Eur. J. Med. Chem. 34 (1999) 83–92

Original article

Automated search for potentially active compounds by using cluster trees

Nereide Stela Santos Magalhães[a], Socrates Cabral De Holanda Cavalcanti[a], Irwin Rose Alencar De Menezes[a], Adriano Antunes De Sousa Araújo[a], Hélio Magalhães De Oliveira[b], Antonio José Alves[a]*

[a]*Departamento de Farmácia, Universidade Federal de Pernambuco, Rua Prof. Artur Sá s/n, Cidade Universitária, CEP 51740-520 Recife-PE, Brazil*
[b]*Departamento de Eletrônica, Universidade Federal de Pernambuco, Caixa Postal 7800, Cidade Universitária, CEP 51711-970 Recife-PE, Brazil*

(Received 26 June 1998; accepted 17 August 1998)

Abstract – A new agglomerative (bottom-up) hierarchical cluster technique referred to as the Adaptive Mean-Linkage algorithm is derived. Cluster algorithms are also offered as a tool to explore the descriptor space knowing the quantitative structure–activity relationship (QSAR). The substituents are clustered building a dendrogram (cluster tree) per site. Choosing appropriate pathways on such cluster trees according to the QSAR equation, an automated search for potentially active substituted compounds can be performed. Applications to a series of substituted phenylguanidines with anticancer activity are focused illustrating this approach. © Elsevier, Paris

cluster analysis / QSAR / adaptive mean-linkage algorithm / phenylguanidines

1. Introduction

One of the most important approaches in medicinal chemistry is drug design. The search for compounds presenting high biological activity in a fast and rational way is highly suitable. Recently, drug design has also been subjected to manifold advances [1, 2]. One way for developing drugs is defining substitution sites on a basic structure, such as a natural or a synthetic molecule presenting some biological activity. The quantitative relationship between the structure and the activity (QSAR) for substituted series is an approach largely adopted in the search of new drugs [3–6]. Beginning with a compound of interest, a well-designed sample of derivatives is synthesised and their activities are determined. The hierarchical cluster analysis has long been used to design such a sample [7, 8]. Another approach to choose substituents in a logical way was a non-mathematical method early suggested by Topliss [9]. The QSAR allows the prediction of compounds' activity without synthesising them. It is therefore possible to select those with greater potential for further investigation. But how can one use QSAR information to choose substituents? This paper introduces a new idea on substituent selection via the computer. When there are only a few sites and up to two parameters prevailing in activity, there are almost no difficulties to choose appropriate substituents. A typical example is a chemical series where the Hansch model is applied [3, 10, 11]. However, things become more unclear if there exist more than two physicochemical properties correlated with the activity and many substituent sites. One of the strategies to find an active compound consists of an exhaustive search for the greatest activity as predicted by the QSAR. The 'brute force' is intended as the random exploring of substituent space in order to look for the substituents having descriptor values consistent with a maximal pharmacological activity, as predicted by the correspondent QSAR. Nevertheless, that is not a good approach to understand the structure–activity relationship. For the purpose of avoiding this process, one should try to gather compounds of which the biological activities do not differ substantially. Compounds with alike physicochemical properties usually present very similar activities although the converse is not true. Therefore substituents with close physicochemical characteristics could be clustered owing to the

*Correspondence and reprints

isometric bioisosterism. In this paper a variant of the classical hierarchical cluster analysis [8, 12] is reported. It can be interpreted as a mean-linkage algorithm of which the threshold value is updated at each interaction. A new method for the selection of active compounds is then proposed, that is, an algorithm that points out which substituents can be used to achieve high activity.

The QSAR and the cluster trees are combined so as to define a pathway on dendrograms where substituents yielding potent compounds are likely to be found. The search for active compounds by finding paths in cluster trees normally furnishes, for the same set of possible substituents per site, similar results to an exhaustive search. However, it allows a systematical exploring of the descriptor space and gives powerful insights.

It is well known [5] that cluster analysis has been used both before (in the selection of a training set with maximal variability in the data space) or after a QSAR procedure (to improve the set of substituent candidates in each site). The novelty here is the use of the cluster tree *combined* with the QSAR where the equation is applied as a 'pathfinder' through the trees.

1.1. Fundamentals on cluster trees

Substituents with similar physicochemical parameters practically render the same contribution to the activity. Thus, one should cluster substituents that have similar physicochemical parameters [3, 12]. This idea is explored in cluster analysis, considering normalised parameters firstly and then defining a metric in order to merge substituents [13, 14]. For each substituent S_i, a number p of physicochemical parameters of interest is considered. For instance, if the Hansch hydrophobic constant π and the Hammet constant σ were the two parameters correlated with the activity of a given chemical series, the normalised parameters for each substituent would be, respectively:

$$X_{i1} = [\pi_i - \text{mean}(\pi_i)]/[\text{s.d.}(\pi_i)],$$

$$X_{i2} = [\sigma_i - \text{mean}(\sigma_i)]/[\text{s.d.}(\sigma_i)]$$

where X_{ik} denotes the value of the k–th normalised parameter for the substituent S_i, $k = 1, 2, ..., p$. The distance from a substituent S_i to a substituent S_j can be computed by means of the Euclidean distance [14] between their respective normalised parameter vectors, that is:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} = \|X_i - X_j\|$$

where p is the number of parameters present in the QSAR equation. A matrix of distances between substituents is then constructed. It is a symmetric matrix with a null main diagonal, henceforth denoted by (d_{ij}). Instead these distance matrices, simimilarity matrices could also be used [15, 16]. The substituents can be clustered based on such a matrix resulting on a cluster tree or dendrogram. The classical examples of hierarchical clustering appear in biological taxonomy but many clustering techniques have been adopted in drug design. There are hierarchical cluster procedures: divisive (top-down) and agglomerative (bottom-up). This work considers the second method. Different agglometative hierarchical clustering data description schemes are used depending on the metric adopted to merge the 'nearest' pair of clusters. Some common distance measures [14] lead to the nearest-neighbour algorithm, the furthest-neighbour algorithm, the average-neighbour algorithm and the mean-neighbour algorithm. These algorithms are often associated with an arbitrary 'distance threshold' so that clustering terminates when the distance between neighbours exceeds such a threshold. This threshold association yields, respectively, the single-linkage algorithm, the complete-linkage algorithm, the average-linkage algorithm and the mean-linkage algorithm. A new cluster method termed 'adaptive mean-linkage algorithm' is introduced in the next section.

2. Results

2.1. The adaptive mean-linkage algorithm

A few definitions and simple results are required so as to understand the proposed method. The cut-off distance is defined as the limit distance between substituents that will be clustered (a threshold). In order to determine the cut-off distance it is necessary to take the lowest value of the Euclidean distance between S_i and any other substituents S_j. This is carried out for each substituent S_i and then the greatest value obtained is defined as the cut-off distance.

Mathematically speaking,

$$d_u = \text{Max}_i \, \text{Min}_j \, d_{ij} .$$

One substituent S_i is said to be within the cut-off distance related to the substituent S_j if and only if the distance d_{ij} between them is less than or equal to the cut-off distance. Denoting by n_i the number of substituents within the cut-off distance from S_i, it is possible to define the following set: A set Γ_i is called a S_i-neighbourhood if it contains all the j-indexes of the substituents S_j within the cut-off distance to S_i, arranged in a non-decreasing order, i.e.

$$\Gamma_i = \{j_1, j_2, ..., j_{ni}\}$$

in such a way that $0 \le d_{ij1} \le d_{ij2} \le ... \le d_{ijni} \le d_u$.

Lemma 1. Each neighbourhood contains at least two distinct elements.

Proof: It is obvious that the substituent S_i belongs to Γ_i so Γ_i is not empty. The cut-off distance is formulated to include at least one closest substituent. If the substituent S_{j^*} ($j^* \neq i$) is the one closest to S_i, then $d_{ij^*} = \mathrm{Min}_j\, d_{ij} \leq d_u$, hence S_{j^*} also belongs to Γ_i.

For each substituent a neighbourhood set is created. From these sets it is possible to create subsets named sub-neighbourhoods. Let Γ_i^{ν} be a sub-neighbourhood from S_i, defined as the subset of Γ_i that contains just its first ν elements. The set Γ_j^{ν} denotes a sub-neighbourhood from S_j. The substituents associated with the indexes i_1, i_2, ..., i_ν are said to be 'extremely close' if and only if all their sub-neighbourhood of ν elements have identical elements. This is denoted as

$$[i_1, i_2, ..., i_\nu] \Leftrightarrow \Gamma_1^{\nu} = \Gamma_2^{\nu} = \Gamma_\nu^{\nu}.$$

Lemma 2. In each set of substituents there is at least one pair of them that are extremely close.

Proof: Let $d_{i^*j^*}$ be the smallest non-zero element in the (d_{ij}) matrix, then j^* belongs to Γ_{i^*} due to the presence of at least two elements in each neighbourhood. In other words, $d_{i^*j^*} = \mathrm{Min}_j\, d_{i^*j}$, and $d_{i^*j^*} = \mathrm{Min}_i\, d_{ij^*}$, so that i^* belongs to Γ_{j^*} due to the symmetry of the matrix and Lemma 1. Consequently $\Gamma_{i^*}{}^2 = \Gamma_{j^*}{}^2$, hence $[i^*, j^*]$ are extremely close.

Linkage algorithms require the choice of a *fixed* but arbitrary threshold. The value of such a threshold is rather empirical: it should not be too large (otherwise all of the points will be assigned to one cluster) or too small (each point will form an isolated cluster). In contrast, the new approach introduces a rational and objective rule to obtain *adaptive* thresholds based on a minimax criterion. Thresholds based on the cut-off distance are sufficiently small to assure the homogeneity and also large enough to guarantee at least a pair of merging sets.

In each step of the tree generation, only the extremely close substituents are clustered. The new 'substituent' formed by merging the extremely close substituents is called pseudosubstituent and its physicochemical parameters are taken as the mean of the clustered substituent parameters. The idea behind such a procedure is that pseudosubstituents built in this way are homogeneous [14]. In other words, given two arbitrary absolutely close points i and i' belonging to $[i_1, i_2, ..., i_\nu]$, then $\forall\, j \notin [i_1, i_2, ..., i_\nu]$, it is true that $d_{ii'} \leq d_{ij}$ and $d_{ii'} \leq d_{i'j}$. This means that two substituents in the same cluster have a greater similarity to each other than to any one outside the cluster. At this point a loop is performed. The procedure iterates using the new pseudosubstituents and the remaining substituents, until only one pseudosubstituent is left. This way, several cluster levels are obtained. The following algorithm might accomplish the cluster tree generation.

2.2. An algorithm for the cluster tree generation

Select the present physicochemical parameters in the QSAR analysis. Then, for a given substitution site perform the following steps:

Step 1: Compute the normalised parameters, after obtaining the substituents and their respective parameter values.

Step 2: Compute the distance matrix and the cut-off distance.

Step 3: Determine the substituents' neighbourhoods and identify 'extremely close' sets.

Step 4: Merge extremely close substituents generating pseudosubstituents of which the parameters are the mean-value of the clustered substituent parameters.

Step 5: Stop if a single pseudosubstituent is found. Otherwise return to *step 2*.

For each different site, a corresponding tree must be generated. The algorithm can cluster at a single step several pairs, triplets, etc. Trees derived from the modified algorithm are therefore compact and less complex than stepwise hierarchical clustering [8]. A naive illustrative example applied to substituted phenylguanidines studied by Yang et al. [17] is discussed in the sequel.

2.3. Selection of potentially active compounds

The search for an active compound is made regarding the cluster trees (i.e. ortho, para and meta trees), beginning with the highest aggregation levels. The QSAR equation is used to predict the biological activity of pseudocompounds obtained through pseudosubstituents. In other words, the activity of a 'representative compound' of the cluster is calculated, firstly taking in account the penultimate pseudosubstituents. The tree branch corresponding to the most active pseudocompound is selected while the remaining ones are thrown out. Only the pseudosubstituents that are related to surviving branches are considered. This procedure continues until a path in each tree directed to the most promising pseudosubstituents is traced. It is especially straitghforward in cases in which the QSAR enables one to calculate the influence of each position separately. The process is split into two parts: the cluster trees generation and the search of a path in these trees. The cluster tree generation requires plenty of computation, but the search of a path is fast and practical. Some cautions for a validation of the mathematical model should be taken into account [18]. Interesting applications of the tree are related to the risk of extrapolation that might occur in

some cases [19]. In many cases the scattergram between the activity and a lipophilic parameter presents a parabolic shape. Initially the greater the lipophilicity is, the greater the activity will be, but as lipophylicity increases, the odds are that the activity will decrease. The biological behaviour may change when the parameter values are not within the range where the multiple regression was established. One way to prevent this phenomenon is defining 'an acceptable range of values' for each physicochemical parameter of the series. In each step, the parameter values are checked for each pseudocompound. If there is any extrapolation, the corresponding branch is not taken as a true survivor. The parameter range in which the regression was established defines a region with no extrapolated physicochemical parameters and sets up a boundary in the tree. Some alternative solutions can be found by slacking the constraints over the acceptable range. This option is crucial in order to look for active compounds. Yet, the pitfall of excessive extrapolation must be kept in mind. The results can be applied to different series presenting different biological activities but with the same physicochemical parameters correlated to the activity. They may also be extended to a series in which the QSAR equation has been derived from several models [20, 21]. Moreover, this clustering approach can also be valuable when toxicity is addressed [22–24].

The method introduced here does not necessarily solve the problem of maximising the biological activity of a chemical series because the selected compound may not exactly be the most active one. However, it grants a practical approach to select compounds with substantially high activity. Two conditions strengthen the likelihood of finding active compounds: First, the activity is normally a 'well-behaved' function of the physicochemical properties. Second, the pseudosubstituents are often representative for the merged substituents owing to their homogeneity.

2.4. Application to a phenylguanidine series

Phenylguanidines are selective inhibitor of proteolytic enzymes. The chemical series focused was synthesised by Yang et al. [17] and presents anticancer activity by inhibiting the urokinase enzyme (UK). It additionally inhibits trypsin, which was also considered in this study. Moreover, they are potential antiviral agents due to their selective viral thymidine kinase. It was recently reported that N^2-phenylguanidines inhibit the Herpes simplex virus HSV1 and HSV2 [25, 26]. Two QSAR analyses have been performed with Yang's series and the corresponding QSAR equations are given below.

– *Para position*:

$$\log 1/K_i(\mathrm{UK}) = 0.41(\pm 0.06)\,\pi_4 - 0.55(\pm 0.26)\,\mathrm{MR}_4 + 2.07(\pm 0.53)\,\mathrm{R}_4 + 4.89(\pm 0.21),$$

$$n = 12,\quad s = 0.344,\quad r = 0.944,\quad F = 21.8; \qquad (1)$$

$$\log 1/K_i(\mathrm{trypsin}) = 0.26(\pm 0.09)\,\pi_4 - 0.82(\pm 0.35)\,\mathrm{MR}_4 + 2.15(\pm 0.72)\,\mathrm{R}_4 + 3.84(\pm 0.30),$$

$$n = 12,\quad s = 0.480,\quad r = 0.839,\quad F = 6.40; \qquad (2)$$

– *Ortho position*:

$$\log 1/K_i(\mathrm{UK}) = -1.27(\pm 0.32)\,\mathrm{B1}_2 + 0.86(\pm 0.19)\,\sigma_{p2} + 5.86(\pm 0.49),$$

$$n = 6,\quad s = 0.190,\quad r = 0.948,\quad F = 13.31; \qquad (3)$$

$$\log 1/K_i(\mathrm{trypsin}) = -1.38(\pm 0.26)\,\mathrm{B2}_2 + 1.57(\pm 0.31)\,\mathrm{F}_2 + 5.10(\pm 0.41),$$

$$n = 6,\quad s = 0.185,\quad r = 0.965,\quad F = 20.24. \qquad (4)$$

The π_4, R_4 and MR_4 values are taken from 12 substituents at the para site *(table I)* and 06 substituents at ortho position *(table II)*.

Cluster trees are built by the procedure shown therein. *Table I* and *table II* show the several steps required for obtaining dendrograms as well as the normalised parameters for substituents and pseudosubstituents. Since the parameters correlating with the activity are similar in both QSAR equations (for urokinase and trypsin enzymes), it is only necessary to build one tree at the para site. Distance matrices *(table III* and *IV)* are calculated using normalised parameters, followed by the computation of the cut-off distance *(table V* and *VI)*. The para-neighbourhoods are shown on the *table VII* where underlined substituents represent the subneighbourhoods. In that case, the substituents labelled {3 and 12}as well as the substituents {4, 7, 9 and 10} are extremely close and therefore are clustered *(table V)*. After defining [3;12] and [4;7;9;10] pseudosubstituents, the new parameter mean values are calculated and the procedure iterates until only one pseudosubstituent is achieved. Results concerning para-substitution are shown on *figure 1*. A pathway on the

para-tree is found by using the QSAR. Pseudocompounds predicted (urokinase) activities are (3.925*, 3.702), (4.456*, 3.393), (4.719*, 4.193), (4.934*, 4.504), (4.833, 5.035*), (4.702, 5.368*), from top-down decision levels, respectively. The pathway according to *table I* is 1 \ 1 \ 1 \ 1 \ 3 \ 12 from loop 6 \ ... \ loop 1, respectively. The asterisk indicates survivor branches. Para-position dendrograms are exactly the same for both the urokinase and trypsin inhibition because their QSAR analyses are based on the same parameters. However, the urokinase-path and the trypsin-path could be different. Numerical results presented the same pathway in both cases. A similar procedure is also carried out at the ortho site *(table VIII)*. Pseudocompounds predict (urokinase) activities are

Table I. Para-position normalised parameters (trypsin and urokinase).

No.	Para substituents (loop 1)	MR_4	π_4	R_4
1	H	0.1030	0.0000	0.0000
2	CO_2H	0.6050	–4.3600	0.1300
3	NO_2	0.7360	–0.2800	0.1600
4	CH_3	0.5650	0.5600	–0.1300
5	F	0.0920	0.1400	–0.3400
6	CO_2CH_3	1.2870	–0.0100	0.1500
7	Cl	0.6030	0.7100	–0.1500
8	OCH_3	0.7870	–0.0200	–0.5100
9	Br	0.8880	0.8600	–0.1700
10	C_2H_5	1.0300	1.0200	–0.1000
11	$CHCHCO_2H$	1.7030	–4.0400	0.2400
12	CF_3	0.5020	0.8800	0.1900
	Para pseudosubstituents (loop 2)			
1	H	0.1030	0.0000	0.0000
2	CO_2H	0.6050	–4.3600	0.1300
3	$NO_2 \setminus CF_3$	0.6190	0.3000	0.1750
4	$CH_3 \setminus Cl \setminus Br \setminus C_2H_5$	0.7715	0.7875	–0.1375
5	F	0.0920	0.1400	–0.3400
6	CO_2CH_3	1.2870	–0.0100	0.1500
7	OCH_3	0.7870	–0.0200	–0.5100
8	$CHCHCO_2H$	1.7030	–4.0400	0.2400
	Para pseudosubstituents (loop 3)			
1	$H \setminus NO_2 \setminus CF_3$	0.3610	0.1500	0.0875
2	$CO_2H \setminus CHCHCO_2H$	1.1540	–4.2000	0.1850
3	$CH_3 \setminus Cl \setminus Br \setminus C_2H_5$	0.7715	0.7875	–0.1375
4	F	0.0920	0.1400	–0.3400
5	CO_2CH_3	1.2870	–0.0100	0.1500
6	OCH_3	0.7870	–0.0200	–0.5100
	Para pseudosubstituents (loop 4)			
1	$H \setminus NO_2 \setminus CF_3 \setminus CH_3 \setminus Cl \setminus Br \setminus C_2H_5$	0.5663	0.4688	–0.0250
2	$CO_2H \setminus CHCHCO_2H$	1.1540	–4.2000	0.1850
3	F	0.0920	0.1400	–0.3400
4	CO_2CH_3	1.2870	–0.0100	0.1500
5	OCH_3	0.7870	–0.0200	–0.5100
	Para pseudosubstituents (loop 5)			
1	$H \setminus NO_2 \setminus CF_3 \setminus CH_3 \setminus Cl \setminus Br \setminus C_2H_5 \setminus F$	0.3291	0.3044	–0.1825
2	$CO_2H \setminus CHCHCO_2H$	1.1540	–4.2000	0.1850
3	CO_2CH_3	1.2870	–0.0100	0.1500
4	OCH_3	0.7870	–0.0200	–0.5100
	Para pseudosubstituents (loop 6)			
1	$H \setminus NO_2 \setminus CF_3 \setminus CH_3 \setminus Cl \setminus Br \setminus C_2H_5 \setminus F \setminus OCH_3$	0.5581	0.1422	–0.3463
2	$CO_2H \setminus CHCHCO_2H \setminus CO_2CH_3$	1.2205	–2.1050	0.1675

Table II. (a) Ortho normalised parameters (urokinase).

No.	Pseudosubstituent (loop 1)	$B1_2$	σ_{p2}
1	H	1.0000	0.0000
2	OCH_3	1.3500	-0.2700
3	CH_3	1.5200	-0.1700
4	NO_2	1.7000	0.7800
5	Cl	1.8000	0.2300
6	NH_2	1.5000	-0.6600
	(loop 2)		
1	H	1.0000	0.0000
2	$OCH_3 \setminus CH_3 \setminus NH_2$	1.4567	-0.3667
3	$NO_2 \setminus Cl$	1.7500	0.5050
	(loop 3)		
1	$H \setminus OCH_3 \setminus CH_3 \setminus NH_2$	1.2283	-0.1833
2	$NO_2 \setminus Cl$	1.7500	0.5050

Table II. (b) Ortho normalised parameters (trypsin).

No.	Pseudosubstituent (loop 1)	$B2_2$	F_2
1	H	1.0000	0.0000
2	OCH_3	1.9000	0.2600
3	CH_3	1.9000	-0.0400
4	NO_2	1.7000	0.6700
5	Cl	1.8000	0.4100
6	NH_2	1.5000	0.0200
	(loop 2)		
1	H	1.0000	0.0000
2	$OCH_3 \setminus Cl$	1.8500	0.3350
3	CH_3	1.9000	-0.0400
4	NO_2	1.7000	0.6700
5	NH_2	1.5000	0.0200
	(loop 3)		
1	H	1.0000	0.0000
2	$OCH_3 \setminus Cl \setminus NO_2$	1.7750	0.5025
3	$CH_3 \setminus NH_2$	1.7000	-0.0100
	(loop 4)		
1	$H \setminus CH_3 \setminus NH_2$	1.3500	-0.0050
2	$OCH_3 \setminus Cl \setminus NO_2$	1.7750	0.5025

(4.142*, 4.072), (4.590*, 3.695). Results concerning the ortho substitution are shown on *figure 2* and *figure 3*.

An appropriate choice for the substituents in this series could be *p*-CF_3, *o*-H for urokinase inhibition and *p*-CF_3, *o*-NO_2 for trypsin inhibition. The urokinase predicted activity values are 5.368 (para position) and 4.590 (ortho position), and for trypsin, 3.806 (ortho position) and 4.066 (para position). This fully agrees with experimental *p*-CF_3 substituted phenylguanidine activities [17]: 5.188 (urokinase) and 4.200 (trypsin). It is worthwhile to remark that substituted *m*-CF_3 also presented the highest activity against both HSV1 and HSV2 among 36 N^2-

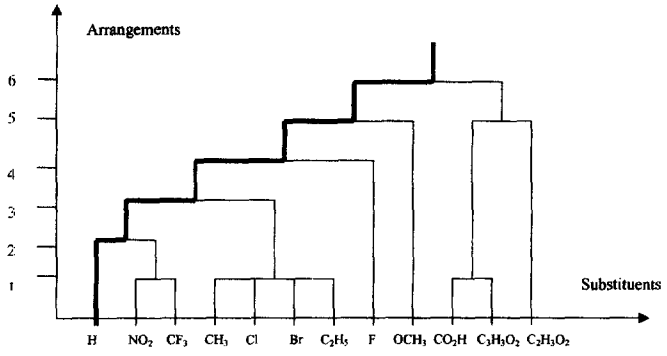

Figure 1. Para position cluster tree (urokinase/trypsin). The dendrogram shows clustered substituents at the para position considering either urokinase or trypsin inhibition. The bold line represents the most active pathway according to the QSAR (either equation (1) or equation (2)).

Phenylguanine derivatives [25]. Further improvements and alternative solutions could be found by expanding the trees.

3. Discussion

The new approach performs a suitable substituents computer-assisted selection and increases the probability of obtaining compounds with a better biological activity. The selection is based upon the QSAR equation and substituent clusters. A few examples have shown how to generate cluster trees and how to select interesting substituents. It can be used to derive consequences from

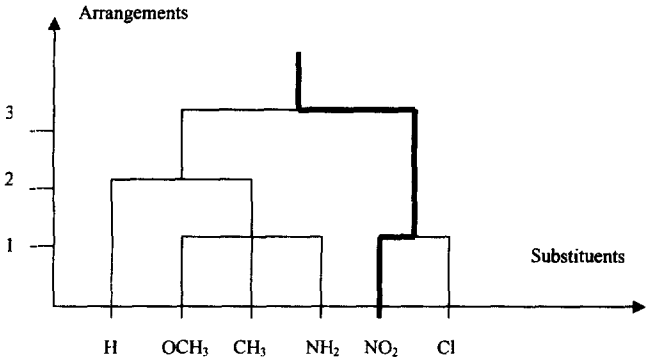

Figure 2. Ortho position cluster tree (urokinase). The dendrogram shows clustered substituents at the ortho position considering urokinase inhibition. The bold line represents the most active pathway according to the QSAR (equation (3)).

Table III. Para distance matrix (trypsin/urokinase).

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
0.0000	2.6772	1.5589	1.1984	1.4711	2.6776	1.3316	2.6660	1.9302	2.1524	4.2711	1.2923
2.6772	**0.0000**	2.2444	2.9068	3.3760	2.8035	3.0147	3.6596	3.1856	3.2315	2.4619	2.8745
1.5589	2.2444	**0.0000**	1.3856	2.5915	1.2187	1.4730	2.9003	1.5904	1.4762	2.9690	0.8244
1.1984	2.9068	1.3856	**0.0000**	1.3974	2.0173	0.1452	1.7414	0.7476	1.0586	3.8816	1.4003
1.4711	3.3760	2.5915	1.3974	**0.0000**	3.3710	1.4239	1.6949	1.9349	2.3539	4.8947	2.4930
2.6776	2.8035	1.2187	2.0173	3.3710	**0.0000**	2.0215	3.0553	1.7037	1.3414	2.4087	1.7977
1.3316	3.0147	1.4730	0.1452	1.4239	2.0215	**0.0000**	1.6554	0.6366	0.9762	3.9189	1.4885
2.6660	3.6596	2.9003	1.7414	1.6949	3.0553	1.6554	**0.0000**	1.5610	1.9347	4.3970	3.1270
1.9302	3.1856	1.5904	0.7476	1.9349	1.7037	0.6366	1.5610	**0.0000**	0.4429	3.6685	1.7711
2.1524	3.2315	1.4762	1.0586	2.3539	1.3414	0.9762	1.9347	0.4429	**0.0000**	3.4544	1.7082
4.2711	2.4619	2.9690	3.8816	4.8947	2.4087	3.9189	4.3970	3.6685	3.4544	**0.0000**	3.7645
1.2923	2.8745	0.8244	1.4003	2.4930	1.7977	1.4885	3.1270	1.7711	1.7082	3.7645	**0.0000**

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0.0000	2.3612	1.1546	1.3797	1.2597	2.2378	2.2645	3.6378
2.3612	**0.0000**	2.2733	2.7098	2.9492	2.4583	3.1897	2.0571
1.1546	2.2733	**0.0000**	1.2127	2.1369	1.2358	2.5573	2.9077
1.3797	2.7098	1.2127	**0.0000**	1.4860	1.4739	1.4331	3.2210
1.2597	2.9492	2.1369	1.4860	**0.0000**	2.8422	1.4216	4.1760
2.2378	2.4583	1.2358	1.4739	2.8422	**0.0000**	2.6075	2.1295
2.2645	3.1897	2.5573	1.4331	1.4216	2.6075	**0.0000**	3.7862
3.6378	2.0571	2.9077	3.2210	4.1760	2.1295	3.7862	**0.0000**

1	*2*	*3*	*4*	*5*	*6*
0.0000	2.9719	1.2482	1.6127	2.0467	2.2983
2.9719	**0.0000**	3.0755	3.8066	2.3181	3.4422
1.2482	3.0755	**0.0000**	1.6902	1.5781	1.3808
1.6127	3.8066	1.6902	**0.0000**	3.1382	1.6407
2.0467	2.3181	1.5781	3.1382	**0.0000**	2.5637
2.2983	3.4422	1.3808	1.6407	2.5637	**0.0000**

1	*2*	*3*	*4*	*5*
0.0000	2.7738	1.4386	1.6296	1.6692
2.7738	**0.0000**	3.5785	2.1663	3.2132
1.4386	3.5785	**0.0000**	2.9673	1.5571
1.6296	2.1663	2.9673	**0.0000**	2.3961
1.6692	3.2132	1.5571	2.3961	**0.0000**

1	*2*	*3*	*4*
0.0000	3.0583	2.4588	1.4739
3.0583	**0.0000**	1.9753	3.0103
2.4588	1.9753	**0.0000**	2.3377
1.4739	3.0103	2.3377	**0.0000**

1	*2*
0.0000	2.4495
2.4495	**0.0000**

QSAR equations by exploring large chemical databases [3, 12, 27]. It specially handles the cases in which the QSAR allows one to calculate the influence of each position separately. It does not avoid the use of cluster analyses in the stage foregoing the QSAR evaluation [13] but the same tree-generator method can be applied before defining parameters by the correlation analysis. The superiority of the Adaptive Clustering with respect to the Average-Linkage Algorithm follows because it achieves an interesting compromise on threshold values.

Table IV. (a) Ortho distance matrix (urokinase).

1	*2*	*3*	*4*	*5*	*6*
0.0000	1.3548	1.8708	2.9435	2.8669	2.2234
1.3548	**0.0000**	0.6348	2.4769	1.8907	0.9573
1.8708	0.6348	**0.0000**	2.0430	1.2837	1.0038
2.9435	2.4769	2.0430	**0.0000**	1.1782	3.0264
2.8669	1.8907	1.2837	1.1782	**0.0000**	2.1053
2.2234	0.9573	1.0038	3.0264	2.1053	**0.0000**

1	*2*	*3*
0.0000	1.4294	2.3778
1.4294	**0.0000**	2.0743
2.3778	2.0743	**0.0000**

1	*2*
0.0000	2.0000
2.0000	**0.0000**

Table IV. (b) Ortho distance matrix (trypsin).

1	*2*	*3*	*4*	*5*	*6*
0.0000	2.7711	2.6165	3.1296	2.7413	1.4532
2.7711	**0.0000**	1.0657	1.5680	0.6068	1.4406
2.6165	1.0657	**0.0000**	2.5882	1.6247	1.1806
3.1296	1.5680	2.5882	**0.0000**	0.9682	2.3810
2.7413	0.6068	1.6247	0.9682	**0.0000**	1.6364
1.4532	1.4406	1.1806	2.3810	1.6364	**0.0000**

1	*2*	*3*	*4*	*5*
0.0000	2.5785	2.4713	2.9239	1.3726
2.5785	**0.0000**	1.2421	1.1770	1.4130
2.4713	1.2421	**0.0000**	2.4008	1.1145
2.9239	1.1770	2.4008	**0.0000**	2.2090
1.3726	1.4130	1.1145	2.2090	**0.0000**

1	*2*	*3*
0.0000	2.4955	1.6380
2.4955	**0.0000**	1.7576
1.6380	1.7576	**0.0000**

1	*2*
0.0000	2.0000
2.0000	**0.0000**

The foundation of the method is independent of a particular QSAR equation. Whatever physicochemical parameters used in the QSAR, an associated cluster tree can be derived and this equation can be applied to find out a pathway. However, it is a bit much complex in cases in which the influence of a substituent is affected by what substituents are present at other positions. Trees can be built off-line, once and for all. Furthermore, the tree generation depends just on the physicochemical parameters. If the tree is built based on a certain set of parameters, it can be stored for later use in the substituent selection of any chemical series having a different QSAR but with the same set of parameters.

This approach does not directly concern the problem of collinearity but correlation matrices derived from sample covariance can be considered in place of the distance matrices. Furthermore, QSAR equations derived from other chemometric methods such as Principal Component Regression (PCR) or Partial Least Square analysis (PLS) techniques can also be used [28]. There are two aspects to

Table V. Cut-off distance at the para site (urokinase and trypsin).

Number of clusters	Clusters	Cut-off distance
8	(1)–(2)–(3,12)–(4,7,9,10)–(5)–(6)–(8)–(11)	d'_u = 2.4087
6	(1,3,12)–(2,11)–(4,7,9,10)–(5)–(6)–(8)	d'_u = 2.0571
5	(1,3,12,4,7,9,10)–(2,11)–(5)–(6)–(8)	d'_u = 2.3181
4	(1,3,12,4,7,9,10,5)–(2,11)–(6)–(8)	d'_u = 2.1663
2	(1,3,12,4,7,9,10,5,8)–(2,11,6)	d'_u = 1.9753
1	(1,3,12,4,7,9,10,5,8,2,11,6)	d'_u = 2.4495

Table VI. Cut-off distance at the ortho site.

Enzyme	Number of clusters	Clusters	Cut-off distance
Urokinase	3	(1)–(2,3,6)–(4,5)	d_u = 1.3548
	2	(1,2,3,6)–(4,5)	d_u = 2.0743
	1	(1,2,3,6,4,5)	d_u = 2.0000
Trypsin	5	(1)–(2,5)–(3)–(4)–(6)	d_u = 1.4532
	3	(1)–(2,5,4)–(3,6)	d_u = 1.3726
	2	(1,3,6)–(2,5,4)	d_u = 1.7576
	1	(1,3,6,2,5,4)	d_u = 2.0000

the extrapolation in the substituent selection. One should be careful not to imply gross extrapolation in the search for new substituents beyond explored data space. Although one wants to be careful one also wants to move as far beyond explored data space as possible. These trade-offs are allowed simply by changing the constraints on the valid range of the physicochemical parameters. This increases the ability of the method to efficiently explore the substituent space and it is perhaps the most powerful tool available in the tree approach. After choosing an extrapolated substituted compound to synthesise, its experimental activity is determined. The activity of such an

Table VIII. Neighborhoods at the ortho site.

Loop 1	Loop 2	Loop 3	(urokinase)
T1 = 1 2	T1 = *1 2*	T1 = *1 2*	
T2 = *2 3 6* 1	T2 = *2 1* 3	T2 = *2 1*	
T3 = *3 2 6* 5	T3 = 3 2		
T4 = *4 5*			
T5 = *5 4* 3			
T6 = *6 2 3*			

Loop 1	Loop 2	Loop 3	Loop 4	(trypsin)
T1 = 1 6	T1 = 1 5	T1 = *1 3*	T1 = *1 2*	
T2 = *2 5* 3 6	T2 = *2 4* 3	T2 = 2 3	T2 = *2 1*	
T3 = 3 2 6	T3 = *3 5* 2	T3 = *3 1* 2		
T4 = 4 5	T4 = *4 2*			
T5 = *5 2* 4	T5 = *5 3* 1			
T6 = 6 3 2 1				

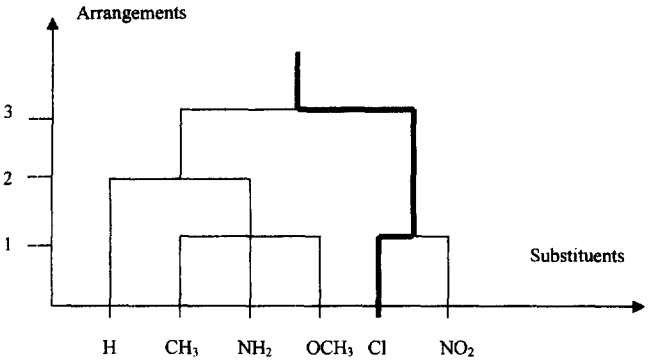

Figure 3. Ortho position cluster tree (trypsin). The dendrogram shows clustered substituents at the ortho position considering trypsin inhibition. The bold line represents the most active pathway according to the QSAR (equation (4)).

Table VII. Neighborhoods at the para site (urokinase/trypsin).

Loop 1	Loop 2	Loop 3	Loop 4	Loop 5	Loop 6
T1 = 1 3 5 4	T1 = *1 3* 5 4	T1 = *1 3* 4 5 6	T1 = *1 3* 4 5	T1 = *1 4*	T1 = *1 2*
T2 = 2 3	T2 = *2 8*	T2 = 2 5	T2 = 2 4	T2 = *2 3*	T2 = *2 1*
T3 = *3 12* 6 4 7 10 1 9 2	T3 = *3 1* 4 6	T3 = *3 1* 6 5 4	T3 = *3 1* 5	T3 = *3 2*	
T4 = *4 7 9 10* 1 3 5 12 8 6	T4 = 4 3 1 7 6 5	T4 = 4 1 6 3	T4 = 4 1 2	T4 = *4 1*	
T5 = 5 4 7 1 8 9 10	T5 = 5 1 7 4	T5 = 5 3 1 2	T5 = 5 3 1		
T6 = 6 3 10 9 12 4 7 11	T6 = 6 3 4	T6 = 6 3 4 1			
T7 = *7 4 9 10* 1 5 3 12 8 6	T7 = 7 5 4				
T8 = 8 9 7 5 4 10	T8 = *8 2*				
T9 = *9 10 7 4* 8 3 6 12 1 5					
T10 = *10 9 7 4* 6 3 12 8 1 5					
T11 = 11 6					
T12 = *12 3* 1 4 7 10 9 6					

extrapolated compound is then compared to the most active compound. It may be higher: in this case, it is interesting to extrapolate further, or lower: then it is advisable to stop extrapolations. Polysubstituted series are interesting because they increase the likelihood of finding active compounds. The higher the number of sites the more advantageous this technique will be. If the constraints on the parameter range are changed, other potentially active compounds can be found. It is obvious that selected compounds must be synthesised in order to corroborate whether they are or not as active as predicted. Cluster trees are flexible and they enable different substituent sets at each site. This allows, for example, taking into account difficulties in the synthesis. Laboratories with large available databases of physicochemical parameters can make a powerful implementation of this algorithm.

Acknowledgements

The authors gratefully thank the Brazilian National Council for Scientific and Technological Development (CNPq) and the Foundation for the Science and Technology (FACEPE) for their partial financial support. They also thank one anonymous referee for his criticism and suggestions that improved this paper.

References

[1] Kubinyi H., Pharmazie 50 (10) (1995) 647–662.

[2] Van DeWaterbeemd H., Drug Des. Discov. 9 (3,4) (1993) 277–285.

[3] Hansch C., Leo A., Substituent Constants for Correlation Analysis in Chemistry and Biology, ACS Professional Reference Book, Washington, DC, 1995.

[4] Martin Y.C., Quantitative Drug Design: A critical Introduction, Marcel Dekker, New York, 1978.

[5] Franke R., Theoretical Drug Design, Vol. 7, Elsevier, New York, 1984.

[6] Taylor J.B., Kennewell P.D., Modern Medicinal Chemistry, Ellis Horwoord, New York, 1993.

[7] Craig P.N., J. Med. Chem. 14 (8) (1971) 680–684.

[8] Hansch C., Unger S.H., Forsythe A.B., J. Med. Chem. 16 (11) (1973) 1217–1222.

[9] Topliss J.G., J. Med. Chem. 15 (10) (1972) 1006–1011.

[10] Hansch C., Clayton J.M., J. Pharm. Sci. 62 (1973) 1–21.

[11] Kubinyi H., Prog. Drug Res. 23 (1979) 97–198.

[12] Norrington F.E., Hydes R.M., Williams S.G., Wooton R., J. Med. Chem. 18 (8) (1975) 604–607.

[13] Wooton R., Crafield R., Sheppey G.C., Goodford P.J., J. Med. Chem. 18 (6) (1975) 607–613.

[14] Diday E., Simon J.C., In: Fu K.S. (Ed.), Unsupervised Learning and Clustering, Springer Verlag, Heidelberg, 1976, pp. 211–252.

[15] Benigni R., Cotta-Ramusino M., Giorgi F., Gallo G., J. Med. Chem. 38 (4) (1995) 629–635.

[16] Good A.C., Peterson S.J., Richards W.G., J. Med. Chem. 36 (20) (1993) 2929–2937.

[17] Yang H., Henkin J., Kim K.H., Greer J., J. Med. Chem. 32 (1990) 2956–2961.

[18] Benigni R., Giuliani A., Mutat. Res. 306 (2) (1994) 181–186.

[19] Wonnacott T.H., Wonnacott R.J., Introductory Statistics (4th ed.), John Wiley, New York, 1990.

[20] Macfarland J.W.L., J. Med. Chem. 20 (5) (1970) 625–629.

[21] Martin Y.C., Prog. Drug Res. 15 (1971) 123–146.

[22] Yapel J.R.A.F., Amer. Chem. Soc. (1972) 183–251.

[23] Lange A.W., Vormann K., QSAR Environ. Res. 3 (3) (1995) 171–177.

[24] Romijn C.A., Luttik R., Van DeMeent D., Slooff W., Canton J.H., Ecotoxicol. Environ. Saf. 26 (1) (1993) 61–85.

[25] Hadjipavlou-Litina D., Pharmazie 50 (12) (1995) 796–798.

[26] Gambino J., Focher F., Hildebrand C., Maga G., Noonan T., Spadari S., Wright G., J. Med. Chem. 35 (1992) 2979–2983.

[27] Leo A., Environ. Health Perspect. 61 (1985) 275–285.

[28] Davis, A.M., In: King F.D. (Ed.), Medicinal Chemistry – Principles and Practice, Royal Society of Chemistry, London, 1994, pp. 98–142.