PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

gCOMBINE: A graphical user interface to perform structure-based comparative binding energy (COMBINE) analysis on a set of ligand-receptor complexes

Rubén Gil-Redondo,[1] Javier Klett,[1] Federico Gago,[2] and Antonio Morreale[1]*

[1] Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa, Campus UAM, c/Nicolás Cabrera 1, Madrid 28049, Spain

[2] Departamento de Farmacología, Universidad de Alcalá, Alcalá de Henares, Madrid 28871, Spain

ABSTRACT

We present gCOMBINE, a Java-written graphical user interface (GUI) for performing comparative binding energy (COMBINE) analysis (Ortiz et al. J Med Chem 1995; 38:2681–2691) on a set of ligand-receptor complexes with the aim of deriving highly informative quantitative structure-activity relationships. The essence of the method is to decompose the ligand-receptor interaction energies into a series of terms, explore the origins of the variance within the set using Principal Component Analysis, and then assign weights to selected ligand-residue interactions using partial least squares analysis to correlate with the experimental activities or binding affinities. The GUI allows plenty of interactivity and provides multiple plots representing the energy descriptors entering the analysis, scores, loadings, experimental versus predicted regression lines, and the evolution of parameters such as r^2 (correlation coefficient), q^2 (cross-validated r^2), and prediction errors as the number of extracted latent variables increases. Other representative features include the implementation of a sigmoidal dielectric function for electrostatic energy calculations, alternative cross-validation procedures (leave-N-out and random groups), drawing of confidence ellipses, and the possibility to carry out several additional tasks such as optional truncation of positive interaction energy values and generation of ready-to-use PDB files containing information related to the importance for activity of individual protein residues. This information can be displayed and color-coded using a standard molecular graphics program such as PyMOL. It is expected that this user-friendly tool will expand the applicability of the COMBINE analysis method and encourage more groups to use it in their drug design research programs.

Proteins 2010; 78:162–172.

Key words: 3D-QSAR; comparative binding energy analysis; drug design.

INTRODUCTION

Accurate prediction of biological activities for newly designed molecules is one of the greatest challenges faced in computer-aided drug research. Because the number of three-dimensional structures for pharmacologically relevant targets is continually increasing, the pioneering quantitative structure-activity relationship (QSAR) methodologies that rely solely on physico-chemical parameters of substituents in congeneric series of compounds[1] or on molecular interaction fields (MIF) calculated at discrete points in a three-dimensional (3D) lattice embedding the spatially aligned compounds, as in the popular comparative molecular field analysis (CoMFA)[2] and comparative molecular similarity indices analysis (CoMSIA),[3] have given way to other computational methods that attempt to derive as much information as possible from the structures of the ligand-receptor complexes. In this regard, continuous advances in the fields of comparative (homology) modeling of proteins of unknown experimental structure[4] and automated ligand docking,[5] as well as in computer power, have made it possible to search for new putative ligands for many different targets out of pools containing millions of candidate compounds (chemical libraries).[6] Although this "virtual screening" approach has met with some success the technique is far from being mature because of several reasons, the most important one possibly being that the scoring functions used to rank the molecules and prioritize the possible hits are not accurate enough. When the number of compounds for a given target is kept under a few hundred, however, and biological activities are known, as is usually the case in a typical medicinal chemistry project, it

The authors state no conflict of interest.
Grant sponsor: Comunidad de Madrid; Grant number: SBIO-0214-2006; Grant sponsor: Ministerio de Ciencia e Innovación (Programa Personal Técnicos de Apoyo 2008)
*Correspondence to: Antonio Morreale Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa, Campus UAM, c/ Nicolás Cabrera 1, Madrid 28049, Spain.
E-mail: amorreale@cbm.uam.es
Received 8 April 2009; Revised 19 June 2009; Accepted 10 July 2009
Published online 20 July 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22543

should be feasible to employ a highly precise energy function to describe the ligand-receptor interactions and then derive an accurate QSAR with predictive ability using a method such as comparative binding energy (COMBINE) analysis.[7]

COMBINE employs a number of residue-based interaction energies (both van der Waals and electrostatic) computed on a set of refined ligand-receptor complexes (rather than MIF calculated on a 3D grid for a set of unbound superimposed molecules, as is done in CoMFA) to build a data matrix that is then subjected to multivariate statistical analysis. The key idea is that partial least squares (PLS) can be used to correlate the computed energy components (plus additional optional terms such as receptor and ligand desolvation energies) with the experimental activities using an expression of the form shown in Eq. (1):

$$\Delta G = \sum_{i=1}^{2n} w_i u_i + C \qquad (1)$$

where, for n protein residues (or protein residues plus selected water molecules), there are $2n$ terms, u_i, each representing either a van de Waals or an electrostatic residue-based interaction energy with the ligand that contributes to the total binding free energy according to a weighting factor, w_i, that is determined from the PLS analysis, and C is a constant. Although other partitioning schemes are also possible,[7,8] only whole protein residues and ligands are supported in the present implementation. Obviously, if one of these contributions does not vary significantly among the complexes it cannot be used to account for activity/affinity differences within the series however important it may be for the constant term, C, and the overall free energy of binding. A pictorial representation of the COMBINE workflow is shown in Figure 1.

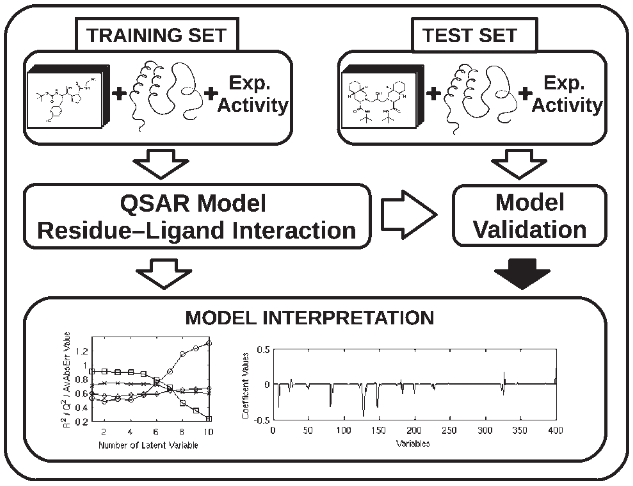


Figure 1

COMBINE workflow.

Since its inception in 1995, when the original approach was developed to account for the differences in activity in a series of human synovial fluid phospholipase A_2 inhibitors,[7] the COMBINE method has been applied to the study of other small molecules binding to different protein targets (HIV-1 protease,[9] human cytochrome P450 1A2,[10] human neutrophil elastase,[11] HIV-1 reverse transcriptase (RT),[12] acetylcholinesterase,[13-15] haloalkane dehalogenases DhlA[16] and LinB[17]) and also to peptide-protein,[18,19] protein-protein,[20] and protein-DNA interactions.[8]

The method has acquired some relevance within the 3D-QSAR field and has been reviewed in detail by Wade et al.,[21,22] Damborsky et al.,[17] and more recently by Lushington et al.,[23] who have also proposed some ideas to enhance COMBINE's capabilities in the future.

An important milestone in the development of the COMBINE methodology was the incorporation of multiple structures into the analysis, which allows the introduction, at least in part, of target flexibility.[8,24-26] All of these studies demonstrate that qualitatively reliable COMBINE models can be obtained using multiple structural representations of the receptor, but care must be taken when attempting to perform a quantitative analysis, due to the conformational dependence of the models. Another interesting issue related to structural variation entering a COMBINE analysis is the joint study of affinity and selectivity by use of different protein targets belonging to the same family, which may provide important guidelines for drug design. Key examples are the study by Wang and Wade[27] of sialic and benzoic acid analogs binding to N2 and N9 subtypes of neuraminidase and the study of ligand binding to three serine proteases (trypsin, thrombin, and coagulation factor Xa) by Murcia et al.[28] This approach can, in principle, be extended to an arbitrary number of receptors from the same protein family.

COMBINE analysis can also be linked to a docking algorithm, as shown by Murcia and Ortiz,[25] when screening virtual libraries to derive more reliable bound conformations of the putative ligands, improve the predictive ability of the regression models, and increase the enrichment factors.

It can be seen therefore that COMBINE analysis occupies a privileged position as an effective tool that can aid in the design and optimization of drug candidates. However, as recently stated by Lushington et al.,[23] the small number of groups worldwide using COMBINE should join their efforts to disseminate the method and make it available to the scientific community. Although our group promoted this initiative some time ago through the free Web-based release of the COMBINE code,[29] a user-friendly graphical interface that allows easy manipulation of data and input/output files was lacking. In this contribution we fill this gap by providing the COMBINE program with a graphical user interface (GUI) written in Java to ensure portability to different operating systems that is being released under a scientific/academic nonprofit and noncommercial license.

MATERIALS AND METHODS

Technical details of the gCOMBINE application

gCOMBINE is the GUI developed as a user-friendly wrapper to the original command-line COMBINE program. The GUI has been written in Java[30] language (v. 1.6.0_10), which ensures platform portability. Graphics functionality and interactivity have been added with Java Foundation Classes (JFC) and Swing components, respectively. For development of the GUI the NetBeans[31] IDE (Integrated Development Environment) 6.1 was used and the Swing Application Framework[32] was included. The different charts were generated using the JFreeChart[33] 1.0.11 and JCommon[34] 1.0.14 libraries. Both are distributed under a GNU Lesser General Public License.[35] These libraries allow gCOMBINE to generate interactive charts with the most relevant data for easy manipulation and analysis. Because the GUI is platform-independent and the COMBINE program is written in standard GNU Fortran, the complete application (COMBINE + gCOMBINE) can be used under Linux, Windows or Mac operating systems with the gcc compiler (v. 3.4.6 was employed in our case). gCOMBINE has an object-oriented design based on the Model-View-Controller (MVC) pattern.[36] The main class for the Model is *CombineModel*. An instance of this class stores the information about a specific model (or configuration) generated from a COMBINE run: name for the model, a description comment, the working folder, the configuration parameters, the output files, and tables and charts. The parameters are stored on an instance of the Parameters class, which uses the *ComplexesListItem* objects to keep the different ligand-receptor complexes related to the COMBINE model under study. The tables and charts are panels generated by the static methods of the classes *CombineTables* and *CombineGraphs*, respectively, taking a *CombineModel* instance as input. The View is launched by the *CombineGUIApp* class that creates an instance of the *CombineGUIView* class. This instance acts as a store for the different graphical objects and also as the Controller for the different actions (including internal validations) that can be performed when interacting with the objects. An internal class (*CombineThread*) is used to run the COMBINE program in a different execution thread to avoid the blockade of the GUI interface while COMBINE is running. *CombineThread* uses an instance of the *CombineWrapper* class to prepare the COMBINE execution: it launches the calculation, controls the process (taking the logs with the *StreamGlobber* class) and loads the results upon completion of the run. Three other classes are used through the life cycle of the application: a) *CombineConstants* (to contain different constants); b) *CombineException* (to propagate customized errors and warnings); and c) *Useful* (to store some common methods).

The statistics behind COMBINE

As in any other method focused on obtaining a quantitative view of structure-activity relationships, the core of COMBINE is a matrix containing structure-related energy descriptors (variables) to be correlated with biological activities or binding energies. The statistical method underlying COMBINE analysis is well known and widely accepted by the scientific community, and therefore it has been intensely reviewed.[37] However, a brief summary of the main steps and ideas follows.

Construction of the X matrix

The *X* matrix contains the entire set of variables describing the interaction energies between each ligand and every protein residue for all the complexes. Usually these are van der Waals interactions calculated using a molecular mechanics force field (typically AMBER) and electrostatic interactions calculated using point charges and either Coulomb's law (and a constant or distance-dependent dielectric definition) or the more elaborate and accurate generalized born (GB)[38] or Poisson-Boltzmann (PB)[39] methods. In addition, desolvation energy terms for both receptor and ligand can also be incorporated as "external variables". For each complex only two AMBER-type files are required, one containing the atomic coordinates (.crd extension) and the other containing the topology (i.e. atom connectivity), atom types and force-field parameters (.top extension). Alternatively, the user can generate the *X* matrix externally in the format described in the User Guide and load it into the program.

Pretreatment of the X matrix

To reduce the number of variables while keeping all the relevant information within the *X* matrix, those interaction energy values with a standard deviation below a user-defined cut-off, which can be safely assumed not to contribute to the overall variance in activity, can be removed (Pretreatment cut-off, see below). Positive energy values, which in some cases could arise from force-field inconsistencies or modeling errors, can optionally be truncated to zero (Pretreatment option, see below). Scaling of the variables can also be performed using two different approaches (Scaling option, see below): (i) standard scaling, where the mean value over the whole set of variables is subtracted from each variable and divided by the standard deviation (it is therefore similar to a Z-score), and (ii) block scaling, by means of which the mean value of the variables is subtracted from the one being scaled and divided by the standard deviation of these variables (again, similar to a Z-score but using a group of variables).

PLS regression

This technique combines and generalizes features from PCA and Multiple Linear Regression (MLR) in the sense that not only orthogonal Principal Components (PC) are extracted, as in PCA, but also a fitting procedure is performed to describe the activities of the compounds (the dependent variable), as in MLR. There are two initial matrices in a COMBINE analysis: (i) a matrix containing the independent variables (interaction energies, and possibly additional variables such as desolvation energies), the X matrix [Eq. (2)], and (ii) a matrix (column vector) with the dependent variable (activities), the Y matrix [Eq. (3)].

$$X = \begin{pmatrix} E_1^1 & E_2^1 & \dots & E_M^1 & V_1^1 & V_2^1 & \dots & V_M^1 & A_1^1 & \dots & A_S^1 \\ E_1^2 & E_2^2 & \dots & E_M^2 & V_1^2 & V_2^2 & \dots & V_M^2 & A_1^2 & \dots & A_S^2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ E_1^N & E_2^N & \dots & E_M^N & V_1^N & V_2^N & \dots & V_M^N & A_1^N & \dots & A_S^N \end{pmatrix} \tag{2}$$

where E_j^i, V_j^i, and A_j^i are the electrostatic, van der Waals, and additional variables, respectively. N is the number of compounds, M is the number of residues in the protein, and S is the number of additional variables.

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} \tag{3}$$

where y_i is the individual activity of compound i. The PLS analysis starts by decomposing the X and Y matrices into one score matrix, T, and two different loading matrices, P and Q [Eq. (4)], using the iterative NIPALS algorithm[40]:

$$X = TP^T \quad Y = TQ^T \tag{4}$$

The loading matrices P and Q contain information about the variables in the so-called LV or PC space. These are orthogonal vectors obtained as linear combinations of the original variables in the X matrix. The coefficients in a given PC provide information on the relative weight of the different terms and can be used to deduce the relevance of each individual ligand–residue interaction to explain the variance in activity/affinity. On the other hand, the score matrix T contains information about the compounds, described in terms of their projections onto the PCs. The PC space is normalized and has a mean of zero, so compounds with high scores should be checked as they could behave as outliers. In addition, clusters of compounds can be detected. A plot of the regression line between the experimentally determined and theoretically calculated activity/affinity values and calculation of the regression coefficient [r^2, Eq. (5)] allow the user to visualize the quality of the fit for the training set compounds, and also for the excluded (not used) or test compounds.

$$r^2 = \frac{\left[\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \langle\hat{y}\rangle)\right]^2}{\sum_{i-1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \langle\hat{y}\rangle)^2} \tag{5}$$

where $\langle\hat{y}\rangle = \frac{\sum_{i=1}^{N}\hat{y}_i}{N}$.

Cross-validation

This method, which is used to check that the derived correlation is not spurious and to assess the robustness of the resulting statistical model, consists of predicting the dependent variable for some complexes that are not included in model derivation. Briefly, if C is the whole set of N compounds ($C = \{C_1,\dots,C_N\}$) with associated activities Y ($Y = \{y_1,\dots,y_N\}$), the method builds a number of subsets of s elements from C (when $s = 1$ it is the commonly employed Leave-One-Out option) and sets them apart for their activities/affinities to be predicted later, thus making up an internal test set. The compounds represented by s can be selected by following a predetermined sequential order or can be randomly assigned to a predetermined number of groups (Cross-validation Method option, see below). In any case, C can be split into k subsets S_i ($S_i = \{S_{i1},\dots,S_{is}\}$) where the subscript i represents any subset number from 1 to k. Numerically, k is the nearest integer greater than or equal to N/s (the ratio between the total number of compounds, N, and the number of those making up each subset, s). Usually, the last group would have less than s elements as it contains the remaining compounds. In the next step, k PLS regression models are built: model 1 with all N compounds except for those in S_1, model 2 with all N compounds except for those in S_2, and so on. In each case, the activities for compounds in S_1, S_2... will be estimated from their respective models ($\hat{y}_1, \dots, \hat{y}_N$) and at the end of the process a list of predicted activities for all the compounds will be obtained. The performance is then quantified by the q^2 cross-validated correlation coefficient [Eq. (6)]:

$$q^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6}$$

where $\bar{y}$ is the average value of the activity ($\bar{y} = \sum_{i=1}^{N} y_i/N$). In simple words, this metric describes the amount of variance in the dataset that is explained by the model. Besides, a standard deviation of error of predictions [$SDEP$, Eq. (7)] and an average absolute error [AAE, Eq. (8)] are also calculated.

$$\mathrm{SDEP} = \sum_{i=1}^{N} \sqrt{\frac{(\hat{y}_i - y_i)^2}{N}} \tag{7}$$

$$\mathrm{AAE} = \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{N} \tag{8}$$

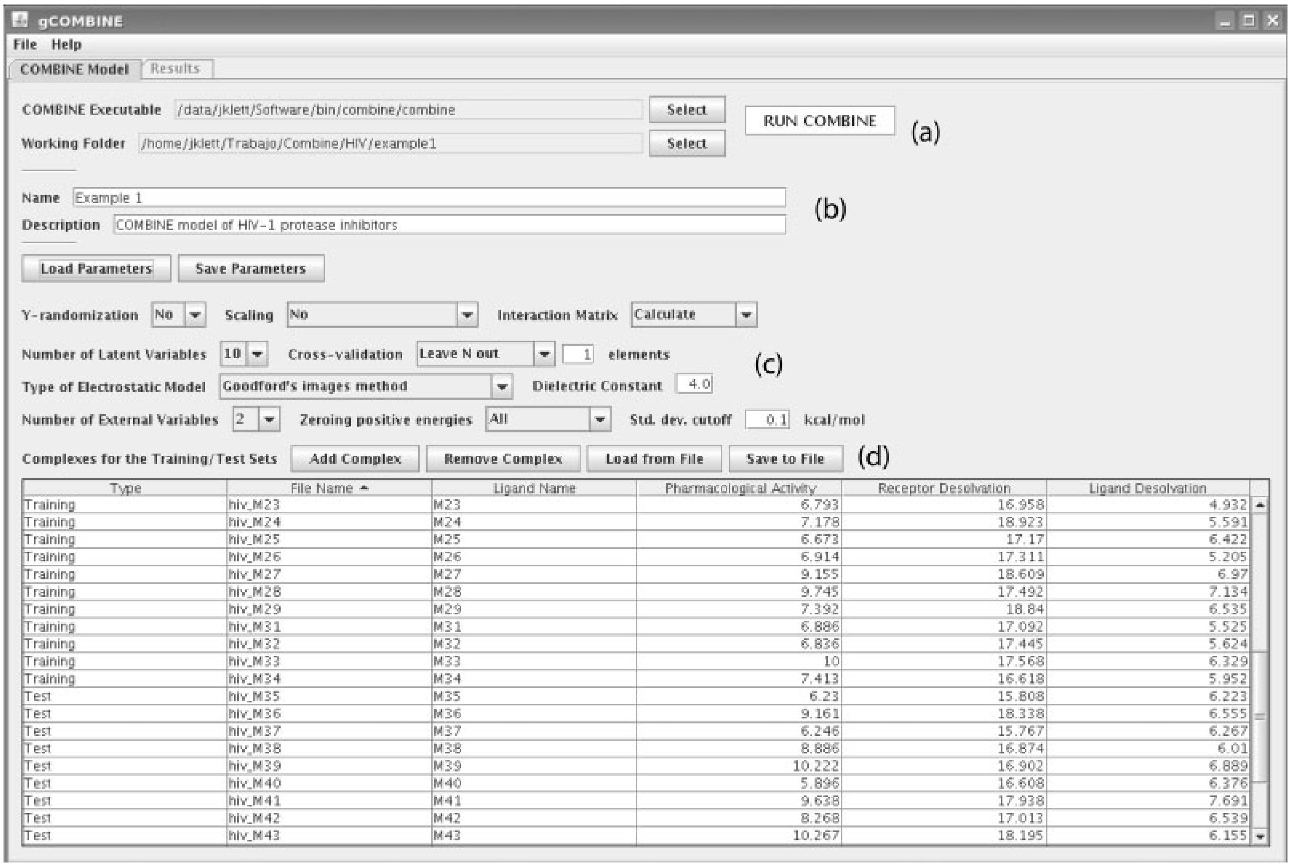
It can be argued that, despite the cross-validation procedure, the resulting model fits the data just by chance due to the selection of a fortuitous equation out of the huge amount of different PLS regression models that can be constructed with the thousands of variables contained in the *X* matrix. To check against this possibility the affinities/activities of the compounds can be randomly reassigned (permutation of activities or Y-randomization) to prove the point that in this case it is usually not possible to derive an acceptable model. gCOMBINE allows the user to carry out this task thorough the Y-randomization option in the main window (see below). This test is performed 100 times and therefore it is quite time-consuming.

Selection of the best model

As successive components are extracted from the *X* matrix, a check is made to estimate the amount of variance that is recovered (it must be borne in mind that the PLS method attempts to explain the variance not only in the *X* matrix, as does PCA, but also in the *Y* matrix). Although there is not a strict rule to select the best model resulting from a PLS analysis, the general guidance is to study the evolution of both the cross-validated correlation coefficient [q^2, Eq. (6)] and the standard deviation of the errors in prediction [SDEP, Eq. (7)]. gCOMBINE provides the user with a graphical description showing the evolution of the main chemometric indices as new components are being extracted (five by default, and up to 10, Number of Latent Variables option, see later) to facilitate the decision on the optimal dimensionality to choose for further analysis.

External validation

The best way to validate a PLS model is to challenge it with an external set of modeled complexes and compare the predicted affinity/activity values for the bound



COMBINE Executable /data/jklett/Software/bin/combine/combine

Working Folder /home/jklett/Trabajo/Combine/HIV/example1

Name Example 1

Description COMBINE model of HIV-1 protease inhibitors

Y-randomization No Scaling No Interaction Matrix Calculate

Number of Latent Variables 10 Cross-validation Leave N out 1 elements

Type of Electrostatic Model Goodford's images method Dielectric Constant 4.0

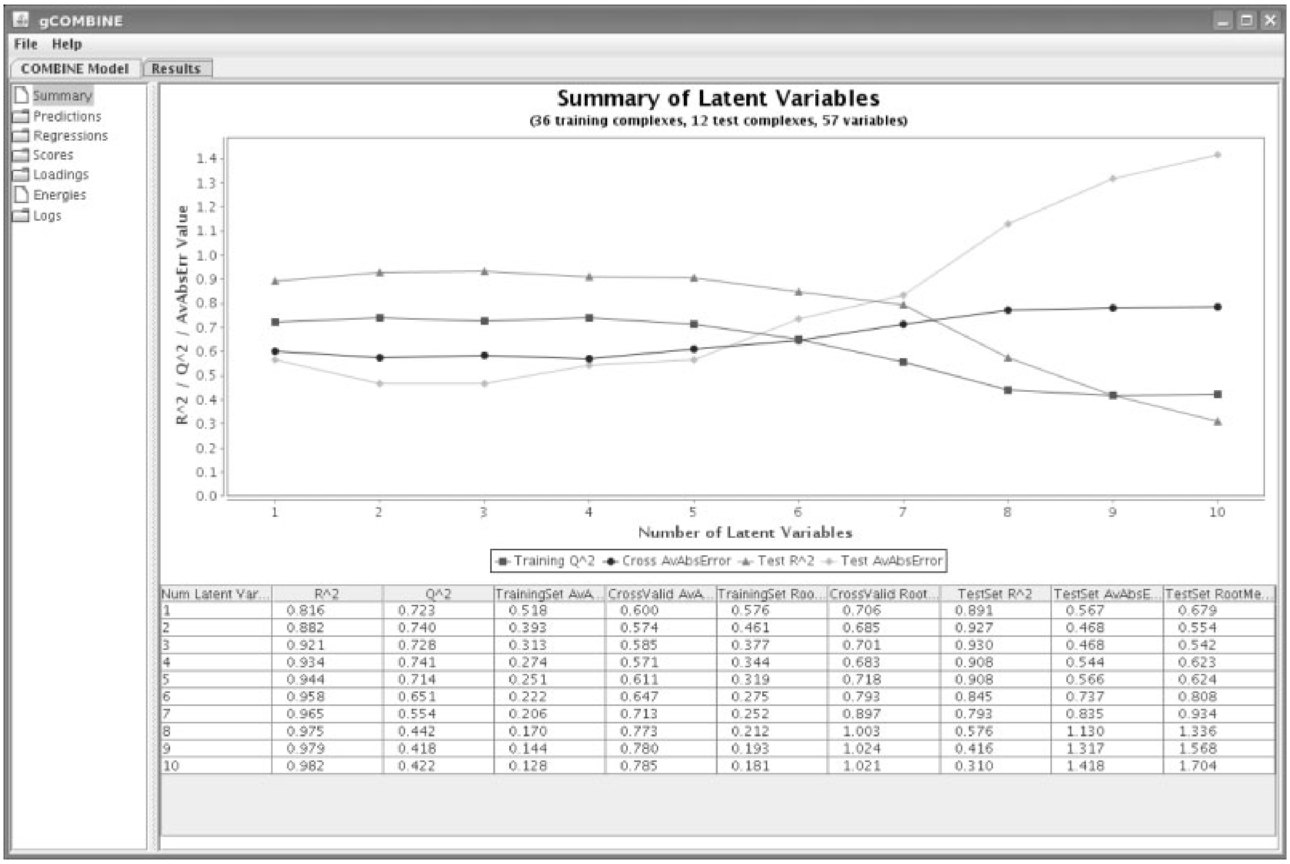
Number of External Variables 2 Zeroing positive energies All Std. dev. cutoff 0.1 kcal/mol

Complexes for the Training/Test Sets

Type	File Name ▲	Ligand Name	Pharmacological Activity	Receptor Desolvation	Ligand Desolvation
Training	hiv_M23	M23	6.793	16.958	4.932
Training	hiv_M24	M24	7.178	18.923	5.591
Training	hiv_M25	M25	6.673	17.17	6.422
Training	hiv_M26	M26	6.914	17.311	5.205
Training	hiv_M27	M27	9.155	18.609	6.97
Training	hiv_M28	M28	9.745	17.492	7.134
Training	hiv_M29	M29	7.392	18.84	6.535
Training	hiv_M31	M31	6.886	17.092	5.525
Training	hiv_M32	M32	6.836	17.445	5.624
Training	hiv_M33	M33	10	17.568	6.329
Training	hiv_M34	M34	7.413	16.618	5.952
Test	hiv_M35	M35	6.23	15.808	6.223
Test	hiv_M36	M36	9.161	18.338	6.555
Test	hiv_M37	M37	6.246	15.767	6.267
Test	hiv_M38	M38	8.886	16.874	6.01
Test	hiv_M39	M39	10.222	16.902	6.889
Test	hiv_M40	M40	5.896	16.608	6.376
Test	hiv_M41	M41	9.638	17.938	7.691
Test	hiv_M42	M42	8.268	17.013	6.539
Test	hiv_M43	M43	10.267	18.195	6.155

Figure 2

gCOMBINE main window. Letters a–d refer to the four main data blocks (see text).



Num Latent Var...	R^2	Q^2	TrainingSet AvA...	CrossValid AvA...	TrainingSet Roo...	CrossValid Root...	TestSet R^2	TestSet AvAbsE...	TestSet RootMe...
1	0.816	0.723	0.518	0.600	0.576	0.706	0.891	0.567	0.679
2	0.882	0.740	0.393	0.574	0.461	0.685	0.927	0.468	0.554
3	0.921	0.728	0.313	0.585	0.377	0.701	0.930	0.468	0.542
4	0.934	0.741	0.274	0.571	0.344	0.683	0.908	0.544	0.623
5	0.944	0.714	0.251	0.611	0.319	0.718	0.908	0.566	0.624
6	0.958	0.651	0.222	0.647	0.275	0.793	0.845	0.737	0.808
7	0.965	0.554	0.206	0.713	0.252	0.897	0.793	0.835	0.934
8	0.975	0.442	0.170	0.773	0.212	1.003	0.576	1.130	1.336
9	0.979	0.418	0.144	0.780	0.193	1.024	0.416	1.317	1.568
10	0.982	0.422	0.128	0.785	0.181	1.021	0.310	1.418	1.704

Figure 3

gCOMBINE *Results* tab showing the evolution of the chemometric indices in graphical (Top) and tabular form (Bottom).

ligands with the actual ones. This is accomplished in gCOMBINE by feeding the program with additional complexes that are marked as “Test” in lieu of “Training” in the Type column (see Main Menu in Fig. 2).

RESULTS AND DISCUSSION

The application

gCOMBINE has been designed bearing in mind the need for a simple and easy-to-use chemometric tool. Once that the program has been launched it displays a menu bar with two submenus (File and Help) and a tabbed panel with two tabs (see Fig. 2): one for the configuration of the model (*COMBINE Model*) and the other one to manage the results (*Results*). The menu bar, under File submenu, contains four options: (a) New Model, to clean the data from the tabs, (b) Load Model, to load a previously saved model into the tabs, (c) Save Model, to save the current model into a specific file, and (d) Exit, to close the application. Help submenu offers information about the original COMBINE publication, the main COMBINE author and contact information. The *COMBINE Model* tab can be divided into four main areas (a through d in Fig. 2): (a) the top part where the user can select the folder for the COMBINE executable and the working folder where the complexes are stored. Clicking on the RUN COMBINE button will start the calculation, (b) a section where the user can introduce commentaries related to the job into two boxes, Name and Description, (c) this section allows the user to load parameters from a previous calculation or to save the current parameters being used thorough the Load/Save Parameters buttons. All the parameters entering the model are configured here (the reader is referred to the “Materials and Methods” section for the definition of the different issues described here): Y-randomization (No by default), Scaling (No by default), Interaction Matrix (it can be calculated and written out by gCOMBINE or read in from an external file), Number of Latent Variables (5 are extracted by default), validation method (Leave N Out or Random Groups), Type of Electrostatic Model (uniform dielectric constant, Goodford’s implementation of the images method,[41] a distance-dependent dielectric constant, PB electrostatic interaction energies read from an external file, and a sigmoidal model[42]), Dielectric

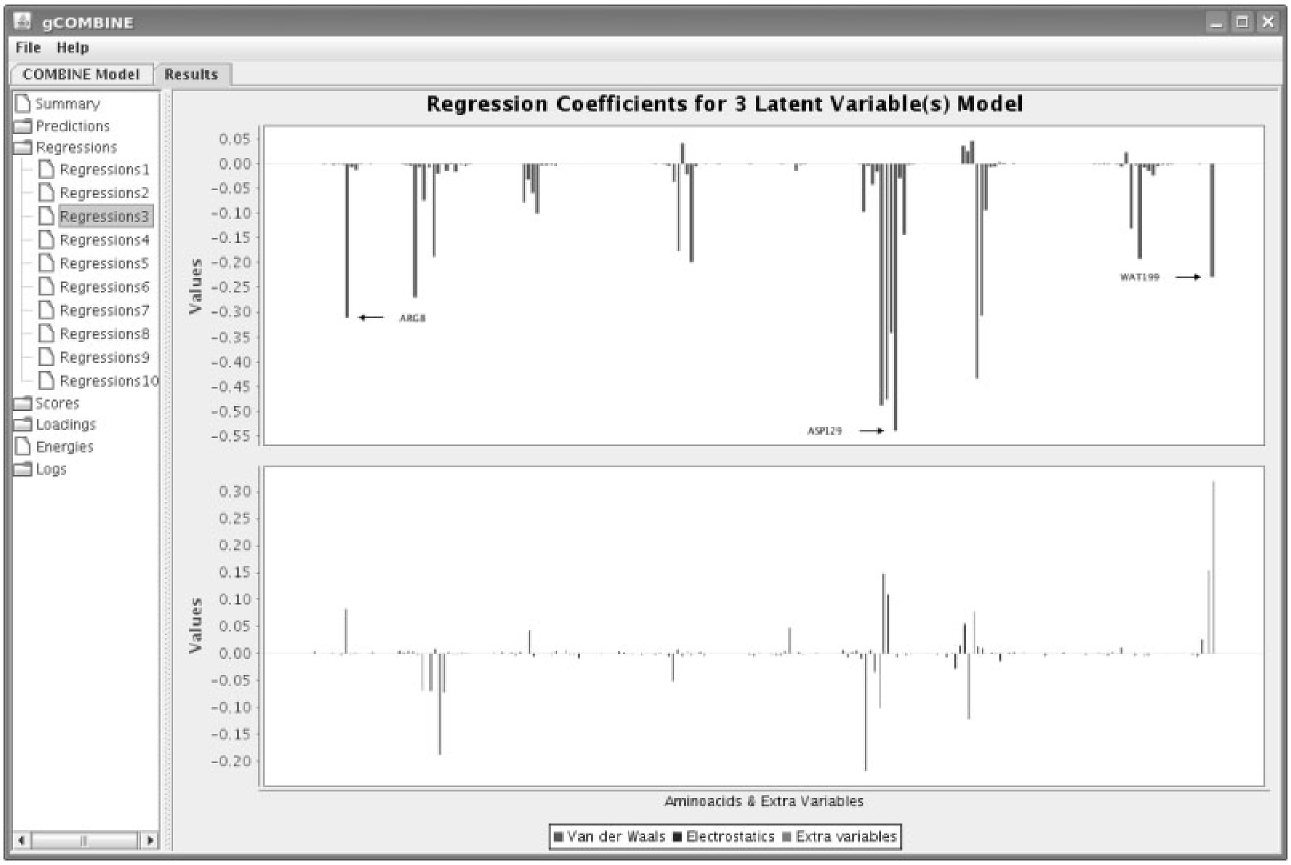


Figure 4

Plots showing the weights assigned to the residue-based van der Waals and electrostatic interaction energy values in a COMBINE model made up of four principal components to account for the differences in activity in the HIV-1 protease inhibitor series.

Constant value (four by default), Number of External Variables, Pretreatment of the data and the Pretreatment cut-off value, and (d) the complex area, where the user can add/remove/load/save complexes from/to a file (Add/Remove Complex and Load/Save from/to File buttons). Below these buttons a table is shown with the following information: the type associated to each complex (Training, Test, or Not Used, as defined by the user), which can be changed at any time to test alternative models, the File Name containing the complex, the Ligand Name, the Pharmacological Activity, and one column for each external variable to be considered. Once the parameters and the complexes have been read in and after the program begins to run, the application checks if all the parameters needed have been supplied, if they have valid values, and if the required files exist.

The user can stop the execution or wait until completion of the run. In this latter case, if no errors are found, the application focuses on the *Results* tab (see Fig. 3) where the user can have access to different tables and charts grouped by *nodes* and *sub-nodes*: a summary of each model (both a graphic showing the evolution of the chemometric indices and a table containing these indices are produced), PLS coefficients plots (Fig. 4), predictions (for each model a graphic of experimental vs. predicted activity values is shown, Fig. 5 panel a), loadings and scores plots (Fig. 5 panels b and c, respectively), and a plot of the interaction energy variables entering the PLS decomposed on a per-residue basis (a very useful plot to detect anomalous energy values, Fig. 5 panel d). The user can interact with the graphs in several ways: zoom in/out, see tooltips for specific data, set tags, change the appearance, save them as images, print, and so forth. Finally, there is a *Logs* node to keep track of the program messages.

Testing gCOMBINE

Among the many publications of successful COMBINE analysis, we have chosen two of them for testing the graphical implementation reported herein. The first one is the set accompanying the original distribution, which employed 48 (32 for the training set and 16 for the test set) inhibitors of human immunodeficiency virus Type 1

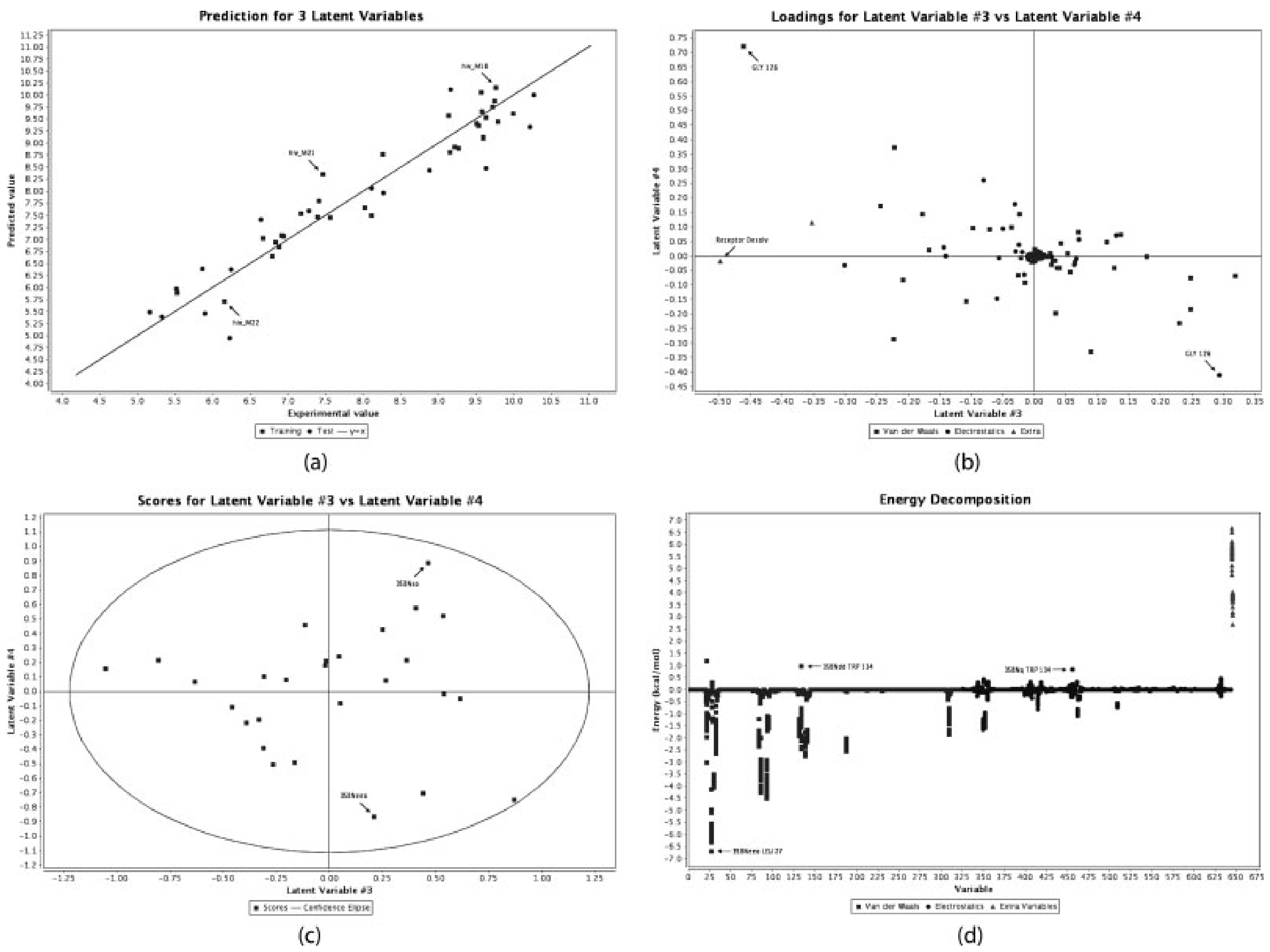

Prediction for 3 Latent Variables
Predicted value
Experimental value
hiv_M18
hiv_M21
hiv_M22
■ Training ● Test — y=x
(a)
Loadings for Latent Variable #3 vs Latent Variable #4
Latent Variable #4
Latent Variable #3
GLY 126
Receptor Desolv
■ Van der Waals ● Electrostatics ▲ Extra
(b)
Scores for Latent Variable #3 vs Latent Variable #4
Latent Variable #4
Latent Variable #3
■ Scores — Confidence Elipse
(c)
Energy Decomposition
Energy (kcal/mol)
Variable
TRP 134
TRP 134
LEU 27
■ Van der Waals ● Electrostatics ▲ Extra Variables
(d)

Figure 5

Selected screenshots from the gCOMBINE program displaying: a) experimental versus predicted activity plot, (b) plot showing the contributions (loadings) of the original variables to the principal components shown, (c) scores plot (the applicability domain is drawn as a confidence ellipse[43]), (d) plot of the original variables entering the PLS analysis following decomposition of the ligand-receptor interaction energy on a per-residue basis.

(HIV-1) protease.[9] In the second example, also related to another pharmacologically relevant HIV-1 target, namely RT,[12] we will show how externally generated electrostatic energy variables, as calculated with the DelPhi software,[44] can be incorporated into the model.

1. Two main issues motivated the first of these two studies: (a) bare ligand-receptor interaction energies (as computed by Merck researchers using the MM2X force field) per se correlated quite well (using linear regression analysis, $r^2 = 0.74$) with experimentally determined enzyme inhibition data (IC_{50} values, that is, compound concentrations giving rise to 50% inhibition of enzyme activity).[44] Moreover, the predictive ability of such a linear model on 16 test set compounds (not included in model derivation) was also remarkable ($q^2 = 0.75$) with an average absolute error around 1 across a range of 5 log units, and (b) the realization that no improvement upon incorporation of either solvation effects (using a continuum description) or using another force field (CHARMM in their case) was achieved. Taking these two issues into account, the two main objectives addressed by the

Table I

Chemometric Indices for the Different Models in the HIV-1 Protease Study Discussed in the Text

Model	Objects	Variables	LV	r^2	q^2	$SDEP_{CV}$	$SDEP_{ex}$
L_{MM2X}	32	1	1	0.74	0.75		1.00
L_{AMBER}	32	1	1	0.81	0.79	0.61	1.08
C_{AMBER}	32	48	2	0.89	0.70	0.72	0.83
gC_{AMBER}[a]	32	48	2	0.89	0.70	0.72	0.80

[a] Calculated with gCOMBINE.

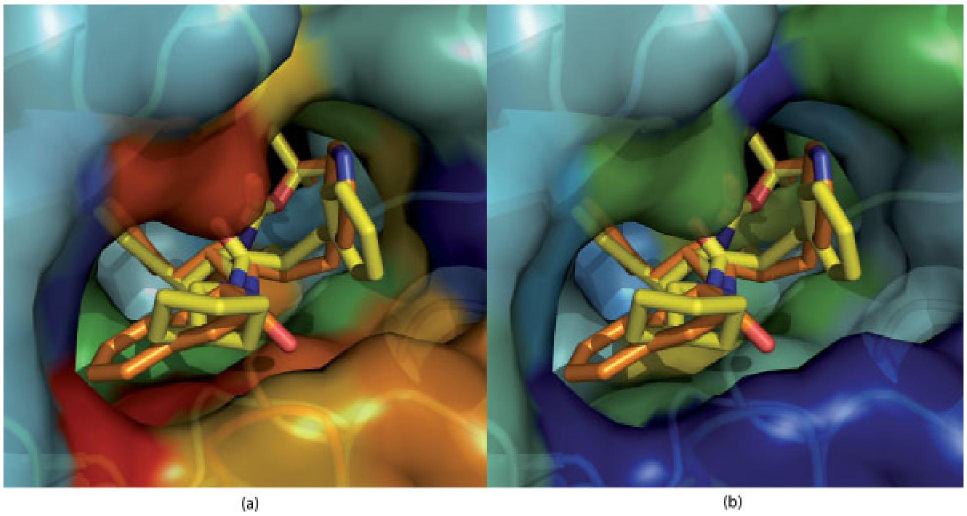


Figure 6

Visualization in PyMOL (http://pymol.sourceforge.net/) of the PLS regression coefficients plotted in Figure 4. The semitransparent surface enveloping the HIV-1 protease target has been spectrum-colored using the van der Waals (a) and electrostatic (b) PLS coefficients from the fourth column (B-factor) in the PDB file generated by gCOMBINE.

COMBINE analysis methodology were (i) to check for possible dependencies of the correlation on the force field used, and (ii) to try and develop more accurate QSAR models. In this exercise, it was shown that similar results could be obtained when comparable models were built within the framework of the AMBER force field, so no force-field dependencies were detected (see the chemometric indices for L_{MM2X} and L_{AMBER} in Table I). On the other hand, remarkable improvements were achieved through the use of COMBINE models (see the original article), especially when partial desolvation effects for ligand and receptor upon complex formation were included using a continuum description (by solving the PB equation) and the standard Coulombic distance-dependent electrostatic term was replaced with solvent-corrected values calculated for each residue. Interestingly, the main conclusion of this work was that simply replacing the Coulombic term with the continuum electrostatics description and including the desolvation effects did not lead to a significant improvement when MLR was used but the performance of the corresponding COMBINE model was dramatically enhanced when the continuum electrostatic interactions were employed. Although different improvements over the standard Coulombic term were included in the original article we will restrict ourselves here to the simplest (and more widely used) case as our intention is to reproduce the data rather than recapitulating the previously published comparison. In particular we show the reproducibility of what was called the C_{AMBER} model, where van der Waals and electrostatic contributions (the latter, a straightforward Coulombic term using a dielectric constant of four) were taken directly from the AMBER force field using the ANAL module. The results cannot be exactly the same, though, because the cross-validation technique contains a random component: the compounds are randomly assigned to one of five groups of approximately the same size, each group in turn is excluded from the analysis, and the whole procedure is repeated 20 times. Nevertheless, clearly comparable results were obtained (see chemometric indices for C_{AMBER} and gC_{AMBER} in Table I).

Finally, the relative weight assigned to individual residue-based interactions (van der Waals and electrostatic) by the COMBINE model can be color-coded and displayed on a surface representation of the protein, as shown in Figure 6.

2. The second article revisited here entailed the study of 27 6-arylsulfonyl-2-aminobenzonitrile derivatives synthesized and tested as second-generation non-nucleoside HIV-1 RT inhibitors (NNRTI).[12] In this case there was a wealth of experimental data including information about activity on RT enzymes bearing different mutations at the NNRTI binding site. COMBINE models were obtained to quantitatively characterize the observed structure-activity data and possibly to account for the effect of some of the mutations. Ligand-residue van der Waals interaction energies were calculated using AMBER parameters (parm99) while their solvent-corrected electrostatic counterparts were obtained by solving the PB equa-

Table II
Chemometric Indices Calculated by gCOMBINE for the Whole Set of 27 6-arylsulfonyl-2-aminobenzonitrile HIV-1 Reverse Transcriptase Inhibitors

LV	r^2	q^2	$SDEP_{CV}$
1	0.84	0.71	0.62
2	0.88	0.78	0.55
3	0.93	0.82	0.49
4	0.94	0.86	0.44
5	0.95	0.86	0.44

tion (as implemented in DelPhi). Also, the desolvation changes incurred by ligands and receptor upon complex formation were included as additional external variables. For testing gCOMBINE, these electrostatic energy calculations were performed using DelPhi in stand-alone mode and then loaded into the application (Type of Dielectric Model menu, Poisson-Boltzmann from.dph files option, see Figure 2 block c). By doing this, the published results were accurately reproduced with only minor variations being obtained, as expected, when random groups were employed (see Table II). Relevant results are presented in Figure 5: loading and scoring plots (panels b and c, respectively), interaction energy variables entering the PLS analysis (panel d), and the evolution of r^2, q^2, and SDEP (Table II) for the whole set of 27 NNRTI complexes as the number of PC being extracted by gCOMBINE increases. As in the previous example, no attempt will be made here to discuss or compare the results, which are already published, but it is clear that gCOMBINE faithfully reproduces the data.

CONCLUSIONS

The objective of this article has been to provide the COMBINE analysis method with an easy-to-use GUI that improves on the original command-line style implementation. The software is written in Java to allow platform portability and is made freely available to academic and/or public research institutions from a public web site (http://ub.cbm.uam.es/gCOMBINE) under an Academic License. This has been done with the idea of disseminating a user-friendly tool among the scientific community to encourage the use of a program that has proven useful in many areas related to ligand binding, structure-activity relationships and drug design.

ACKNOWLEDGMENTS

The authors thank Claire Coderch for help in testing the application. This work is dedicated to the memory of Ángel R. Ortiz, who looked forward to making this tool widely available but passed away before it was fully operative.

REFERENCES

1. Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD. Chembioinformatics: comparative QSAR at the interface between chemistry and biology. Chem Rev 2002;102:783–812.
2. Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. JACS 1988;110:5959–5967.
3. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 1994;37:4130–4146.
4. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol 2006;16:172–177.
5. Warren GL, Andrews CW, Capelli AM, Clarke B, Lalonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem 2006;49: 5912–5931.
6. Shoichet BK. Virtual screening of chemical libraries. Nature 2004;432:862–865.
7. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. J Med Chem 1995;38:2681–2691.
8. Tomic S, Nilsson L, Wade RC. Nuclear receptor-DNA binding specificity: a COMBINE and free-Wilson QSAR analysis. J Med Chem 2000;43:1780–1792.
9. Perez C, Pastor M, Ortiz AR, Gago F. Comparative binding energy 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.
10. Lozano JJ, Pastor M, Cruciani G, Gaedt K, Centeno NB, Gago F, Sanz F. 3D-QSAR methods on the basis of ligand-receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands. J Comput Aided Mol Des 2000;14: 341–353.
11. Cuevas C, Pastor M, Perez C, Gago F. Comparative binding energy (COMBINE) analysis of human neutrophile elastase inhibition by pyridone-containing trifluoromethylketones. Comb Chem High Throughput Screen 2001;4:627–642.
12. Rodríguez-Barrios F, Gago F. Chemometrical Identification of mutations in HIV-1 reverse transcriptase conferring resistance or enhanced sensitivity to arylsulfonylbenzonitriles. JACS 2004;126: 2718–2719.
13. Guo J, Hurley MM, Wright JB, Lushington GH. A docking score function for estimating ligand-protein interactions: application to acetylcholinesterase inhibition. J Med Chem 2004;47:5492–5500.
14. Lushington GH, Wallace NM, Guo JX. Reliable Prescreening of Candidate NerveAgent Prophylaxes via 3D QSAR. DTIC Monitor Series, 2005. 1–28.
15. Martin-Santamaria S, Munoz-Muriedas J, Luque FJ, Gago F. Modulation of binding strength in several classes of active site inhibitors of acetylcholinesterase studied by comparative binding energy analysis. J Med Chem 2004;47:4471–4482.
16. Kmunicek J, Luengo S, Gago F, Ortiz AR, Wade RC, Damborsky J. Comparative binding energy analysis of the substrate specificity of haloalkane dehalogenase from Xanthobacter autotrophicus GJ10. Biochemistry 2001;40:8905–8917.
17. Damborsky J, Kmunicek J, Jedlicka T, Luengo S, Gago F, Ortiz AR, Wade RC. Rational redesign of haloalkane dehalogenases guided by comparative binding energy analysis. In: Svendsen A, Dekker M, editors. Enzyme functionality: design, engineering and screening. New York: Marcel Dekker; 2004. pp 79–96.

18. Schleinkofer K, Wiedemann U, Otte L, Wang T, Krause G, Oschkinat H, Wade RC. Comparative structural and energetic analysis of WW domain-peptide interactions. J Mol Biol 2004;344: 865–881.
19. Wang T, Wade RC. Comparative binding energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. J Med Chem 2002;45:4828–4837.
20. Tomic S, Bertosa B, Wang T, Wade RC. COMBINE analysis of the specificity of binding of Ras proteins to their effectors. Proteins 2007;67:435–447.
21. Wade RC, Henrich S, Wang T. Using 3D protein structures to derive 3D-QSARs. Drug Discovery Today: Technol 2004;1:241–246.
22. Wade RC, Ortiz AR, Gago F. Comparative binding energy analysis. In: Kubinyi H, Folkers G, Martin Y, editors. 3D-QSAR in drug design, Vol. 2. Dordrecht (Netherland): Kluwer-ESCOM; 1998. pp 19–34.
23. Lushington GH, Guo JX, Wang JL. Whither combine? New opportunities for receptor-based QSAR. Curr Med Chem 2007;14: 1863–1877.
24. Mou TC, Gille A, Suryanarayana S, Richter M, Seifert R, Sprang SR. Broad specificity of mammalian adenylyl cyclase for interaction with 2′,3′-substituted purine- and pyrimidine nucleotide inhibitors. Mol Pharmacol 2006;70:878–886.
25. Murcia M, Ortiz AR. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. J Med Chem 2004;47:805–820.
26. Pastor M, Perez C, Gago F. Simulation of alternative binding modes in a structure-based QSAR study of HIV-1 protease inhibitors. J Mol Graph Model 1997;15:364–371.
27. Wang T, Wade RC. Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. J Med Chem 2001;44:961–971.
28. Murcia M, Morreale A, Ortiz AR. Comparative binding energy analysis considering multiple receptors: a step toward 3D-QSAR models for multiple targets. J Med Chem 2006;49:6241–6253.
29. COMBINE homepage. http://ub.cbm.uam.es/software.php. April 2009.
30. Java homepage. http://java.sun.com/. April 2009.
31. NetBeans homepage. http://www.netbeans.org. April 2009.
32. Swing Application Framework homepage. https://appframework.dev.java.net. April 2009.
33. JFreeChart homepage. http://www.jfree.org/jfreechart. April 2009.
34. JCommon homepage. http://www.jfree.org/jcommon. April 2009.
35. GNU Lesser General Public License homepage. http://www.gnu.org/licenses/lgpl.html. April 2009.
36. Model-View-Controller (MVC) pattern. http://java.sun.com/blueprints/patterns/MVC.html. April 2009.
37. Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. J Sci Stat Comp 1984;5:735–743.
38. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. JACS 1990;112:6127–6129.
39. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
40. Wold H. Path models with latent variables: the NIPALS approach. In: Blalock HM, editor. Quantitative sociology: international perspectives on mathematical and statistical model building. New York: Academic Press; 1975. pp 307–357.
41. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28:849–857.
42. Mehler EL, Solmajer T. Electrostatic effects in proteins: comparison of dielectric and charge models. Protein Eng 1991;4:903–910.
43. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 2002;23:128–137.
44. Holloway MK, Wai JM, Halgren TA, Fitzgerald PM, Vacca JP, Dorsey BD, Levin RB, Thompson WJ, Chen LJ, Desolms SJ. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. J Med Chem 1995;38:305–317.
45. Eriksson L, Jaworska J, Worth AP, Cronin MT, Mcdowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 2003;111:1361–1375.