

2D QSAR and similarity studies on cruzain inhibitors aimed at improving selectivity over cathepsin L

Renato F. Freitas,^a Tudor I. Oprea^b and Carlos A. Montanari^{a,*}

^a*Grupo de Química Medicinal de Produtos Naturais, NEQUIMED-PN, Departamento de Química e Física Molecular, Instituto de Química de São Carlos, Universidade de São Paulo, Av. Trabalhador Sancarlense, 400, 13560-970 São Carlos, SP, Brazil*

^b*Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA*

Received 15 July 2007; revised 24 September 2007; accepted 10 October 2007

Available online 22 October 2007

Abstract—Hologram quantitative structure–activity relationships (HQSAR) were applied to a data set of 41 cruzain inhibitors. The best HQSAR model ($Q^2 = 0.77$; $R^2 = 0.90$) employing Surflex-Sim, as training and test sets generator, was obtained using atoms, bonds, and connections as fragment distinctions and 4–7 as fragment size. This model was then used to predict the potencies of 12 test set compounds, giving satisfactory predictive R^2 value of 0.88. The contribution maps obtained from the best HQSAR model are in agreement with the biological activities of the study compounds. The *Trypanosoma cruzi* cruzain shares high similarity with the mammalian homolog cathepsin L. The selectivity toward cruzain was checked by a database of 123 compounds, which corresponds to the 41 cruzain inhibitors used in the HQSAR model development plus 82 cathepsin L inhibitors. We screened these compounds by ROCS (Rapid Overlay of Chemical Structures), a Gaussian-shape volume overlap filter that can rapidly identify shapes that match the query molecule. Remarkably, ROCS was able to rank the first 37 hits as being only cruzain inhibitors. In addition, the area under the curve (AUC) obtained with ROCS was 0.96, indicating that the method was very efficient to distinguishing between cruzain and cathepsin L inhibitors.

© 2007 Elsevier Ltd. All rights reserved.

1. Introduction

Parasitic protozoa infect hundreds of millions of people every year and are collectively some of the most important causes of human misery.^{1–3} There are many species of *Trypanosoma*, which belongs to the genus *Trypanosoma*. In sub-Saharan Africa, *Trypanosoma brucei* causes sleeping sickness, and in Latin America, *Trypanosoma cruzi* is the etiological agent of Chagas' disease, the most important parasitic disease of this region, with an estimated infection toll of ~18 million people. This is the leading cause of heart disease in Latin American countries.⁴

In spite of the impressive advances in the understanding of *T. cruzi* biology, the existing chemotherapy for Chagas' disease is not satisfactory in terms of its lack of effectiveness and due to the toxicity associated with long-term treatments. The drugs currently available

against *T. cruzi* were introduced more than 30 years ago: nifurtimox and benznidazole.² These compounds are active only in the acute stage of the disease. In addition, both drugs bring along important side effects such as anorexia, loss of weight, vomiting, nausea, etc.⁵

The drugs used to treat Chagas' disease are far from ideal, and new compounds are needed as antiparasitic drug candidates. Because of this, there have been increasing efforts to identify novel drug targets for *T. cruzi*. The major cysteine protease found in *T. cruzi*, cruzain, has received considerable attention as a validated target for drug development. Cruzain is essential for infection of host cells, replication, metabolism, and evasion of host defense mechanisms throughout the life cycle of *T. cruzi* parasite.⁶ Target validation studies have shown that irreversible inhibitors specifically bind to the cysteine protease targets thought to represent the major activity within the parasite. In the case of *T. cruzi*, the effect of inhibitors appears to be predominantly in blocking protease processing.⁷ These results indicate that the inhibition of cruzain will provide new basis for antichagasic drugs.⁸

Keywords: HQSAR; ROCS; Cruzain; Cathepsin L; Chagas' disease.

* Corresponding author. Tel.: +55 16 3373 9986; fax: +55 16 3373 9985; e-mail: montana@iqsc.usp.br

Quantitative structure–activity relationships (QSAR) play an important role in lead structure optimization.^{9,10} QSAR methods attempt to capture the relationship between structural features of molecules and their biological activities.¹¹ In the present study, we have employed the hologram QSAR (HQSAR) method to generate predictive 2D QSAR models for a set of 41 cruzain inhibitors.¹² HQSAR attempts to correlate molecular structure with biological activity for a series of compounds using molecular holograms constructed from counts of sub-structural molecular fragments.¹³ Since HQSAR requires no conformational analysis or structural alignment, it is straightforward to use and lends itself readily to the rapid screening of large numbers of compounds.¹³ In addition, due to the shared high sequence identities between cruzain and cathepsin L, we employed the ROCS (Rapid Overlay of Chemical Structures) method to perceive inhibitor similarities, based on their three-dimensional molecular shapes. ROCS approximates their volumes with Gaussian functions instead of hard spheres,¹⁴ thereby resulting in analytic and differentiable mathematical equations that allow for fast and robust global optimization of volume overlap by varying their relative orientations. A similarity function then measures the ‘shape distance’ between the pair of molecules at optimal overlap of volumes.

2. Molecular modeling

2.1. Data set

All calculations were carried out on a Linux-based PC workstation, using the software package SYBYL 7.3, and running on Red Hat Enterprise Linux 4.0.

The data set used for HQSAR analysis contains 41 inhibitors of cruzain, and was taken from the original work by Ellman et al.,¹² as retrieved from The World of Molecular Bioactivity Database (Wombat).¹⁵ The chemical structure and the biological data for the complete set of compounds are listed in Table 1. The data set was divided into training (29 compounds, 1–29, Table 1) and test (12 compounds, 30–41, Table 1) sets. The K_i values were converted to the corresponding pK_i ($-\log K_i$) before being used as dependent variables in the HQSAR analysis.

2.2. Data set preparation

The program OMEGA v.2.1 was used to convert all compounds to 3D multiconformer structures.¹⁶ The algorithm implemented in OMEGA dissects the molecules into fragments, reassembles and regenerates many possible combinations, and then submits each conformer to a simplified energy evaluation. Then, all conformers below an energy threshold are compared and those within a certain RMS distance are clustered into one single representation. Default parameters were used with the following exceptions: (1) *ewindow* (a value used to discard high-energy conformations), this parameter was set to 10.0 kcal mol⁻¹

(default = 25.0 kcal mol⁻¹); (2) *maxconfs* (sets the maximum number of conformations to be generated), this parameter was set to 100 (default = 400). This yielded a total of 3949 structures with an average of 96 conformations per unique compound. Another data set, with only one low-energy conformation (*maxconfs* was set to 1) of each molecule was generated to be used in the HQSAR analysis.

2.3. Calculation of shape similarity

A single low-energy conformation of the most potent inhibitor of the series (compound **23**) was used as the target structure for ROCS (Rapid Overlay of Chemical Structures).¹⁷ ROCS performs shape-based overlays of conformers of a candidate molecule to a query molecule in one or more conformations. The overlays can be performed very quickly based on a description of the molecules as atom-centered Gaussian functions. ROCS maximizes the rigid overlap of these Gaussian functions and thereby maximizes the shared volume between a query molecule and a single conformation of a database molecule. In default operation ROCS compares molecules based purely on their best shape overlap, quantitated by their shape Tanimoto. It was quickly found that adding to the shape Tanimoto the score for the appropriate overlap of groups with like properties (donor, acceptor, hydrophobes, cation, anion, and ring), the so-called color score, and then ranking on this summed score improved virtual screening performance considerably. In this mode, ROCS optimizes the molecular overlay to maximize both the shape overlap and the color overlap obtained by aligning groups with the same properties that are contained in the color force field file. This overlay is then subsequently scored using the sum of shape Tanimoto for the overlay and the color score (the so-called combo score). Customization or target-specific information can be incorporated by adding a term to the color force field file that rewards overlay of specific functional groups.¹⁷

The resulting database of cruzain inhibitors, generated by Omega, was initially screened and scored using the ROCS algorithm in order to generate and score 3D overlays of the library of molecules. To quantify the similarity of two molecules, combo score was used to compare two compounds, and can vary from 0.0 to 2.0, with 2.0 representing an exact match.

2.4. Comparative molecular field analysis (CoMFA)

The best 3D overlay of each molecule in the data set, obtained previously by ROCS, was used as input for CoMFA calculations.¹⁸ CoMFA steric and electrostatic fields were generated at each grid point, of a 3D grid box, using a sp³ carbon atom probe carrying a +1 net charge. The CoMFA grid spacing extends at least 4 Å beyond every molecule in all directions, and has a 2 Å spacing in the *x*, *y*, and *z* directions. The default value of 30 kcal mol⁻¹ was set as the maximum steric and electrostatic energy cutoff.

Table 1. Chemical structures and corresponding K_i values (nM) for a series of cruzain inhibitors

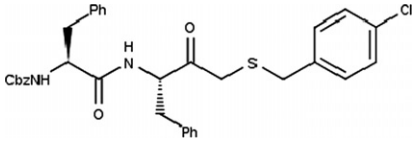
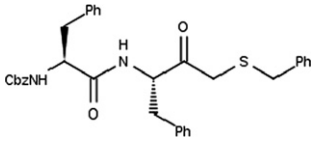
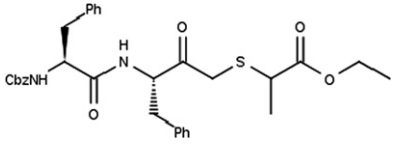
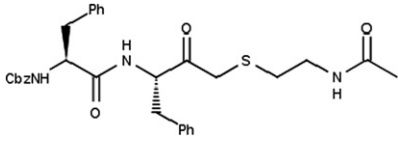
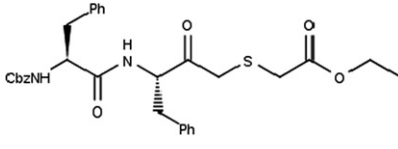
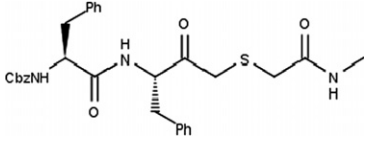
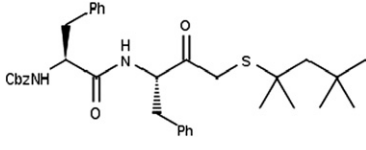
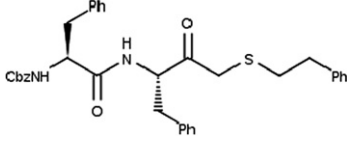
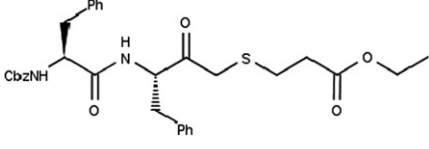
Compound	Structure	K_i
1		82.7 ± 6.4
2		44.0 ± 2.6
3		37.0 ± 6.6
4		34.3 ± 1.6
5		31.2 ± 1.5
6		15.0 ± 0.9
7		10.4 ± 0.6
8		4.2 ± 0.2
9		2.9 ± 0.3

Table 1 (continued)

Compound	Structure	K_i
10		2.0 ± 0.2
11		8.0 ± 0.8
12		7.8 ± 0.6
13		1.0 ± 0.1
14		59.7 ± 6.4
15		54.6 ± 6.3
16		69.6 ± 5.6
17		59.1 ± 6.5
18		130.7 ± 13.3

(continued on next page)

Table 1 (continued)

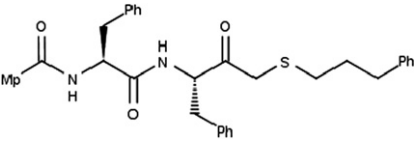
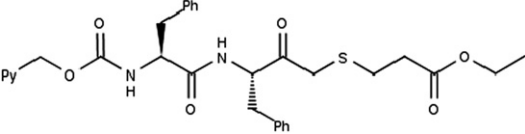
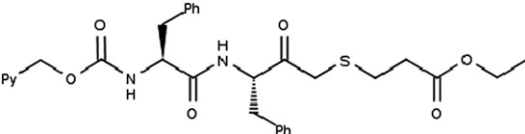
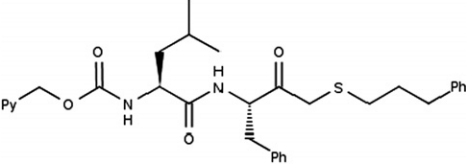
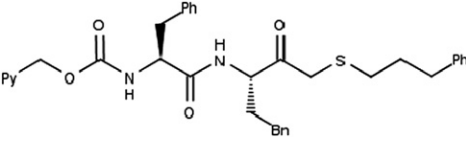
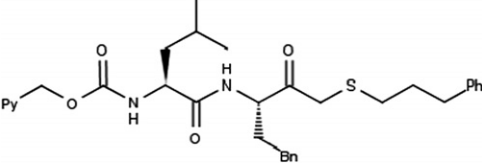
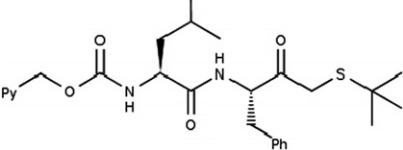
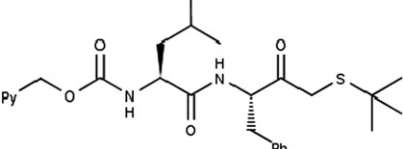
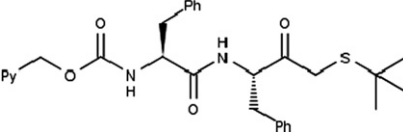
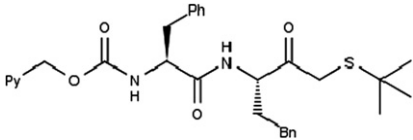
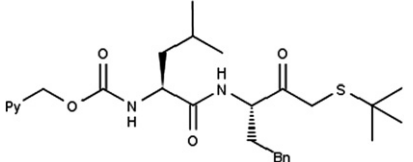
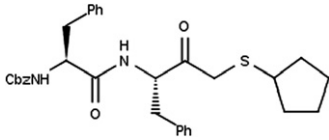
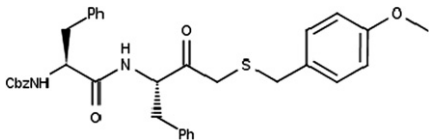
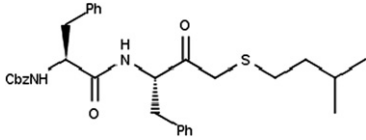
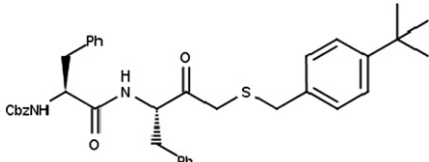
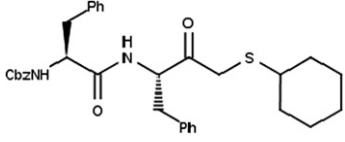
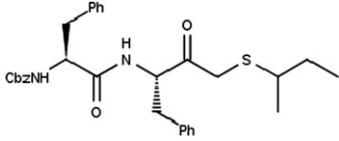
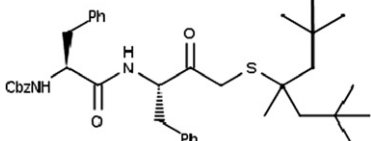
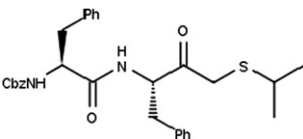
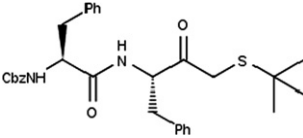
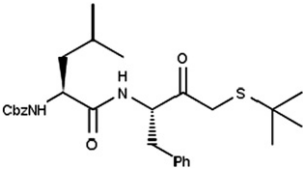
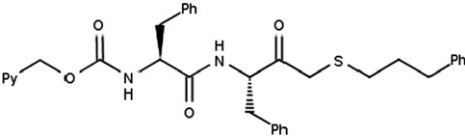
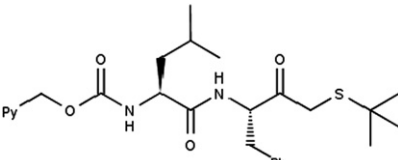
Compound	Structure	K_i
19		216.3 ± 18.9
20		30.4 ± 2.4
21		22.7 ± 2.0
22		5.4 ± 0.2
23		0.9 ± 0.1
24		1.1 ± 0.1
25		13.8 ± 0.8
26		14.1 ± 0.9
27		4.6 ± 0.2

Table 1 (continued)

Compound	Structure	K_i
28		1.4 ± 0.1
29		2.0 ± 0.1
30		34.3 ± 1.5
31		32.4 ± 2.3
32		30.1 ± 3.2
33		22.0 ± 1.6
34		7.5 ± 0.5
35		5.8 ± 0.5
36		5.0 ± 0.7

(continued on next page)

Table 1 (continued)

Compound	Structure	K_i
37		4.6 ± 0.4
38		2.5 ± 0.2
39		1.3 ± 0.1
40		5.5 ± 0.4
41		4.8 ± 0.2

CbzNH, carboxybenzylamide; Ph, phenyl; Mp, morpholine; Py, pyridine.

2.5. HCA

The chemometric technique of hierarchical cluster analysis (HCA) was the first approach used to select the training and test sets based on structural similarities and CoMFA steric and electrostatic fields. As the name suggests, HCA attempts to find groupings within a set of data. At the beginning of the analysis, each row of the data table is a cluster. The nearest pair of clusters is merged, then the next nearest, and so forth until there is only one cluster containing all the rows.¹⁹

The course of this hierarchical clustering process is depicted as a dendrogram.¹⁹ The analysis moves from the bottom of the dendrogram to the top, with each node at the bottom representing a row in the table, and the central branch at the top representing the entire table. Notice however that the left-to-right order of nodes at the bottom of the dendrogram is determined only by the requirement that branches may not cross, and is thus incidentally related to the row order in the original table. The lengths of the vertical lines in the dendrogram provide qualitative informa-

tion about the separation, or linkage distance, between various clusters. Clusters represented by long unbranched strands are strongly separated from other clusters.

This method was applied as implemented in SYBYL 7.3 (Tripos). The distance between the clusters was calculated by the complete-linkage method, that is, the distance between the most distant pair of data points in both clusters is taken into account.²⁰ The thereby obtained clustering dendrogram shows the most similar compounds clustered together at the lowest levels.

2.6. Surflex-Sim

The second approach used to select the training and test sets was Surflex-Sim.²¹ This method employs a molecular similarity function, coupled with quantitative pressure to minimize overall molecular volume, which forms an effective objective function for generating hypotheses of bioactive conformations of sets of small molecules binding to their cognate proteins.

The methodology of Surflex-Sim takes a small number of input molecules and yields a superposition of the molecules that optimizes an objective function that is composed of a molecular similarity component, a restraint against excessive volume that enforces parsimony, and a restraint against ligand self-clashing (inappropriate atomic overlap between non-bonded atoms).²¹

2.7. HQSAR analysis

Hologram QSAR (HQSAR), recently introduced by Tripos, Inc., is a novel QSAR method that eliminates the need for determination of 3D structure, putative binding conformations, and molecular alignment.¹³ In HQSAR, each molecule in the data set is divided into structural fragments that are then counted in the bins of a fixed length array to form a molecular hologram. The bin occupancies of the molecular hologram are structural descriptors (independent variables) encoding compositional and topological molecular information. A linear regression equation that correlates variation in structural information (as encoded in the hologram for each molecule) with variation in activity data is derived through PLS regression analysis to produce a QSAR model. Unlike other fragment-based methods, HQSAR encodes all possible molecular fragments (linear, branched, and overlapping). Optionally, additional 3D information, such as hybridization and chirality, may be encoded in the molecular holograms. Molecular holograms are generated in the same manner as hashed fingerprints where different unique fragments may populate the same holographic bin, allowing the use of a fixed length hologram fingerprint. This hashing procedure emphasizes the importance of patterns of fragment distribution within the hologram bins, which represents the nature of chemical structures more appropriately.

Since HQSAR models can be affected by a number of parameters concerning hologram generation: *hologram length*, *fragment size*, and *fragment distinction*, several combinations of these parameters were considered during the HQSAR modeling runs. Holograms were generated using the six fragment sizes. HQSAR analysis was performed by screening the 12 default series of holograms length values from 51 to 401 bins. The fragment patterns counts from the training set compounds were then related to the measured inhibition constant (K_i).

2.8. Statistical analysis

All models in this work were produced using partial least squares (PLS) analysis. The optimum number of principal components (PCs) corresponding to the smallest standard error of prediction (SEP) was determined by the Leave-One-Out (LOO) cross-validation procedure. By this procedure, each compound is systematically excluded once from the data set, after which its activity is predicted by a model derived from the remaining compounds. The predicted activities of the 'left out'

compounds allow the calculation of Q^2 and cross-validated standard error.

3. Results and discussion

An important factor for validating the quality of a QSAR model is the range of biological activity within a data set, which should be considered during the comparison of the quality of QSAR models across different data sets.¹³ The K_i values have a normal distribution, in the range of 0.9–216.3 nM, and the pK_i values used in this work span approximately two orders of magnitude, which indicate that the data are appropriate to derive a good QSAR model.

3.1. Training and test set selection

A crucial requisite in the development of a QSAR model is the division of the whole data set in training and test sets in order to maximize the diversity of the test set and to examine the predictive accuracy of the model when extrapolating outside the training set. Bearing this in mind, we evaluated two approaches, in terms of their ability to generate reliable training and test sets. The first one was the hierarchical cluster analysis (HCA), and the other was the molecular similarity module of Surflex, Surflex-Sim.

The results of HQSAR analysis for the 29 training set compounds, chosen according to HCA analysis, are reported in Table 2. The dendrogram with the highlighted compounds used as training set is displayed in Figure 1.

As it can be seen from Table 2, the best statistical result was obtained for model 7 ($Q^2 = 0.73$, $R^2 = 0.93$, with six components), using atoms (A), bonds (B), connections (C), and chirality (Ch) as fragment distinction. Incorporating hydrogens (H) and donor and acceptor (DA) containing fragments into molecular hologram did not provide any general improvement in the basic model as evaluated by R^2 and Q^2 . This result is in agreement with previous studies, which suggest that donor and acceptor fragment generation should not be used simultaneously with hydrogen atoms.²² This is due to the substantial increase in the number of fragments generated when both of these options are considered in the model construction. On the other hand, considering chirality in fragment distinction dramatically improved the model (compare 2 and 7). This outcome is expected since all compounds in our data set are chiral.

The same analysis was performed using the training set selected by Surflex-Sim (compounds 1–29, Table 1), and the fragment size default (4–7). As it can be seen from Table 3, model 14 gives the best statistical result ($Q^2 = 0.77$, $R^2 = 0.90$, with five components), using atoms (A), bonds (B), and connections (C) as fragment distinctions. The hologram that gives the lowest standard error has a length of 53. Inclusion of other fragment distinction for generation of molecular hologram fails to improve the quality of the model. Additionally, the insertion of hydrogen atoms as fragment distinctions

Table 2. Results of HQSAR analyses for various fragment distinctions on the key statistical parameters using fragment size default (4–7)

Model	Fragment distinction	Q^2	R^2	SEE	SEP	N	HL
1	A/B	0.64	0.79	0.297	0.389	3	71
2	A/B/C	0.63	0.81	0.280	0.387	2	353
3	A/B/H	0.61	0.84	0.279	0.435	6	59
4	A/B/Ch	0.59	0.78	0.302	0.415	3	353
5	A/B/D	0.62	0.79	0.301	0.401	4	53
6	A/B/C/H	0.63	0.84	0.268	0.412	5	151
7	A/B/C/Ch	0.73	0.93	0.182	0.360	6	53
8	A/B/C/D	0.62	0.76	0.317	0.400	3	151
9	A/B/H/Ch	0.63	0.89	0.225	0.424	6	257
10	A/B/H/D	0.64	0.82	0.290	0.406	5	71
11	A/B/Ch/D	0.63	0.85	0.256	0.406	4	53
12	A/B/C/H/Ch/D	0.65	0.83	0.279	0.401	5	53

Q^2 , cross-validated correlation coefficient; R^2 , non-cross-validated correlation coefficient; SEE, non-cross-validated standard error; SEP, cross-validated standard error; HL, hologram length; N , optimum number of components. Fragment distinction: A, atoms; B, bonds; C, connections; H, hydrogens; Ch, chirality; DA, donor and acceptor.

The training set used in these analyses was obtained by HCA analysis.

decreases the quality of the statistical parameters, since the models that present this parameter show the worst Q^2 values (models 15, 18, 21, 22 and 24, respectively).

Selecting an appropriate training set is of crucial importance in QSAR studies. An essential characteristic of a

training set is that the molecules must be orthogonal (i.e. dissimilar from each other). If the molecules in the training set are highly similar, there is little additional discriminatory information added for each additional molecule.²³ Surflex-Sim was able to provide the most orthogonal and diverse set of molecules to be included

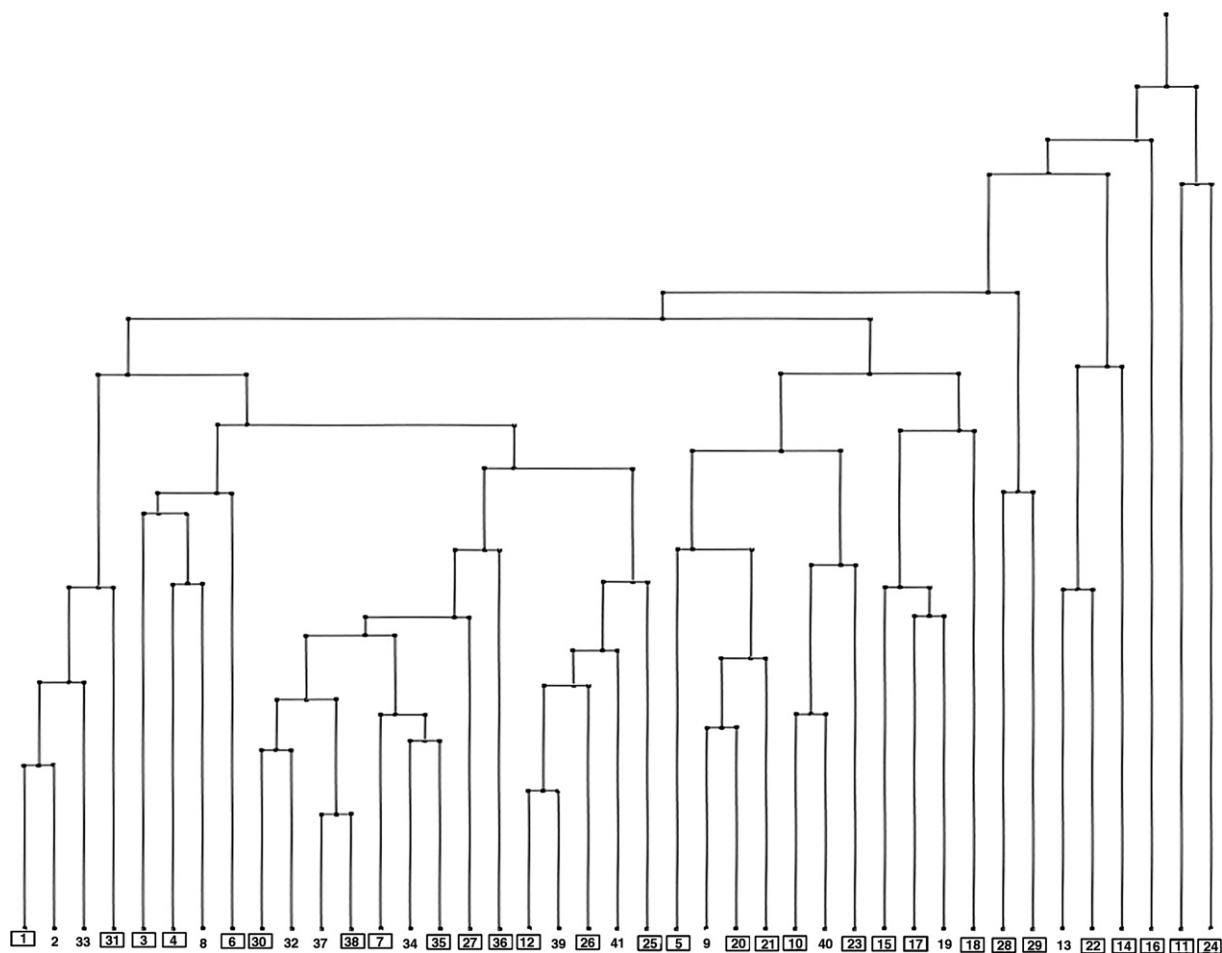
**Figure 1.** Clustering of the 41 cruzain inhibitors. Training set compounds are marked with squares.

Table 3. Results of HQSAR analyses for various fragment distinctions on the key statistical parameters using fragment size default (4–7)

Model	Fragment distinction	Q^2	R^2	SEE	SEP	N	HL
13	A/B	0.73	0.91	0.233	0.402	6	53
14	A/B/C	0.77	0.90	0.235	0.363	5	53
15	A/B/H	0.65	0.82	0.325	0.455	6	83
16	A/B/Ch	0.71	0.90	0.239	0.417	6	53
17	A/B/D	0.70	0.91	0.238	0.422	6	151
18	A/B/C/H	0.63	0.81	0.338	0.467	6	59
19	A/B/C/Ch	0.75	0.90	0.244	0.381	5	53
20	A/B/C/D	0.66	0.90	0.249	0.451	6	53
21	A/B/H/Ch	0.66	0.79	0.346	0.440	5	59
22	A/B/H/D	0.66	0.83	0.312	0.443	5	97
23	A/B/Ch/D	0.72	0.91	0.231	0.405	6	151
24	A/B/C/H/Ch/D	0.62	0.82	0.323	0.466	5	353

The training set used in these analyses was obtained by Surfex-Sim.

Table 4. Influence of fragment sizes on the statistical parameters using the best fragment distinction (atoms, bonds, and connections)

Model	Fragment size	Q^2	R^2	SEE	SEP	N	HL
25	2–5	0.67	0.88	0.258	0.425	4	97
26	3–6	0.75	0.91	0.227	0.389	6	61
27	4–7	0.77	0.90	0.235	0.363	5	53
28	5–8	0.68	0.89	0.246	0.426	5	151
29	6–9	0.68	0.90	0.237	0.425	5	353
30	7–10	0.61	0.79	0.332	0.451	3	199

in the training set. Moreover, Surfex-Sim yields a more consistent training set than HCA analysis, as it can be seen in Table 3, where six Q^2 values are ≥ 0.70 . Based on this result, the training set selected for the continuation of the work was that generated by Surfex-Sim.

3.2. Model development

Six HQSAR models were generated to evaluate the effect of different fragment sizes on the statistical parameters (Table 4), using the best combination of fragment distinctions (A/B/C, model 14) found previously. Fragment size controls the minimum and maximum length of fragments to be included in the hologram fingerprint. As mentioned previously, molecular holograms are produced by the generation of all linear and branched fragments between M and N atoms in size. The parameters M and N can be changed to include smaller or larger fragments in the holograms.

A close analysis of Table 4 reveals that the variation of the fragment size led to a decrease in the quality of the models generated as measured by statistical parameters. Once again, the fragment size default (4–7) provides the best statistical (model 27) results when compared with other fragment sizes (Table 4).

In the previous analysis employing HCA as training set generator, the introduction of chirality improved the statistical parameters (model 7). On the other hand, when the Surfex-Sim was used to generate the training set, models 14 and 19 showed similar statistics, but the best model 14 did not include the chirality as fragment distinction. This prompted us to check if the inclusion

of chirality would result in a better model. The effect of different fragment sizes in the statistical parameters was evaluated using atoms (A), bonds (B), connections (C), and chirality (Ch) as fragment distinctions (model 19). The introduction of chirality failed to improve the quality of the model, since the best one ($Q^2 = 0.75$, $R^2 = 0.91$, with six components) obtained with a fragment size of 3–6 was slightly inferior to that of the final model 27.

3.3. Validation of the HQSAR model

One of the most important characteristics of QSAR models is their predictive power. The latter can be defined as the ability of a model to predict accurately the target property (e.g., biological activity) of compounds

Table 5. Experimental and predicted affinities (pK_i) with residual values for the test set compounds

Test set	Experimental	Predicted	Residual
30	7.46	8.22	−0.76
31	7.49	7.31	0.18
32	7.52	8.20	−0.68
33	7.66	7.32	0.34
34	8.12	8.67	−0.55
35	8.24	8.38	−0.14
36	8.30	7.56	0.74
37	8.34	8.39	−0.05
38	8.60	8.61	−0.01
39	8.89	8.62	0.28
40	8.26	8.17	0.09
41	8.32	8.08	0.24

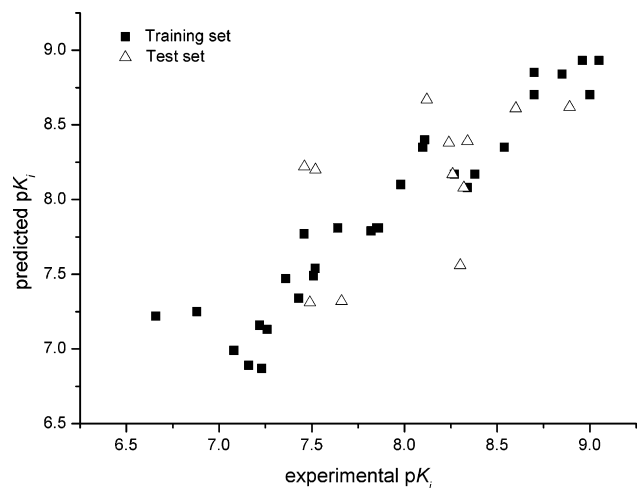


Figure 2. Plot of predicted versus experimental values of pK_i for the training and test sets.

that were not used for model development.²⁴ To estimate the true predictive power of our QSAR model we compared the predicted and observed activities of an external test set of 12 compounds that were not used to generate the model. The results are presented in Table 5, and the graphic result for the experimental versus predicted affinities of both training and test sets is illustrated in Figure 2.

The results show that a satisfactory linear correlation ($R^2 = 0.88$, $SD = 0.285$) was obtained between the experimental and predicted activities of the compounds in the training and test sets. The predicted values fall close to the experimental pK_i values, deviating no more than 0.4 log units. The only exceptions are compounds **30**, **32**, **34**, and **36**, for which the predicted values deviate more (from 0.55 to 0.76 log units).

The results of a HQSAR analysis can be graphically displayed as a color-coded structure diagram in which the color of each atom reflects the contribution of that atom to the molecule's overall activity. The colors at the red end of the spectrum (red, red orange, and orange) reflect poor (or negative) contributions, while colors at the green end (yellow, green blue, and green) reflect favorable (positive) contributions. Atoms with intermediate contributions are colored white.²² The individual atomic contributions of the compound **23** (the most potent inhibitor of the data set) are displayed in Figure 3.

As it can be seen in Figure 3, the side chain homophenylalanine was found to be strongly correlated to the biological activity of compound **23**. This is supported by the fact that when homophenylalanine is replaced by phenylalanine there is a decrease in potency (compare **23**, **24**, **28**, and **29** with **40**, **22**, **27**, and **41**, respectively). This result is in agreement with the work of Ellman et al.,¹² who found that the inhibitors presenting the fragment homophenylalanine were the most potent of their series.

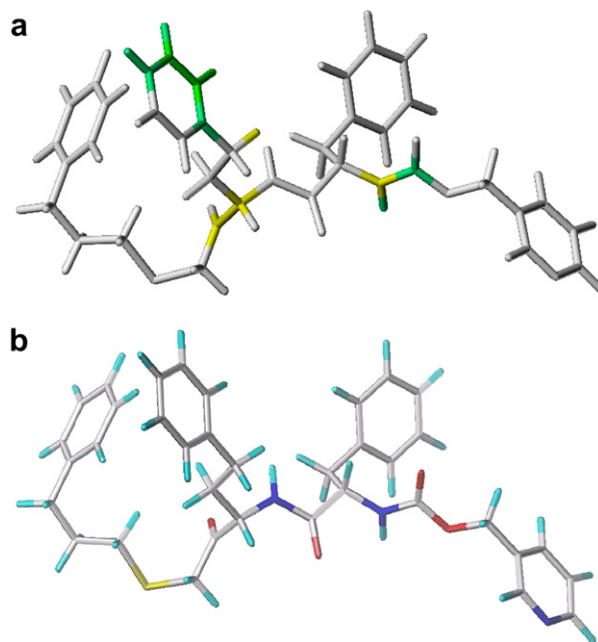


Figure 3. (a) Individual atomic contributions for the activity of the most potent cruzain inhibitor of the series; (b) structure of the compound **23** colored by atom type.

3.4. Shape similarity

Cruzain, a cathepsin L-like member of the C1 (papain) family of endopeptidases, is the major cysteine protease of *T. cruzi*.²⁵ It shares a high degree of sequence identity with cathepsin L (35.7%), according to the Align program.²⁶ Therefore, potency and selectivity need to be accounted for in the design of these cysteine protease inhibitors. Bearing this in mind, inhibitors known to cause no problems with host proteases have been designed. For instance, an irreversible inhibitor of cruzain, which is well-tolerated and orally bioavailable, is currently in the late stages of preclinical development for Chagas' disease.²⁷

In order to identify similarities between cruzain and cathepsin L, ROCS calculations were carried out to obtain shape Tanimoto values for the series of cruzain inhibitors, herein named *actives*, and a data set of 82 cathepsin L inhibitors, now called *inactives*. This set of compounds was also retrieved from the Wombat database.¹⁵ They were prepared in the same way as that for cruzain inhibitors (see Section 2), yielding 7107 3D structures with an average of 86 conformations per unique compound. So, all molecules in the final database (123 compounds, *actives* + *inactives*) were screened by ROCS, which performs overlays of conformers of a candidate molecule to a query molecule (in this case, compound **23** was used as a query).

ROCS requires that suitable conformers be generated for each ligand. To assess if the conformational ensemble, herein generated by Omega, was reasonable to be used in ROCS calculations, we evaluated the generated conformations for two ligands (compounds **124** and

125, see text below) complexed with the cruzain enzyme, whose coordinates were determined by high-resolution (≤ 2.0 Å) X-ray crystallography.²⁸ All conformations of these compounds were overlaid on the corresponding unmodified X-ray structure by using ROCS. Only non-hydrogen atoms were matched. **Figure 4a** shows the superposition of the X-ray structure of compound **124** (yellow) and the corresponding best overlaid conformation (cored by atom type) found by ROCS. The shape Tanimoto is 0.894 and the corresponding RMSD value is 0.591 Å. The superposition of the X-ray structure of compound **125** (yellow) and the corresponding best overlaid conformation (cored by atom type) found by ROCS are displayed in **Figure 4b**. The shape Tanimoto is 0.877 and the corresponding RMSD value is 0.708 Å. These results suggest that Omega settings were able to generate bioactive conformations for these two compounds, and the RMSD values are close to the value (RMSD = 0.500 Å) proposed to a conformation to reproduce an experimental structure.²⁹ It is worth to be considered that the conformations were generated by Omega without any bias, since the input conforma-

tions were based on the 2D structure retrieved from Wombat.

Table 6 shows the results of the ROCS analysis, ranking the compounds by their combo score (shape + color score). Molecules that are ‘actives’ are shown in bold.

It can be seen that the active compounds cluster almost exclusively toward the top of the ranking by combo score, with a mean shape Tanimoto for actives and inactive compounds equal to 0.750 and 0.530, respectively (**Fig. 5a**). The mean color score for the ‘actives’ compounds is 0.650, whereas the inactives have a mean color score of 0.269 (**Fig. 5b**).

Additionally, it should be noted that the first 37 hits identified by ROCS are only active compounds, which corresponds to 90% of all cruzain inhibitors present in the database. Also, it can be observed in **Figure 6** that these compounds align very well to the reference compound.

3.5. Evaluation of ROCS method

To evaluate the performance of the ROCS method, the receiver operating characteristic (ROC) curve was used.³⁰ A ROC curve consists of reporting the evolutions of the true positive rate (TPR, *Y* axis) plotted versus the false positive rate (FPR, *X* axis) at all possible detection thresholds. For ideal distributions, where active compounds are completely separated from the inactives, the curve climbs vertically to the upper-left corner (TPR = 1, FPR = 0) and then joins the upper-right corner horizontally (TPR = 1, FPR = 1). Since the relative positions of ROC plots give an insight into the respective accuracies, the *area under the curve* (AUC)³⁰ is a practical way of measuring the overall performance of the test. If the AUC is close to 0.5 (random test), the test is said to be poor; the highest possible AUC is 1, corresponding to an ideal case. In general, the greater the AUC, the more effective the method is in discriminating active from inactive compounds. **Figure 7** shows the ROC curve using the background of cathepsin L inhibitors, which quantify model selectivity. The first observation to be made from this figure is that the ROC curve remains above the diagonal representing a random distribution. This curve overlaps the ideal graph along the left side from a true positive rate of 0.0–0.902 for low false positive rates. This result indicates that ROCS was at least capable of discriminating between cruzain and non-cruzain inhibitors. This affirmation is supported by the fact that the three most selective cruzain inhibitors are found in the top 37 hits, both in decreasing order of shape Tanimoto score and selectivity against cathepsin L (compounds **24**, **29**, and **26**, respectively).

The area under the curve for the ROC curve shown in **Figure 7** is 0.963. This means that ROCS is actually able to give a higher score to a randomly selected active compound than to a randomly selected inactive in more than nine trials out of 10. Experience shows that an AUC of 0.900 or greater is an indication of a useful measure.³¹

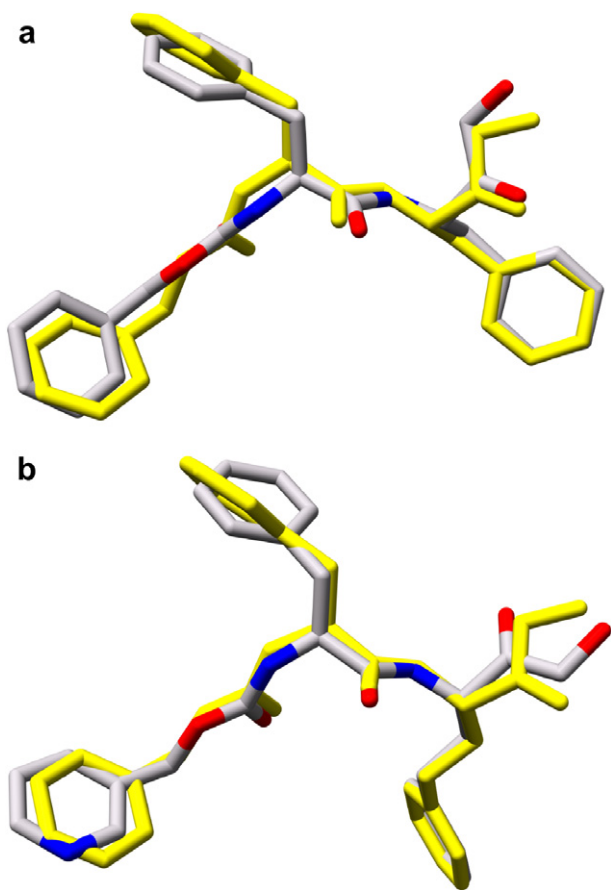


Figure 4. (a) Superposition of the X-ray structure of compound **124** (yellow) and the corresponding best overlaid conformation (colored by atom type) found by ROCS. The Tanimoto shape index is 0.894 and the corresponding RMSD value is 0.591 Å. (b) Superposition of the X-ray structure of compound **125** (yellow) and the corresponding best overlaid conformation (colored by atom type) found by ROCS. The shape Tanimoto is 0.877 and the corresponding RMSD value is 0.708 Å.

Table 6. ROCS results, sorted by combo score

Compound	Rank	Shape ^a	Color ^b	Combo ^c	Compound	Rank	Shape	Color	Combo
23	1	1.000	1.000	2.000	78	64	0.498	0.300	0.798
40	2	0.870	0.869	1.739	82	65	0.517	0.276	0.793
10	3	0.870	0.779	1.648	84	66	0.542	0.249	0.791
24	4	0.908	0.735	1.643	81	67	0.516	0.273	0.790
21	5	0.844	0.765	1.608	105	68	0.374	0.407	0.781
28	6	0.769	0.810	1.579	106	69	0.382	0.397	0.780
20	7	0.843	0.695	1.539	69	70	0.544	0.233	0.778
9	8	0.843	0.677	1.520	114	71	0.457	0.313	0.769
27	9	0.740	0.754	1.493	11	72	0.510	0.257	0.766
19	10	0.780	0.698	1.478	72	73	0.542	0.222	0.763
32	11	0.776	0.666	1.442	79	74	0.461	0.296	0.757
8	12	0.778	0.646	1.424	101	75	0.498	0.255	0.753
38	13	0.739	0.674	1.412	88	76	0.501	0.250	0.751
5	14	0.796	0.615	1.410	83	77	0.520	0.227	0.747
35	15	0.725	0.677	1.401	67	78	0.409	0.335	0.744
37	16	0.740	0.656	1.395	90	79	0.505	0.233	0.738
30	17	0.768	0.618	1.386	113	80	0.458	0.271	0.729
7	18	0.692	0.688	1.380	50	81	0.548	0.181	0.729
4	19	0.737	0.625	1.362	14	82	0.525	0.204	0.729
15	20	0.757	0.595	1.351	80	83	0.391	0.337	0.728
34	21	0.723	0.628	1.351	92	84	0.508	0.220	0.728
6	22	0.731	0.618	1.349	85	85	0.535	0.192	0.726
36	23	0.644	0.680	1.324	87	86	0.509	0.212	0.720
29	24	0.706	0.617	1.323	44	87	0.537	0.180	0.717
2	25	0.727	0.595	1.322	99	88	0.463	0.235	0.697
3	26	0.704	0.614	1.318	119	89	0.374	0.319	0.693
33	27	0.737	0.571	1.308	51	90	0.429	0.264	0.693
1	28	0.731	0.569	1.300	57	91	0.471	0.219	0.690
31	29	0.718	0.571	1.288	56	92	0.495	0.193	0.688
15	30	0.658	0.599	1.256	49	93	0.501	0.186	0.687
18	31	0.689	0.564	1.253	118	94	0.419	0.266	0.685
41	32	0.659	0.582	1.241	121	95	0.357	0.322	0.679
25	33	0.658	0.552	1.210	117	96	0.357	0.321	0.678
26	34	0.658	0.541	1.199	43	97	0.541	0.127	0.668
39	35	0.658	0.533	1.191	108	98	0.452	0.216	0.667
12	36	0.658	0.524	1.182	58	99	0.387	0.280	0.667
16	37	0.678	0.443	1.121	65	100	0.491	0.174	0.665
66	38	0.591	0.419	1.009	115	101	0.448	0.213	0.661
95	39	0.637	0.361	0.998	48	102	0.443	0.216	0.659
63	40	0.558	0.405	0.963	116	103	0.343	0.314	0.657
68	41	0.592	0.347	0.938	107	104	0.454	0.198	0.652
13	42	0.456	0.469	0.925	109	105	0.389	0.262	0.650
22	43	0.454	0.466	0.920	120	106	0.363	0.278	0.641
62	44	0.494	0.416	0.910	123	107	0.393	0.246	0.639
104	45	0.402	0.506	0.907	53	108	0.425	0.206	0.631
61	46	0.516	0.380	0.896	111	109	0.382	0.248	0.630
96	47	0.559	0.325	0.884	100	110	0.454	0.174	0.629
71	48	0.540	0.338	0.878	112	111	0.379	0.249	0.628
97	49	0.540	0.327	0.867	77	112	0.334	0.289	0.623
94	50	0.518	0.344	0.861	42	113	0.393	0.200	0.593
86	51	0.558	0.287	0.846	46	114	0.376	0.207	0.583
102	52	0.518	0.327	0.845	122	115	0.406	0.175	0.581
75	53	0.532	0.310	0.842	55	116	0.359	0.220	0.579
93	54	0.514	0.317	0.831	70	117	0.343	0.230	0.573
64	55	0.509	0.318	0.827	47	118	0.357	0.203	0.561
91	56	0.539	0.285	0.824	45	119	0.403	0.148	0.551
103	57	0.546	0.278	0.823	52	120	0.343	0.202	0.545
73	58	0.497	0.323	0.820	59	121	0.365	0.178	0.543
98	59	0.571	0.246	0.816	54	122	0.321	0.154	0.475
89	60	0.566	0.249	0.815	60	123	0.306	0.142	0.447
74	61	0.497	0.310	0.808					
110	62	0.421	0.387	0.807					
76	63	0.496	0.310	0.806					

^a Shape Tanimoto score.^b Color score.^c Combo score.

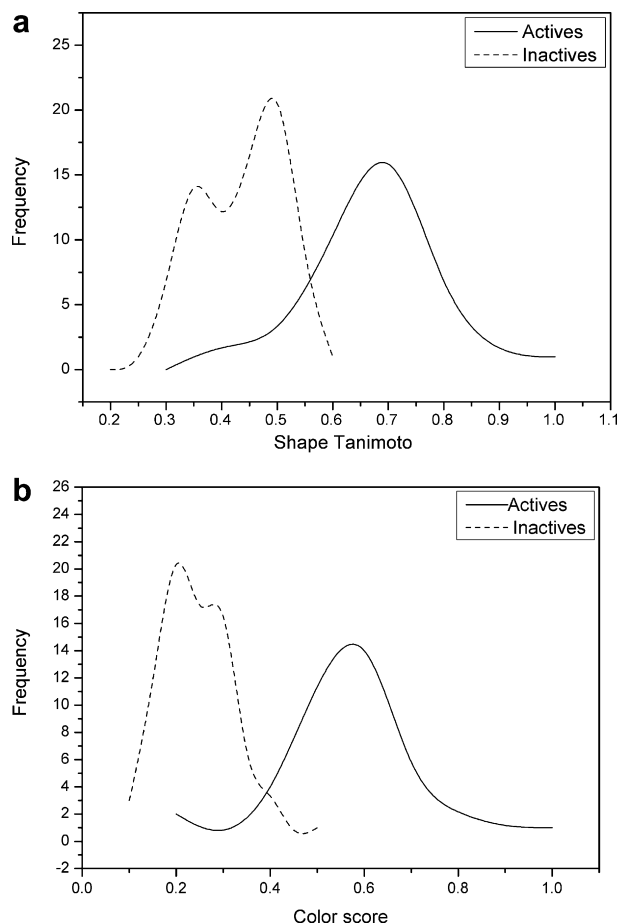


Figure 5. A plot showing the distribution of the shape Tanimoto (a) and color score (b) for the actives and inactives, respectively.

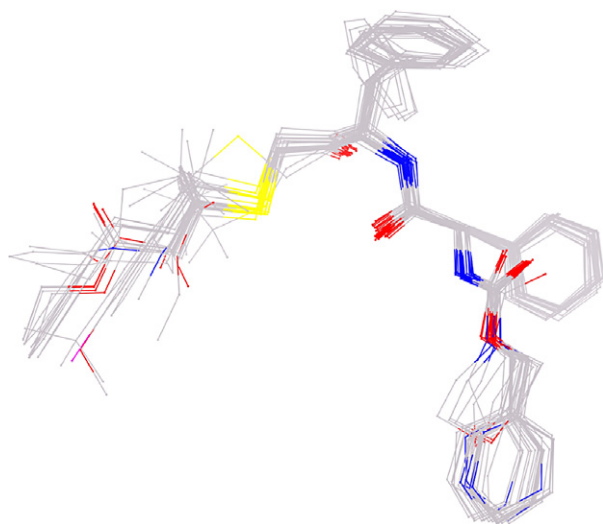


Figure 6. Alignment of the top 37 hits to the reference compound. Atoms and bonds are represented in the wire model. Hydrogen atoms were removed for clarity.

The performance of the ROCS method was evaluated using a data set of six cruzain inhibitors (compounds

124–129, Table 7) as an external validation set.²⁸ This data set was prepared as previously described and screened by ROCS using compound **23** as a target. Table 7 shows the ranking, 2D structure, shape Tanimoto, color, and combo score for this data set.

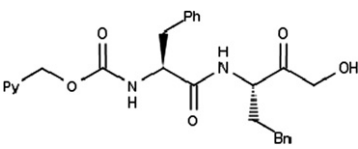
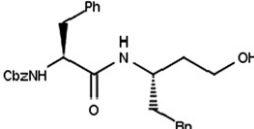
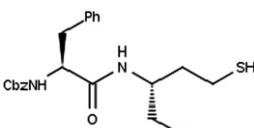
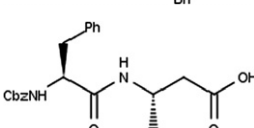
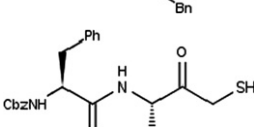
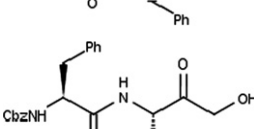
The mean shape Tanimoto and color score for the external set were equal to 0.733 and 0.656, respectively, which indicates that the six compounds are clustered with the active compounds. This is true, since after screening all database including the external set, ROCS was able to retrieve again in the first 43 positions not only the same 37 active compounds, but also all the six compounds along with these molecules.

The results achieved with ROCS seem excellent, in view of the fact that the predicted pK_i values by HQSAR analysis are far from the experimental values (residual > 1.30 log units) for two of three compounds (**124** and **125**). Only compound **126** has a predicted pK_i value near the experimental value. The other three compounds do not have an appropriate affinity measurement (values are reported as higher than), but their predicted pK_i values appear to be reasonable considering that all values are predicted to be higher than 9.00. The good predicted pK_i value for the compound **126** could be explained by the fact that this compound is one of the two in the external validation set which presents a sulfur atom in its basic scaffold structure, whereas in the other compounds this atom is substituted by an oxygen atom. Because this fragment is present in all compounds used to construct the training and test sets, it could be pointed as being essential for the HQSAR model to predict a reliable biological activity of external compounds.

4. Conclusion

A predictive HQSAR model was developed for a series of cruzain inhibitors with statistical significance ($Q^2 = 0.77$, $R^2 = 0.90$). The predictive power of the model was further validated with a test set, showing satisfactory predictive R^2 value of 0.88. It is worth noting that the selection of the training and test sets should be treated with caution. In this work we have shown that distinct approaches used to generate training and test sets provided different HQSAR models. It was demonstrated that Surflex-Sim was able to provide the most orthogonal and diverse set of molecules to be included in the training set. In addition, Surflex-Sim yields a more consistent training set than HCA analysis. The individual atomic contributions of the compound **23** are in agreement with the experimental work, where the inhibitors presenting the homophenylalanine fragment were the most potent of the series. Selectivity needs to be accounted in the development of cruzain inhibitors because cruzain presents a high identity degree with mammalian cysteine proteases (cathepsin L). To evaluate this question, we have employed ROCS to screen a database presenting both cruzain and cathepsin L inhibitors. The use of ROCS led to the discrimination between cruzain and cathepsin L inhibitors, and showed an area under the curve (AUC) of 0.963. The perfor-

Table 7. ROCS results, sorted by combo score

Compound	2D structure	Rank	Shape	Color	Combo	pK_i^a	pK_i^b
125		6	0.787	0.818	1.605	7.19	8.50
127		18	0.758	0.642	1.400	>5.00	9.20
129		19	0.758	0.635	1.392	>5.30	9.22
128		24	0.739	0.623	1.362	>5.00	9.26
126		30	0.704	0.640	1.344	8.36	8.33
124		39	0.652	0.578	1.229	6.84	8.31

The pK_i values are reported, too.

^a Experimental pK_i .

^b Predicted pK_i by HQSAR.

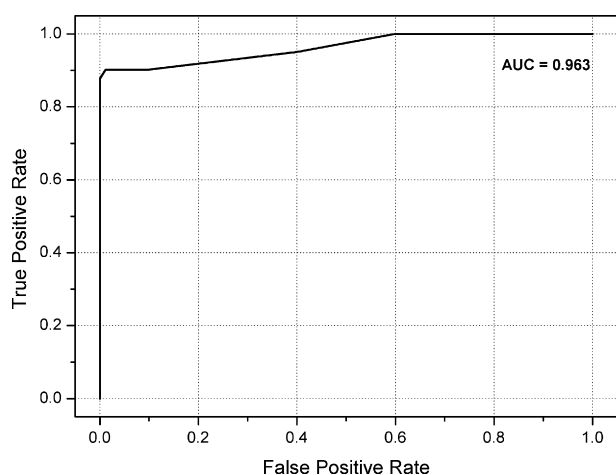


Figure 7. ROC graph provided by ROCS method. The AUC corresponding to this curve is displayed on the upper-right corner in bold black.

mance of ROCS method was further validated by using an external set of six cruzain inhibitors. ROCS was able to retrieve all the six compounds used as external set to-

gether with the previous 37 (out of 41) cruzain inhibitors in the first 43 positions.

Acknowledgments

We are indebted to CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), for financial support and all provided scholarships. Financial Support is also acknowledged from the New Mexico Tobacco Settlement Fund (T.I.O.).

References and notes

1. Nwaka, S.; Hudson, A. *Nat. Rev. Drug Discov.* **2006**, *5*, 941.
2. Croft, S. L.; Barrett, M. P.; Urbina, J. A. *Trends Parasitol.* **2005**, *21*, 508.
3. Barrett, M. P.; Burchmore, R. J. S.; Stich, A.; Lazzari, J. O.; Frasch, A. C.; Cazzulo, J. J.; Krishna, S. *The Lancet* **2003**, *362*, 1469.
4. Maya, J. D.; Cassels, B. K.; Vasques, P. I.; Ferreira, J.; Faúndez, M.; Galanti, N.; Ferreira, A.; Morrelo, A. *Comp. Biochem. Physiol. A: Physiol.* **2007**, *146*, 601.

5. Linares, G. E. G.; Ravaschino, E. L.; Rodriguez, J. B. *Cur. Med. Chem.* **2006**, *13*, 335.
6. Urbina, J. A. *Expert Opin. Ther. Patents* **2003**, *13*, 661.
7. McKerrow, J. H.; Engel, J. C.; Caffrey, C. R. *J. Exp. Med.* **1999**, *7*, 639.
8. Polticelli, F.; Zaini, G.; Bolli, A.; Antonini, G.; Gradoni, L.; Ascenzi, P. *Biochemistry* **2005**, *44*, 2781.
9. Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.* **2003**, *2*, 192.
10. Kubinyi, H. *Drug Discov. Today* **1997**, *2*, 538.
11. Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. *J. Med. Chem.* **2004**, *47*, 5541.
12. Huang, L.; Lee, A.; Ellman, J. A. *J. Med. Chem.* **2002**, *45*, 676.
13. Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669.
14. Talafous, J. *Drug Discov. Today* **2005**, *10*, 737.
15. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T., Ed.; Wiley-VHC: New York, 2004; pp 223–239.
16. Greenwood, B. J.; Gottfries, J. J. *Mol. Graph. Model* **2003**, *21*, 449.
17. Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. *J. Med. Chem.* **2005**, *48*, 1489.
18. Cramer, R. D., III; Patterson, D. E.; Bunce, J. J. *Am. Chem. Soc.* **1988**, *110*, 5959.
19. HCA™ Manual, SYBYL 7.3, Tripos, St. Louis, MO.
20. Stoyanov, K.; Walmsley, A. D. Classification and Pattern Recognition. In *Practical Guide to Chemometrics*; Gemperline, P. J., Ed.; CRC-Taylor and Francis: Oxford, 2006; pp 347–351.
21. Jain, A. N. *J. Med. Chem.* **2004**, *47*, 947.
22. HQSAR™ Manual, SYBYL 7.3, Tripos, St. Louis, MO.
23. Surflex Manual: Docking and Similarity, BioPharmics LLC, San Mateo, CA.
24. Golbraikh, A.; Tropsha, A. *Mol. Diversity* **2000**, *5*, 231.
25. Lecaille, F.; Authié, E.; Moreau, T.; Serveau, C.; Gauthier, F.; Lalmanach, G. *Eur. J. Biochem.* **2001**, *268*, 2733.
26. <http://www.ebi.ac.uk/emboss/align/index.html>. Query carried on 06/20/2007.
27. Abdulla, M. H.; Lim, K. C.; Sajid, M.; McKerrow, J. H.; Caffrey, C. R. *PLoS Med.* **2007**, *4*, 130.
28. Huang, L.; Brinen, L. S.; Ellman, J. A. *Bioorg. Med. Chem.* **2003**, *11*, 21.
29. Boström, J.; Greenwood, J. R.; Gottfries, J. J. *Mol. Graphics Modell.* **2003**, *21*, 449.
30. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. *J. Med. Chem.* **2005**, *48*, 2534.
31. Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. *J. Chem. Inf. Model* **2005**, *45*, 673.