Journal of Mathematical Chemistry 7(1991)273–285

THE HYPOTHETICAL ACTIVE SITE LATTICE – IN VITRO AND IN VIVO EXPLORATIONS USING A THREE-DIMENSIONAL QSAR TECHNIQUE

A.M. DOWEYKO
CIBA–GEIGY Corporation, Environmental Health Center, Farmington, CT 06032, USA

Abstract

A novel three-dimensional QSAR technique (HASL) is described. HASL (Hypothetical Active Site Lattice) represents a composite description of a macromolecular binding site obtained from ligand molecule structures and activities. The scope and limitations of the method are discussed by means of its application to in vitro and in vivo data.

1. Introduction

The creative use of computers has led to a variety of ways to generate quantitative structure–activity relationships (QSAR) [1–16]. Most techniques provide a retrospective assessment of QSAR and are limited in scope to a relatively narrow spectrum of closely related structures.

The real merit of QSAR should be to provide a prospective assessment of structure–activity relationships which would provide for some degree of prediction and creation of promising structural types. This feature is not normally found in the classical "phenyl-substituted series". What is needed is a three-dimensional QSAR technique which takes into account the space a molecule occupies, the fragments/atoms present in each molecule, and the activity of the molecule. In principle, this type of information can be "merged" to produce a model of a binding site capable of predicting the binding affinity of new molecules.

The Hypothetical Active Site Lattice (HASL) represents an effective approach to the complex problem of relating activity to structure in three dimensions [17,18]. The method has been successfully demonstrated by the in vitro analysis of *E. coli* dihydrofolate reductase (DHFR) inhibitors [17] and *E. coli* glutamine synthetase inhibitors [19]. In order to gain further insight into the scope and limitations of the method, it was applied on three data sets. In vitro data were obtained from Wiese, and consisted of 20 *M. lufu* DHFR inhibitors [20]. Two sets of in vivo data were utilized: parainfluenza antiviral activities for a set of 28 nucleosides [21] and *Botrytis cinerea* fungicide activities for 16 cyclic alkanoxy structures [22]. Comparing HASL analyses of in vitro and in vivo data would be expected to serve as a means to evaluate model predictivity using data from isolated enzyme systems as well as systems having some degree of transport involved (antiviral data and fungicide data).

2. Theory

Although detailed descriptions of HASL methodology have been published elsewhere [17,18], a brief discussion of the procedures utilized herein is warranted. The creation of a site model is based on spatial molecular comparisons which embody atom types. Each molecule used to create the model is first converted to a lattice of points separated by an arbitrary resolution (typically 2–3 Å). Each lattice point (x, y, z, H) also contains an atomic character descriptor (H) which can be arbitrarily assigned (e.g. $H = 1$ for electron-rich atoms, $H = -1$ for electron-poor atoms). The atomic character descriptor (H) has two roles: (1) it provides the means to systematically compare and superpose molecules, and (2) it acts to mathematically anchor partial binding values at specific lattice points (discussed separately). When a second molecule is converted to a lattice, the two molecules can be compared by means of their respective lattices. The degree of correspondence between the two is quantifiable by means of the number of common lattice points (overlap). The second molecule is optimally aligned on the first by translational and rotational movements, with the degree of overlap calculated for each movement through the intermediacy of lattices. Conformationally flexible molecules are treated as rigid structures, which may have been obtained from X-ray crystallographic data, energy-minimization algorithms, or molecular modelling.

A composite model is created when two or more molecular lattices are merged. This procedure simply involves adding any new lattice points obtained from the incoming molecule to the target set of points. After each molecule is fitted to the current model, new lattice points are added. This is repeated until all the molecular data has been merged.

In order to be a predictive tool, the HASL needs to have binding data incorporated. This is most appropriate when the activity data are available in a form that is proportional to the linear free energy of binding. For example, K_i, the enzyme-inhibitor dissociation constant, can be expressed as pK_i. In this form, a molecular binding constant can be spread out as partial pK_i values (ppK_i) in the molecular lattice (x, y, z, H, ppK_i). As the model is created by the addition of new molecular lattices, each common lattice point carries with it an averaged ppK_i. Prediction of activity is then obtained by fitting a molecule to the model lattice and adding up the ppK_i at each common point. In order to obtain accurate predictions, the ppK_i values at each point in the model are adjusted so that optimal predictions are obtained for each molecule in the learning set. The iterative method by which this is done is discussed in detail in the cited references.

The prospective power of HASL provides the investigator with a means to assess the current model by predicting activities for structures outside the learning set. These new structure–activity data can then be added to the model to increase its predictivity. Thus, HASL represents a continually evolving model of ligand–molecule interaction. In addition, the model permits estimation of relative molecular binding as kcal/atom, so that portions of the molecule with significant contributions to the

overall binding are identified. Binding prediction and molecular fragment binding estimates provide the investigator with a valuable tool in future synthetic strategy.

3. Experimental

Assessment of HASL predictivity was carried out using an approach sometimes described as cross-validation. Thus, for a limited series made up of N compounds, it was possible to construct a HASL from $N-1$ compounds and test the Nth compound. This was done N times. The results of such an analysis were visually assessed by plotting N (predicted versus actual activity) point values and calculating the linear correlation coefficient (r) between them. For each of the three sets of compounds in this study, correlations are expressed in terms of r^2, forcing the regression line through the origin.

Further assessment of predictivity was made for each set of compounds by examining r^2 dependence on N. Typically, for a set of N compounds for which a cross-validation was conducted, r^2 was calculated for $N-1$ points, excluding the compound predicted most poorly. This procedure was repeated (r^2 at $N-2$, r^2 at $N-3, \ldots$) until two compounds remained. Examining plots of r^2 versus number of compounds compared provides a visual estimate of HASL predictivity. In order to ascertain whether the HASL is significantly predictive, an additional analysis was conducted. A set of random numbers spanning the range of compound activities was used in place of the predicted set. The same $r^2/N-1, N-2, \ldots/$ analysis yielded the dependence of r^2 on random data. Comparing the two analyses on the same graph provides a visualization of HASL predictivity versus random noise. Five random number sets were used for each set of compounds in this study.

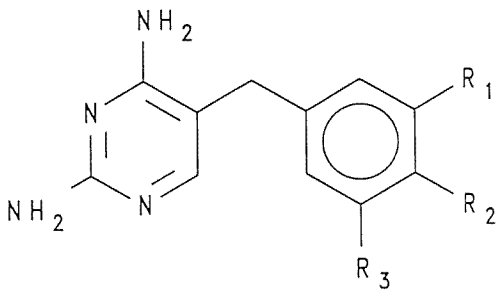
HASL modelling was conducted using HASL v 3.16, a program written by the author for the IBM personal computer in FORTRAN and BASIC. H values were assigned as $+1$, 0, and -1 according to MM2 atom type [17]. Statistical analyses and plotting were conducted using Quattro Pro (Borland International). Negative r^2 values were obtained when linear regression was forced through the 0, 0 origin of predicted versus actual binding data. Structure building was conducted either on an AMIGA 1000 using an author-written modelling program or on an IBM personal computer using Chemcad (C_graph). The Cartesian coordinate data used in these analyses can be obtained from the author.

3.1. *M. LUFU* DIHYDROFOLATE REDUCTASE INHIBITION

The twenty-compound data set was obtained from data provided by Wiese (Forschungsinstitut Borstel, Borstel, Germany) [20]. The compound structures are listed in table 1 as **W1–W20**. Their syntheses have been published previously [23,24]. Dihydrofolate reductase (DHFR) was isolated as a partially purified, cell-free extract from *M. lufu*. DHFR activity was determined spectrophotometrically at 340 nm (NADPH E_{340} 12300 M^{-1} cm^{-1}). The reaction mixture (1.0 ml) contained

Table 1

M. lufu dihydrofolate reductase inhibition data[a]. Cross-validation results: actual pI_{50} compared to predicted pI_{50} using 19-compound HASL at 3.0 Å.



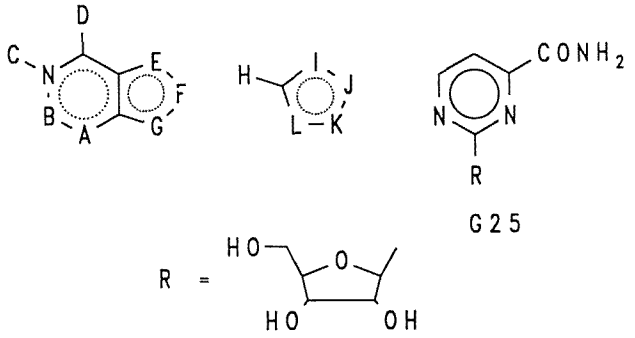
W1–W20

	R_1	R_2	R_3	Actual pI_{50}	Predicted pI_{50}
W1	OCH_3	OCH_3	OCH_3	0.49	0.77
W2	OCH_3	Br	OCH_3	0.82	0.71
W3	OCH_3	SCH_3	OCH_3	0.86	0.46
W4	OCH_3	$OCH_2CH_2OCH_3$	OCH_3	0.02	−0.01
W5	OCH_3	OCH_3	H	0.37	0.28
W6	Cl	NHC_2H_5	Cl	−0.15	0.33
W7	Br	NH_2	Br	0.61	−0.17
W8	H	OCH_3	H	−0.48	−0.46
W9	OCH_3	H	H	−0.35	−0.03
W10	OCH_3	$NHCONHC_3H_7$	OCH_3	−0.91	0.01
W11	$O(CH_2)_7CH_3$	H	H	0.01	0.04
W12	H	OC_7H_{15}	H	−0.60	−0.40
W13	H	OCyclohexane	H	−0.33	−0.24
W14	OCH_3	$O(CH_2)_5Br$	OCH_3	0.71	0.97
W15	OCH_3	$O(CH_2)_4COOC_2H_5$	OCH_3	0.56	0.44
W16	OCH_3	$OCH_2PhSO_2PhNH_2$	OCH_3	0.58	0.73
W17	OCH_3	$OCH_2CH_2NHPhSO_2PhNH_2$	OCH_3	0.80	0.72
W18	OCH_3	$OCH_2PhSO_2PhNH_2$	H	0.65	0.42
W19	$OCH_2PhSO_2PhNH_2$	OCH_3	H	1.04	0.37
W20	OCH_3	$O(CH_2)_3NHPhSO_2PhNH_2$	OCH_3	1.47	0.45

[a]Data obtained from Wiese [20] (I_{50} expressed as μM).

Table 2

Parainfluenza inhibition data[a]. Cross-validation results: actual virus rating (Actual*VR*) compared to predicted virus rating (Pred*VR*) using 27-compound HASL at 2.5 Å



	A	B	C	D	E	F	G	Actual*VR*	Pred*VR*
G1	N	CH	–	SCH_3	N	CH	NR	2.20	1.75
G4	N	CH	–	NH_2	NH	N	CR	1.40	1.22
G5	N	CH	–	SCH_3	NH	N	CR	1.35	1.44
G8	NCH_3	CH	–	NH_2	N	N	CR	1.28	0.55
G9	NH	CO	–	NH_2	NH	N	CR	1.22	1.04
G11	N	CH	CH_3	NH_2	N	N	CR	1.16	0.76
G12	N	CH	H	=S	CH	CH	NR	1.15	0.33
G13	N	CH	H	=S	NH	H	CR	1.10	0.79
G14	N	CH	–	SCH_3	CH	N	NR	1.05	1.06
G15	N	CH	H	=O	NH	H	CR	0.94	0.98
G16	N	CH	–	SCH_3	CH	CH	NR	0.90	1.33
G17	NH	=O	–	NH_2	NCH_3	N	CR	0.88	0.42
G18	N	CH	–	NH_2	NCH_3	N	CR	0.86	1.01
G19	CH	CNH_2	H	=O	N	CH	NR	0.80	0.25
G20	N	CH	H	=S	CH	N	NR	0.70	0.42
G21	NH	=O	H	=O	NH	N	CR	0.68	0.82
G22	CH	CCL	–	NH_2	N	CH	NR	0.35	0.97
G23	N	CH	H	=O	CNH_2	N	NR	0.28	0.14
G24	N	CCL	H	=O	CH	NR	N	0.25	0.65
G26	N	CCL	–	NH_2	CH	N	NR	0.15	0.08
G28	N	CH	H	=O	CH	CH	NR	0.08	0.43
	H	I	J	K	L				
G2	$CONH_2$	NH	N	CR	COH			1.65	0.68
G3	$CONH_2$	NCH_3	N	CR	$COCH_3$			1.65	1.45
G6	$CONH_2$	N	C	NR	N			1.30	0.87
G7	$CONH_2$	CH	S	CR	N			1.30	0.78
G10	$C(NH)NH_2$	N	CH	NR	N			1.20	1.45
G27	$CONH_2$	COH	N	NR	CH			0.15	0.15

[a]Data obtained from Ghose [21].

100 mM Tris at pH 7.2, 0.1 mM NADPH, 0.02 mM dihydrofolic acid (H_2folate), DHFR (0.1 units) and varying amounts of inhibitor. All components except H_2folate were incubated at 25 °C for 5 min prior to initiation of the reaction with H_2folate. Inhibition for each compound was determined as I_{50} (μM, 50% inhibition) using a nonlinear least-squares program and is reported in table 1 as pI_{50}.

Molecular conformations of each of the 20-compound set were kept consistent in that the *R* groups attached to the phenyl group were modelled linearly away from the 2, 4-diamino moiety. Thus, the highly variable long chains were constructed in a systematic manner to minimize arbitrary spatial variation. The HASL was developed using a lattice spacing of 3.0 Å and atom types as defined by the standard MM2 (1977) program and described previously [17].

3.2. PARAINFLUENZA ANTIVIRAL SCREEN

The 28-compound data set (**G1–G28**) was obtained from Ghose [21]. The structures of these nucleoside analogs and their antiviral activities are listed in table 2. The biological activities were determined using parainfluenza virus type 3 in Vero cells and are expressed as virus ratings. The details of the viral assay system have been published [25]. Virus ratings are numbers that are based on a cell culture microplate in which a log scale concentration gradient of compound is used.

The structures of this set of nucleoside analogs were provided by Ghose as Biograf Cartesian coordinate files, wherein each structure represented an optimized superpositioning. The energy-minimized conformation of **G1** was used by Ghose et al. as a reference structure in the model-building process, which ultimately led to energy-minimized structures for all 28 compounds. The HASL was developed using a lattice spacing of 2.5 Å.

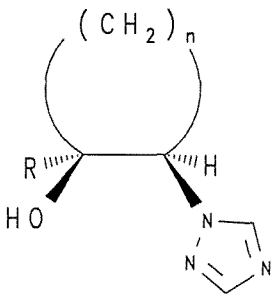
3.3. *BOTRYTIS CINEREA* INHIBITION

The 16-compound antifungal data set (**F1–F16**) was obtained from the recent work of Fujita et al. [22] and is listed in table 3. The original set of cyclic alkanoxy structures contained four benzylic analogs which were exluded from this analysis in order to limit the HASL investigation to one conformational type. The excluded benzylic compounds (where *R* = subst. $PhCH_2$) represented a limited series with an extra degree of conformational freedom. The biological data were obtained using a cell culture system wherein I_{50} (M, 50% inhibition of mycelial growth of *Botrytis cinerea*) was measured using an agar medium dilution method [26]. Inhibition data are reported in table 3 as pI_{50}.

The structures and conformations of this series of compounds were modelled on standard cycloalkane geometries. Both bulky substituents, i.e. the substituted phenyl and heterocyclic moieties, were arranged equatorially with the planes of both ring systems roughly parallel and at a 90° torsional angle to the plane of the cycloalkanol ring. These conformations were used in the HASL modelling without energy minimization, with lattice point spacing set at 2.5 Å.

Table 3

B. cinerea fungal inhibition data[a]. Cross-validation results: actual pI_{50} compared to predicted pI_{50} using 15-compound HASL at 2.5 Å



F1–F16

	R	n	Actual pI_{50}	Predicted pI_{50}
F1	4-Cl-Ph	3	3.66	3.48
F2	Ph	4	4.28	3.95
F3	2-Cl-Ph	4	5.28	3.30
F4	3-Cl-Ph	4	4.01	3.78
F5	4-F-Ph	4	4.93	5.06
F6	4-Cl-Ph	4	5.23	4.77
F7	2, 5-Cl_2-Ph	4	3.45	4.59
F8	3, 4-Cl_2-Ph	4	3.91	4.74
F9	1-Naphthyl	4	3.78	4.61
F10	2-Cl-Thiophen-2-yl	4	5.04	5.20
F11	3, 4, 5-Cl_3-Thiophen-2-yl	4	4.43	4.62
F12	4-F-Ph	5	5.70	5.80
F13	4-Cl-Ph	5	5.97	5.63
F14	3, 4-Cl_2-Ph	5	4.57	5.59
F15	Thiophen-2-yl	5	4.83	5.98
F16	5-Cl-Thiophen-2-yl	5	6.01	4.38

[a]Data obtained from Fujita et al. [22].

4. Results and conclusions

4.1. IN VITRO *M. LUFU* DHFR INHIBITION

The HASL obtained using DHFR I_{50} data for compounds **W1–W20** was examined by cross-validation and the results plotted in fig. 1. The predictivity of the HASL

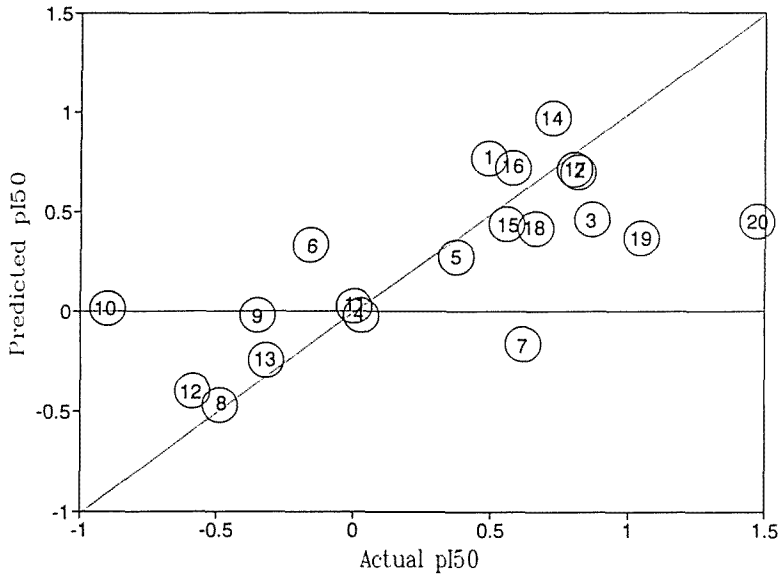


Fig. 1. Plot of actual pI_{50} versus predicted pI_{50} for *M. lufu* DHFR inhibitors (**W1–W20**, data from table 1). Predicted pI_{50} for each compound was obtained using HASL at 3.0 Å constructed from the remaining 19 compounds of the data set.

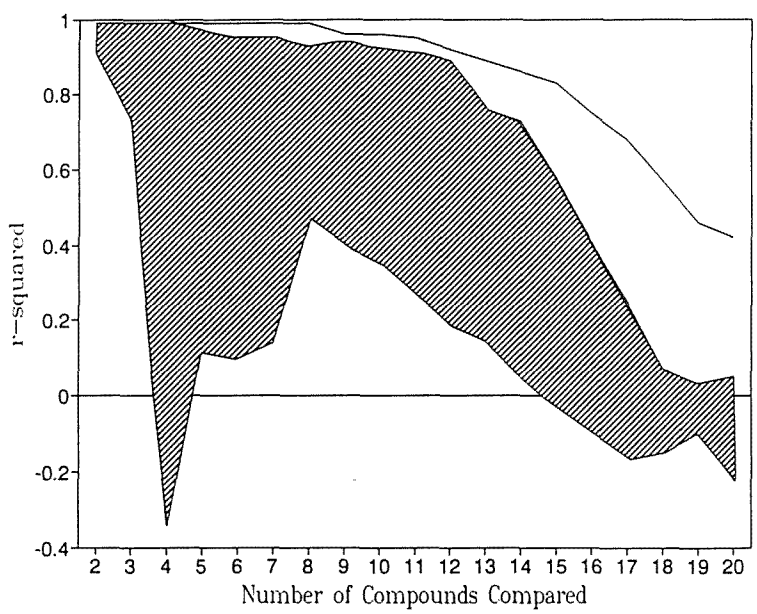


Fig. 2. Predictivity assessment of *M. lufu* DHFR 19-compound HASL: r^2 versus number of compounds compared. Solid line represents the HASL-predicted r^2 set. The striped area represents the r^2 range found for 5 sets of random predictions made within the actual pI_{50} range of −0.911 to 1.47.

model was found to be fair (r^2 = 0.42) using all 20 compounds in the data set, and quite good if considering only the best 16 (r^2 = 0.75). When examined against a background of five sets of random data (I_{50} values of −0.911 to 1.47), the HASL correlation curve appeared well above the random noise level (fig. 2). These results suggest that a HASL can be constructed from I_{50} data that is linearly proportional to the free energy of binding. This would be expected in the case of an in vitro, isolated enzyme system, where I_{50} is theoretically proportional to K_i.

4.2. IN VIVO PARAINFLUENZA INHIBITION

Cross-validation of HASL predictivity applied to compounds **G1–G28** (fig. 3) yielded results statistically similar to those observed in the preceding in vitro DHFR case. The correlation between actual and predicted virus ratings (*VR*) was found to

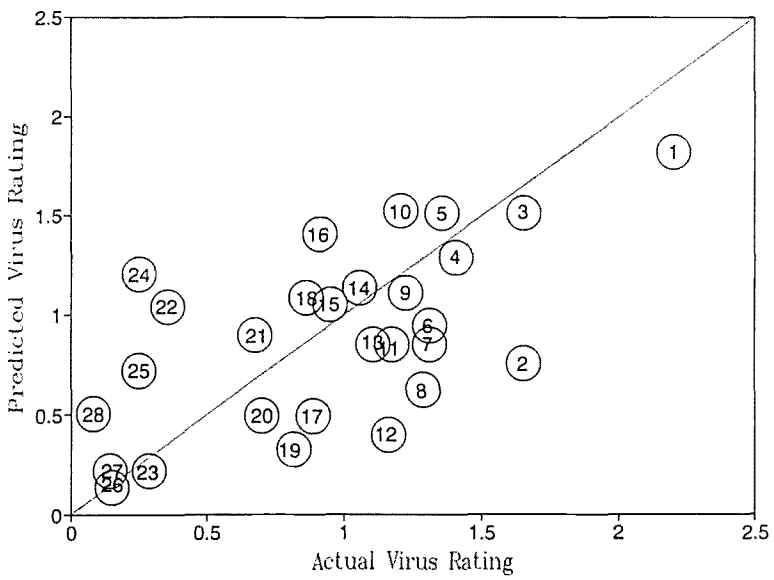


Fig. 3. Plot of actual virus rating versus predicted virus rating for parainfluenza inhibitors (**G1–G28**, data from table 2). Predicted virus rating for each compound was obtained using HASL at 2.5 Å constructed from the remaining 27 compounds of the data set.

be weak (r^2 = 0.24) when all 28 analogs were included, and good (r^2 = 0.66) with the five worst fitting analogs excluded. Again, it was of interest to see how HASL predictivity compared with random data. The data in fig. 4 illustrate r^2 dependence on the number of compounds used as compared with a similar analysis using five sets of random *VR* values (each ranging from 0.8 to 2.20). The HASL-predicted line

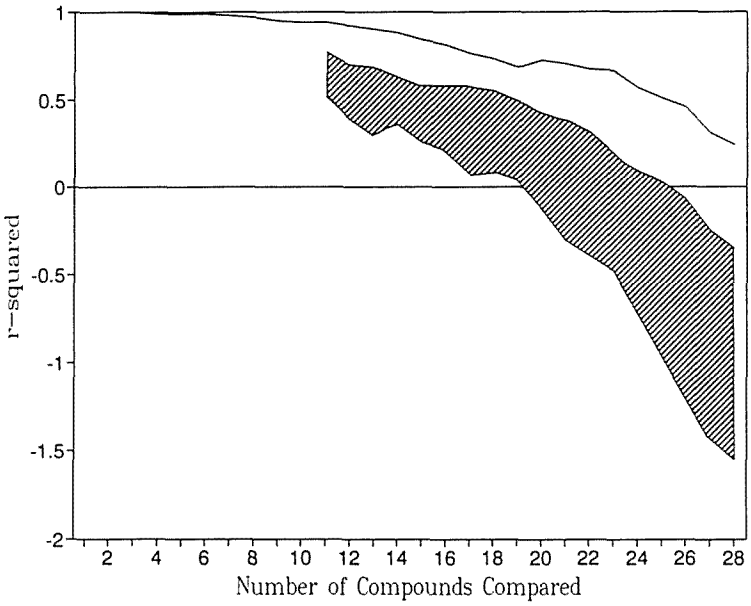


Fig. 4. Predictivity assessment of parainfluenza 27-compound HASL: r^2 versus number of compounds compared. Solid line represents the HASL-predicted r^2 set. The striped area represents the r^2 range found for 5 sets of random predictions made within the actual virus rating range of 0.8 to 2.20.

is observed well above the noise level, and serves to confirm that a potentially useful HASL can be constructed from such an in vivo assay system. Interestingly, these results would suggest that the virus rating numbers are proportional to a binding process typically found in receptor/ligand interactions. Therefore, any transport process taking place for this series of nucleoside analogs in this in vivo (cell culture) assay does not appear to contribute significantly to the virus rating number.

4.3. IN VIVO *B. CINEREA* INHIBITION

A greater degree of scatter is evident in the predictions made by the HASL model in the case of these in vivo data. This is clearly observed in the actual versus predicted pI_{50} values plotted in fig. 5. When all 16 compounds were examined, the correlation was poor ($r^2 = -0.22$). Excluding the five worst fitting compounds yielded a better correlation ($r^2 = 0.63$), but of questionable significance. Comparing HASL predictivity (r^2) to that obtained by five random sets of predictions (random pI_{50} 3.45–6.01), produced the data shown in fig. 6. Using this analysis, the HASL was found to give predictions somewhat above the random noise level. Thus, in this second in vivo test system, using a HASL modelling approach appeared to provide

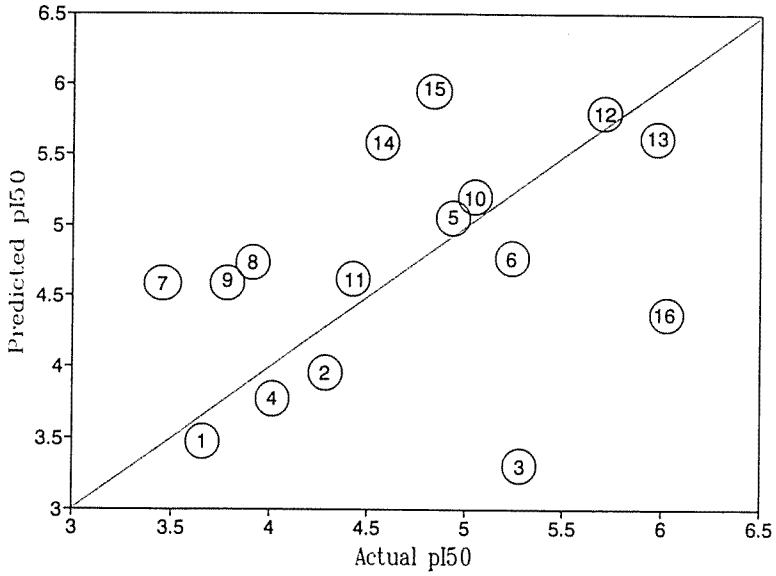


Fig. 5. Plot of actual pI_{50} versus predicted pI_{50} for *B. cinerea* inhibitors (**F1–F16**, data from table 3). Predicted pI_{50} for each compound was obtained using HASL at 2.5 Å constructed from the remaining 15 compounds in the data set.

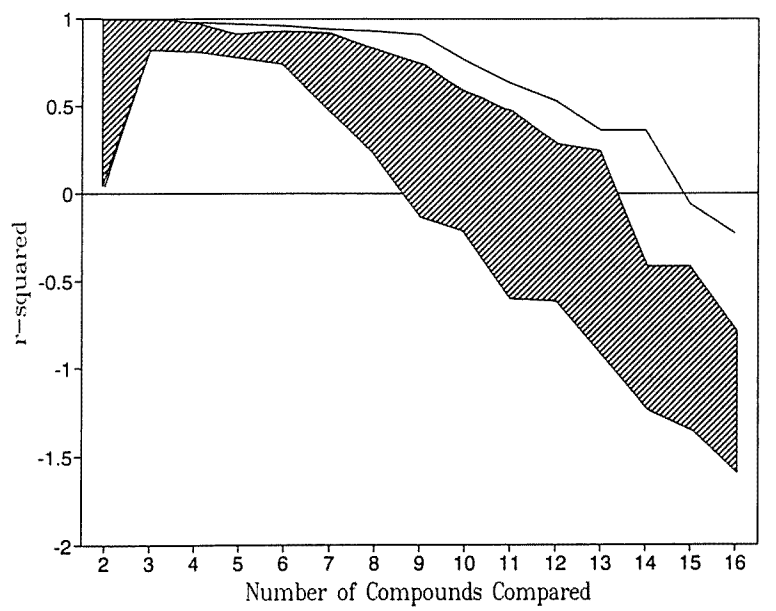


Fig. 6. Predictivity assessment of *B. cinerea* 15-compound HASL: r^2 versus number of compounds compared. Solid line represents the HASL-predicted r^2 set. The striped area represents the r^2 range found for 5 sets of random predictions made within the actual pI_{50} range of 3.45 to 6.01.

meaningful predictions. This is a significant observation, since these data were generated using a fungal cell culture system in which transport phenomena would be expected to contribute to activity. The data in fig. 6 would also suggest that the structural variation in the data set may be rather large, since each compound's activity was found difficult to predict using a model based on the others.

5. Final remarks

A comparison of in vitro and in vivo applications of HASL was achieved. The comparison highlighted differences in each application which included the nature of the activity, and the number and type of structures used. It was possible to gauge the performances of the three HASL models using cross-validation techniques. Furthermore, comparing HASL performance against random data was found to be a way to assess predictivity using a limited series of compounds. By assessing a HASL in this way, it was possible to ascertain whether the activity values used in the assay would allow the ultimate construction of a HASL capable of making meaningful predictions. The analyses suggest that HASL is an effective tool in the prospective three-dimensional QSAR of a variety of structural types using activity data derived from enzyme kinetics and cell culture.

References

[1] C. Silipo and C. Hansch, J. Amer. Chem. Soc. 97(1975)6849.
[2] P.C. Jurs and T. Isenhour, *Chemical Applications of Pattern Recognition* (Wiley–Interscience, New York, 1975).
[3] A. Cammarata and G.K. Menon, J. Med. Chem. 19(1976)739.
[4] D.R. Henry and J.H. Block, J. Med. Chem. 22(1979)465.
[5] G.R. Marshall, *Computer-Aided Molecular Design, Proc. 2-Day Conf.* (1984), p. 1.
[6] P.C. Jurs, T.R. Stouch, M. Czerwinski and J.N. Narvaez, J. Chem. Inf. Comput. Sci. 25(1985)295.
[7] A.J. Hopfinger, J. Med. Chem. 28(1985)1133.
[8] Z. Simon, H. Bojarska-Dahlig and T. Glabski, Pol. J. Pharmacol. Pharm. 33(1981)359; I. Gergen, M. Bohl, H. Simon and Z. Simon, Rev. Roum. Chim. 34(1989)995.
[9] A.T. Brint and P. Willett, J. Chem. Inf. Comput. Sci. 27(1987)152.
[10] S.H. Unger, Drug. Inf. J. 21(1987)267.
[11] T. Gund and P. Gund, in: *Three-Dimensional Molecular Modeling by Computer* (VCH Publ., New York, 1987), p. 319.
[12] J.P. Tollenaere and P.A. Jansen, J. Med. Res. Rev. 8(1988)1.
[13] A.J. Hopfinger, J. Amer. Chem. Soc. 102(1980)7196.
[14] G.M. Crippen, J. Med. Chem. 22(1979)988.
[15] R.D. Cramer, J. Amer. Chem. Soc. 110(1988)5959.
[16] Y. Kato, A. Itai and Y. Iitaka, Tetrahedron 43(1987)5229.
[17] A.M. Doweyko, J. Med. Chem. 31(1988)1396.
[18] A.M. Doweyko, in: *Probing Bioactive Mechanisms*, ed. P.S. Magee, D.R. Henry and J.H. Block (American Chemical Society, Washington, DC, 1989), p. 82.
[19] M. Nowakowski, M. Tishler and A.M. Doweyko, Phosphorus, Sulfur, and Silicon 45(1989)183.

[20] M. Wiese (Forschungsinstitut Borstel, Institut für Experimentelle Biologie und Medizin, Park Allee 1–40, D-2061 Borstel, Germany), private communication.
[21] A.K. Ghose, G.M. Crippen, G.R. Revankar, P.A. McKernan, D.F. Smee and R.K. Robins, J. Med. Chem. 32(1989)746, and private communication.
[22] T. Kataoka, Y. Hayase, T. Hatta, Y Hayashi, A. Murabayashi, Y. Makisumi and T. Fujita, Pest. Biochem. Physiol. 34(1989)228.
[23] K.-H. Czaplinsky, M. Kansy, J.K. Seydel and R. Haller, Quant. Struct.–Act. Relat. 6(1987)70.
[24] G. Hachtel, R. Haller and J.K. Seydel, Arzneimittel-Forschung/Drug Res. 38(1988)1778.
[25] R.W. Sidwell and J.H. Huffman, Appl. Microbiol. 22(1971)797.
[26] T. Kataoka, H. Kai, I. Ishizuka, T. Hatta and M. Ogata, J. Pestic. Sci. 12(1987)445.