusters, *etc*. in the objects. *Principal components analysis* (PCA) is a technique extremely useful to *summarize* all the information contained in the **X**-matrix and put it in a form understandable by human beings.[56–61] The PCA works by decomposing the **X**-matrix as the product of two smaller matrices **P** and **T**. The loading matrix (**P**) with information about the variables contains a few vectors, the so-called Principal Components (PC), which are obtained as linear combinations of the original *X*-variables. The score matrix (**T**), with information about the objects, is such that each object is described in terms of their projections onto the PCs, instead of the original variables: $\mathbf{X} = \mathbf{TP}' + \mathbf{E}$. The information not contained in the matrices remains as *unexplained* X-*variance* in a residual matrix (**E**). Every PC_i is a new coordinate expressed as a linear combination of the old features x_j: $PC_i = \sum_j b_{ij} x_j$. The new coordinates PC_i are called scores or factors while coefficients b_{ij} are called loadings. The scores are ordered according to their information content with regard to the total variance among all objects. The score–score plots show the positions of compounds in the new coordinate system, while loading–loading plots show the position of features that represent compounds in the new coordinate system. PCs have two interesting properties. (1) They are extracted in decreasing order of importance. The first PC always contains more information than the second does, the second more than the third, *etc*. (2) Every PC is orthogonal to each other. There is absolutely no correlation between the information contained in different PCs.

A PCA was carried out for the TCs. The importance of the PCA factors F_1–F_4 for $\{i_1,i_2,i_3,i_4\}$ is calculated. In particular the use of only the first factor F_1 explains 32% of the variance (68% of the error); the combined use of the first two factors $F_{1–2}$ explains 62% of the variance (38% error); the use of the first three factors $F_{1–3}$ explains 85% of the variance (15% error).

The PCA factor loadings are computed.

The PCA $F_{1–4}$ profile for the vector property is calculated. In particular for F_1 and F_4, variable i_2 has the greatest weight in the profile; however F_1 cannot be reduced to two variables $\{i_2,i_4\}$ without a 13% error. For F_2 variable i_1 has the greatest weight; notwithstanding F_2 cannot be reduced to two variables

$\{i_1,i_3\}$ without a 28% error. For F_3 variable i_1 has the greatest weight; nevertheless F_3 cannot be reduced to two variables $\{i_1,i_3\}$ without an 8% error. Factors $F_{1–4}$ can be considered as linear combinations of $\{i_2,i_4\}$, $\{i_1,i_3\}$, $\{i_1,i_3\}$ and $\{i_2,i_4\}$ with 13%, 28%, 8% and 23% errors, respectively.

In the PCA F_2–F_1 scores plot, the TCs with the same vector property appear superimposed. Three classes of TCs are clearly distinguished in agreement with the reference clustering, *viz*. class 1 with 16 compounds ($F_1 > F_2 > 0$), class 2 with 23 substances ($F_1 < F_2$), and class 3 with 23 molecules ($0 \approx F_1 > F_2$). The classification is in qualitative agreement with the partial correlation diagram, dendrograms, radial tree, splits graph and previous results.

From the PCA factor loadings of the TCs, the F_2–F_1 loadings plot depicts the four properties. In addition as a complement to the scores plot for the loadings, it is confirmed that the TCs in class 1 present a contribution of R_1 = –H, situated on the same side. The TCs in class 2 have more pronounced contributions from $B_{1/2} = B_1$. Finally TCs in class 3 present a contribution of R = 4-substitution and R_2 = –H. Two classes of properties are clearly distinguished in the loadings plot, *viz*. class 1 $\{R,R_2\}$ ($F_1 > F_2$), and class 2 $\{B_{1/2},R_1\}$ ($F_1 < F_2$).

Instead of 62 TCs in the $\Re^4$ space of four vector properties consider four properties in the $\Re^{62}$ space of 62 TCs. The dendrogram for the vector properties separates first properties R and R_2 (class 1) and, finally, properties $B_{1/2}$ and R_1 (class 2), in agreement with PCA the loadings plot.

The radial tree for the vector properties separates the same classes as the PCA loadings plot and dendrogram.

The splits graph for the properties reveals no conflicting relationship between the vector components and is in agreement with the PCA loadings plot, binary and radial trees.

A PCA was performed for the vector properties. The use of only the first factor F_1 explains 43% of the variance (57% of the error); the combined use of the first two factors $F_{1–2}$ explains 64% of the variance (36% error); the use of the first three factors $F_{1–3}$ explains 82% of the variance (18% error), *etc*. In the PCA F_2–F_1 scores plot, R_2 (class 1) appears superimposed on R_1 (class 2). Two classes of

properties are distinguished, *viz*. class 1 {R, R_2} ($F_1 < F_2$), and class 2 {$B_{1/2}$, R_1} ($F_1 > F_2$), in agreement with the PCA loadings plot, binary and radial trees and splits graph.

In the recommended format for the periodic table (PT) of the TCs they are classified first by i_4, then by i_3, i_2 and, finally, by i_1. Periods of four units are assumed; *e.g*., group g00 stands for $<i_1,i_2> = <00>$, *viz*. <0011> (–B_2 –H –H –H, *etc*.), *etc*. Those inhibitors in the same column appear close in the partial correlation diagram, dendrograms, radial tree, splits graph, PCA scores and previous results.

It is calculated the variation of property P (cycloprotection activity against HIV-1) of vector $<i_1,i_2,i_3,i_4>$ *vs*. structural parameters $\{i_1,i_2,i_3,i_4\}$ for the TCs. This property was not used in the development of the PT and serves to validate it. The results agree with a PT of properties with vertical groups defined by $\{i_1,i_2\}$ and horizontal periods described by $\{i_3,i_4\}$.

The variation of property P of vector $<i_1,i_2,i_3,i_4>$ *vs*. the number of the group in PT for the TCs reveals or extrapolates minima corresponding to TCs with $<i_1,i_2>$ *ca*. <00> (group g00). The p1, p10 and p11 represent rows 1, 2 and 3. The $P(i_1,i_2,i_3,i_4)$ corresponding function denotes a series of *waves* clearly limited by maxima or minima, which suggest a periodic behaviour that recalls the form of a trigonometric function. For $<i_1,i_2,i_3,i_4>$ a minimum is clearly shown. The distance in $<i_1,i_2,i_3,i_4>$ units between each pair of consecutive minima is four, which coincides with the TC sets in the successive periods. The minima occupy analogous positions in the curve and are in phase. The representative points in phase should correspond to the elements of the same group in PT. For the $<i_1,i_2,i_3,i_4>$ minima there is coherence between the two representations; however the consistency is not general. The comparison of the waves shows two differences: (1) periods 1–2 are incomplete and (2) period 3 is somewhat sawtooth-like. The most characteristic points of the plot are the minima, which lie about group g00. The values of $<i_1,i_2,i_3,i_4>$ are repeated as the periodic law (PL) states.

An empirical function $P(p)$ reproduces the different $<i_1,i_2,i_3,i_4>$ values. A minimum of $P(p)$ has meaning only if it is compared with the former $P(p–1)$ and later $P(p+1)$ points, needing to fulfil:

$$P_{\min}(p) < P(p-1)$$
$$P_{\min}(p) < P(p+1) \qquad (8)$$

Order relations (8) should repeat at determined intervals equal to the period size and are equivalent to:

$$P_{\min}(p) - P(p-1) < 0$$

$$P(p+1) - P_{\min}(p) > 0 \tag{9}$$

As relations (9) are valid only for minima more general others are desired for all values of p. The $D(p) = P(p+1) - P(p)$ differences are calculated by assigning each of their values to TC p:

$$D(p) = P(p+1) - P(p) \tag{10}$$

Instead of $D(p)$ the values of $R(p) = P(p+1)/P(p)$ can be taken by assigning them to TC p. If PL were general the elements in the same group in analogous positions in different waves would satisfy:

$$D(p) > 0 \text{ or } D(p) < 0 \tag{11}$$

$$R(p) > 1 \text{ or } R(p) < 1 \tag{12}$$

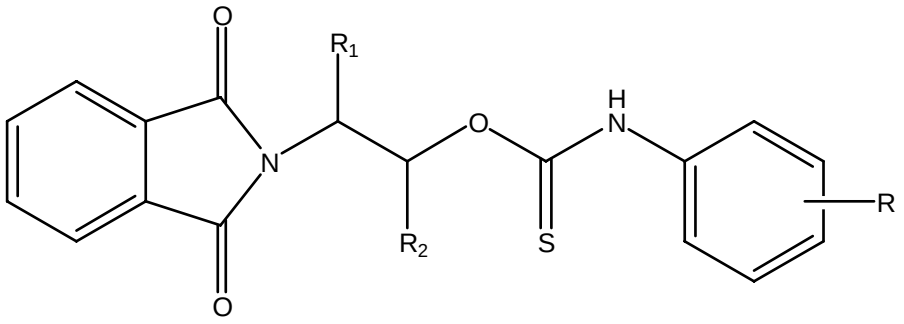
However the results show that this is not the case so that PL is not general existing some anomalies; *e.g.*, the variation of $D(p)$ *vs*. group number presents lack of coherence between the $<i_1,i_2,i_3,i_4>$ Cartesian and PT representations. If consistency were rigorous all the points in each period would have the same sign. In general, there is a trend in the points to give $D(p) < 0$ for the lower groups but not for the greater groups. In detail, however, there are irregularities in which the TCs for successive periods are not always in phase.

The change of $R(p)$ *vs*. group number confirms the lack of constancy between the Cartesian and PT charts. If steadiness were exact, all the points in each period would show $R(p)$ either lesser or greater than one. There is a trend in the points to give $R(p) < 1$ for the lower groups but not for the greater groups. Notwithstanding, there are confirmed incongruities in which the TCs for successive waves are not always in phase.

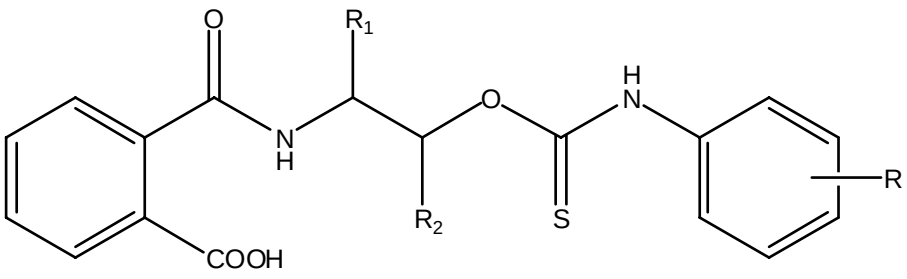
7. Computational Method

The key problem in classification studies is to define *similarity indices*, when several criteria of comparison are involved. The first step in quantifying the concept of similarity, for molecules of TCs, is

to list the most important portions of such molecules. Furthermore the *vector of properties* $\bar{i} = \langle i_1, i_2, \ldots i_k, \ldots \rangle$ should be associated with each inhibitor i, whose components correspond to different characteristic groups in the TC molecule, in a hierarchical order according to the expected importance of their pharmacological potency. If the *m-th* portion of the molecule is pharmacologically more significant for the inhibitory effect than the *k-th* portion, then $m < k$. The components i_k are "1" or "0", according to whether a similar (or identical) portion of rank k is present or absent in TC i, compared with the reference TC. The analysis includes four regions of structural variations in the TC molecules: one is the R position on the phenyl ring (showing diverse substitution pattern), and the remaining are R_1, R_2 and $B_{1/2}$ positions (showing limited substitution pattern, *cf*. Fig. 3). It is assumed that the *structural elements* of a TC molecule can be *ranked*, according to their contribution to inhibitory potency in the following order of decreasing importance: $B_{1/2} > R > R_1 > R_2$. Index $i_1 = 1$ denotes $B_{1/2} = B_1$ (0 for $B_{1/2} = B_2$), $i_2 = 1$ denotes 4-substitution on the phenyl ring, $i_3 = 1$ denotes $R_1 = H$ and $i_4 = 1$ denotes $R_2 = H$. In some inhibitors $B_{1/2} = B_1$, in some others $B_{1/2} = B_2$. In TC 17 $B_{1/2} = B_1$, $R = 4\text{–}CH_3$ and $R_1 = R_2 = H$. Obviously its associated vector is <1111>. In this study, TC 17 was selected as a *reference* because of its maximum cytoprotection activity against HIV-1.



(a)



(b)

Fig. 3. Molecular structure of an anti-HIV thiocarbamate molecule: (a) B_1 and (b) B_2.

Table 1 contains the vectors associated with 62 TCs. Vector <1011> is associated with TC 1 since $B_{1/2} = B_1$, and $R = R_1 = R_2 = H$. Let us denote by r_{ij} $(0 \le r_{ij} \le 1)$ the similarity index of two TCs associated with the $\bar{i}$ and $\bar{j}$ vectors, respectively. The relation of similitude is characterized by a *similarity matrix* $\mathbf{R} = [r_{ij}]$. The similarity index between two TCs $\bar{i} = \langle i_1, i_2, \ldots i_k \ldots \rangle$ and $\bar{j} = \langle j_1, j_2, \ldots j_k \ldots \rangle$ is defined as:

$$r_{ij} = \sum_k t_k (a_k)^k \quad (k = 1, 2, \ldots) \tag{1}$$

where $0 \le a_k \le 1$ and $t_k = 1$ if $i_k = j_k$ but $t_k = 0$ if $i_k \ne j_k$. The definition assigns a weight $(a_k)^k$ to any property involved in the description of molecule i or j.

8. Conclusions

From the present results and discussion the following conclusions can be drawn.

1. Several criteria, selected to reduce the analysis to a manageable quantity of structures from the set of thiocarbamates, refer to the structural parameters related with the R position on the phenyl ring and the R_1, R_2 and $B_{1/2}$ positions. Many algorithms for classification are based on *information entropy*. For sets of moderate size an excessive number of results appear compatible with the data, and the number suffers a combinatorial explosion. However after the *equipartition conjecture*, one has a selection criterion between different variants resulting from classification between hierarchical trees. According to the conjecture, the best configuration of a flowsheet is the one in which the entropy production is most uniformly distributed. The method avoids the problem of other methods of continuum variables because, for the 16 compounds with constant <1111> vector, the null standard deviation always causes a Pearson correlation coefficient of $r = 1$. The lower-level classification processes show lower entropy.

2. Program MolClas is a simple, reliable, efficient and fast procedure for molecular classification, based on the equipartition conjecture of entropy production. It was written not only to analyze the equipartition conjecture of entropy production, but also to explore the world of molecular classification.

3. In this study we classified a new class of non-nucleoside reverse transcriptase inhibitor thiocarbamate isosteres of phenethylthiazolylthiourea derivatives. The biological results show that the ring-closed thiocarbamates bearing *para* substituents on the *N*-phenyl ring, *e.g.*, methyl, iodo, chloro, bromo, nitro and methoxy, were potent inhibitors, but maximum potency was reached by introducing an additional methyl group at the 4-position of the phthalimide framework in a *p*-nitro ring-closed thiocarbamate. In terms of resistance against the clinically relevant mutations, the major molecular flexibility of the thiocarbamates with regard to phenethylthiazolylthiourea derivatives did not give the eagerly awaited results. Nevertheless, the significant activity of a thiocarbamate (50% inhibitory concentration 2.3μM) against the K103R mutant in enzyme assays and of five thiocarbamates against the Y181C in cell-based assays offers a stimulus for the design of new thiocarbamate analogues with better resistance profile.

4. The good comparison of our classification results, with other clustering taken as *good*, confirm the adequacy of the cytoprotection activity for the molecular structures of the thiocarbamates. Information entropy and principal component analyses permit classifying the compounds and agree. The substances are grouped into different classes. In general the three classical clusters are recognized.

5. Classification algorithms are proposed based on *information entropy*. The 62 thiocarbamates are classified by structural chemical properties. The analysis includes four regions of structural variations in the thiocarbamate molecules: the R position on the phenyl ring and the R_1, R_2 and $B_{1/2}$ positions. The *structural elements* of a thiocarbamate molecule can be *ranked* according to their cytoprotection activity in the order: $B_{1/2} > R > R_1 > R_2$. In thiocarbamate 17, $B_{1/2} = B_1$, R = 4–CH_3 and $R_1 = R_2$ = –H; its associated vector is <1111>. Thiocarbamate 17 was selected as a *reference*. The examination is in agreement with principal component analysis, comparing well with other clustering taken as *good*.

6. The periodic law has not the rank of the laws of physics: (1) the cytoprotection activities of the thiocarbamates are not repeated; perhaps their chemical character; (2) the order relationships are repeated with exceptions. The analysis forces the statement: The relationships that any thiocarbamate *p* has with its neighbour $p + 1$ are approximately repeated for each period. Periodicity is not general;

however if a natural order of the compounds is accepted the law must be phenomenological. The cytoprotection activity was not used in the generation of the periodic table and serves to validate it.

7. The representation of other properties of the thiocarbamates in the periodic table would give an insight into the possible generality of this table.

References

[1] H. Jonckheere, J. Anne and E. de Clercq, Med. Res. Rev., 20 (2000) 129.

[2] E. de Clercq, Farmaco, 56 (2001) 3.

[3] M. Artico, Farmaco, 51 (1996) 305.

[4] T. J. Tucker, W. C. Lumma and J. C. Culberson, Methods Enzymol., 275 (1996) 440.

[5] E. de Clercq, J. Med. Chem., 38 (1995) 2491.

[6] E. de Clercq, Clin. Microbiol. Rev., 8 (1995) 200.

[7] E. de Clercq, Antiviral Res., 38 (1998) 153.

[8] E. de Clercq, Collect. Czech. Chem. Commun., 63 (1998) 449.

[9] E. de Clercq, Expert Opin. Invest. Drugs, 3 (1994) 253.

[10] O. S. Pedersen and E. B. Pedersen, Antiviral Chem. Chemother., 10 (1999) 285.

[11] M. Artico, Drugs Future, 27 (2002) 159.

[12] O. S. Pedersen and E. B. Pedersen, Synthesis, 4 (2000) 479.

[13] E. de Clercq, Chem. Biodiversity, 1 (2004) 44.

[14] J. Balzarini, Curr. Top. Med. Chem, 4 (2004) 921.

[15] G. Tachedjian y S. P. Goff, Curr. Opin. Invest. Drugs, 4 (2003) 966.

[16] J. Lindberg, S. Sigurdsson, S. Lowgren, H. O. Andersson, C. Sahlberg, R. Noreen, K. Fridborg, H. Zhang and T. Unge, Eur. J. Biochem., 269 (2002) 1670.

[17] J. Ren, J. Diprose, J. Warren, R. M. Esnouf, L. E. Bird, S. Ikemizu, M. Slater, J. Milton, J. Balzarini, D. I. Stuart and D. K. Stammers, J. Biol. Chem., 275 (2000) 5633.

[18] J. Ren, J. Milton, K. L. Weaver, S. A. Short, D. I. Stuart and D. K. Stammers, Structure, 8 (2000) 1089.

[19] J. Ren, R. M. Esnouf, A. L. Hopkins, D. I. Stuart and D. K. Stammers, J. Med. Chem., 42 (1999) 3845.

[20] A. L. Hopkins, J. Ren, R. M. Esnouf, B. E. Willcox, E. Y. Jones, C. Ross, T. Miyasaka, R. T. Walker, H. Tanaka, D. K. Stammers and D. I. Stuart, J. Med. Chem., 39 (1996) 1589.

[21] J. Ding, K. Das, C. Tantillo, W. Zhang, A. D. J. Clark, S. Jessen, V. Lu, Y. Hsiou, A. Jacobo-Molina, K. Andries, R. Pauwels, H. Moereels, L. Koymans, P. A. J. Janssen, R. H. J. Smith, M. K. Koepke, C. J. Michejda, S. H. Hughes and E. Arnold, Structure, 3 (1995) 365.

[22] J. Ding, K. Das, H. Moereels, L. Koymans, K. Andries, P. A. Janssen, S. H. Hughes and E. Arnold, Nat. Struct. Biol., 2 (1995) 407.

[23] J. Ren, R. Esnouf, E. Garman, D. Somers, C. Ross, I. Kirby, J. Keeling, G. Darby, Y. Jones, D. Stuart and D. Stammers, Nat. Struct. Biol., 2 (1995) 293.

[24] J. Ren, R. Esnouf, A. Hopkins, C. Ross, Y. Jones, D. Stammers and D. Stuart, Structure, 3 (1995) 915.

[25] D. D. Richman, D. Havlir, J. Corbeil, D. Looney, C. Ignacio, S. A. Spector, J. Sullivan, S. Cheeseman, K. Barringer, D. Pauletti, C. K. Shih, M. Mayers, J. Griffin, J. Virol., 68 (1994) 1660.

[26] R. F. Schinazi, B. A. Larder and J. W. Mellors, Int. Antiviral News, 5 (1997) 129.

[27] S. D. Young, S. F. Britcher, L. O. Tran, L. S. Payne, W. C. Lumma, T. A. Lyle, J. R. Huff, P. S. Anderson, D. B. Olsen, S. S. Carroll and E. A. Emini, Antimicrob. Agents Chemother., 39 (1995) 2602.

[28] J. Balzarini, H. Polemans, S. Aquaro, C. F. Perno, M. Witvrouw, D. Schols, E. de Clercq and A. Karlsson, Mol. Pharmacol., 50 (1996) 394.

[29] J. P. Kleim, R. Bender, U. M. Billhardt, C. Meichsner, G. Riess, M. Rosner, I. Winkler and A. Paessens, Antimicrob. Agents Chemother., 37 (1993) 1659.

[30] A. Ranise, A. Spallarossa, S. Schenone, O. Bruno, F. Bondavalli, L. Vargiu, T. Marceddu, M. Mura, P. La Colla and A. Pani, J. Med. Chem., 46 (2003) 768.

[31] A. S. Cantrell, P. Engelhardt, M. Hogberg, S. R. Jaskunas, N. G. Johansson, C. L. Jordan, J. Kangasmetsa, M. D. Kinnick, P. Lind, J. M. Morin Jr., M. A. Muesing, R. Noreen, B. Oberg, P. Pranc, C. Sahlberg, R. J. Ternansky, R. T. Vasileff, L. Vrang, S. J. West and H. Zhang, J. Med. Chem., 39 (1996) 4261.

[32] F. W. Bell, A. S. Cantrell, M. Hogberg, S. R. Jaskunas, N. G. Johansson, C. L. Jordan, M. D. Kinnick, P. Lind, J. M. Morin Jr., R. Noreen, B. Oberg, J. A. Palkowitz, C. A. Parrish, P. Pranc, C. Sahlberg, R. J. Ternansky, R. T. Vasileff, L. Vrang, S. J. West, H. Zhang and X.-X, Zhou, J. Med. Chem., 38 (1995) 4929.

[33] R. G. Strickley and B. D. Anderson, Pharm. Res., 10 (1993) 1076.

[34] A. Ranise, A. Spallarossa, S. Cesarini, F. Bondavalli, S. Schenone, O. Bruno, G. Menozzi, P. Fossa, L. Mosti, M. La Colla, G. Sanna, M. Morreddu, G. Collu, B. Busonera, M. E. Marongiu, A. Pani, P. La Colla and R. Loddo, J. Med. Chem., 48 (2005) 3858.

[35] Varmuza, K., 1980, Pattern Recognition in Chemistry, Springer, New York.

[36] Benzecri, J.-P., 1984, L'Analyse des Données, Dunod, Paris, Vol. 1.

[37] Tondeur, D., and Kvaalen, E., 1987, "Equipartition of Entropy Production. An Optimality Criterion for Transfer and Separation Processes," Ind. Eng. Chem., Fundam., 26, pp. 50-56.

[38] Torrens, F., and Castellano, G., 2006, "Periodic Classification of Local Anaesthetics (Procaine Analogues)," Int. J. Mol. Sci., 7, pp. 12-34.

[39] Torrens, F., and Castellano, G., 2009, "Periodic Classification of Human Immunodeficiency Virus Inhibitors," Biomedical Data and Applications, A. S. Sidhu, T. Dillon and M. Bellgard, eds., Stud. Comput. Intelligence No. 224, Springer, Berlin, in press.

[40] F. Torrens and G. Castellano, Table of periodic properties of human immunodeficiency virus inhibitors, J. Comput. Intelligence Bioinformatics, in press.

[41] Torrens, F., and Castellano, G., 2009, "Classification of complex molecules," Foundations of Computational Intelligence Vol. 5, A.-E. Hassanien and A. Abraham, eds., Stud. Comput. Intelligence No. 205, Springer, Berlin, pp. 243-315.

[42] Kaufmann, A., 1975, Introduction à la Théorie des Sous-ensembles Flous, Masson, Paris, Vol. 3.

[43] Cox, E., 1994, The Fuzzy Systems Handbook, Academic, New York.

[44] Kundu, S., 1998, "The Min–Max Composition Rule and its Superiority over the Usual Max–Min Composition Rule," Fuzzy Sets Sys., 93, pp. 319-329.

[45] Lambert-Torres, G., Pereira Pinto, J. O., and Borges da Silva, L. E., 1999, "Minmax Techniques," Wiley Encyclopedia of Electrical and Electronics Engineering, Wiley, New York.

[46] Shannon, C. E., 1948, "A Mathematical Theory of Communication: Part I, Discrete Noiseless Systems," Bell Syst. Tech. J., 27, pp. 379-423.

[47] Shannon, C. E., 1948, "A Mathematical Theory of Communication: Part II, the Discrete Channel with Noise," Bell Syst. Tech. J., 27, pp. 623-656.

[48] White, H., 1989, AI Expert, 12, pp. 48-48.

[49] Kullback, S., 1959, Information Theory and Statistics, Wiley, New York.

[50] Iordache, O., Corriou, J. P., Garrido-Sánchez, L., Fonteix, C., and Tondeur, D., 1993, "Neural Network Frames. Application to Biochemical Kinetic Diagnosis," Comput. Chem. Eng., 17, pp. 1101-1113.

[51] IMSL, 1989, Integrated Mathematical Statistical Library (IMSL), IMSL, Houston.

[52] Tryon, R. C., 1939, J. Chronic Dis., 20, pp. 511-524.

[53] Jarvis, R. A., and Patrick, E. A., 1973, "Clustering Using a Similarity Measure Based on Shared Nearest Neighbors," IEEE Trans. Comput., C22, pp. 1025-1034.

[54] Page, R. D. M., 2000, Program TreeView, Universiy of Glasgow.

[55] Huson, D. H., 1998, "SplitsTree: Analizing and Visualizing Evolutionary Data," Bioinformatics, 14, pp. 68-73.

[56] Hotelling, H., 1933, "Analysis of a Complex of Statistical Variables into Principal Components," J. Educ. Psychol., 24, pp. 417-441.

[57] Kramer, R., 1998, "Chemometric Techniques for Quantitative Analysis," Marcel Dekker, New York.

[58] Patra, S. K., Mandal, A. K., and Pal, M. K., 1999, "State of Aggregation of Bilirubin in Aqueous Solution: Principal Component Analysis Approach," J. Photochem. Photobiol., Sect. A, 122, pp. 23-31.

[59] Jolliffe, I. T., 2002, "Principal Component Analysis," Springer, New York.

[60] Xu, J., and Hagler, A., 2002, "Chemoinformatics and Drug Discovery," Molecules, 7, pp. 566-600.

[61] Shaw, P. J. A., 2003, "Multivariate Statistics for the Environmental Sciences," Hodder-Arnold, New York.