ELSEVIER

Simulation of alternative binding modes in a structure-based QSAR study of HIV-1 protease inhibitors

Manuel Pastor, Carlos Pérez, and Federico Gago

Department of Pharmacology, University of Alcalá, Alcalá de Henares, Spain

We have used a published set of inhibitors of HIV-1 protease[1] *to build a COMBINE-type structure-based QSAR model with good predictive ability* ($r^2 = 0.90$, $q^2 = 0.69$).[2] *Since the compounds in the training series exhibit most of their structural variability on one-half of the pseudosymmetrical binding cavity and only one binding orientation was explored for each molecule, the model describes mainly the effect of the structural changes on interactions involving only one-half of the binding cavity (pockets S1′ and S2′). Thus, the model cannot be expected to give accurate predictions for new compounds exhibiting structural variation in both halves. The model does in fact show a tendency to underpredict slightly the biological activity of the molecules in the external test set. In an attempt to improve the quality of the model, both possible orientations of the ligands are now considered so that structural variation takes place in all binding pockets. One possibility would have been to build an additional set of complexes with the inhibitors docked in a reversed orientation. The alternative we have explored, however, consists of manipulating the data matrix describing the interaction energies so that each row is duplicated and the order of the variables in the duplicated rows is swapped between subunits. This simple approach has produced a new model that is similar in quality to the original model* ($r^2 = 0.89$, $q^2 = 0.64$) *but lacks the tendency to underpredict the activity of the compounds in the external set. Moreover, since equivalent residues are assigned equivalent weights, the model is insensitive to ligand orientation and is easier to interpret. © 1998 by Elsevier Science Inc.*

INTRODUCTION

The biological activity of a compound largely depends on its interaction energy with the receptor. Accurate estimations of this interaction energy would have important uses for the prediction of the activity of novel compounds in advance of their synthesis. Unfortunately, even when large amounts of structural information are available, such interaction energies are not easy to calculate accurately.

When dealing with series of closely related compounds, it is possible to build simple models that relate biological activity to energy of interaction computed with molecular mechanics. An

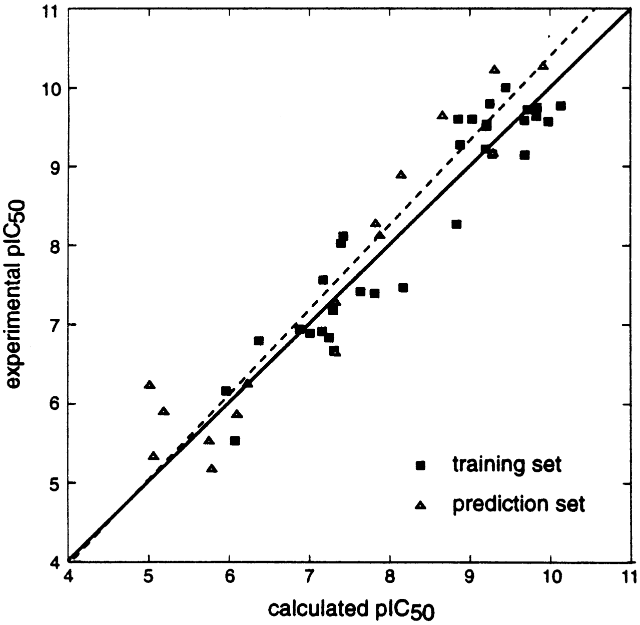


Figure 1. Correlation between experimental and calculated activities (pIC$_{50}$ values) for the training set (filled squares) and the prediction set (open triangles), using model C$_{single}$. The increase in slope of the line representing the least-squares regression fit for the compounds in the prediction set (dashed) with respect to a perfect fit (solid) is indicative of a tendency to underpredict.

Color Plate for this article appears on page 389.

Address reprint requests to: F. Gago, Department of Pharmacology, University of Alcalá, E-28871 Alcalá de Henares, Spain.

Journal of Molecular Graphics and Modelling 15, 364–371, 1997
© 1998 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/97/$17.00
PIIS1093-3263(98)00007-2

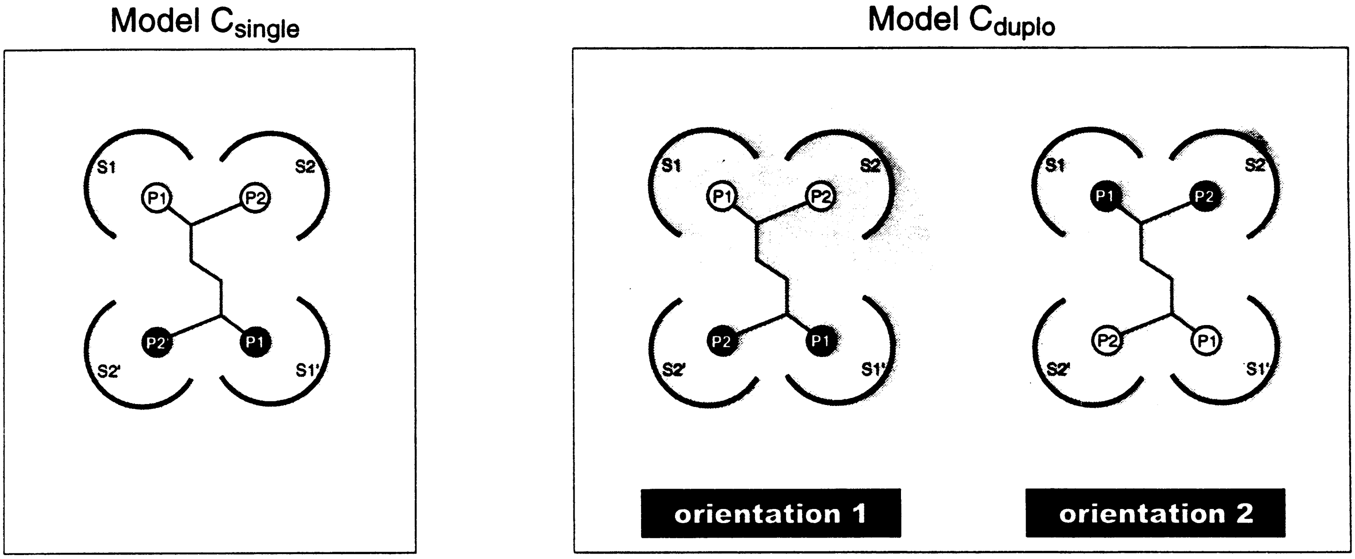


Figure 2. Schematic representation of models C_{single} and C_{duplo}. Binding sites on the enzyme are denoted S1, S2, S1′, and S2′ whereas substituents on the inhibitors are labeled P1, P2, P1′, and P2′. Note that the two alternative orientations of the molecules in C_{duplo} are related by a 180° rotation.

interesting example is the series of HIV-1 protease inhibitors published by Holloway et al.[1] for which a linear regression model was reported between inhibitory potencies (pIC_{50}) and ligand–receptor interaction energies calculated with a simple force field. More recently, we have used the same data set to build a model, also relating biological activities and interaction energies, which partitions these energies into a number of residue-based contributions.[2] This methodology, which we term COMBINE (COMparative BINding Energy)[2–4] analysis, is essentially a 3D QSAR method that uses structures of the ligand–receptor complexes. Ligand–receptor interaction energies, computed within the framework of the AMBER suite of programs[5] and broken down into residue contributions, were supplemented with estimations of the electrostatic contribution to the desolvation of ligands and receptors using a continuum method.[6] All these energy values were then correlated with the biological activity using partial least squares (PLS) regression analysis.

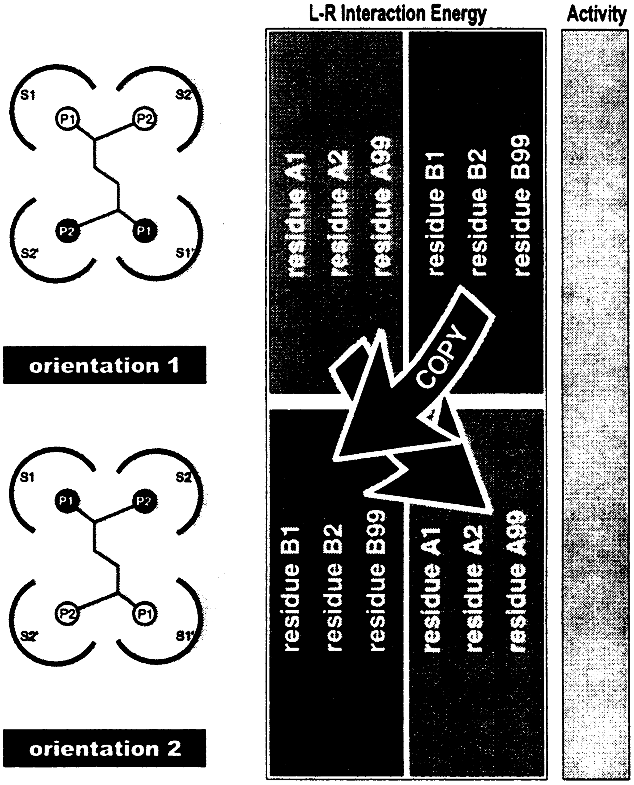
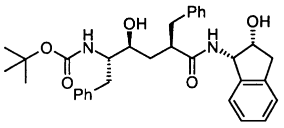
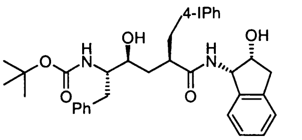
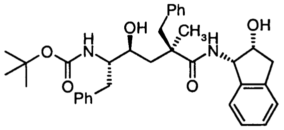
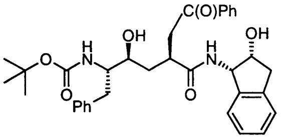
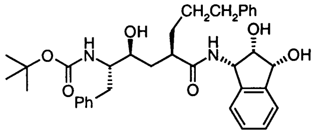
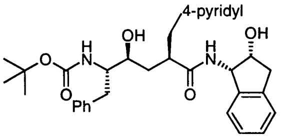
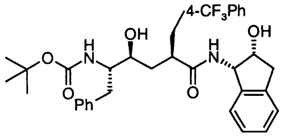
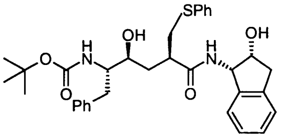
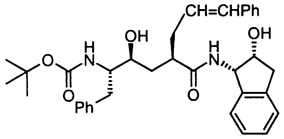
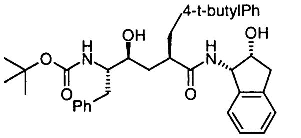
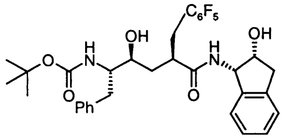
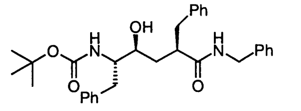
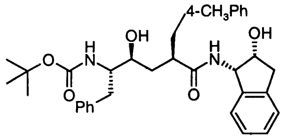
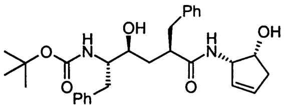
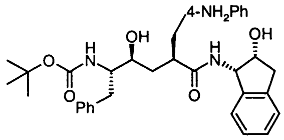
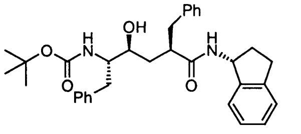
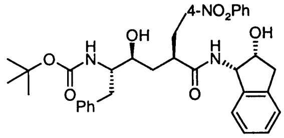
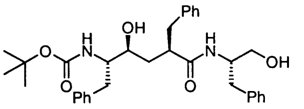
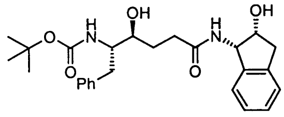
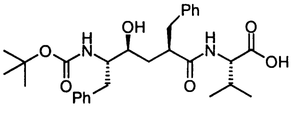
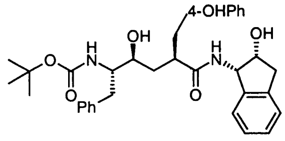
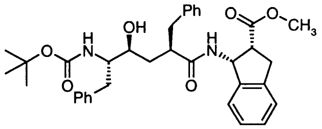
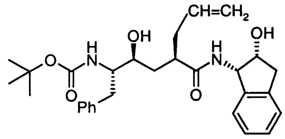
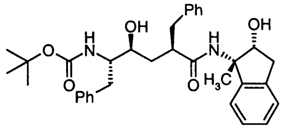


Figure 3. Scheme representing the “copy and paste” procedure used to derive C_{duplo} from the original C_{single} energy decomposition matrix. See Methods for details.

HIV-1 protease is a homodimeric enzyme that, in its complex with many inhibitors, has the interesting peculiarity of being pseudosymmetrical.[7] In such a situation, the ligands can be inserted into the binding site in two alternative and nearly equivalent orientations, but usually just one of the two possibilities is considered. The COMBINE model obtained using only one orientation (C_{single}) is reasonably good but is limited in two respects. First, since the model computes the contribution of each ligand–residue interaction to the activity, unequal importance is assigned to equivalent residues in subunits A and B, which are only arbitrarily distinguishable. Second, when C_{single} is used to predict the activity of novel compounds, their biological activity tends to be slightly underpredicted (Figure 1).

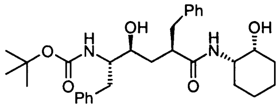
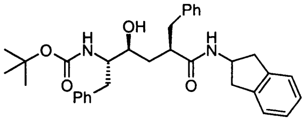
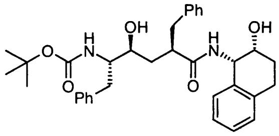
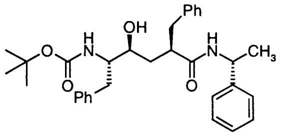
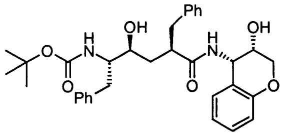
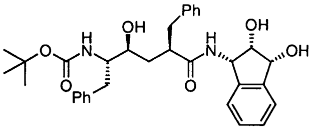
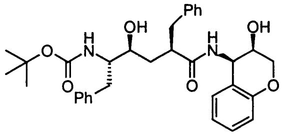
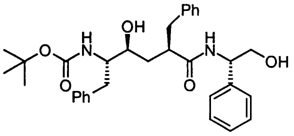
These limitations arise from the fact that the compounds in the training series exhibit structural variation mainly on one side of the structure (P1′ and P2′ substituents) (Color Plate 1a). Consequently, the model is trained to recognize the influence on the activity of structural changes taking place on only one-half of the molecules but receives no information about the other half. However, as mentioned above, this class of ligands can be docked into the binding site in two alternative and equivalent orientations. Thus, if the ligands were considered in both orientations, the model would “learn” the effect of intro-

Table 1. HIV-1 protease inhibitors included in the training set[a]

No.	Chemical structure	Exp. pIC_{50}	No.	Chemical structure	Exp. pIC_{50}
1		9.602	**14**		9.143
3		8.113	**15**		8.266
4		9.721	**16**		9.276
5		9.585	**17**		9.602
6		9.638	**18**		9.770
7		9.222	**19**		6.943
8		9.538	**20**		8.021
9		9.509	**21**		7.465
10		9.569	**22**		6.161
11		5.532	**23**		6.793
12		9.796	**24**		7.179
13		7.561	**25**		6.673

(*continued*)

Table 1. ***(Continued)***

No.	Chemical structure	Exp. pIC_{50}	No.	Chemical structure	Exp. pIC_{50}
26		6.914	**31**		6.886
27		9.155	**32**		6.836
28		9.745	**33**		10.000
29		7.392	**34**		7.413

[a] See Refs. 1 and 2.

ducing structural variability on both sides of the binding cleft. In our example, this means that the model would adequately represent the interactions of the ligand substituents with equivalent residues from each of the two HIV-1 protease subunits (Color Plate 1 and Fig. 2).

The aim of this work is to obtain a model that adequately considers the duality of the potential interaction in order to check its sensitivity to ligand orientation within the receptor-binding site and its predictive ability.

METHODS

To consider both alternative orientations of n ligands within the binding site, the most straightforward approach would be to actually build $2n$ complexes. Every compound would then be included in the analysis twice. However, since the target is symmetrical, it can be argued that the interactions of the variable part of the ligands with subunit A in one orientation can be considered equivalent to the interactions of the same

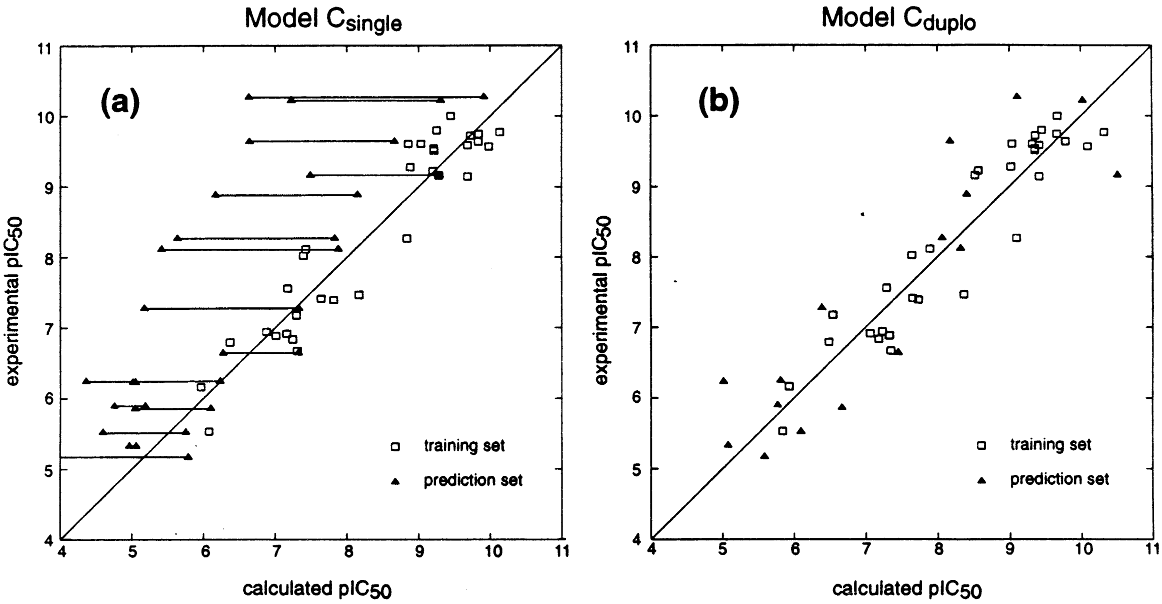
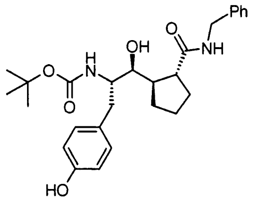
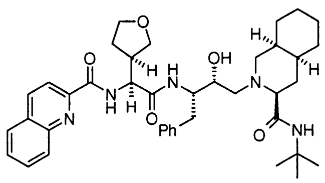
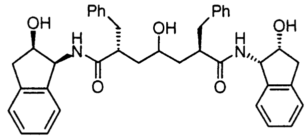
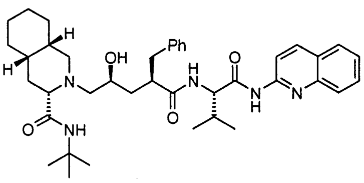
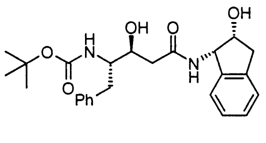
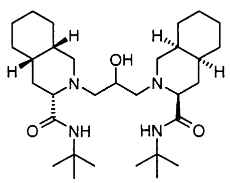
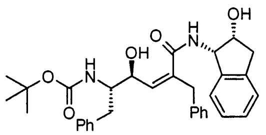
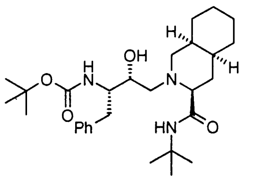
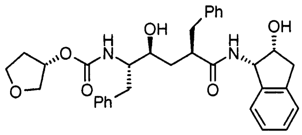
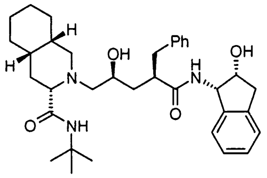
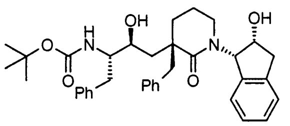
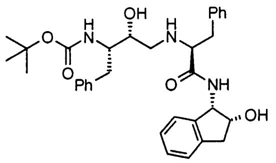
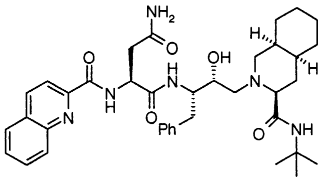
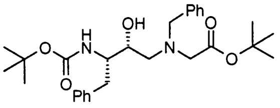
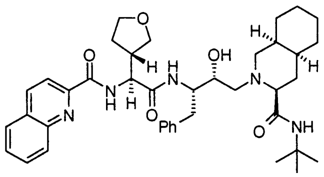
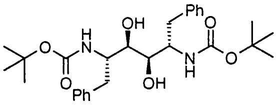


Figure 4. Plots of experimental versus calculated inhibitory activities (pIC_{50} values) for the compounds belonging to the training set (open squares) and those in the prediction set (filled triangles). Two possible orientations are considered for the compounds in the prediction set. Model C_{single} (a) provides two different sets of predicted activities for each orientation, linked together by a horizontal line, whereas model C_{duplo} (b) yields identical values for both orientations.

Table 2. HIV-1 protease inhibitors included in the prediction set[a]

No.	Chemical structure	Exp. pIC_{50}	No.	Chemical structure	Exp. pIC_{50}
35		6.230	**43**		10.267
36		9.161	**44**		7.277
37		6.246	**45**		5.168
38		8.886	**46**		5.523
39		10.222	**47**		8.116
40		5.897	**48**		6.640
41		9.638	**49**		5.328
42		8.268	**50**		5.862

[a] See Refs. 1 and 2.

substituents with subunit B when the ligand binds in the alternative orientation. Therefore, both alternative binding modes can be represented with a simple "copy and paste" of the matrix describing the interaction (Figure 3). The alternative orientation is thus simulated by duplicating and swapping the blocks of variables that describe the electrostatic and steric contributions of every residue in each subunit to the ligand–receptor interaction energy. The descriptors involving interactions with

Table 3. Comparison of models C_{single} and C_{duplo}[a]

Model	Objects	Variables	LV	r^2	q^2	$SDEP_{cv}$	$SDEP_{ext}$	$SDEP_{dup}$
C_{single}	32	47	2	0.90	0.73	0.69	0.59	1.54
C_{duplo}	64	56	2	0.89	0.72	0.69	0.79	0.79

[a] r^2, Squared correlation coefficient; q^2, squared cross-validated correlation coefficient (using five randomly assigned groups); $SDEP_{cv}$, standard deviation error of the predictions, as obtained from the cross-validation analysis; $SDEP_{ext}$, standard deviation error of the predictions, obtained from the prediction of the external prediction set; $SDEP_{dup}$, standard deviation error of the predictions, obtained from the prediction of the external prediction set, but considering both orientations for each ligand.

residues Asp A25 and Asp B25 were not duplicated as these aspartic acid residues cannot be considered equivalent owing to their different protonation states.[8] As in C_{single}, two extra terms were included to describe the electrostatic contribution to the desolvation of both the ligand and the receptor.[2]

The new model (C_{duplo}) was built using the same methodology employed to derive C_{single}, as described in Ref. 2. Both models were produced using compounds **1–34** (Table 1) as the training set and compounds **35–50** (Table 2) as the prediction set. The steric contributions to the ligand–receptor interaction energies were computed using the AMBER force field,[9] whereas the electrostatic contributions to both the ligand–receptor interaction energies and the desolvation of ligands and receptor were computed using DelPhi.[6] Data pretreatment and PLS model building and validation were performed using GOLPE 3.0.[10] The quality of models C_{single} and C_{duplo} is summarized in Table 3.

RESULTS

Advantages of the new model: Predictive ability

Table 3 shows that the quality of models C_{single} and C_{duplo} is comparable. The ability to predict the activities of the external set, in the binding orientation studied, is better for C_{single} ($SDEP_{ext}$ = 0.59 compared with $SDEP_{ext}$ = 0.79 in C_{duplo}). However, if the compounds in the external prediction set are considered in both possible binding orientations, so that the number of objects and activity values is duplicated, C_{single} yields a larger error in the predictions ($SDEP_{dup}$ = 1.54) whereas C_{duplo} is insensitive to ligand orientation ($SDEP_{dup}$ = 0.79).

It must be borne in mind that the training set consists of molecules in which substituent variation is limited to just half of the ligand, so that when only one binding orientation is

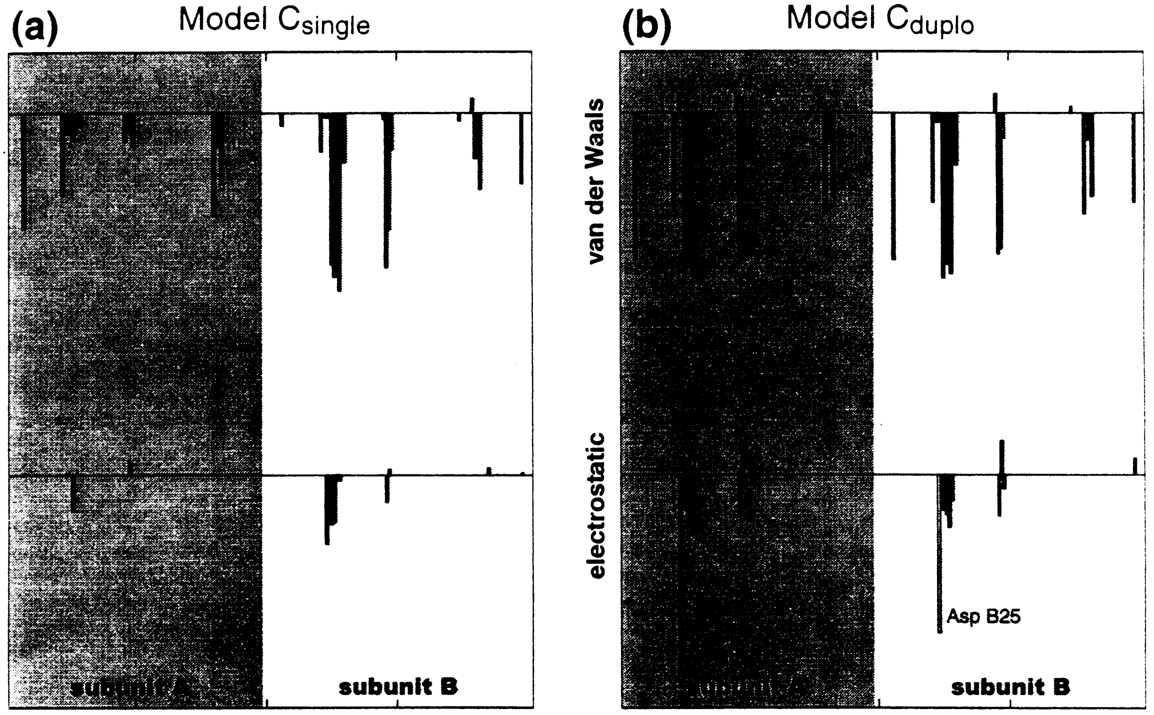


Figure 5. Weighted PLS pseudocoefficients for each of the van der Waals and electrostatic energy variables studied for models C_{single} (a) and C_{duplo} (b). On the horizontal axis, the variables are ordered sequentially within subunit A (1–99, darker background) and subunit B (100–198, lighter background), followed by the interactions with the water molecule (199). Note the equivalence between residues from both subunits in model C_{duplo} except for residues Asp A25 and Asp B25.

considered the resulting QSAR model (C_{single}) has learned about the effect of differences from interactions with just one side of the receptor. In the C_{duplo} model, these differences in the ligand are considered as making contributions to interactions with both halves of the receptor. In this way, the QSAR model has been trained to recognize changes in both sides of the ligand, and this is why C_{duplo} performs better on the external prediction set in which there are changes on both sides of the ligand. As can be seen from the plots represented in Figure 4, model C_{single} gives two different activity values to each compound, one for each orientation (Figure 4a). One of them is consistently much lower than the experimental value whereas the other is much closer, even though a tendency to underpredict is also observable. The reason for this behavior is related to the characteristics of the training series: the model is able to recognize the influence on the activity of just one-half of the molecules. Therefore, when challenged with molecules displaying variation on both halves, the predictions are more accurate when the part of the ligand showing more variation is oriented toward the best represented part of the receptor, but even so the model will not recognize the increase in affinity provided by the other part. On the other hand, C_{duplo} produces unique, consistent data of biological activity (Figure 4b). Moreover, the predicted residuals do not show any significant bias of the model to overpredicting or underpredicting the activity of external compounds.

Advantages of the new model: Interpretability

COMBINE models are useful to highlight those receptor residues whose interaction with the ligand contribute most to increasing the biological activity.[2–4] This information can be represented in a simple fashion in the form of weighted PLS pseudocoefficients, as shown in Figure 5. It is apparent that model C_{single} (Figure 5a) assigns different importance to equivalent residues of subunits A and B (except for Asp A25 and Asp B25, which are not equivalent),[8] whereas in the new C_{duplo} model (Figure 5b) a unique value is given.

DISCUSSION

The rationale behind a COMBINE analysis is to accurately translate distinct ligand–receptor interaction energies from a set of complexes into a large number of informative variables in order to find a model that correlates these descriptors and the biological activity. The hope is that the data matrix can capture the essence of all the structural changes in the compounds studied ("training set"), and that the resulting quantitative model will highlight those variables that are more important for improving the activity.[2–4] Once this information is gained, it can be used to advantage in the design of new structural changes. However, the quality of a QSAR model is limited by the quality of the training series, and in the original work,[1,2] the arbitrary assumption of studying only one of the two alternative binding modes may have limited the ability of the model to give accurate predictions for new compounds.

In the present study, since each compound is included in the training set twice (in both orientations and with the same activity value), the method is forced to produce an answer that respects the pseudosymmetry of the target. Thus, the solution is constrained to reproduce an already known characteristic of the ligand–receptor complex. Such restrained solutions have the advantages of being less sensitive to the noise in the descriptor variables and of producing more robust models. In addition, since the number of objects is duplicated, the variables-to-objects ratio is decreased.

It should also be emphasized that the method described here requires neither additional model building nor extra experimental work, as it simply manipulates already existing data. The numerical manipulation of the data matrix is simple and can be carried out on a standard spreadsheet or using small purpose-written programs.

The method reported here has a limited range of application because there are few biological receptors showing similar symmetric characteristics. However, it can be applied to different series of HIV-1 protease inhibitors. Also, it should be noted that the method described is not limited to COMBINE models and can be applied in the context of other 3D QSAR methodologies,[11] such as CoMFA, GRID/GOLPE, etc.

CONCLUSIONS

The procedure described represents the incorporation of symmetry constraints into a COMBINE 3D QSAR model. The method is performed by a simple manipulation of the data matrix already obtained and results in a model with comparably good predictive ability that is easier to interpret.

REFERENCES

1 Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscount, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L., and Zugay, J.A. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* 1995, **38,** 305–317

2 Pérez, C., Pastor, M., Ortiz, A.R., and Gago, F. Comparative binding energy (COMBINE) analysis of HIV-1 protease inhibitors: Incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.* 1998, **41,** 836–852

3 Ortiz, A.R., Pisabarro, M.T., Gago, F., and Wade, R.C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* 1995, **38,** 2681–2691

4 Wade, R.C., Ortiz, A.R., and Gago, F. Comparative binding energy analysis. In: *3D QSAR in Drug Design,* Vol. 2: *Ligand–Protein Interactions and Molecular Similarity* (H. Kubinyi, G. Folkers, and Y.C. Martin, eds.). Kluwer Academic Publishers, Dordrecht, 1998, pp. 19–34

5 Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., Ferguson, D.M., Seibel, G.L., Singh, U.C., Weiner, P., and Kollman, P.A. *AMBER (UCSF): Assisted Model Building with Energy Refinement,* version 4.1. Department of Pharmaceutical Chemistry, University of California, San Francisco, 1995

6 Nicholls, A., and Honig, B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *J. Comput. Chem.* 1991, **12,** 435–445

7 Wlodawer, A., and Erickson, J.W. Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* 1993, **62,** 543–585

8 Wang, Y.-X., Freedberg, D.I., Yamazaki, T., Wingfield, P.T., Stahl, S.J., Kaufman, J.D., Kiso, Y., and Torchia, D.A. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry* 1996, **35,** 9945–9950

9 Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 1995, **117,** 5179–5197

10 Pastor, M. *GOLPE*, version 3.0. *Multivariate Infometric Analysis (MIA)*. Perugia, Italy, 1996

11 Kubinyi, H. (ed.). *3D QSAR in Drug Design: Theory, Methods and Applications*. ESCOM Science Publishers, Leiden, 1993