Journal of Computer-Aided Molecular Design, **17:** 53–64, 2003.
KLUWER/ESCOM
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

In silico rationalization of the structural and physicochemical requirements for photobiological activity in angelicine derivatives and their heteroanalogues

Fabrizio Giordanetto[a], Paola Fossa[b,*], Giulia Menozzi[b] & Luisa Mosti[b]
[a]*Centre for Computational Science, Department of Chemistry, Queen Mary, University of London, Mile End Road, London E1 4NS, United Kingdom;* [b]*Dipartimento di Scienze Farmaceutiche, Università degli Studi di Genova, Viale Benedetto XV, n.3 -16132 Genova, Italy;*

Received 18 July 2002; Accepted 2 March 2003

Key words: angelicine, angelicine derivatives, molecular descriptors, photobiological activity, neural networks, QSAR model

Summary

In PUVA (Psoralen plus UVA) chemotherapy 8-methoxypsoralen is the most widely used compound, although its efficacy is endowed with undesired side effects. In order to have an evident anti-proliferative activity with a reduced phototoxicity, many linear and angular derivatives have been synthesised. In this paper we describe a QSAR study in which, by means of the neural networks methodology, a useful model for predicting biological activity, expressed as ID_{50} (the UVA dose that reduces to 50% the DNA synthesis in Ehrlich cells), has been derived. A decision tree that is able to discriminate between active and inactive compounds has been built based on recursive partitioning. The study shows the key structural features responsible for the activity and could be a helpful tool in the rational design of new, less toxic, photochemotherapeuthic agents.

Introduction

Psoralens are photoactive drugs used in PUVA (psoralen plus UVA) therapy to cure several skin diseases [1], in photopheresis to prevent organ transplants rejection and to treat T-cell lymphoma [2].

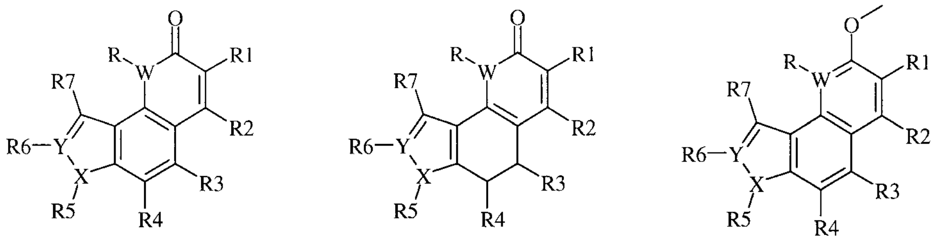
In spite of the high efficacy of PUVA therapy, this treatment shows some severe side effects, such as skin erythemas [1], genotoxicity [3–5] and carcinogenicity [6], mostly attributed to the lesions induced in DNA by furocoumarin sensitisation. Indeed, various kinds of lesions are formed, i.e. covalent mono- and diadducts with pyrimidine bases [7], and covalent DNA-protein cross-links (DPC) [8]. In particular, inter-strand cross-links (ISC) are regarded as the main cause responsible for the furocoumarin toxicity.

Therefore, to obtain less toxic compounds, various authors have prepared and studied several new compounds characterized by a prevalent ability of forming monofunctional damage. Among these, angelicine derivatives and isosters represent the main research line, since their angular structure prevents the formation of ISC for geometrical reasons [9–10].

Our interest to rationalise the quantitative structure-activity relationships (QSAR) on these compounds, led us to perform the present study on a number of angelicine derivatives previously synthesised in our laboratory [11–15], together with some literature compounds [9, 16–18]. This will serve as basis for the design of better and less toxic photochemotherapeutic agents. The selected molecular set was chosen in order to cover the experimental space of the majority of possible chemical modifications on the angular skeleton of angelicin. Several molecular descriptors have been calculated for each compound, and a correlation analysis with their photobiological activity, expressed as inhibition of DNA synthesis in Ehrlich cells (ID_{50}), has been performed using different statis-

[*] To whom correspondence should be addressed.
E-mail: fossap@unige.it

Table 1. Molecular structures and biological activities for the data set.



Compound	W	X	Y	R	R1	R2	R3	R4	R5	R6	R7	ID$_{50}$^{a}
Angelicine	O	O	C	-	H	H	H	H	-	H	H	25.3
1	O	O	C	–	H	CH_3	H	CH_3	–	H	CH_3	0.06
2	N	O	C	CH_3	H	H	H	H	–	H	H	0.70
3	N	S	C	CH_3	H	H	H	H	–	H	H	0.89
4	N	O	C	H	CO_2CH_3	H	H	H	–	H	H	1.34
5	O	S	C	–	H	H	H	H	–	H	H	1.48
6	N	S	C	H	H	H	H	H	–	H	H	2.09
7	N	O	C	H	H	H	H	H	–	H	H	2.17
8	O	O	C	–	H	H	H	CH_3	–	H	CH_3	2.20
9	O	O	C	–	H	CH_3	H	CH_3	–	CH_3	H	3.87
10	O	O	C	–	H	CH_3	H	CH_3	–	H	H	4.00
11	O	O	C	–	Cl	$N(CH_3)C_6H_5$	H	H	–	H	H	6.85
12	N	O	C	CH_3	CO_2CH_3	H	H	H	–	H	H	7.51
13	O	O	C	–	H	CH_3	H	H	–	CH_3	$CH_2O(CH_2)_2N(CH_3)_2$	8.67
14	O	O	C	–	$CO_2C_2H_5$	H	H	H	–	H	H	9.00
15	O	O	C	–	H	H	CH_3	H	–	H	H	9.10
16	O	O	C	–	H	CH_3	H	H	–	CH_3	CH_2OH	9.20
17	O	O	C	–	H	CH_3	H	H	–	CH_3	H	10.80
18	O	S	C	–	Cl	morpholin-4-yl	H	H	–	H	H	10.80
19	O	O	C	–	H	CH_3	H	H	–	CH_3	H	10.90
20	N	O	C	H	$CO_2C_2H_5$	H	H	H	–	H	H	11.80
21	O	O	C	–	H	H	H	CH_3	–	CH_3	H	11.86
22	N	S	C	H	CO_2H	H	H	H	–	H	H	13.08
23	O	O	C	–	Cl	morpholin-4-yl	H	H	–	H	H	13.50
24	O	O	C	–	H	H	CH_3	H	–	CH_3	H	13.80
25	O	O	C	–	H	H	H	H	–	CH_3	H	16.60
26	N	O	C	H	CO_2H	H	H	H	–	H	H	17.62
27	O	O	C	–	CO_2CH_3	H	H	H	–	H	H	18.00
28	N	S	C	H	CO_2CH_3	H	H	H	–	H	H	18.06
29	O	O	C	–	Cl	$N(CH_3)_2$	H	H	–	H	H	20.20
30	O	S	C	–	Cl	$N(CH_3)C_6H_5$	H	H	–	H	H	22.30
31	O	O	C	–	H	CH_3	H	H	–	CH_3	CH_2OCH_3	22.70
32	O	S	C	–	Cl	$N(CH_3)_2$	H	H	–	H	H	23.00
33	O	O	C	–	H	CH_3	H	H	–	H	H	24.00
34	O	O	C	–	H	H	H	CH_3	–	H	H	26.30
35	N	S	C	H	$CO_2C_2H_5$	H	H	H	–	H	H	27.73
36	O	O	C	–	H	CH_3	CH_3	CH_3	–	H	H	30.00
37	O	O	C	–	H	CH_3	CH_3	CH_3	–	CH_3	H	30.00
38	N	O	C	H	$CO_2C_2H_5$	H	H	H	–	H	H	38.10
39	O	S	C	–	CO_2H	H	H	H	–	H	H	45.00
40	N	O	C	H	CO_2CH_3	H	H	H	–	H	H	49.70

Table 1 continued.

Compound	W	X	Y	R	R1	R2	R3	R4	R5	R6	R7	ID_{50}[a]
41	O	O	C	–	$N(CH_3)_2$	H	H	H	–	H	H	50.90
42	O	O	C	–	CO_2H	H	H	H	–	H	H	51.90
43	O	O	C	–	morpholin-4-yl	H	H	H	–	H	H	57.50
44	O	O	C	–	NH_2	H	H	H	–	H	H	76.30
45	O	O	C	–	$COCH_3$	H	H	H	–	H	H	> 250
46	O	O	C	–	COC_6H_5	H	H	H	–	H	H	> 250
47	O	O	C	–	$N(C_2H_5)_2$	H	H	H	–	H	H	> 250
48	O	O	C	–	$NH(CH_2)_2OH$	H	H	H	–	H	H	> 250
49	O	O	C	–	$NH(CH_2)_2OC_2H_5$	H	H	H	–	H	H	> 250
50	O	O	C	–	$NH(CH_2)_2N(CH3)_2$	H	H	H	–	H	H	> 250
51	O	O	C	–	$N(C_2H_4OH)_2$	H	H	H	–	H	H	> 250
52	O	O	C	–	4-methyl-piperazin-1-yl	H	H	H	–	H	H	> 250
53	O	S	C	–	CO_2CH_3	H	H	H	–	H	H	> 250
54	N	S	C	CH_3	CO_2CH_3	H	H	H	–	H	H	> 250
55	O	N	N	–	CO_2CH_3	H	H	H	CH_3	H	H	> 250
56	N	S	C	H	CO_2CH_3	H	H	H	–	H	H	> 250
57	N	S	C	H	$CO_2C_2H_5$	H	H	H	–	H	H	> 250
58	N	S	C	–	H	H	H	H	–	H	H	> 250
59	N	S	C	–	CO_2CH_3	H	H	H	–	H	H	> 250

[a] The UVA dose that reduces to 50% DNA synthesis in Ehrlich cells at 20 μM drug concentration.

tical methods. Among these, the application of neural networks has proved to yield better results. This is due to the fact that structure-activity relationships are often non-linear and very complex and neural networks are able to approximate any kind of analytical continuous function, according to Kolmogorov's theorem [19]. Another important characteristic of neural networks is their 'parsimony' [20]. This quality allows the networks to yield better results, with the same numbers of parameters, than other data modelling tools.

The result of this study is a satisfactory *in silico* screening method for predicting the photobiological activity of new synthesised compound before carrying out *in vitro* and *in vivo* evaluations.

Data and methods

Table 1 reports the chemical structures and the photobiological activities for the 60 photosensitisers [9, 11–18] employed in this study. The Cerius2™ [21] environment served for structural manipulation and for the development and analysis of QSAR models. Conformational analysis has been carried out using the Metropolis sampling and allowing the generation of 1000 different rotamers. The best 20 conformers, according to energy criteria, have been retained for further calculations.

Several molecular descriptors have been computed for each ligand. The different classes of descriptors employed are presented in Table 2 whereas a comprehensive listing is provided in the Appendix section.

The numerical values of conformation-dependent descriptors (e.g. molecular surface descriptors) corresponded to the descriptor means over all the conformations for a given molecule.

Different statistical methods have been applied to derive QSAR models. Among these, the selection of useful predictors by means of genetic algorithms has proved to yield better results in this study. Using the genetic function approximation (GFA) module available within Cerius2™ [21] an initial population of 1000 randomly built models has been evolved for 50000 generations. The initial equation length was set to 3 terms. Term creation types included linear (a*x, where 'a' is the coefficient to be fitted and 'x' is the independent variable), spline (a*(x — c), where 'c' is a constant), quadratic (a*x^2) and offset quadratic (a*(x^2 — c)) terms. During the evolution, the probability of creating a new term, reducing or extending the equation, was given a value of 50%. The resulting models have been evaluated by leave-one-out cross-validation

Table 2. Different classes of descriptors employed to derive quantitative structure-activity relationships. See the Appendix section for a detailed list.

Class	N. of descriptors
Electronic	2
Electrotopological	10
Information-content	14
Quantum mechanical	4
Spatial	48
Structural	5
Thermodynamic	4
Topological	21

and compared to the models built using randomised activity values.

Recursive partitioning has also been used to effectively include the compounds (**45–59**) whose biological activity was not precisely defined (Table 1). Classes have been weighted equally, the 'twoing' rule has been applied as splitting scoring and a moderate pruning has been performed on the obtained trees.

Machine learning techniques have been employed to develop QSAR models: back-propagated feed forward neural networks have been built using Matlab™ [22].

The architecture of the back-propagated networks involved three different levels: an input layer, a hidden and an output layer. All 108 molecular descriptors previously computed have been scaled using normalisation so that they had zero mean and unity variance. Moreover, principal component analysis has been performed on the normalised descriptors eliminating those principal components that contribute less than 5% to the total variation in the descriptor set. Input variables for the networks consisted of the resulting 6 principal components (see the Appendix section), whereas the output (target) variables were represented by the normalised biological activities.

The number of neurons in the hidden layer has been iteratively changed monitoring the predictive power of the resulting network. The hidden neurons displayed a scalar bias and a '*tanh*' activation function. The single output neuron showed a linear transfer function instead. Angelicin and compounds **1**–**44** have been divided into a training set (80% of the data), a validation (10%) and a test set (10%). This partition scheme has been used to optimise the number of neurons in the hidden layer which has been finally set to 16. Training has been performed using the Levenberg–Marquardt algorithm [23]. Here, the network performance for both the training and the validation set is monitored throughout the training process. Whenever the validation error increases for a specified number of iterations, the training is stopped (early stopping) and the resulting neural network at the minimum of the validation error is returned.

In order to validate the neural network, we carried out an extensive cross-validation procedure: among angelicine and compounds **1**–**44**, we randomly picked 35 molecules for the training set, 5 for the validation set and another 5 for the test set. This procedure was repeated 500 times in order to cover a large number of compounds combinations in the different sets. Here, for each cross-validation round, the network was trained using the Levenberg–Marquardt algorithm [23]. The validation set has been used to monitor the training process using early stopping and the predictive power of the resulting network has been evaluated on the corresponding test set. The test set used to optimise the number of neurons in the hidden layer has not been employed during the cross-validation procedure.

60 different compounds, which have been tested for their photobiological activity, represented the data set used to derive quantitative structure–activity relationships (Table 1). Different ways to evaluate the photochemotherapeutic efficacy of molecules are currently available. The UVA dose that reduces to 50% the DNA synthesis in Ehrlich cells (ID_{50}) has been employed here. This ranges from a minimum of 0.06 kJ/m^2 for the best photosensitiser, **1**, to a maximum of 73.6 kJ/m^2 for **44** (Table 1). Compounds ranging from **45** to **59** displayed ID_{50} values larger than 250 kJ/m^2 (Table 1). In those cases the molecules are inactive and their ID_{50} could not have been precisely measured. Therefore, these compounds were not included in the development of QSAR models based on regression techniques. However, they served to validate the predictivity of the resulting models and it was possible to use them in classification-based procedures (see below).

A total of 108 molecular descriptors (see Appendix), belonging to different classes (Table 2), have been computed. Analysis of the correlation matrix did not reveal any strong linear relationship between the biological activity and a particular descriptor. The 8 independent variables that showed the largest correlation with the activity for angelicine and compounds **1**–**44** are shown in Table 3. A slight improvement can be gained if we consider only the molecules **1**–**28** whose

Table 3. Correlation matrix for angelicine and compounds **1–44**. Only the best 8 descriptors are shown.

	HBA	AlogP	LUMO	DIPOLE	HOMO	S-XL	FH_2O	S_dO	Activity
HBA	1								
AlogP	−0.45	1							
LUMO	−0.81	0.43	1						
DIPOLE	0.53	−0.37	−0.74	1					
HOMO	−0.63	0.22	0.61	−0.48	1				
S-XL	0.72	−0.43	−0.56	0.34	−0.26	1			
FH_2O	−0.76	0.41	0.70	−0.42	0.53	−0.50	1		
S_dO	0.59	−0.66	−0.74	0.78	−0.43	0.61	−0.55	1	
Activity	0.42	−0.41	−0.49	0.43	−0.46	0.41	−0.49	0.51	1

HBA, hydrogen-bond acceptors count; **AlogP**, logarithm of the water/octanol partition coefficient; **LUMO**, lowest unoccupied molecular orbital energy; **DIPOLE**, dipole moment; **HOMO**, highest occupied molecular orbital energy; **S-XL**, molecular length in the X dimension; **FH_2O**, desolvation free energy for water; **S_dO**, electrotopological index for the carbonyl oxygen.

Table 4. Best QSAR models, as obtained using the genetic function approximation.

Model 1 Y =	−2215.18 −1590.53***K-2-AM** + 373.65***logZ** + 1563.48***PHI** + 223.26***FPSA-2** +1 5.91***S-XL** + 13.19***S_aasC** −192.52***HOMO** + 4442.32*((**BIC** −0.785058) 2) −101.13***SC-3_C** + 0.25***WPSA-2**
Model 2 Y =	−575.84 + 0.34***WPSA-2** −62.49***HOMO** + 2.74***RNCS** −93.79***CHI-V-1** + 439.21***RNCG** + 94.96***CHI-V-3_P** +8.05***S-XL** + 52.13*1.53-**S_aasC**
Model 3 Y =	−479.20 +6.14***HOMO**∧2 +498.49***RNCG**-96.88***CHI-V-1**+10.33***S-XL**+0.36***WPSA-2**−0.97***MW** +118.04***CHI-V-3_P** +16.72***CHI-2**

	r^2	F-test	XV r^2	BS r^2	BS Error	N obs	N vars	Outliers	LOF	R^2 adj	LSE	r
Model 1	0.867	22.247	0.834	0.895	0.002	45	11	2	138.137	0.828	39.292	0.931
Model 2	0.840	23.589	0.747	0.840	0.004	45	9	1	122.647	0.804	47.484	0.916
Model 3	0.826	21.312	0.777	0.861	0.003	45	9	2	124.419	0.787	51.672	0.909

K-2-AM, second order Kier's alpha-modified shape index; **logZ**, logarithm of the Zagreb index; **PHI**, molecular flexibility index; **FPSA-2**, fractional positive partial surface area; **S-XL**, length of the molecule in the X dimension; **S_aasC**, electrotopological state for the substituted aromatic carbon; **HOMO**, highest occupied molecular orbital energy; **BIC**, bonding information content index; **SC-3_C**, Kier and Hall third order path subgraph count index; **WPSA-2**, surface-weighted positive partial surface area; **RNCS**, relative negative charge surface area; **CHI-V-1**, first order Kier & Hall valence-modified chi index; **RCNG**, relative negative charge; **CHI-V-3_P**, third order Kier & Hall valence-modified chi index; **MW**, molecular weight; **CHI-2**, second order Kier & Hall chi index, **XV r^2**, cross validated r^2, **BS r^2**, bootstrap r^2; **BS error**, bootstrap error; **N obs**, number of observations; **N vars**, number of independent variables; **LOF**, lack of fit score; **r^2 adj**, adjusted r^2; **LSE**, least squared error.

activity falls below 20 kJ/m^2. Enrichment is obtained for HBA, LUMO, HOMO, S-XL and FH_2O whereas AlogP, DIPOLE and S-dO displayed a decreased correlation. The correlation matrix indicated high degrees of collinearity between several descriptors, thus leading to an overparameterised set. It was therefore necessary, prior to model building, to actively select the most relevant independent variables.

Results and discussion

We used genetic algorithms to efficiently select the descriptors during the model building process. This method produced interesting results and the best models are displayed in Table 4. These equations differ for the number of terms employed, for their complexity and for their statistical significance. However, a common feature among them is the use of the HOMO, the S-XL and the WPSA-2 molecular descriptors (Table 4). According to the evolutionary selection process, we can infer that these descriptors are the most interesting independent variables.

The energy of the highest occupied molecular orbital (HOMO), as computed with MOPAC, displayed both a linear (Model 1 and 2, Table 4) and a quadratic (Model 3, Table 4) relationship with the photobiological activity. The corresponding coefficient is negative

Table 5. Values of the most relevant descriptors for the biological activity, computed for the whole data set.

Compound	Activity	Class	HOMO	S-XL	WPSA-2
Angelicine	25.3	0	−9.00	8.93	92.49
1	0.06	1	−8.74	10.32	226.96
2	0.7	1	−8.50	10.20	138.49
3	0.89	1	−8.28	10.63	212.65
4	1.34	1	−8.98	12.16	216.29
5	1.48	1	−8.65	9.82	145.94
6	2.09	1	−8.38	10.07	170.99
7	2.17	1	−8.61	9.17	112.27
8	2.2	1	−8.77	10.28	171.78
9	3.87	1	−8.72	11.46	239.19
10	4	1	−8.83	10.03	178.13
11	6.85	1	−8.91	14.59	402.56
12	7.51	1	−8.86	12.30	258.64
13	8.67	1	−8.77	13.94	595.86
14	9	1	−9.32	13.25	260.32
15	9.1	1	−8.93	10.07	130.61
16	9.2	1	−8.76	12.33	253.95
17	10.8	0	−8.70	13.14	436.23
18	10.8	0	−8.83	11.91	185.63
19	10.9	0	−8.83	11.92	180.80
20	11.8	0	−8.86	15.10	337.86
21	11.86	0	−8.76	11.08	180.96
22	13.08	0	−8.64	12.67	253.83
23	13.5	0	−9.05	13.15	315.20
24	13.8	0	−8.79	11.57	172.75
25	16.6	0	−8.87	11.48	131.06
26	17.62	0	−8.88	12.53	182.32
27	18	0	−9.33	12.04	184.63
28	18.06	0	−8.72	12.34	302.15
29	20.2	0	−9.03	12.00	231.59
30	22.3	0	−8.72	14.61	523.95
31	22.7	0	−8.82	12.35	295.03
32	23	0	−8.69	11.91	332.07
33	24	0	−8.96	10.56	131.34
34	26.3	0	−8.86	10.32	131.97
35	27.73	0	−8.62	15.41	451.10
36	30	0	−8.63	11.79	279.66
37	30	0	−8.75	10.43	213.88
38	38.1	0	−8.74	15.12	424.31
39	45	0	−8.89	12.64	222.19
40	49.7	0	−8.75	13.92	325.49
41	50.9	0	−9.13	13.77	286.00
42	51.9	0	−9.25	12.47	150.90
43	57.5	0	−9.21	14.97	394.75
44	76.3	0	−9.15	12.55	185.74
45	> 250	0	−9.19	15.58	486.77
46	> 250	0	−9.25	15.39	344.23
47	> 250	0	−9.17	13.37	242.61
48	> 250	0	−9.16	15.21	493.54
49	> 250	0	−9.22	18.05	555.57
50	> 250	0	−9.12	15.70	550.92
51	> 250	0	−9.13	14.98	410.10
52	> 250	0	−8.53	14.12	400.37
53	> 250	0	−9.15	12.39	179.16
54	> 250	0	−8.97	12.27	266.35
55	> 250	0	−9.14	15.33	306.16
56	> 250	0	−8.67	15.45	544.77
57	> 250	0	−8.68	14.15	425.86
58	> 250	0	−8.36	12.31	207.06
59	> 250	0	−8.68	14.02	360.74

HOMO, highest occupied molecular orbital energy; **S-XL**, molecular length in the X dimension; **WPSA-2**, surface weighted positive partial surface area.

for Model 1 and 2 whilst is positive for Model 3. Energies of the HOMO for the compounds under study have negative values (Table 5). Therefore, higher HOMO values (less negative energies) would positively contribute to the biological activity: this would reflect a more pronounced electronic reactivity. According to our results, powerful photosensitisers are more likely to donate their electrons.

These molecules exert their pharmacological action by intercalating within a specific sequence of DNA bases; after complexation with the nucleic acids and upon UVA irradiation they form a covalent complex with DNA through cycloaddition with adjacent pyrimidine bases. Therefore, the energy level of the HOMO could play an important role in controlling the electronic reactivity of the photosensitiser. This could possibly govern the correct interaction with the lowest unoccupied molecular orbital (LUMO) on the thymidine base, thus leading to the occurrence of the cycloaddition reaction.

Another important descriptor is S-XL, which represents the length of the molecule in the X dimension. The compounds of the data set possess a common coumarin skeleton that is diversely substituted (Table 1). This descriptor measures the maximum extension of the molecule on the plane individuated by the coumarin moiety. The positive sign of the coefficients in our models creates a negative dependence between this variable and the biological activity. This indicates that larger values would negatively contribute to the pharmacological efficacy (Table 4). These molecules exploit the planarity of their coumarin part to intercalate between two stacked base pairs. The resulting stacking creates a suitable orientation for the conden-

Table 6. Average predicted values obtained from the ensemble of 500 cross-validated neural networks for all the compounds in the data set.

Compound	Activity	Predicted	Residual
Angelicine	25.3	27.18	1.88
1	0.06	0.29	0.23
2	0.7	0.46	0.24
3	0.89	0.71	0.18
4	1.34	1.72	0.38
5	1.48	1.8	0.32
6	2.09	3.71	1.62
7	2.17	3.88	1.71
8	2.2	3.62	1.42
9	3.87	5.69	1.82
10	4	6.21	2.21
11	6.85	7.27	0.42
12	7.51	8.85	1.34
13	8.67	9.33	0.66
14	9	8.81	0.19
15	9.1	9.79	0.69
16	9.2	9.88	0.68
17	10.8	11.21	0.41
18	10.8	9.63	1.17
19	10.9	11.75	0.85
20	11.8	12.3	0.5
21	11.86	11.62	0.24
22	13.08	12.06	1.02
23	13.5	12.88	0.62
24	13.8	15.11	1.31
25	16.6	17.45	0.85
26	17.62	17.79	0.17
27	18	17.68	0.32
28	18.06	18.44	0.38
29	20.2	19.82	0.38
30	22.3	23.31	1.01
31	22.7	22.33	0.37
32	23	23.58	0.58
33	24	25.67	1.67
34	26.3	28.23	1.93
35	27.73	28.51	0.78
36	30	29.29	0.71
37	30	29.44	0.56
38	38.1	37.88	0.22
39	45	45.52	0.52
40	49.7	49.32	0.38
41	50.9	48.35	2.55
42	51.9	53.11	1.21
43	57.5	55.52	1.98
44	76.3	73.83	2.47
45	> 250	165.21	–
46	> 250	102.07	–
47	> 250	118.52	–
48	> 250	121.92	–
49	> 250	99.08	–
50	> 250	95.43	–
51	> 250	127.88	–
52	> 250	125.42	–
53	> 250	142.67	–
54	> 250	92.51	–
55	> 250	97.14	–
56	> 250	112.37	–
57	> 250	123.76	–
58	> 250	105.29	–
59	> 250	109.96	–

sation reaction with the pyrimidine base. Therefore, the extension of the core planar scaffold represents a key structural element for the correct interaction with the DNA.

WPSA-2 is the third common descriptor displayed by the best QSAR equations in this study (Table 4). It is a spatial descriptor and represents the surface- and charge-weighted positively charged molecular area. Higher values of this variable reflect a larger number of solvent-accessible positively charged atoms. This descriptor is present in all the equations with a positive coefficient (Table 4). Larger values would thus reduce the biological activity. A possible interpretation for this behaviour can lie on the physico-chemical nature of the biological target. The double stranded DNA displays a large negatively charged backbone. The phosphate groups are located on the surface of the DNA molecule, thus shielding the core of the double helix. A highly positively charged molecule would be more likely to interact with the DNA phosphate groups than to intercalate inside the DNA strands. These different interactions could possibly reduce the rate of the proper cycloaddition reaction, thus diminishing the pharmacological efficiency.

The equations obtained with the predictors selected by means of genetic algorithms have been proved very useful in individuating independent variables that are more likely to be related with the biological activity. This could serve to rationalise the differences in the potency values of these photobiological agents and could also form the basis for intelligent drug design. However, the resulting QSAR equations display poor

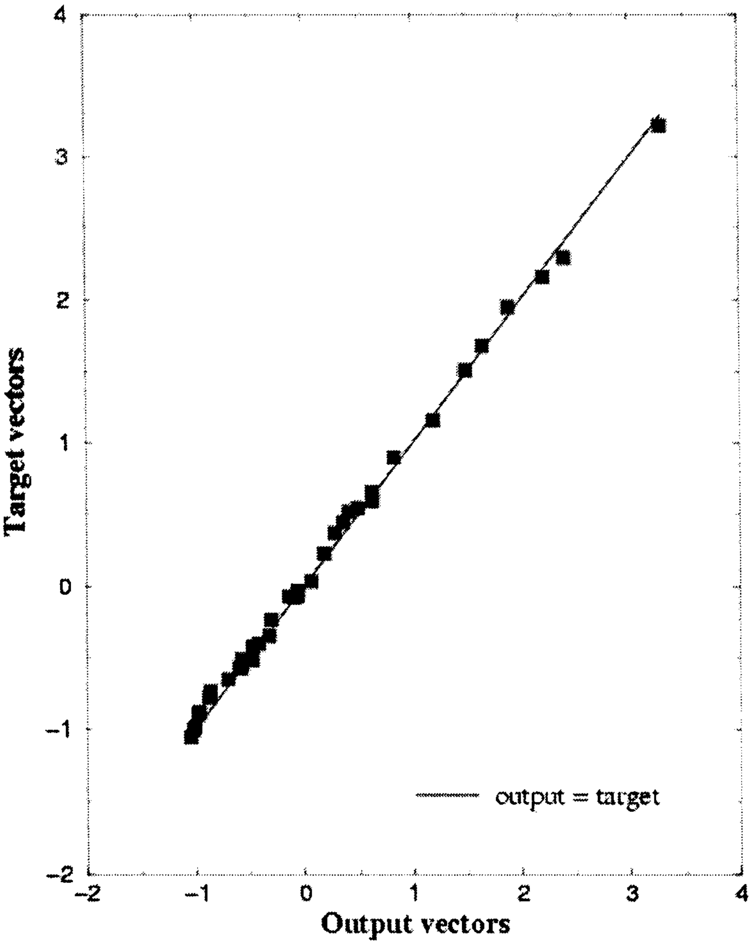


Figure 1. Training results obtained from the ensemble of 500 back-propagated feed forward neural networks during the cross-validation procedure. Average output vectors are plotted against their corresponding target vectors.

predictive character with large errors in the predictions. Specifically, the models predicted compound **54** to have an ID_{50} of 29.45 kJ/m^2 (model 1), 22.21 kJ/m^2 (model 2) and 36.02 kJ/m^2 (model 3), compared to an experimental value that is larger than 250 kJ/m^2. We believe this is due to the high molecular similarity that compound **54** shares with **12** (ID_{50} 7.51 kJ/m^2), **22** (ID_{50} 18.06 kJ/m^2) and **28** (ID_{50} 18.06 kJ/m^2), thus avoiding a correct discrimination. Although all the other compounds displaying ID_{50} greater than 250 kJ/m^2 have been predicted to lie on a higher ID_{50} range, 68.71–121.45 kJ/m^2 (model 1), 82.36–100.22 kJ/m^2 (model 2), 61.19–153.74 kJ/m^2 (model 3), this is clearly unacceptable for new compound selection or for screening purposes since one faces the risk of including uninteresting molecules.

We applied machine-learning techniques to possibly obtain better models. The average predictions obtained from the ensemble of 500 back-propagated feed forward neural networks during the cross-validation procedure are presented in Table 6. During the training process the networks were able to learn the hidden relationships between the molecular input values and the biological activities as shown in Figure 1. The results were also satisfactory for the randomly picked test sets with an average root mean square error of 1.22 kJ/m^2. As a comparison, the experimental error in ID_{50} determinations has been estimated to be 0.85 kJ/m^2. The relatively high number of neurons in the hidden layer (16) resulted in a number of adjustable parameters in the network that is almost twice greater than the molecules in the data set. In standard data modelling, this could have easily produced

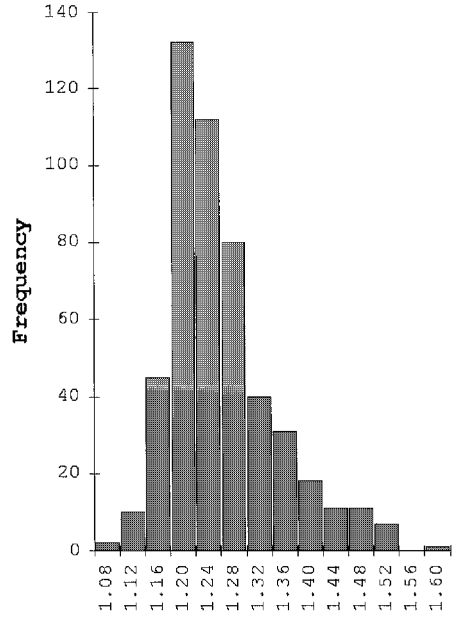


Figure 2. Frequency distribution of the root mean square error (RMSE) for the predictions of the different test sets during the 500-fold leave-one-out cross-validation process.

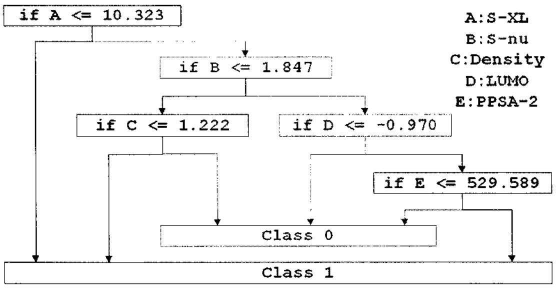


Figure 3. Decision tree obtained using recursive partitioning.

on overfitting problem. However, it has already been shown [24] that the use of an ensemble of neural networks combined with early stopping training enables the production of highly predictive models even if the number of adjustable parameters exceeds the number of observations in the data set.

Figure 2 displays a frequency distribution for the RMSE of the predictions for the different test sets during the 500-fold leave-five-out cross-validation procedure. The histogram shows a unimodal distribution with a positive skewness. The highest populated bin displays RMSE values in the 1.16–1.20 kJ/m^2 range. The absence of additional peaks in the distribution indicates that the average RMSE of 1.22 kJ/m^2 for the cross-validated networks is not biased and could support the robustness of the ensemble of networks obtained. This is supported by the predictions for compounds **45–59**, not included in the model building process, whose calculated ID$_{50}$ values were larger than 92.51 kJ/m^2, as shown in Table 6. Here, compound **54** (ID$_{50}$ larger than 250 kJ/m^2) has been predicted to have an average ID$_{50}$ of 92.51 kJ/m^2. This represents an improvement compared to the results obtained with the genetic models (36.02 kJ/m^2 in the best case). The more sophisticated nature of the mathematical functions used in the neural network compared to the ones employed by genetic algorithms could be held responsible for a better modeling of the hidden relationships between biological activity and molecular structure and therefore for the higher accuracy in the predictions. Moreover, it is likely that the principal components used as inputs for the neural networks contained less noise and redundant information than the simple molecular descriptors used with genetic algorithms and this could have helped the discrimination between active and inactive compounds.

Although compared to the previous QSAR models we lost the interpretability of the results, a relevant improvement has been obtained in the prediction ability. This enhanced predictivity will prove very useful when the network will be employed as a filtering tool for new photosensitisers.

Molecules **45–59** lack precise photobiological values and their activities (ID$_{50}$) have been measured to be greater than 250 kJ/m^2. We applied recursive partitioning techniques in order to include these compounds in our QSAR studies. The whole data set has been divided into different classes, according to the observed biological activity. Different class combinations have been tried to specifically cluster the compounds. However, the best results have been obtained for a simple two-class separation: the first class holds the best photosensitisers (ID$_{50}$ $<$ 10 kJ/m^2) whereas the second contains all the remaining molecules.

The best decision-tree obtained with recursive partitioning is displayed in Figure 3 while Table 7 shows the results gained for the data set. The compounds are firstly discriminated on the basis of the S-XL descriptor. Interestingly, this variable has already been highlighted by genetic algorithms as important for the biological activity. The first node accounts for half of the most active compounds. All these molecules have S-XL values smaller than 10.32 Å. The distance between two DNA strands in a canonical B form, computed using the C1′ atoms of the ribose rings ranges from 10 to 11 Å. Therefore, the biological target im-

Table 7. Classification results using the recursive partitioning (RP) decision tree of Figure 3. Class 1 holds the compounds with ID_{50} lower than 10 kJ/m^2, whilst class 0 contains the compounds whose ID_{50} is greater than 10 kJ/m^2.

Compound	Activity	Class	RP Class	Class 0 Probability Assignment	Class 1 Probability Assignment
Angelicine	25.3	0	1	0.11	0.89
1	0.06	1	1	0.2	0.8
2	0.7	1	1	0.11	0.89
3	0.89	1	1	0.2	0.8
4	1.34	1	1	0	1
5	1.48	1	1	0.11	0.89
6	2.09	1	1	0.11	0.89
7	2.17	1	1	0.11	0.89
8	2.2	1	1	0.11	0.89
9	3.87	1	1	0.2	0.8
10	4	1	1	0.11	0.89
11	6.85	1	1	0	1
12	7.51	1	1	0	1
13	8.67	1	1	0.2	0.8
14	9	1	1	0	1
15	9.1	1	1	0.11	0.89
16	9.2	1	1	0.2	0.8
17	10.8	0	0	1	0
18	10.8	0	0	1	0
19	10.9	0	0	1	0
20	11.8	0	0	1	0
21	11.86	0	0	1	0
22	13.08	0	0	1	0
23	13.5	0	0	1	0
24	13.8	0	0	1	0
25	16.6	0	0	1	0
26	17.62	0	0	1	0
27	18	0	0	1	0
28	18.06	0	0	1	0
29	20.2	0	0	1	0
30	22.3	0	0	1	0
31	22.7	0	0	1	0
32	23	0	0	1	0
33	24	0	0	1	0
34	26.3	0	0	1	0
35	27.73	0	0	1	0
36	30	0	1	0.2	0.8
37	30	0	0	1	0
38	38.1	0	0	1	0
39	45	0	0	1	0
40	49.7	0	0	1	0
41	50.9	0	0	1	0
42	51.9	0	0	1	0
43	57.5	0	0	1	0
44	76.3	0	0	1	0
45	250	0	0	1	0
46	250	0	0	1	0
47	250	0	0	1	0

Table 7 continued. Classification results using the recursive partitioning (RP) decision tree of Figure 3. Class 1 holds the compounds with ID_{50} lower than 10 kJ/m^2, whilst class 0 contains the compounds whose ID_{50} is greater than 10 kJ/m^2.

Compound	Activity	Class	RP Class	Class 0 Probability Assignment	Class 1 Probability Assignment
48	250	0	0	1	0
49	250	0	0	1	0
50	250	0	0	1	0
51	250	0	0	1	0
52	250	0	0	1	0
53	250	0	0	1	0
54	250	0	0	1	0
55	250	0	0	1	0
56	250	0	0	1	0
57	250	0	0	1	0
58	250	0	0	1	0
59	250	0	0	1	0

poses a severe constraint on the molecular extension of its ligands.

The second most discriminant decision is made on the basis of a combination of two descriptors (S-nu and LUMO) (Figure 3). The first one is a spatial descriptor that measures the symmetry of a molecule (ratio between the largest and the smallest dimension) whereas the second is the energy of the lowest-unoccupied molecular orbital as computed with MOPAC. 32 out of 44 weak photosensitisers show both high degrees of shape symmetry (S-nu < 1.847) and low LUMO energies (LUMO < -0.970 eV). The first independent variable stresses the importance of the molecular shape for the correct intercalation into DNA. The second descriptor reflects the electrophilic character of weak molecules compared to the nucleophilic nature of the most effective agents. Since electron mobility is a key factor for the establishment of the cycloaddition reaction with the pyrimidine bases on the DNA, molecules that are less likely to donate their electrons (low LUMO energies) will prove weaker in the covalent complexation with nucleic acids.

The terminal nodes of the decision tree divide molecules into weak and potent compounds on the basis of Density and PPSA-2 values (Figure 3). Density could indicate the importance of the molecular volume for the activity whereas the total charge weighted positive area (PPSA-2) highlights again the negative contribution of the positively charged ligand surface to the molecular recognition event.

The built decision tree is able to correctly classify 58 out of 60 compounds and it provides interesting guidelines for the interpretation of structure-activity relationships.

Conclusions

The present study investigated the complex relationships between the ID_{50} and the molecular structure in a set of 60 photochemotherapeutic agents. Different data modelling techniques have been exploited in order to derive essentially two types of results. Firstly, it was important to identify the molecular descriptors that play an important role in the structure-activity relationships. This will provide the ‘rationale’ for the design of new compounds. Secondly, the creation of predictive models based on machine learning applications will support the subsequent virtual evaluation of the ligands. This combined approach will be integrated in the quest for new photosensitisers.

References

1. Parrish, J.A., Stern, R.S., Pathak, M.A., Fitzpatrick, T.B., In Regan, J.D. and Parrish, J.A. (Eds.), The Science of Photomedicine, Plenum Press: New York, NY, 1982, pp. 595-624.
2. Gasparro, F.P., Extracorporeal Photochemotherapy Clinical Aspects and the Molecular Basis for Efficacy, Landes Press, Georgetown, TX, 1994.
3. Kirkland, D.J., Creed, K.L., Mannisto, P., Mutat. Res., 116 (1983) 73.
4. Abel, G., Mutat. Res., 190 (1987) 63.
5. Stivala, L.A., Pizzala, R., Rossi, R., Melli, R., Verri, M.G., Bianchi, L., Mutat. Res., 327 (1995) 227.

6. Stern, R.S., Lange, R., J. Invest. Dermatol., 91 (1988) 120.
7. Ben-Hur, E., Song, P.S., Adv. Radiat. Biol., 11 (1984) 131.
8. Bordin, F., Carlassare, F., Busulini, L., Baccichetti, F., Photochem. Photobiol., 58 (1993) 133.
9. Guiotto, A., Rodighiero, P., Manzini, P., Pastorini, G., Bordin, F., Baccichetti, F., Carlassare, F., Vedaldi, D., Dall'Acqua, F., Tamaro, M., Recchia, G., Cristofolini, M., J. Med. Chem., 27 (1984) 959.
10. Bordin, F., Dall'Acqua, F., Guiotto, A., Pharmac. Ther., 52 (1991) 331.
11. Mosti L., Schenone, P., Menozzi G., Romussi, G., Baccichetti, F., Carlassare F., Vedaldi D., Bordin, F., Eur. J. Med. Chem., 18 (1983) 113.
12. Mosti L., Schenone, P., Menozzi G., Romussi, G., Baccichetti, F., Carlassare F., Bordin, F., Farmaco, 39 (1984) 81.
13. Iester, M., Fossa, P., Menozzi, G., Mosti, L., Baccichetti F., Marzano, C., Simonato, M., Farmaco, 50 (1995) 669.
14. Mosti, L., Lo Presti, E., Menozzi, G., Marzano, C., Baccichetti, F., Falcone, G., Filippelli, W. and Piucci, B., Farmaco, 53 (1998) 602.
15. Fossa, P., Mosti, L., Menozzi, G., Marzano, C., Baccichetti, F., Bordin, F., Bioorg. Med. Chem. Lett., 10 (2002) 743.
16. Dall'Acqua, F., Vedaldi, D., Caffieri, S., Guiotto, A., Rodighiero, P., Baccichetti, F., Carlassare, F., Bordin, F., J. Med. Chem., 24 (1981) 178.
17. Dall'Acqua, F., Vedaldi, D., Guiotto, A., Rodighiero, P., Carlassare, F., Baccichetti, F., Bordin, F., J. Med. Chem., 24 (1981) 806.
18. Guiotto, A., Rodighiero, P., Pastorini, G., Bordin, F., Baccichetti, F., Carlassare, F., Vedaldi, D., Dall'Acqua, F., Farmaco, 36 (1981) 537.
19. Kolomogorov, A.N., Doklady Akademiia Nauk SSSR, 114 (1957) 953.
20. Duprat, A.F., Huynh, T., Dreyfus, G., J. Chem. Inf. Comput. Sci., 38 (1998) 586.
21. Cerius2™ http://www.accelrys.com
22. MATLAB™ http://www.mathworks.com
23. Hagan, M.T., Menhaj, M., IEEE Transactions on Neural Networks, 5 (1994) 989.
24. Tetko, I.V., Livingstone, D.J., Luik, A.I., J. Chem. Inf. Comput. Sci., 35 (1995) 826.