Journal of Computer-Aided Molecular Design **18:** 577–586, 2004.
DOI 10.1007/s10822-004-4076-0

Classification of a large anticancer data set by Adaptive Fuzzy Partition

Nadège Piclin[a], Marco Pintore[a], Christophe Wechman[b] & Jacques R. Chrétien[b,*]
[a]*BioChemics Consulting, 16 rue Leonard de Vinci, F-45074 Orleans Cedex 2, France;* [b]*University of Orleans, LBLGC/CBI, UPRES EA 1207, F-45067 Orleans Cedex 2, France*

Received 19 July 2004; accepted in revised form 23 August 2004

Key words: Adaptive Fuzzy Partition, anticancer activity, data mining, drug discovery, Genetic Algorithms, model validation

Summary

An Adaptive Fuzzy Partition (AFP) algorithm, derived from Fuzzy Logic concepts, was used to classify an anticancer data set, including about 1300 compounds subdivided into eight mechanisms of action. AFP classification builds relationships between molecular descriptors and bio-activities by dynamically dividing the descriptor hyperspace into a set of fuzzy subspaces. These subspaces are described by simple linguistic rules, from which scores ranging between 0 and 1 can be derived. The latter values define, for each compound, the degrees of membership of the different mechanisms analyzed. A particular attention was devoted to develop structure–activity relations that have a real utility. Then, well-defined and widely accepted protocols were used to validate the models by defining their robustness and prediction ability. More particularly, after selecting the most relevant descriptors with help of a genetic algorithm, a training set of 640 compounds was isolated by a rational procedure based on Self-Organizing Maps. The related AFP model was then validated with help of a validation set and, above all, of cross-validation and Y-randomization procedures. Good validation scores of about 80% were obtained, underlining the robustness of the model. Moreover, the prediction ability was evaluated with 374 test compounds that had not been used to establish the model and 77% of them were predicted correctly.

Introduction

The search for new drugs against cancer plays a central role in the research programs of pharmaceutical companies but also those of many governmental organizations. Since 1990, for example, the National Cancer Institute (NCI) has screened more than 70,000 natural and synthetic compounds against a panel of 60 different human tumor cell lines, concerning leukemia, melanomas, lung cancers as well as colon, central nervous system, ovary, kidney, prostate and breast cancers [1].

A second example is given by the MDL Drug Data Report (MDDR) database [2], covering patent literature, journals, and congress reports, which includes more than 21,000 compounds associating general antineoplastic characteristics. A subset of about 15,800 compounds is identified by a specific activity index and 4500 of them can be represented by a well-identified mechanism of action.

Moreover, in the next few years, Combinatorial Chemistry (CC) and High Throughput Screening (HTS) technologies are expected to increase in an exponential way the number of new compounds tested on a large panel of cancer targets. The industry and the academic world are now faced with the new challenge of managing and analyzing this huge amount of high dimensional data. This requires to develop even more efficient data mining tools able to classify large chemical libraries and to get enhanced informational contents. More

*To whom correspondence should be addressed. Tel: +33-2-3841-7076; Fax: +33-2-3841-7221; E-mail: jacques.chretien@biochemics-consulting.com

particularly, the diversity of these databases can be exploited with help of automated and multivariate data analysis methods [3–5]. The latter relate the molecular structures with their biological properties by establishing computational models able to assign activity values to new untested compounds [6, 7]. This allows then to better rationalize the CC/HTS programs by pre-selecting the compounds likely to have the desirable biological activities.

The first models of antineoplastic targets were established on the NCI database. Several computational methods were used to relate the 60 cell lines with subgroups of compounds characterized by the same mechanism of action, mode of drug resistance or structural similarity [8–14]. Amidst them, one of the most interesting contributions is probably the work of Weinstein et al., that used a back-propagation neural network to classify 134 chemotherapeutic agents according to six mechanisms of action [10]. The model was able to predict correctly 91% of the agents, with satisfactory cross-validation scores. But no validation by test set being given, it is impossible to assess the real prediction power of the model. Moreover, a data set of 134 compounds is poorly representative of the drug population. A similar work by Koutsoukos et al. [12], performed on the same data set by Principal Component Analysis, gave slightly lower results but, once more, no test set validation was enclosed.

Later, another computational approach was used to build antineoplastic models, consisting in relating the activity values or the mechanisms of action with 2D and 3D molecular descriptors [6]. These models do not need to determine experimental properties, but only *in silico* parameters, which can be computed on a personal computer in a few minutes for several thousands of compounds.

Firstly, several quantitative structure–activity relationships (QSAR) were established, e.g., on series of ellipticine and camptothecin analogues, for predicting cell line activities [15, 16]. Several modeling methods were applied, such as genetic function approximations (GFA), partial least squares, stepwise linear regression etc.; it turned out that GFA allowed to obtain the best results and quite robust models. But these models are derived from congeneric and very reduced data sets, including less than 170 compounds. Then, the field of application of such QSAR approaches is clearly inappropriate if the aim is to screen large combinatorial databases to obtain focused libraries with increased diversity and anticancer properties.

With the latter objective, Blower et al. presented an interesting method based on Recursive Partitioning and Simulated Annealing (RP/SA) to identify, in the NCI database, groups of active compounds having some key features in common [17]. More particularly, about 28,300 compounds were distributed in a regression tree that models a cell line activity on the space of LeadScope structural features. Each node was characterized by a specific combination of descriptors and the terminal nodes with high mean activities roughly corresponded to different classes of compounds. However, at present, no real and validated RP/SA model is proposed to predict the activity of new compounds; as a result it is impossible to evaluate its efficiency in a virtual screening strategy.

Finally, another attractive solution is proposed by PASS. This software can predict, for each compound, hundreds of different mechanisms of action, 50 of which could be related to antineoplastic actions [18]. It was specifically tested on the angiogenesis inhibitors of the NCI database, but at present drawing statistically valid conclusions about its prediction ability is difficult. Moreover, such a screening may remain inconclusive for assessing an antineoplastic activity; actually, an angiogenesis inhibitor can be related to such other pharmacological areas as cardiovascular or rheumatoid arthritis therapies [19], and the software does not seem to give any information about that.

The aim of this paper is to propose a further improvement, consisting in classifying about 1300 antineoplastic compounds derived from the Cipsline® database [20] and subdivided into eight mechanisms of action. A large set of molecular descriptors was computed on the 2D structures and the most relevant parameters were selected by a procedure combining the genetic algorithm concepts and a stepwise technique [21]. Then, structure–activity relationships were established with help of a recursive partitioning method derived from Fuzzy Logic [22] and named Adaptive Fuzzy Partition (AFP) [23]. Fuzzy Logic concepts are worth using to handle the ‘concept of partial truth’ and to represent the boundaries between classes as continuous, assigning to the compounds a degree of membership of each class. Then, they provide suitable solutions to problems within the context of imprecise categories, in which antitumor activities can be included.

In terms of classification ability, AFP allowed to obtain better results than other standard techniques, like classification and regression tree (CART), learning vector quantization, backpropagation neural network, discriminant analysis, when complex data sets concerning Central Nervous System (CNS) and environmental toxicity were analyzed [23, 24]. More generally, this technique was also successfully applied to ADME and olfaction fields [25, 26].

Another main object of this work consisted in establishing a strict and 'exhaustive' procedure to validate the AFP model proposed. Actually, most models available in literature, above all in the classification context, lack rigorous validation and remain an 'academic' exercise that is of no use in real applications [27]. More particularly, the following steps were to be evaluated carefully:

i. defining a 'good quality' data set and selecting the most relevant molecular descriptors to model the biological activity under analysis;
ii. selecting training and test sets in a rational way, in order to enclose all the molecular diversity covered by the global data set;
iii. validating the model by several and simultaneous robustness approaches such as cross-validation procedures, Y-scrambling test, etc.;
iv. evaluating the prediction power of the model with help of a test set.

Materials and methods

Compound selection

A data set of antineoplastic compounds was selected from the Cipsline® database by Prous [20]. A first analysis of this database allowed to retrieve about 4500 compounds. But after discarding the compounds presenting undetermined activities or metal atoms, incomplete structures, complex molecules and numerically negligible mechanisms, only 1294 compounds were included in the final work data set. This data set regroups eight activities, characterized by different mechanisms of action: (i) antimetabolites, (ii) alkylating agents, (iii) antracyclines, (iv) topoisomerase inhibitors, (v) tubulin-active antimitotic agents, (vi) tyrosine kinase inhibitors, (vii) farnesyl transferase inhibitors, and (viii) aromatase inhibitors.

The data set compounds were split into three sets: training, validation and test sets. The test set included molecules that were never used for developing the model. The validation set was used during the development of the model, based on the training set, to optimize the parameters and to validate the model.

Molecular descriptor selection

General molecular descriptors have proved to be a good compromise for data mining in large databases, as they are able to account for the main structural feature of each molecule. The antineoplastic data set was then distributed within a hyperspace defined by 167 descriptors, computed on 2D structures by the software MDL QSAR [28]. This descriptor set included constitutional, informational, topological, physicochemical, and electronic parameters. More details about the different molecular descriptors used can be found in refs. [23, 29].

To select, amidst the 167 descriptors, the best parameters for classifying the data set compounds, an Hybrid Selection Algorithm (HSA) was used, based on genetic algorithm (GA) concepts [21, 30]. GA, inspired by population genetics, is very effective for exploratory search, applicable to problems where little information is available, but it is not particularly suitable for local search. So a stepwise approach was also implemented in combination with GA in order to reach local convergence [21, 31], as it is quick and adapted to find solutions in 'promising' areas already identified.

Finally, a specific classification index was derived by the fuzzy clustering method [29] to evaluate the fitness function of HSA. This index has the advantage to be calculated quite quickly and to give an estimate of the descriptor relevance also by analyzing complex molecular distributions, in which finding separating edges between the different categories is difficult.

Furthermore, to prevent over-fitting and a poor generalization, a cross-validation procedure was included in the algorithm during the selection procedure, randomly dividing the data set into training and test sets. The fitness score of each set of descriptors derives from the combination of the scores of both the training and test sets.

All details about the strategy proposed for molecular descriptor selection and the proprietary software used can be found in ref. [21].

The following parameters were used to process the antineoplastic data set by HSA:

(i) Fuzzy parameters: weighting coefficient = 1.5; tolerance convergence = 0.001; number of iterations = 30; cluster number = 6.
(ii) Genetic parameters: chromosome number = 10; chromosome size = 167 (number of descriptors used); initial active descriptors in each chromosome = 8; crossover point number = 1, percentage of rejections = 0.1, percentage of crossover = 0.8, percentage of mutation = 0.05, number of generations = 10.
(iii) Stepwise parameters: ascending coefficient = 0.02; descending coefficient = −0.02.

Self-Organizing Maps (SOM)

SOM is a non-linear mapping technique which gives a 2D space representation of a given set of points from a multidimensional space derived from a series of molecular descriptors [32]. Each point of this set is related to a SOM node, which is characterized by N weighted connections varying between 0 and 1.

Training SOM consists in rearranging the layer nodes by gradually adjusting their weights. After selecting a first hyperspace point, the distances between its coordinates and each node of the SOM layer are calculated. The nearest node is called 'winner' and the hyperspace point is 'projected' on this node of the map. Then, the weights of the winning node and its neighbors are modified according to the equation:

$$w_{ij}(t+1) = w_{ji}(t) + \alpha(t)\gamma(t,r)(x_j - w_{ij}(t)), \tag{1}$$

where x_j is the component j of input vector x; w_{ij} represents the weight vector of the node i for the descriptor j; t and $\alpha(t)$ are respectively the iteration number and the learning rate; $\gamma(t, r)$ is the triangular neighborhood function depending on the iteration number and the distance r between the node i and the winning unit.

The learning rate $\alpha(t)$ is linearly decreased during the training process from $\alpha(0)$ to zero. The triangular function $\gamma(t, r)$ works on the whole map and it is discretely decreased with an increase in the distance and the number of iterations.

The same procedure is successively repeated for all the hyperspace vectors and each point is associated with a node in the SOM layer. The points which are close in the descriptor hyperspace remain close in the SOM layer, occupying the same nodes or the neighboring ones. When SOM is applied on a chemical data set, the maps can then reveal similar compounds or natural regroupings, if the Euclidean distance is accepted as a similarity measure.

The calculations were performed using proprietary software and the following parameters: number of columns = 10, number of rows = 10, coefficient γ for bias calculations = 0.01, number of iterations for the training phase = 50,000, coefficient β for frequency calculations = 0.001.

Adaptive Fuzzy Partition (AFP)

AFP is a supervised classification method implementing a fuzzy partition algorithm [33]; it was already fully presented [23] and validated elsewhere [23–26]. It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces defined by fuzzy rules. The aim of the algorithm is then to select the descriptor and the cut position which allow to get the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace A and in its neighboring subspaces.

Let us assume that the working space is an n-dimension hyperspace defined by n molecular descriptors, each dimension i can be partitioned into L intervals I_{ij}, where j represents an interval in the partition selected. Indicating with $P(x_1, \ldots, x_n)$ a molecular vector in an n-dimensional hyperspace, a *rule* for a subspace S_k, derived by combining n intervals I_{ij}, is defined by [34]:

$$\begin{aligned} &\text{if } x_1 \text{ is associated with } \mu_{1k}(x_1)\\ &\text{and } x_2 \text{ is associated with } \mu_{2k}(x_2)\ldots \text{ and}\\ &x_N \text{ is associated with } \mu_{Nk}(x_N) \rightarrow\\ &\text{the score of the activity } O \text{ for } P \text{ is } O_{kP}, \end{aligned} \tag{2}$$

where x_i represents the value of the ith descriptor for the molecule P, μ_{ik} is a trapezoidal membership function related to the descriptor i for the subspace k, and O_{kP} is the biochemical activity value related to the subspace S_k. The 'and' of the fuzzy rule is represented by the *Min operator* [35], which selects the minimal value amidst all the μ_{ik} components.

All the rules created during the fuzzy procedure are considered to establish the model between descriptor hyperspace and biochemical activities. The global score in the subspace S_k can be represented by

$$O_k = \frac{\sum_{j=1}^{M}(\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j}).(A_{P_j})}{\sum_{j=1}^{M}(\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j})}. \quad (3)$$

M is the number of molecular vectors in a given subspace, N is the total number of descriptors, $\mu_{ik}(x_i)_{p_j}$ is the fuzzy membership function related to the descriptor i for the molecular vector P_j, and A_{Pj} is the experimental activity of the compound P_j. A classic centroid defuzzification procedure [36] is implemented to determine the chemical activity of a new test molecule. All the subspaces k are considered and the general formula to compute the score of the activity O for a generic molecule P_j is

$$O(P_j) = \frac{\sum_{k=1}^{\mathrm{N\_subsp}}(\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j}) \cdot (O_k)}{\sum_{k=1}^{\mathrm{N\_subsp}}(\mathrm{Min}_i^N \mu_{ik}(x_i)_{P_j})} \quad (4)$$

where N_subsp represents the total number of subspaces.

The following AFP parameters were used to process the antineoplastic data set: maximal number of rules for each chemical activity = 30; minimal number of compounds for a given rule = 5; maximal number of cuts for each axis = 9.

Validation tools

The robustness of the AFP models was evaluated on the training set by two main techniques, leave-several-out (LSO) and Y-randomization test. Actually, these methods, derived from QSAR works [6, 27], were slightly modified to be adapted to the classification aims. LSO is a cross-validation method consisting in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. A classification LSO coefficient is computed by evaluating the percentage of these 'test' compounds rightly predicted. The procedure is iterated many times and the related model should be reasonably robust if a high average LSO coefficient is obtained. In this work, the training set was subdivided into 10 groups and, then, the LSO procedure was performed 10 times before computing the final LSO value.

The Y-randomization test is another widely used method in which the dependent-variable vector, Y-vector, is randomly shuffled and a new model is established by using the same original independent-variable matrix. After repeating this test several times, the average LSO value is expected to be low. If a high score is obtained, the original model is not acceptable, as due to a chance correlation or a structural redundancy in the training set [27].

Results and discussion

Molecular descriptor and training set selection

The first step of the data mining strategy consisted in selecting the most relevant molecular descriptors, by HSA, amidst the global set of 167 parameters (Table 1). The nine descriptors selected cover a wide diversity and include a constitutional parameter (ncirc), topological (xvch3, xvch9) and electrotopological indices (SssCH2, SdO, SssO, SHsSH, SHvin), and, finally, a lipophilicity parameter (log *P*). The latter descriptor plays a fundamental role in developing classification models also in many other biological fields, as proved in previous works concerning, for example, CNS mechanisms, ADME properties, and environmental toxicity [23, 25, 37].

These nine descriptors were firstly used to derive the SOM chart reported in Figure 1, which represents the 2D distribution of the eight compound classes included in the antineoplastic data set. A 10 × 10 map was employed, including 100 cells. The histogram heights define the number of compounds of each class in any cell of the map. Several regions of this chart can be clearly associated with specific antineoplastic activities. For example, anthracyclines, topoisomerase inhibitors, and farnesyl transferase inhibitors are defined

Table 1. List of the nine most relevant descriptors selected by HSA.

Symbol	Definition	Descriptor family
Ncirc	Number of graph circuits	Constitutional
Xvch3	Valence 3rd order chain chi indices	Topological
Xvch9	Valence 7th order chain chi indices	Topological
SssCH2	Sum of all (—CH_2—) E-state values	Electrotopological
SdO	Sum of all (=O) E-state values	Electrotopological
SssO	Sum of all (—O—) E-state values	Electrotopological
SHsSH	Sum of all hydrogen E-state values for (—SH)	Electrotopological
Shvin	Carbon atoms in the vinyl group (=CH)	Electrotopological
log *P*	Lipophilicity at pH = 7	Physicochemical

nearly univocally by three main 'clusters' represented in blue, yellow, and magenta, respectively. Analogously, antimetabolites and aromatase inhibitors can be regrouped into two couples of clusters, colored in red and black, respectively. The other properties in return, e.g. those of alkylating agents, are distributed throughout many regions. They include a lot of different chemical families that act by different submechanisms and explain the widely scattered distribution. Globally, this SOM map suggests the AFP method should be able to define robust structure–activity relationships by working directly on the hyperspace.

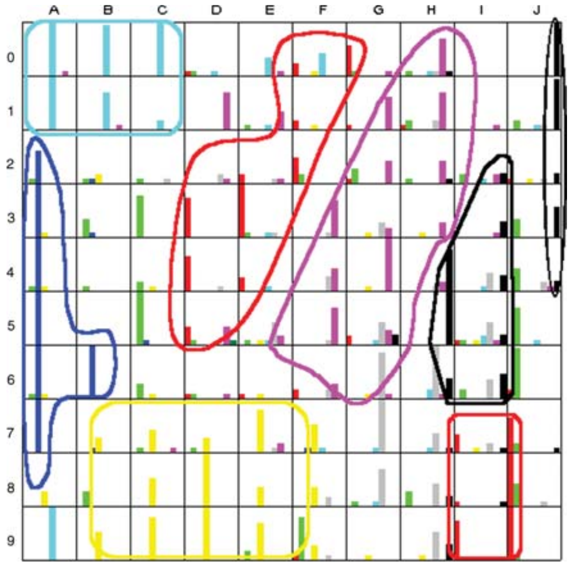

Figure 1. SOM chart derived from the 9D hyperspace in which the antineoplastic data set was distributed. The nine descriptors used are listed in Table 1. A 10 × 10 map was employed, including 100 cells. The histogram heights represent the number of compounds of each class in any cell of the map. This same chart was also used to define the training and test sets, selecting the compounds in accord to the molecular frequency and the mechanism of action. 🟥 = antimetabolites; 🟩 = alkylating agents; 🟦 = anthracyclines; 🟨 = topoisomerase inhibitors; 🟦 = tubulin-active antimitotic agents; ⬜ = tyrosine kinase inhibitors; 🟪 = farnesyl transferase inhibitors; ⬛ = aromatase inhibitors.

Besides giving a very useful representation of the compounds distribution, SOM is also a useful tool to select the training and test sets, maximizing, for each activity, the molecular diversity. The compounds were selected in each of the 100 cells of the map, according to the molecular frequency and the mechanism of action. The molecular subdivision of these three sets, within the eight activities, is summarized in Table 2.

Building and validating the classification model

The AFP model was established on the 640 training set compounds distributed in the 9D descriptor hyperspace. This model allowed to define eight descriptor–activity relationships, one for each antineoplastic activity, and 30 rules were implemented to define each of them. The parameters to use for developing the best model were tuned with help of the validation set.

Figure 2 shows the descriptor composition for each activity, evaluated by computing the frequency of the molecular parameters in the rules. Each descriptor contributes in a different way to the model development, according to the mechanism of action considered. For example, the electrotopological parameter SssCH2 shows a relevant contribution only in the rules concerning the antimetabolites compounds, so this descriptor can be considered as fundamental to individuate new compounds with selective antimetabolites activity.

Table 2. Compound repartition of the antineoplastic data set in the training, validation, and test sets.

Activity	Training set	Validation set	Test set
Antimetabolites	86	39	99
Alkylating agents	79	36	38
Intercalator antracyclines	80	35	48
Topoisomerase inhibitors	80	35	46
Tubulin-active antimitotic agents	80	35	30
Tyrosine kinase inhibitors	80	35	31
Farnesyl transferase inhibitors	80	35	55
Aromatase inhibitors	75	30	27
Total	640	280	374

Moreover, the lipophilicity parameter (log *P*) and the number of graph circuits (ncirc) are important to discriminate all antineoplastic activities, although their relative weight is quite different, by varying the mechanism of action.

The AFP method allows to get the degrees of membership of the different activities for each compound, within an 0–1 range. Then, a compound is attributed to a given mechanism if its degree of membership is the highest amidst all eight values and superior to 0.3. The detailed validation results for the best model are shown in Table 3. The experimental antineoplastic activity was predicted correctly for 86% of the validation compounds and a similar score was obtained when testing the training compounds, showing the model developed has a general application. Moreover, the prediction scores on the single mechanisms of action do not present relevant gaps.

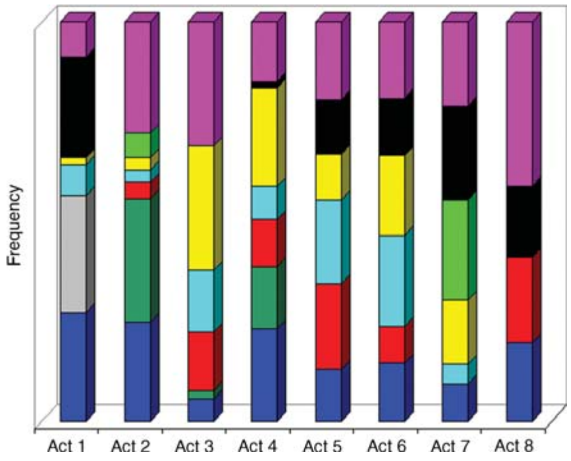

Figure 2. Representation of the descriptor contribution for each activity, evaluated by computing descriptor frequency in the AFP rules associated with the best model developed. 🟦 = ncirc; ■ = xvch3; 🟥 = xvch9; ▨ = SssCH2; □ = SdO; 🟨 = SssO; 🟩 = SHsSH; ⬛ = SHvin; 🟪 = log *P*. Act1 = antimetabolites; Act2 = alkylating agents; Act3 = antracyclines; Act4 = topoisomerase inhibitors; Act5 = tubulin-active antimitotic agents; Act6 = tyrosine kinase inhibitors; Act7 = farnesyl transferase inhibitors; Act8 = aromatase inhibitors.

Furthermore, the robustness of this AFP model is chiefly confirmed by the LSO score that reaches quite impressively 75%! Even more important, this value is quite similar to the prediction scores associated with the training and validation sets; this means the model is not significantly perturbed by eliminating 10% of training set information and represents well all different structure–mechanism relationships associated with the data set.

Finally, a Y-randomization test was performed by shuffling 50% of the training set compounds; this process was repeated five times. The final average LSO score falls down to 30% and underlines, once more, that no chance correlation or compound redundancy is present in the model.

Evaluating the prediction ability of the best AFP model

As recently underlined, a satisfactory model robustness is a necessary condition to have a high prediction power, but it is not a sufficient condition [38]. The prediction ability of a model can be assessed only with help of an external test set never used to build or validate the model. Table 3 shows the percentage of test compounds rightly predicted, globally and mechanism by mechanism. The good average score of 77% indicates that the AFP model is undoubtedly predictive. Moreover, also the last two columns of Table 3, reporting the

Table 3. Statistical validation scores derived from the best AFP model developed on the antineoplastic data set.

Activity	ID	Training set (%)	Validation set (%)	Test set (%)	PPV (%)	NPV (%)
Antimetabolites	Act 1	84	90	74	92	94
Alkylating agents	Act 2	81	89	71	73	94
Antracyclines	Act 3	96	97	88	91	96
Topoisomerase inhibitors	Act 4	89	94	94	70	91
Tubulin-active antimitotic agents	Act 5	83	89	83	81	95
Tyrosine kinase inhibitors	Act 6	80	80	68	51	90
Farnesyl transferase inhibitors	Act 7	80	74	60	83	95
Aromatase inhibitors	Act 8	87	73	78	81	96
Total		86	86	77	76	94

The columns under ‘validation’ report the percentages of compounds whose activities are predicted correctly. The last two columns, positive prediction values (PPV) and negative prediction values (NPV), are associated with the test set and indicate which is the probability that a positive (negative) prediction is really positive (or negative).

Table 4. Percentages of correct (in bold) and incorrect predictions for each antineoplastic mechanism of the test. The last column, *None*, means the related compounds were attributed to no mechanism.

Experimental activity	Predicted activity								
	Act1	Act2	Act3	Act4	Act5	Act6	Act7	Act8	None
Act1	**74**	5	2	4	3	8	0	1	3
Act2	7	**71**	3	5	5	3	0	5	0
Act3	0	0	**88**	4	2	2	2	0	2
Act4	0	2	0	**94**	0	2	1	0	0
Act5	0	0	0	3	**83**	0	3	3	7
Act6	0	10	0	6	0	**68**	13	3	0
Act7	2	0	2	11	0	20	**60**	0	5
Act8	7	4	0	4	0	4	0	**78**	4

Act1 = antimetabolites; Act2 = alkylating agents; Act3 = antracyclines; Act4 = topoisomerase inhibitors; Act5 = tubulin-active antimitotic agents; Act6 = tyrosine kinase inhibitors; Act7 = farnesyl transferase inhibitors; Act8 = aromatase inhibitors.

positive and negative prediction values (PPV and NPV), confirm that the model is globally satisfactory. More precisely, these values indicate the probability that a positive (or negative) prediction is really positive (or negative).

By analyzing more deeply the single mechanism predictions (Table 4), it has to be underlined that the scores vary sensibly, from 60% to 94%. Beside the obvious limits and errors of any models developed, these results can be explained by the multi-mechanism behavior of many compounds. For example, alkylating agents are among those whose activities have the worst predictability. They include chemical families such as nitrogen mustards, nitrosoureas, triazenes, and ethylene imines which act by different mechanisms of action and exhibit specific effects against various types of cells and experimental tumors. All this information is not included in the training set here proposed and the prediction ability is inevitably weakened.

Conclusions

Pharmaceutical companies have developed ambitious projects for accelerating the drug discovery process, more particularly in the anticancer field. This has involved the combined use of CC and HTS strategies that have generated huge amounts of screening information on a lot of disease targets. All this information now has to be rightly managed. This requires to develop efficient data mining tools able to classify large anticancer data bases and to design focused libraries with enhanced informational contents.

Amidst these data mining methods, Fuzzy Logic concepts constitute an interesting solution to derive general classification models. An AFP algorithm was applied on a data set of about 1300 antineoplastic compounds, subdivided into eight mechanisms of action. An important part of the work consisted also in implementing several procedures of validation to assess the robustness and prediction ability of the models developed. Actually, many models available in literature often lack such requirements and their practical use becomes negligible.

After selecting the most relevant descriptors by a procedure based on genetic algorithms, the AFP model was established on a training set of 640 compounds selected rationally by SOM. A first internal validation was performed with help of a validation set of 280 compounds, used also to optimize the best AFP parameters. The experimental mechanism of action was predicted correctly for 86% of the validation compounds. But the robustness of the model developed was chiefly confirmed by cross-validation and Y-randomization tests. The first one gives a very satisfactory score of 75%; the second one shows the model is significant with a high degree of certitude, and no chance correlation or training compound redundancy is hidden in it.

Finally, the prediction ability of the model was evaluated on a set of 374 test compounds never used to establish the model. A good validation score of 77% was obtained. Moreover, it has to be also underlined that with the AFP method it takes only a few minutes to test several thousands of molecules. Then, it could be used to rapidly screen large series of compounds in order to find new potential anticancer hits.

The perspectives of this work are now orientated towards two main goals:

i. Generating binary models able to make a preliminary and fast discrimination active/inactive, and generic/specific antineoplastic targets. The model here proposed will represent then the final step of mechanism optimization.
ii. Adding models that will pre-screen ADME-Tox properties. This would allow to integrate all the factors involved in a global Drug Discovery strategy. The ability of AFP in facing such complex issues was already proved elsewhere [25, 37].

References

1. Boyd, M.R., In Teicher B.A. (Ed.), Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval. Humana Press, Totowa, NJ, 1997, pp. 23–42.
2. MDL® Drug Data Report. MDL Information Systems Inc., San Leandro, CA, 2004.
3. Willett, P., Perspect. Drug Discov. Des., 7/8 (1997) 1.
4. Bayada, D.M., Hamersma, H. and van Geerestein, V.J., J. Chem. Inf. Comput. Sci., 39 (1999) 1.
5. Gordon, E.M. and Kerwin, J.F. Jr. (Eds.), Combinatorial Chemistry and Molecular Diversity in Drug Discovery. Wiley, New York, 1998.
6. van de Waterbeemd, H. (Ed.), Chemometric Methods in Molecular Design. VCH, Weinheim, Germany, 1995.
7. Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.), 3D QSAR in Drug Design, Recent Advances. Kluwer/Escom, Dordrecht, The Netherlands, 1998.
8. Paull, K., Shoemaker, R.H., Hodes, L., Monks, A., Scudiero, D.A., Rubinstein, L., Plowman, J. and Boyd, M.R., J. Natl. Cancer Inst., 81 (1989) 1088.
9. Paull, K., Hamel, E. and Malspeis, L., In Foye, W.O. (Ed.), Cancer Chemotherapeutic Agents. American Chemical Society Books, Washington, DC, 1995, pp. 9–45.
10. Weinstein, J.N., Kohn, K.W., Grever, M.R., Viswanadhan, V.N., Rubinstein, L.V., Monks, A.P., Scudiero, D.A., Welch, L., Koutsoukos, A.D., Chiausa, A.J. and Paull, K.D., Science, 258 (1992) 447.
11. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. and Weinstein, J.N., J. Natl. Cancer Inst., 86 (1994) 1853.
12. Koutsoukos, A.D., Rubinstein, L.V., Faraggi, D., Simon, R.M., Kalyandrug, S., Weinstein, J.N., Kohn, K.W. and Paull, K.D., Stat Med., 13 (1994) 719.
13. Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J. Jr., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johnson, G.S., Wittes, R.E. and Paull, K.D., Science, 275 (1997) 343.
14. Shi, L.M., Fan, Y., Lee, J.K., Waltham, M., Andrews, D.T., Scherf, U., Paull, K.D. and Weinstein, J.N., J. Chem. Inf. Comput. Sci., 40 (2000) 367.
15. Shi, L.M., Myers, T.G., Fan, Y., O'Connor, P.M., Paull, K.D., Friend, S.H. and Weinstein, J.N., Mol. Pharmacol., 53 (1998) 241.
16. Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y. and Weinstein, J.N., J. Med. Chem., 44 (2001) 3254.
17. Blower, P., Fligner, M., Verducci, J. and Bjoraker, J., J. Chem. Inf. Comput. Sci., 42 (2002) 393.
18. Poroikov, V.V., Filimonov, D.A., Ihlenfeldt, W.D., Gloriozova, T.A., Lagunin, A.A., Borodina, Y.V., Stepanchikova, A.V. and Nicklaus, M.C., J. Chem. Inf. Comput. Sci., 43 (2003) 228.
19. Folkman, J., Nat. Med., 1 (1995) 27.
20. CIPSLINE ®. Prous Science, Barcelona, Spain, 2003.
21. Ros, F., Pintore, M. and Chrétien, J.R., Chemometr. Intell. Lab. Syst., 63 (2002) 15.
22. Zadeh, L.A., In Van Ryzin, J. (Ed.), Classification and Clustering. Academic Press, New York, 1977, pp. 251–299.

23. Ros, F., Taboureau, O., Pintore, M. and Chrétien, J.R., Chemometr. Intell. Lab. Syst., 67 (2003) 29.
24. Pintore, M., Piclin, N., Benfenati, E., Gini, G. and Chrétien, J.R., Environ. Toxicol. Chem., 22 (2003) 983.
25. Pintore, M., van de Waterbeemd, H., Piclin, N. and Chrétien, J.R., Eur. J. Med. Chem., 38 (2003) 427.
26. Pintore, M., Audouze, K., Ros, F. and Chretien, J.R., Data Sci. J., 1 (2002) 99.
27. Tropsha, A., Grammatica, P. and Gombar, V.K., QSAR, Comb. Sci., 22 (2003) 69.
28. MDL® QSAR version 2.2. MDL Information Systems Inc., San Leandro, CA, 2003.
29. Pintore, M., Taboureau, O., Ros, F. and Chrétien, J.R., Eur. J. Med. Chem., 36 (2001) 349.
30. Haupt, R.L. and Haupt, S.E. (Eds), Practical Genetic Algorithms. Wiley, New York, 1998.
31. Leardi, R. and Gonzales, A.L., Chemometr. Intell. Lab. Syst., 41 (1998) 195.
32. Kohonen, T. (Ed.), Self-Organizing Maps. Springer-Verlag, Berlin, Germany, 2001.
33. Lin, Y. and Cunninghan, G.J., J. Intell. Fuzzy Syst., 2 (1994) 243.
34. Sugeno, M. and Yasakawa, T.A., IEEE T. Fuzzy Syst., 1 (1993) 7.
35. Dubois, D. and Prade, H., In Shafer, G. and Pearl, J. (Eds.), Readings in Uncertain Reasoning. Morgan Kaufmann, San Francisco, CA, 1990, pp. 742–761.
36. Gupta, M.M. and Qi, J., Fuzzy Set Syst., 40 (1991) 431.
37. Pintore, M., Piclin, N., Benfenati, E., Gini, G. and Chrétien, J.R., QSAR Comb. Sci., 22 (2003) 210.
38. Goldbraikh, A. and Tropsha, A., J. Mol. Graph. Model., 20 (2002) 269.