BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 1 2010, pages 53–60
doi:10.1093/bioinformatics/btp623

Structural bioinformatics

Automatic clustering of docking poses in virtual screening process using self-organizing map

Guillaume Bouvier, Nathalie Evrard-Todeschi, Jean-Pierre Girault and Gildas Bertho*

Laboratoire de Chimie et de Biochimie Pharmacologiques et Toxicologiques, Unité Mixte de Recherche 8601, Centre National de la Recherche Scientifique (CNRS), Université Paris Descartes, 45 rue des Saints-Pères, 75006 Paris, France

Received on June 2, 2009; revised on September 22, 2009; accepted on October 25, 2009

Advance Access publication November 12, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Scoring functions provided by the docking software are still a major limiting factor in virtual screening (VS) process to classify compounds. Score analysis of the docking is not able to find out all active compounds. This is due to a bad estimation of the ligand binding energies. Making the assumption that active compounds should have specific contacts with their target to display activity, it would be possible to discriminate active compounds from inactive ones with careful analysis of interatomic contacts between the molecule and the target. However, compounds clustering is very tedious due to the large number of contacts extracted from the different conformations proposed by docking experiments.

Results: Structural analysis of docked structures is processed in three steps: (i) a Kohonen self-organizing map (SOM) training phase using drug–protein contact descriptors followed by (ii) an unsupervised cluster analysis and (iii) a Newick file generation for results visualization as a tree. The docking poses are then analysed and classified quickly and automatically by AuPosSOM (Automatic analysis of Poses using SOM). AuPosSOM can be integrated into strategies for VS currently employed. We demonstrate that it is possible to discriminate active compounds from inactive ones using only mean protein contacts' footprints calculation from the multiple conformations given by the docking software. Chemical structure of the compound and key binding residues information are not necessary to find out active molecules. Thus, contact–activity relationship can be employed as a new VS process.

Availability: AuPosSOM is available at http://www.aupossom.com.

Contact: contact@aupossom.com; gildas.bertho@parisdescartes.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Virtual screening (VS) is usually described as a cascade of filter approaches to narrow down a set of compounds to be tested for biological activity against the intended drug target. Starting with a fast evaluation of the drug-likeness of compounds, VS is often followed by ligand-based approaches and/or structure-based approaches if the target structure is available (Muegge, 2008).

In structure-based approaches, dockings of molecules onto 3D structure of receptor are made. The improvement of docking algorithms allows to generate ligand conformations similar to crystallographically determined protein/ligand complex structures in most cases. However, scoring functions were less successful at distinguishing the crystallographic conformation from the set of docked poses (Warren *et al.*, 2006). Improvement of scoring functions remains a significant challenge in the application of structure-based VS. However, de-selection of inappropriate compounds is more easily achieved than selection by current scoring schemes. This negative selection is sufficient to reduce the initial large compound database to a shortlist of preferred candidates, but the relative ranking of elements belonging to these shortlist cannot be done using previous criteria. It is a rather common practise to subject a 'reasonably' small number of preselected candidates to visual inspection (Kitchen *et al.*, 2004). This last step is very tedious when the filters used are not restrictive enough. It is dependent on a human interpretation, and user's experience in the field. Such a visual inspection can only be applied to a data sample of perhaps 300–500 compounds (Alvarez and Shoichet, 2005), but it remains very subjective and time consuming. When the number of molecules is very important (more than 500 compounds, where the filters used are not restrictive enough) visual inspection of the complexes is out of question. Accordingly, a reliable filtering has to be performed at the previous steps.

This article demonstrates that it is possible to cluster potentially active molecules using mathematical vectors describing interactions between compounds and targets. Data are clustered using Kohonen self-organizing map (SOM; Kohonen, 2001). This artificial neural network can map high-dimensional data onto a low-dimensional grid such that similar data elements are placed close together. Of particular interest is the method developed by Renner *et al.* (2008). This method is based on a consensus docking scoring using interaction fingerprints comparison and triplet ranking. A SOM analysis of the distribution of docking poses is proposed but not goes any further towards an unsupervised SOM cluster analysis.

The hierarchical clustering of the SOM proposed here bypasses the problem of alternate binding mode identification.

Furthermore, the current article describes the possibility to overcome the problem of scoring function using a statistical analysis of different poses of docking of a same compound onto the target. Iterative learning of the data may neutralize the statistical noise

*To whom correspondence should be addressed.

produced by this analysis. We implement this method in a software called AuPosSOM (Automatic analysis of Poses using SOM). We use for this new software a recent method to perform unsupervised SOM cluster analysis, determine cluster confidence and process result visualization as a tree (Samsonova *et al.*, 2006).

Thus, molecules are clustered according to their binding mode to the macromolecular target. Active compounds and potentially active compounds are then clustered in the same group as their way to explore the macromolecular surface is homologous. AuPosSOM is a flexible tool that can be integrated into strategies for VS currently employed.

2 APPROACH

The goal of VS is to select a small number of potential binders to the receptor of interest from a large source library. Database docking is an approach to solve the problem of identifying compounds in a database of small organic compounds that display favourable interactions (hydrogen bonds, hydrophobic contacts, electrostatic interactions) to the target binding site. A docking program is composed of two elements : an algorithm that explore the conformational space of the small molecule within the binding site and a scoring function that estimate the relative binding energy. This function calculates a crude measure of binding affinity. An ideal scoring function must be able to find out the most favourable docking solutions (called poses) for active molecules in accordance with experimental results (e.g. X-ray structures) and should be able to separate active compounds from inactive ones.

Classical approaches use scoring function in order to discriminate between active compounds and inactive ones. It has been demonstrated that combining multiple scoring functions (consensus scoring) improves the enrichment of true positives (Yang *et al.*, 2005). However, recent validation studies have highlighted the poor performance of currently used scoring functions in estimating binding affinity and hence in ranking large datasets of docked ligands (Waszkowycz, 2008). Energy functions need to be fairly 'soft' so that ligands are not heavily penalized for small errors in the binding geometry. Thus, non-active compounds can obtain a good rank with scoring function but can be excluded using position descriptors.

Another way to classify active molecules and inactive ones is to compare poses between all molecules of the screened database. The process we have developed is shown in Figure 1. The ensemble of active compounds—K compounds (Known)—is defined as $\mathbb{K}$. The ensemble of database compounds with unknown activity—U compounds—is defined as $\mathbb{U}$. For each molecule, an interleaved mean vector is calculated. This vector results from two mean vectors describing hydrogen bonds and hydrophobic contacts involved in protein–molecule interactions. If n docking poses were calculated for each molecule, these vectors describe mean interactions for these n poses. Cardinality of each vector is equal to the number of amino acids in the protein. A subensemble of vectors, composed of all vectors describing known ligands and a random set of i ($i \in [0, a]$, where $a = \text{card}(\mathbb{U})$) vectors from the database, is used for training the SOM. SOM for the Kohonen neural network and TreeSOM algorithm from Samsonova *et al.*, (2006) have been implemented in Python and included in AuPosSOM software. The resulting SOM is then used for clustering all the molecules according to their way to interact with the protein. Mapping of data on a trained SOM is called calibration. Each vector is assigned to the

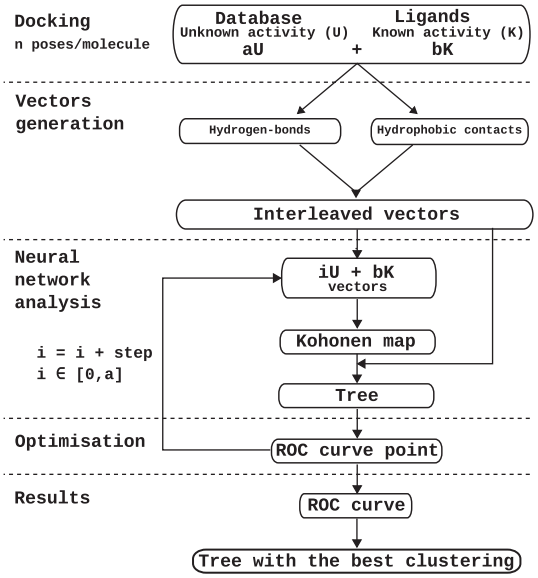

Fig. 1. Diagram illustrating the overall workflow described in this article. From docking results mean vectors describing hydrogen bonds and hydrophobic contacts involved in protein/molecule interactions are computed. Each two vectors are interleaved in order to obtained one vector per molecule in the database (a vector is the mean of the n poses of the docking process). Incremental subensemble are then used to train random SOM. For each trained SOM, all vectors are calibrated on it and then clustered according to the SOM. Each clustering can be visualized as a tree. The best subensemble of vectors to train the SOM is found by plotting a ROC curve.

SOM's node that is most similar to it, so that some nodes may get many data elements, and others none at all. Cluster discovery procedure define clusters as groups of nodes with short distances between them and long distances to the other nodes for a given distance threshold. Python TreeSOM implementation automatically find the best threshold corresponding to the best clustering. For the best clustering, distances between vectors in the same group are minimized and distances to the data in other groups are maximized. AuPosSOM program is able to represent SOM as a tree in Newick file format. At each threshold in the clustering series one or more clusters are split into several subclusters that are represented as a node in the tree. The resultant tree allows a classification of the molecules docked according to their similarity of contacts. Molecules which are clustered with active ones (K molecules) should be probably active too. The best group containing the maximum number of known ligands (K) and minimum number of molecule with unknown activities (U) is found by scanning all the tree, from leaves to root, to calculate for each group sensitivity and specificity. For a given ensemble $\mathbb{E}$, sensitivity (Se) is defined as the number of K compounds which belong to $\mathbb{E}$ divided by the total number of truly active compounds (number of true positive and false negative). If $\mathbb{K}$ is the ensemble of truly active compounds:

$$\text{Se} = \frac{\text{card}(\mathbb{K} \cap \mathbb{E})}{\text{card}(\mathbb{K})} \quad (1)$$

Specificity (Sp) is the number of truly inactive compounds which does not belong to the ensemble $\mathbb{E}$ divided by the total number of molecules with unknown activities. If $\mathbb{U}$ is the ensemble of

molecules with unknown activities:

$$\mathrm{Sp} = \frac{\mathrm{card}(\mathbb{U} \setminus \mathbb{E})}{\mathrm{card}(\mathbb{U})} = 1 - \frac{\mathrm{card}(\mathbb{U} \cap \mathbb{E})}{\mathrm{card}(\mathbb{U})} \quad (2)$$

If Se is plotted against $1-\mathrm{Sp}$ a plot that curves above the diagonal rising from the origin to the upper right corner is obtained. This plot is known as a 'Receiver Operating Characteristic' curve, or ROC curve (Triballeau *et al.*, 2005). On such a graph, a random classification of the compounds would be represented by a diagonal rising from the origin to the upper right corner, whereas a test capable of detecting the correct signal would have an ROC plot that curves above that diagonal. This curve can be used for finding a consensus group with maximum sensitivity and specificity. The best group is the one with the smallest γ, where γ is the norm of the vector between $(0,1)$ point and an ROC curve point :

$$\gamma = \sqrt{(1-\mathrm{Sp})^2 + (1-\mathrm{Se})^2} \quad (3)$$

γ is equal to the Euclidean distance between the characteristics of an ideal test with sensitivity and specificity equal to 1 and the given test studied.

While i value is less than $\mathrm{card}(\mathbb{U})$, i value is incremented by *step* (Fig. 1). The subensemble of vectors obtained is used for training a new random SOM. The resulting SOM is then used for clustering all vectors and the group with minimal γ value is found. The Se and Sp values calculated for the best group for each iteration are then plotted in a new ROC curve (ROC plot in Fig. 1). A γ coefficient can be calculated for each point of the ROC diagram obtained. The best subensemble of vectors used for training the SOM is defined as the subensemble corresponding to the ROC point with minimal γ value.

3 METHODS

3.1 Dataset

AuPosSOM program was tested with three different models: (i) HIV-1 protease, (ii) HIV-1 reverse transcriptase and (iii) human thrombin. For each model, known ligands were used in order to test the ability of AuPosSOM program to discriminate known inhibitors of an enzyme from randomly chosen molecules (decoys). Decoys came from a sub-database of 1108 molecules built from the French National Chemical Library (http://chimiotheque-nationale.enscm.fr). Known ligands with lowest possible pairwise ligand similarity (as measured by pairwise Tanimoto index values) were extracted from the binding database (Liu *et al.*, 2007). For HIV-1 protease, HIV-1 reverse transcriptase and human thrombin, 22, 14 and 20 known ligands were used, respectively. For HIV-1 reverse transcriptase model, non-nucleoside reverse transcriptase inhibitors (NNRTI) were used. For each model ligands lists were built with the respective known ligands and the 1108 decoys.

3.2 Docking procedure

The AutoDock 4.0 package was used for docking simulation (Huey *et al.*, 2007; Morris *et al.*, 1998). Each docking experiment was performed 20 times, yielding 20 docked conformations. Parameters for the docking are as follows: population size of 150; random starting position and conformation; maximal mutation of 2Å in translation and 50° in rotations; elitism of 1; mutation rate of 0.02 and crossover rate of 0.8; and local search rate of 0.06. Simulations were performed with a maximum of 2.5 million energy evaluations and a maximum of 27 000 generations. Cubic grid, centred on the active site, with 108 points spaced by 0.375Å was generated using AutoGrid 4 for each structure. VS with docking was performed on a Linux Cluster Platform (INRA MIGALE bioinformatics platform; http://migale.jouy.inra.fr) which contains 160 CPU (80 Intel Quad Core 5340 2.33GHz, 40 Intel Dual Core 5140 2.33GHz and 40 Xeon 3.2GHz) on 40 computing nodes. PDB files used for docking were 2bpw (Chang *et al.*, 2007), 1uwb (Wang *et al.*, 2005) and 1afe (Schafferhans and Klebe, 2001) for HIV-1 protease, HIV-1 reverse transcriptase and thrombin, respectively. The ligand and crystallographic waters were removed. Polar hydrogens were added. Receptors were prepared with Python AutoDockTools scripts. Polar hydrogens were added to small molecules and energy minimization was computed using PRODRG (Schüttelkopf and van Aalten, 2004).

3.3 Contacts analysis

3.3.1 Vectors generation AuPosSOM program processes contacts analysis (hydrogen bonds and hydrophobic contacts) for each docked molecule. Contacts are represented as interleaved vectors. The cardinality of each vector is the double of the number of amino acids in the protein targeted. The hydrogen bonds are computed over all possible donor–acceptor pairs, such that one atom belongs to the protein and the other to the ligand. An interaction is considered as an hydrogen bond when the D-H...A distance is 1.85 ± 0.65Å and the D-H...A angle is $180 \pm 80°$. Interactions are considered as hydrophobic ones when two hydrophobic atoms are closer than 3.9Å. Bio.PDB module (Hamelryck and Manderick, 2003) from Biopython project (http://biopython.org) was used to measure angle and distances from coordinates contained in PDB files of the docked structures obtained. The AuPosSOM performance was tested on 20 000 structures (1000 molecules with 20 poses per molecule). On one Linux PC (quadricore Intel 2.66 GHz, 1GB RAM) this calculation takes about 58 min, or on average 0.174s per structure.

3.3.2 Neural network analysis With vectors described in contacts analysis section, we used a Euclidean SOM, trained in two phases with the following parameters: map size 5×4, exponential decrease of learning rate and radius; phase 1: starting learning rate 0.2, starting radius 6, 1000 iterations; phase 2: starting learning rate 0.02, starting radius 3, 10 000 iterations. Such training length ensured that cluster tree topology remained unchanged with any additional training. Consensus trees were made with different map sizes in order to test cluster confidence. A map size of 20 neurons gives confident clusters. As shown by Samsonova *et al.*, (2006) smaller SOMs tend to yield more confident clusters. We further test that the tree resulting from a randomly initialized map is not dependent on the order of the data elements. Furthermore, the fact to tag molecules as known or unknown does not influence in anyway SOM analysis, but these data are just used for plotting ROC curves.

AuPosSOM distribution includes new python implementations of Kohonen SOM and unsupervised clustering.

3.3.3 k-means clustering In order to compare SOM algorithm with another clustering method, a Python implementation of k-means calculation has been made and included in the AuPosSOM software. The algorithm proceeds by alternating between two steps: (i) assignment step: assign each vector to the cluster with the closest mean (centroid), (ii) update step: calculate the new means (centroids) of the vectors in the cluster. To initialize the process, k initial centroids are randomly chosen from the dataset. In addition, the UPGMA (Unweighted Pair Group Method with Arithmetic mean) algorithm has been written to transform the pairwise distance matrix between clusters into a rooted tree for the results visualization.

4 RESULTS

The ROC plot provides not only a way to fine-tune the parameters of VS processes but also allows a direct comparison of different VS workflows (Triballeau *et al.*, 2005): the closer a curve comes to the upper-left corner of the graph, the better the screening process is.

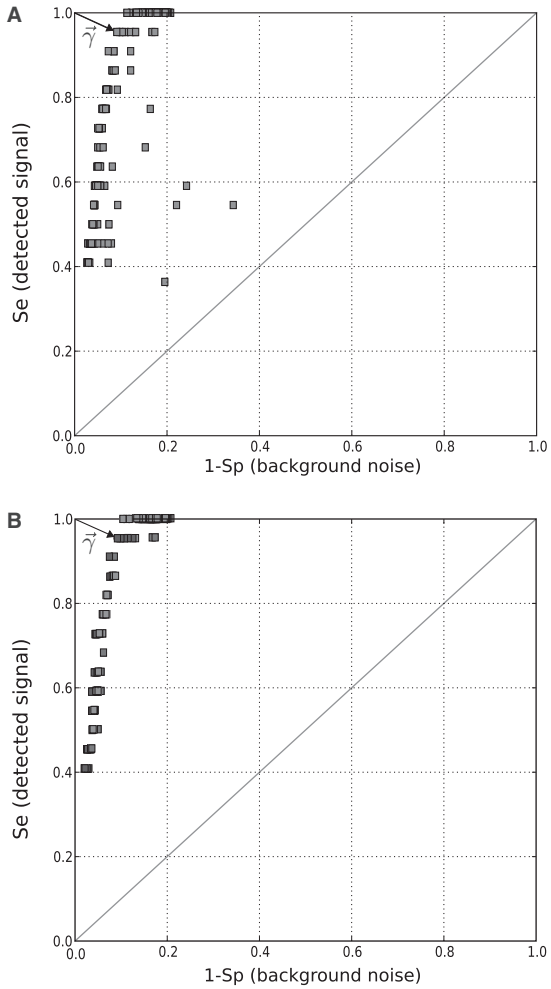

Fig. 2. ROC plots for HIV-1 protease model. For each plot, $\vec{\gamma}$ with minimal norm is represented. (**A**) ROC plot obtained using AuPosSOM software with various subensembles of vectors to train the Kohonen map. The number of elements for each subensemble used is <25% of the total number of molecules in the database. (**B**) Effect of enrichment of the dataset ($\mathbb{L}$ ensemble) used for training the SOM. The number of elements for each subensemble used is >25% of the total number of molecules in the database.

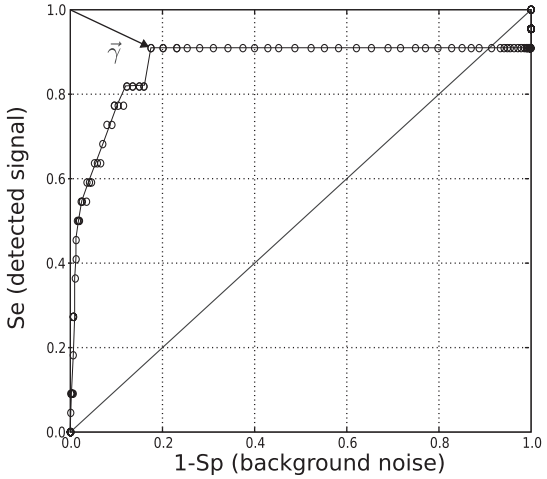

Fig. 3. ROC plot for HIV-1 protease model using the AutoDock 4.0 scoring function with various score thresholds.

Find out minimal γ factor described in Equation (3) is a way to find out the closest point to the upper-left corner of the graph. Consequently, this minimal γ is a quantification of the performance of the test. The smaller the γ factor is, the better the test is at isolating signal from background noise. Another way to measure the overall performance of the computer test is to calculate the area under the curve (AUC). This calculation is possible for simple tests (e.g. using score-filter), but is much more difficult for ROC plot resulting from neural network analysis. ROC plot obtained from neural network analysis of poses and from different scoring threshold are given in Figure 2 and 3, respectively.

The first observation to be made from these two figures is that the ROC curves remain above the diagonal representing a random distribution. The ROC curve derived from neural network analysis is plotted according to the workflow described in Sections 2 and 3. Points of the curve result from different training sets used for training the map. Training sets are subensemble of vectors defined as $\mathbb{L}$ containing all vectors of the ensemble $\mathbb{K}$ and a random set of vectors of the ensemble $\mathbb{U}$ (each element of the different ensembles is a vector describing corresponding compound). Effect of enrichment of the training set $\mathbb{L}$ is shown in Figure 2A and B. Figure 2A reports the overall result of the screening process with different sets of training vectors; for each set the number of vectors is <25% of the total ensemble of vectors. Effect of enrichment of the number of training vectors is shown in Figure 2B. The increase of the number of vectors in the training set yields less dispersion in the ROC plot. When the number of training vectors is too small the effect due to the addition of new randomly chosen vectors is important leading to outliers in Figure 2A. All true positive compounds are detected (Se = 1) with a small proportion of false positives ($1-\mathrm{Sp} = 0.10$). The point with the smallest γ value is not the most sensible point but a point with a sensitivity of 0.95 and a background noise ($1-\mathrm{Sp}$ value) of 0.09 (Table 1).

Score analysis of the docking is not able to find out all active compounds. This is due to a bad estimation of the ligand binding energies for two ligands. This problem is overcome by the contact analysis approach developed. However, the point with the smallest γ value shows a sensitivity of 0.91 and a background noise of 0.18 (Table 1). The ROC plot obtained with AutoDock 4.0 scoring function with 22 known ligands is in accordance with the result obtained by Chang *et al.* (2007) with 11 known ligands.

ROC analysis for two other models (Fig. 4)—HIV-1 reverse transcriptase and human thrombin—shows that contact clustering gives more sensible and more specific results than score analysis provided by AutoDock 4.0 software. However, results obtained with human thrombin show no significant amelioration. For reverse transcriptase, score analysis is not able to discriminate active compounds. This is either due to a bad estimation of protein ligand binding energy or incorrect poses conformation. Surprisingly, contact clustering analysis is able to find out 10 active compounds among 14 (71%). Therefore, AutoDock 4.0 is able to generate good conformations for HIV-1 reverse transcriptase inhibitors.

For all models tested so far, AuPosSOM analysis yields more sensible and more specific results than score analysis provided by

Table 1. Sensitivity [Se, Equation (1)] and specificity [Sp, Equation (2)] with smallest γ value [Equation (3)] for each model

		SOM clustering			*k*-means clustering			Score analysis		
Model	card($\mathbb{K}$)	Se	Sp	γ	Se	Sp	γ	Se	Sp	γ
HIV-1 protease (2bpw)	22	0.95	0.91	0.10	0.82	0.94	0.19	0.91	0.82	0.20
HIV-1 reverse transcriptase (1uwb)	14	0.57	0.90	0.44	0.71	0.70	0.42	0.50	0.63	0.62
Human thrombin (1afe)	20	0.85	0.85	0.21	0.80	0.90	0.22	0.80	0.85	0.25

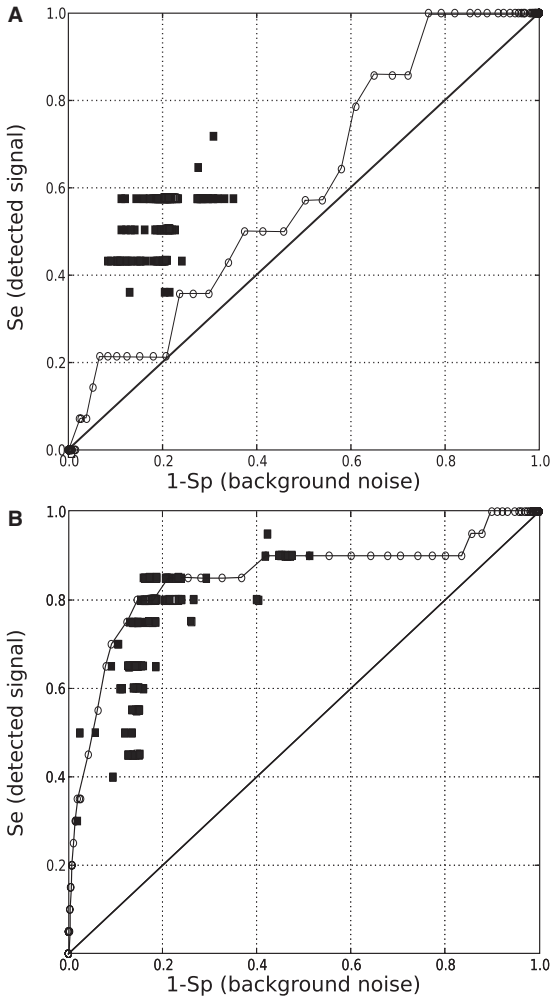

Fig. 4. ROC plots obtained from score analysis using AutoDock 4.0 scoring function (open circle) and using contact analysis using AuPosSOM software (filled points). (**A**) ROC plots obtained from HIV-1 reverse transcriptase screening. (**B**) ROC plots obtained from human thrombin screening.

AutoDock 4.0 package. The resultant γ values are consequently lower (Table 1) for contact analysis processes.

Unrooted tree representation of contact footprints clustering for protease screening experiment with the biggest sensitivity (Se = 1) and the smallest background noise ($1-\mathrm{Sp}=0.1$) is given in Figure 5. $\mathbb{E}$ is the ensemble with the smallest γ value [see Equation (3)]. A leave is defined as a subensemble of elements (compounds of the database screened) belonging to $\mathbb{K}$ or $\mathbb{U}$. A tree is composed of three different leaves: (i) leaves composed of U elements exclusively, notated as $\{U\}_{U\in\mathbb{U}}$, (ii) leaves composed of K elements exclusively, notated as $\{K\}_{K\in\mathbb{K}}$ and (iii) leaves composed of K and U elements, notated as $\{K\}_{K\in\mathbb{K}}\cup\{U\}_{U\in\mathbb{U}}$. Branch length is proportional to divergence of contacts made by molecules belonging to the different leaves. The first observation to be made from this tree is that the ensemble $\mathbb{E}$ is isolated from the rest of the tree ($\mathbb{U}\setminus\mathbb{E}$ ensemble). This means that the Kohonen neural network found that compounds which belong to $\mathbb{E}$ interact with their target with distinct contact patterns. Thus, active compounds can be isolated from inactive ones using only mean protein contact footprints. This result demonstrates the existence of a contact–activity relationship (CAR). Furthermore, this CAR can be employed as a new VS process.

A comparison between the SOM clustering and a baseline clustering method such as *k*-means has been done for the three models. Due to a random initialization for both SOM and *k*-means clusterings, different results can be obtained with the same set of parameters. Furthermore, the k value or the map size influences the results. ROC calculations were performed 10 times, with k variations ($k=\{2,\ldots,30\}$) for *k*-means method or map size variations ($(X,Y)=\{(i,j)\in\mathbb{N}^2 : 2\leq i\leq j\leq 6\}$) for SOM. The performance of the SOM network is higher for the HIV-1 protease model (Fig. 6): SOM clustering method yields more sensible results (Se = 1.00) than a simpler method such as *k*-means (Se = 0.82). SOM clustering method leads to less dispersion for sensitivity and specificity than *k*-means algorithm. According to γ values (Table 1), no significant amelioration is brought by the SOM clustering for HIV-1 reverse transcriptase and human thrombin in comparison with the *k*-means method. However, the ROC plots obtained for HIV-1 reverse transcriptase and human thrombin (See Figs S1 and S2 in Supplementary Material) show less dispersion for SOM clustering.

To understand these results, dispersion of the input data has been studied. Shannon's entropy has been calculated (Supplementary Material) for all the vectors for the three models and results reported in a histogram (see Fig. S3 in Supplementary Material). The dispersion calculated for protease data ($\bar{\rho}=0.84$; $\sigma=0.04$, defined in Supplementary Material) is more important than the dispersion of the data for reverse transcriptase ($\bar{\rho}=0.93$; $\sigma=0.02$) and thrombin ($\bar{\rho}=0.93$; $\sigma=0.01$). $\bar{\rho}$ value gives the mean dispersion within the vectors. σ value gives the mean dispersion between vectors. As the level of dispersion in the data increases, the performance advantage of the SOM network relative to the *k*-means clustering method increases to a dominant level.

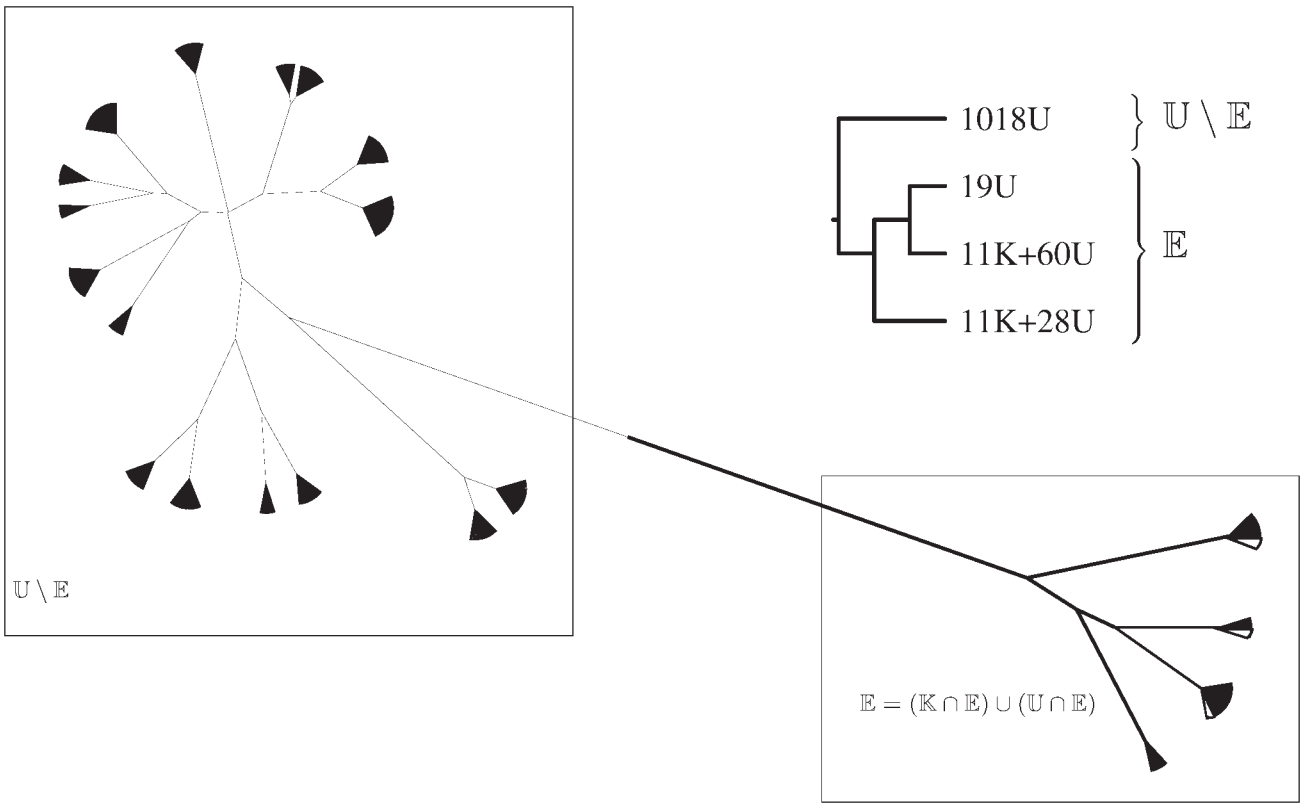

Fig. 5. Tree representation of contact footprints clustering for protease screening experiment. The tips stand for the different molecules of the database screened. Branch length is proportional to divergence of contacts made by ligands belonging to the different clusters (leaves). Molecules with known activities (K compounds) are represented with white tips. The ensemble $\mathbb{E}$ with the smallest γ value according to the ROC plot is shown in bold. A simplified tree (inset) shows the number of K compounds and U compounds in homogeneous leaves ensembles.

In addition, important structural informations are found in this analysis of atomic interactions. Contact clustering allows identification of key residues implicated in ligand binding. These residues can be mapped onto the target structure. For example, with the HIV-1 protease (Fig. 7), a clear relationship is found between footprints of the ensemble $\mathbb{E}$ and amino acids thought to be essential to the interaction of antiproteases. Major point mutations associated with resistance to protease inhibitors (Johnson *et al.*, 2005) and located in the binding site correlates with the contact pattern characteristic of compounds found to be active by the automatic contact analysis. Only one amino acid, namely Ala-28, is found to make essential interaction without point mutation at this position. Interestingly, Ala-28 is highly conserved and important for the structuration of HIV-1 protease. This amino acid can be exploited to design resistance-evading drugs (Wang and Kollman, 2001).

AuPosSOM analysis can be performed with docking results given by other docking programs. This strategy has been employed for a VS process to search new β-TrCP inhibitors with a docking protocol described by Evrard-Todeschi *et al.* (2008) using the Surflex-Dock 2.0 program (Jain, 2003) from Sybyl 7.3 molecular modelling program package (Tripos Inc., France) (data not shown).

Docking on HIV-1 protease has also been performed using Surflex-Dock 2.0 with the same dataset presented in Section 3. Score analysis with Surflex-Dock 2.0 scoring function gives more sensible results than thus obtained with AutoDock 4.0 program. However, AuPosSOM analysis on docking poses proposed by Surflex–Dock yields more specific results than Surflex–Dock scoring (see docking procedure and Fig. S4 in the Supplementary Material).

5 DISCUSSION

Incorrect estimation of ligand–protein free energy of binding leads to the apparition of false negative and even false positive compounds in score-based VS process. Chang *et al.* (2008) observed that incorrect lowest energy conformations will be found in ∼1% of docking experiments and the correct conformation will be found in 25–100% of the experiments. Thus, a simple procedure that chooses the conformation of best energy from a set of multiple docking experiments will yield an incorrect conformation. With the process exposed here, all poses per molecule are processed since mean vectors for all poses are calculated. Thus, distribution of the digits in the resulting vectors gives information about poses cluster size, and conformation frequencies: amino acids which are contacted with high frequency for the 20 poses calculated for a ligand are represented in the resulting vector by a higher figure than the other ones. As the frequency of finding a given conformation is providing information on the energy landscape of binding, and that a high frequency is a measure of favourable entropy in the binding process (Chang *et al.*, 2008), our statistical

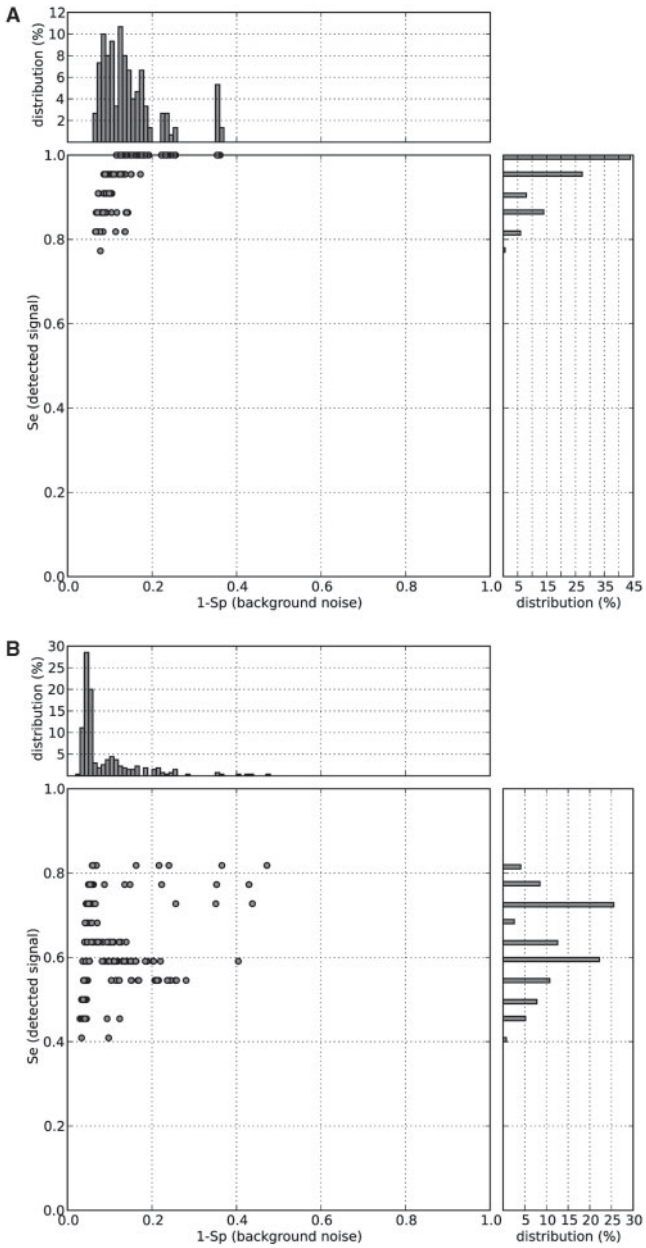

Fig. 6. ROC plots obtained from (**A**) Kohonen SOM and (**B**) *k*-means clustering methods for different Kohonen map sizes and different *k*-values for the HIV-1 protease dataset. For each map size and each *k*-values, calculations are repeated 10 times. The histograms show points distribution. The values of the most represented point are: Se = 1, Sp = 0.87and Se = 0.73, Sp = 0.95 for SOM and *k*-means, respectively.

analysis of poses takes into account the vibrational entropy factor.

TreeSOM algorithm developed by Samsonova *et al.* (2006) is an unsupervised method for cluster analysis and confidence testing for SOMs. This process allows to find reliable clusters. Interestingly, learning samples influence final clustering. Iteration over learning samples coupled with ROC plot, allows to find out the most appropriate dataset to solve classification problem between active

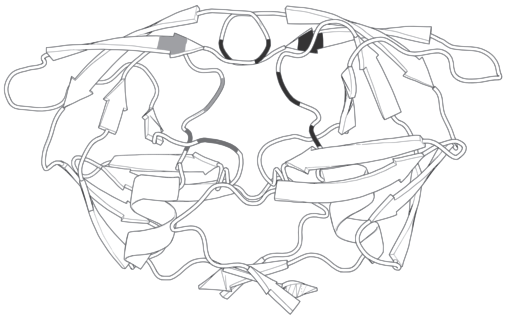

Fig. 7. HIV-1 protease structure. Left monomer shows contact footprints characteristic of the ensemble $\mathbb{E}$. Greyscale gradient indicates contact frequency. Right monomer highlights point mutations associated with HIV-1 protease resistance to inhibitor (Johnson *et al.*, 2005).

compounds and inactive ones. However, default learning with all vectors gives good results for the three models tested. Furthermore, active compounds are isolated from the other leaves of the tree without taking into account experimental data such as key residues or activities. Thus, good classification can be obtained from default run without plotting ROC curves.

k-means clustering (implemented in AuPosSOM) could give interesting results when input data are less dispersed. Compared with *k*-means clustering, SOM clustering is more efficient when input data are dispersed (e.g. for the protease dataset) (Mangiameli *et al.*, 1996). Furthermore, the performance of the SOM network is shown to be robust across a wide range of data imperfection inherent to outlier poses proposed by the docking software. This advantage is explained by the iterative learning of the Kohonen map which may neutralize the noise produced by incorrect poses of docking. Moreover, SOM network is less sensitive to its set of parameters, and cluster results are therefore more repeatable.

Another approach based on structural interaction fingerprints (SIFt) has been described to analyse and organize protein-small molecule complexes (Deng *et al.*, 2004). In this method, similarity measurement between vectors is based on the Tanimoto coefficient and a hierarchical clustering is done using an agglomerative hierarchical clustering. Compared with the method proposed in this article, this process has the same advantages: it allows an unbiased docking protocol and a bypass to scoring functions. Furthermore, interaction descriptor-based methods enable scaffold-hopping. However, the use of the Tanimoto coefficient as a similarity measurement leads to two major drawbacks: (i) due to slight errors in the binding geometry, missing components in interaction vectors may lead to incorrect clustering; (ii) comparison of macro-ligands (e.g. peptide) and small organic compounds could yield locally similar patterns but globally different vectors, which results in high Tanimoto coefficient whereas local binding mode may be the same. Furthermore, SIFt method needs known complex structures as referenced fingerprints, thus only expected binding modes are selected. Finally, one correct pose must be selected, according to the reference, which sets aside the statistical entropic analysis.

In order to check the influence of the number of known active compounds included in the database for the benchmarks presented here, the probability of clustering at least all active compounds (true positives) with an undefined number of false positive was calculated.

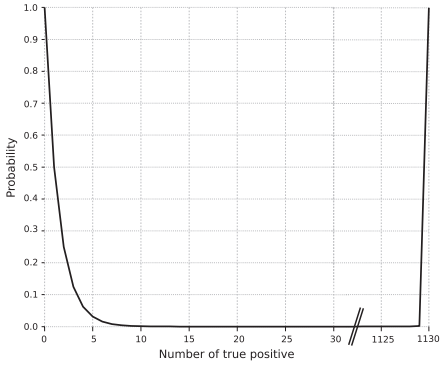

Fig. 8. Effect of the number of true positive in a database of 1130 compounds in the probability of clustering at least all active compounds [see Equation (4)].

This probability P is given by :

$$P = \frac{\sum_{i=0}^{n-p} C_{n-p}^{i}}{\sum_{j=p}^{n} C_{n}^{j}} \quad (4)$$

where p is the number of true positive and n the total number of compounds. For the HIV-1 protease screening, $p = 22$ and $n = 1130$. Thus, the probability P of clustering at least all active compounds is $2.38 \cdot 10^{-7}$. For a database composed of 1130 molecules, the probability $P < 1\%$ when the number of true positive is more than 6. This probability becomes very small ($P < 10^{-3}$) when the number of true positive is more than 10 (Fig. 8). Therefore, the number of active compounds put into the total number of compounds of the database is a good benchmark to check the reliability of the contact clustering process presented in this article.

From a dataset of compounds, contact clustering is able to discriminate active compounds based only on CAR. We observed that using first this analysis of the position of the docked molecules is more successful to find out active compounds than a classical energy analysis. This new method can be integrated into VS process currently available to rapidly visualize results. Furthermore, contact clustering process allows identification of key residues implicated in ligand binding. Hierarchical classification of compounds according to their contacts can be employed as a starting point to achieve *in silico* drug-design process.

ACKNOWLEDGEMENTS

We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources. We thank Dr Daniele Marinazzo from the Laboratory of Neurophysics and Physiology, CNRS UMR 8119, at Université Paris Descartes for his helpful discussions on the clustering methods.

Conflict of Interest: none declared.

REFERENCES

Alvarez,J. and Shoichet,B. (2005) *Virtual Screening In Drug Discovery*. CRC Press, Boca Raton, USA.

Chang,M. *et al.* (2007)Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries. *J. Chem. Inf. Model.*, **47**, 1258–1262.

Chang,M. *et al.* (2008)Empirical entropic contributions in computational docking: evaluation in APS reductase complexes. *J. Comput. Chem.*, **29**, 1753–1761.

Deng,Z. *et al.* (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.*, **47**, 337–344.

Evrard-Todeschi,N. *et al.* (2008) Structure of the complex between phosphorylated substrates and the SCF β-TrCP ubiquitin ligase receptor: a combined NMR, molecular modeling, and docking approach. *J. Chem. Inf. Model.*, **48**, 2350–2361.

Hamelryck,T. and Manderick,B. (2003)PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.

Huey,R. *et al.* (2007) A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, **28**, 1145–1152.

Jain,A. (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, **46**, 499–511.

Johnson,V. *et al.* (2005) Update of the drug resistance mutations in HIV-1: fall 2005. *Top HIV Med*., **13**, 125–131.

Kitchen,D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, **3**, 935–949.

Kohonen,T. (2001) *Self-Organizing Maps*. Springer Series in Information Sciences, Heidelberg, Germany.

Liu,T. *et al.* (2007)BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

Mangiameli,P. *et al.* (1996) A comparison of SOM neural network and hierarchical clustering methods. *Eur. J. Oper. Res.*, **93**, 402–417.

Morris,G. *et al.* (1998)Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.

Muegge,I. (2008) Synergies of virtual screening approaches. *Mini Rev. Med. Chem.*, **8**, 927–933.

Renner,S. *et al.* (2008)Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints. *J. Chem. Inf. Model.*, **48**, 319–332.

Samsonova,E.V. *et al.* (2006) TreeSOM: cluster analysis in the self-organizing map. *Neural Netw.*, **19**, 935–949.

Schafferhans,A. and Klebe,G. (2001) Docking ligands onto binding site representations derived from proteins built by homology modelling. *J. Mol. Biol.*, **307**, 407–427.

Schüttelkopf,A. and van Aalten,D. (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Cryst.*, **60**, 1355–1363.

Triballeau,N. *et al.* (2005)Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.*, **48**, 2534–2547.

Wang,J. *et al.* (2005)Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MMPB/SA. *J. Med. Chem.*, **48**, 2432–2444.

Wang,W. and Kollman,P. (2001)Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. *Proc. Natl Acad. Sci. USA*, **98**, 14937–14942.

Warren,G.L. *et al.* (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **49**, 5912–5931.

Waszkowycz,B. (2008) Towards improving compound selection in structure-based virtual screening. *Drug Discov. Today*, **13**, 219–226.

Yang,J. *et al.* (2005)Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.*, **45**, 1134–1146.