Mol Divers (2016) 20:421–438
DOI 10.1007/s11030-015-9648-5

FULL-LENGTH PAPER

Integrating sampling techniques and inverse virtual screening: toward the discovery of artificial peptide-based receptors for ligands

Germán M. Pérez[1,2] · **Luis A. Salomón**[3,4] · **Luis A. Montero-Cabrera**[2] · **José M. García de la Vega**[5] · **Marcello Mascini**[6]

Received: 27 May 2015 / Accepted: 2 November 2015 / Published online: 9 November 2015

Abstract A novel heuristic using an iterative select-and-purge strategy is proposed. It combines statistical techniques for sampling and classification by rigid molecular docking through an inverse virtual screening scheme. This approach aims to the de novo discovery of short peptides that may act as docking receptors for small target molecules when there are no data available about known association complexes between them. The algorithm performs an unbiased stochastic exploration of the sample space, acting as a binary classifier when analyzing the entire peptides population. It uses a novel and effective criterion for weighting the likelihood of a given peptide to form an association complex with a particular ligand molecule based on amino acid sequences. The exploratory analysis relies on chemical information of peptides composition, sequence patterns, and association free energies (docking scores) in order to converge to those peptides forming the association complexes with higher affinities. Statistical estimations support these results providing an association probability by improving predictions accuracy even in cases where only a fraction of all possible combinations are sampled. False positives/false negatives ratio was also improved with this method. A simple rigid-body docking approach together with the proper information about amino acid sequences was used. The methodology was applied in a retrospective docking study to all 8000 possible tripeptide combinations using the 20 natural amino acids, screened against a training set of 77 different ligands with diverse functional groups. Afterward, all tripeptides were screened against a test set of 82 ligands, also containing different functional groups. Results show that our integrated methodology is capable of finding a representative group of the top-scoring tripeptides. The associated probability of identifying the best receptor or a group of the top-ranked receptors is more than double and about 10 times higher, respectively, when compared to classical random sampling methods.

Keywords Molecular modeling · Inverse virtual screening · Sampling techniques · Molecular docking

Electronic supplementary material The online version of this article (doi:10.1007/s11030-015-9648-5) contains supplementary material, which is available to authorized users.

Germán M. Pérez and Luis A. Salomón have contributed equally.

✉ Germán M. Pérez
german@fq.uh.cu

1 General Chemistry Department, Chemistry Faculty, University of Havana, Zapata y G, Havana 10400, Cuba

2 Laboratory of Theoretic and Computational Chemistry, Chemistry Faculty, University of Havana, Havana, Cuba

3 Applied Mathematics Department, Mathematics and Computer Science Faculty, University of Havana, Havana, Cuba

4 Department of Mathematical Sciences, EAFIT University, Medellín, Colombia

5 Department of applied chemical-physics, Autonomous University of Madrid, Teramo, Spain

6 Faculty of Bioscience and Technology for Food, Agriculture and Environment, University of Teramo, Teramo, Italy

Introduction

For the last two decades, the virtual screening (VS) of chemical libraries has been established as a standard technique among many computational and experimental research teams, and it is mostly used for the identification of bioactive compounds through computational means [1,2]. This technique aims to increase the probability of identifying bioactive molecules by maximizing a true-positive rate. It enables the

user to considerably reduce the number of compounds to be synthesized and tested experimentally, offering an alternative to costly experimental high-throughput screenings [3–5].

Virtual screening (VS) uses two main approaches: structure-based, when structural information about a protein target (i.e., molecular receptor) is known [6–8], or ligand-based when known bioactive ligands exist [9,10]. VS was designed to analyze libraries of hundreds to millions of molecules, and rank them according to their likelihood of exhibiting affinity toward a disease-driving biological target. Among the most traditional techniques integrated in VS are pharmacophore modeling [11–13] and molecular docking [8,14–17], frequently used in combination [18–20]. In particular, docking has become a primary component in many VS-lead discovery efforts [21–25] and it is often combined with homology modeling [24], QSAR studies [26,28], and molecular dynamics [29–31].

In general, the most used docking approach in VS is receptor centric. It essentially focuses on screening databases of prospective ligands, using a comparative analysis of their structural shape and chemical complementarity to a known receptor. However, the opposite approach can also be used to identify possible receptors for a known ligand [32,33] involving the screening of a database of prospective receptors against a known ligand. This approach, known as inverse VS (iVS), has demonstrated its effectiveness in the in-silico identification of targets for small ligands [34]. In recent studies, iVS procedures have been largely studied [35,36], used for identification of potential anti-cancer targets [37], drug repurposing, and target fishing experiments [33,35,38–40].

Most techniques available, however, are focused on matching actual compounds where the structure of both ligands and receptor molecules is known. In such cases, the purpose is to identify the possible association complexes they might form, or the poses a ligand might adopt when binding to the active site in a receptor. These receptors are mostly macromolecules of high molecular weight, like proteins or DNA, but in other fields (i.e., diagnostics, biosensors) smaller receptors are often preferred as an alternative to large proteins such as enzymes or antibodies [41,42].

The computational design of peptide-based receptors for specific target ligands from scratch has been used in previous works, and experimental results have demonstrated the convenience of this methodology [43–46]. In these studies, a combination of two major semi-combinatorial approaches has been used aiming to reduce the number of simulations: knowledge-based approach and incremental construction, although other methods are also available and discussed [47–49]. The knowledge-based approach consists of reducing the amount of prospective receptors by constraining the number, types, and location of amino acids in a peptide sequence. This can be achieved by mimicking the geometry and chemical environment present in the binding site of a biological receptor [44]. The incremental construction approach uses a consensus criterion, selecting amino acids found to interact with different zones of the target ligand throughout different known receptors [43]. The affinity for a ligand can be increased with both approaches, either by the selective increase of peptide length or mutation of residues that do not participate in the interaction.

However, even when using a combined approach, the exploratory analysis of the sample space is rather small and it is intrinsically biased. It targets only a minimal, not representative, and not necessarily correct fraction of all possible sequence arrangements, discarding amino acid combinations that might result in good solutions for the molecular association problem. For this reason, these approaches cannot estimate how accurate their predictions are when considering all possible combinations; they can only establish a local discrete sub-space in which the receptors analyzed are ranked according to their estimated binding parameters. False positives or negatives are hardly detected, and there is no tangible evidence that the optimal receptors identified are even within the above-average solutions found when considering the entire sample space.

To address these issues, a new methodology is presented in this work. It uses the estimated affinity of a receptor for a target ligand to define a heuristic algorithm. Specifically, the algorithm analyzes the relationship between amino acids and their positions in peptide sequences, and how this affects the calculated binding score when forming the association complex with a given ligand. In particular, it tries to identify which amino acids and in which positions in the sequence yield either the best or worst scores. At the same time, the algorithm estimates the probability of sampling among the top scoring receptors, ranked by their association scores. The main goal of the algorithm is to improve sampling by increasing the proportion of true positives in the sampled compounds while reducing false positive rates.

This work describes a new VS methodology that performs a stochastic exploration of the sample space, and presents the results obtained in the computational identification of tripeptides that could act as molecular receptors for selected ligands. The selected ligands covered different functional groups commonly found in organic compounds. Tripeptides, on the other hand, were obtained using a combinatorial approach and grouped using their amino acid composition as descriptor.

Methods

Preparation of structures

In this work, 159 ligands were used, covering 9 of the most representative functional groups (Alcohols, Ketones, Esters,

etc.) found in organic compounds as well as some of their combinations (Phenols, Benzaldehydes, Nitrophenols, etc.). These molecules were selected based on their diversity in functional groups, molecular volumes, and atoms connectivity. Molecular weights ranged from 40 to 480 Da and solvent accessible surfaces from 170 to over 700 $Å^2$.

The 3D structures of ligands were automatically generated based on their IUPAC names using the LEXICHEM package from OpenEye [50]. An initial molecular mechanics geometry optimization was carried out with SZYBKI 1.5.1 using the Merck molecular force field (MMFF) [51]. A second geometry optimization step using an AM1 semiempirical method was performed on the ligands with MOPAC 7.01-4, which is included in VEGA ZZ 2.4.0 [52]. Final structures were inspected to correct errors such as valences, angles and bond distances, atoms connectivity, atoms clashes, etc. For each ligand, a group of conformers was generated with OMEGA 2.4.3 [53] within 10 kcal from the local minimum and an RMSD ≥ 0.5 $Å^2$.

For the methodology we propose in this work it should be noted that its performance is not determined by the type of complexes analyzed. Instead, the performance depends on the ability of the energy assessment method to maintain a stable trend and a correlation among experimental determinations and estimated binding affinities of different receptors for a given ligand. Because of this, it was not necessary to use a larger set of compounds to validate the VS methodology.

The initial population of 8000 tripeptides was generated in *zwitterionic* form based on their primary structures. Geometry optimizations were carried out with up to 500 iterations of the Steepest Descent algorithm, followed by a maximum of 2000 iterations of a Polak–Ribiere conjugated gradient algorithm and a convergence condition of 0.01 kcal/(Åmol). Peptides were re-optimized with SZYBKI 1.5.1, followed by a conformational analysis using the same parameterization as for the ligands. For each receptor, the three conformers with higher entropic contributions were identified and subsequently used in the docking simulations.

Docking simulations

Molecular docking simulations were carried out with two different software packages: OEDocking 3.0.0 [54] and AutoDock Vina 1.1.2 [55]. The former is an upgraded version of a rigid-body docking program based on molecular mechanics and with consensus scoring, used by the authors in previous works with excellent results [43–46]. The latter is one of the most used programs for flexible docking and virtual screening studies with proteins, parameterized for protein receptors, and using a completely empiric scoring function named X-Score [55].

For rigid-body docking a receptor was built using the entire peptide surface. The boxes were generated by adding 4 Å to frontier coordinates of each peptide. Box sizes were in the range of 4000 to 7000 $Å^3$. All 159 multi-conformer ligand files were docked against each of the receptors created for selected conformers of all 8000 tripeptides, using default OEDocking parameterization. The systems were modeled in vacuo using a MMFF94 charge model. All structures were energy-minimized using the MMFF94 force field.

The sole purpose of using AutoDock Vina was to show how the performance of the algorithm was independent of the energy assessment method used. In this case, only three ligands (Ethanol, Decanal, Ethyl Propanoate) were considered against all 8000 tripeptides. These ligands were selected as representatives of their corresponding groups and as worst case scenario since they showed the highest deviations from the overall trend observed in the algorithm performance. In this case, the local minimum conformation for each ligand (enabling rotation of all possible rotatable bonds) was docked against each MMFF-optimized peptide with flexible side chains, and exhaustiveness = 15 for a good speed/accuracy ratio. In both cases, the energy window was 5 kcal/mol from the minimum.

Select and purge (SP) algorithm

Objectives and definition

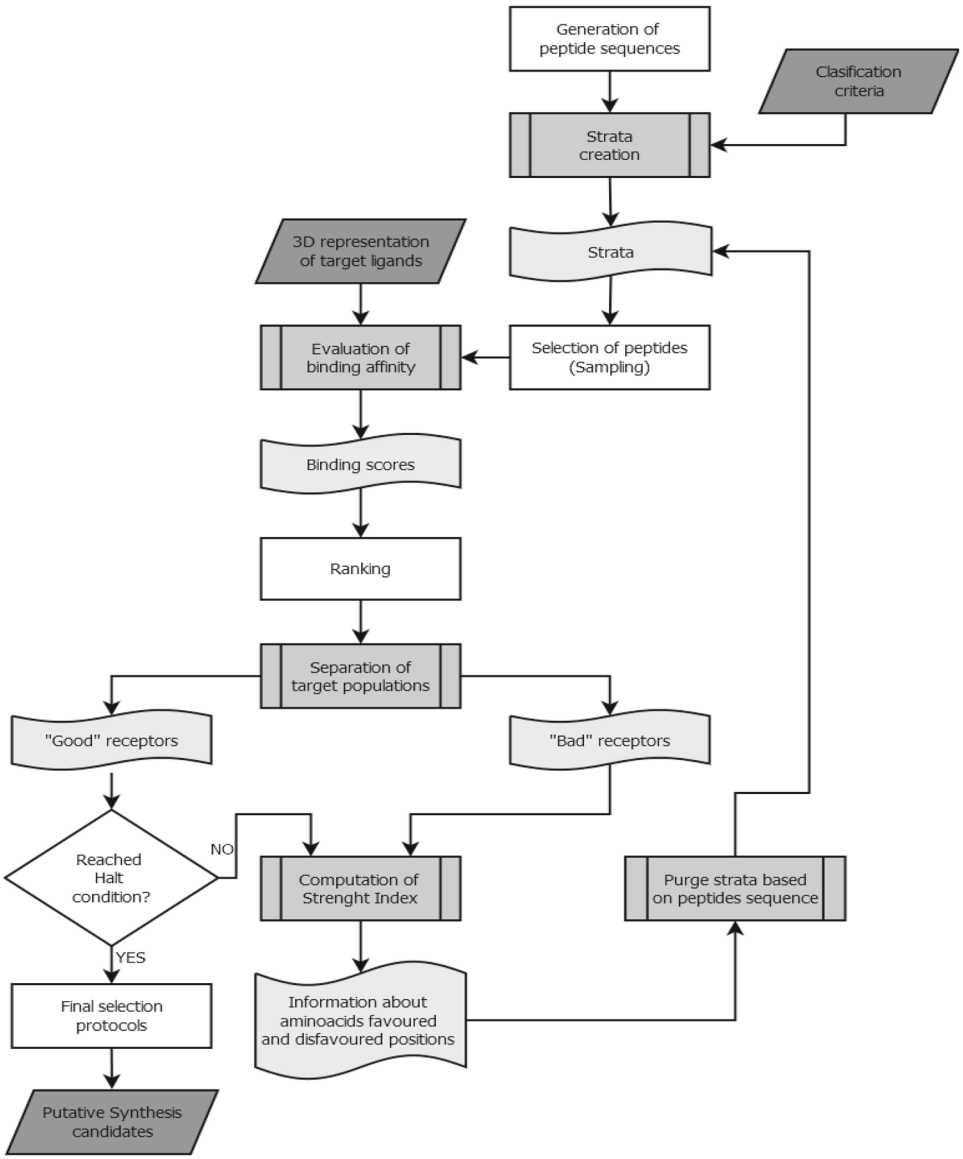
A two-stage stratified sampling algorithm was used, aiming to evaluate the effects of amino acid composition on tripeptides and their propensity to form association complexes with specific target ligands. It introduces the *strength index* (*SI*), a measurement that allows the algorithm to differentiate two target populations from sampled data: the first one related with the tripeptides showing the best scores and the second, those with worst scores.

The algorithm goes through an iterative process in which, after reaching the stop condition, a group of probable receptors is identified from the stock of peptides analyzed. The selection of suitable candidates and elimination of false positives are supported by the information of favorable and unfavorable pairs amino acid-position (PAP) in peptide sequences. The general formulation and workflow of the algorithm are shown in Fig. 1.

Sample size

Sample size n was representative of the entire sample space (the 8000 tripeptides), avoiding small values that otherwise could have been misleading and large values that would have rendered impracticable the collection of data under analysis. Since the magnitude of interest is a proportion, the following formula was applied to determine the smallest value of n for which the sample could be considered as representative for this proportion [55,56]:

Fig. 1 Flowchart of SP scheme



$$n = \frac{Z_{1-\alpha}^2 N p q}{\varepsilon^2 (N-1) + Z_{1-\alpha}^2 p q}, \tag{1}$$

where N is the population size, ε is the desired error, p is the expected probability of the characteristic under analysis, $q = 1 - p$, and Z_α is the α-percentile (typically 0.01, 0.05 or 0.1) of the standard Gaussian distribution, according to:

$$\Phi(Z_\alpha) = P(Z \leq Z_\alpha) = \alpha, \text{ with } Z \sim N(0,1). \tag{2}$$

Equation (1) gives an idea about the correct sample size, showing that for fixed values of α, N, and ε, the value of n entirely depends on p. In situations where this magnitude is unknown a worst case scenario can be used where n takes the value for $p = 0.5$ [56].

For very large populations ($N \to +\infty$), Eq. (1) takes the following form:

$$n = \frac{Z_{1-\alpha}^2 p q}{\varepsilon^2 + \frac{Z_{1-\alpha}^2 p q - \varepsilon^2}{N}} \xrightarrow{N \to +\infty} \frac{Z_{1-\alpha}^2 p q}{\varepsilon^2}, \tag{3}$$

This new equation shows that even for large values of population size N, the sample size n exhibits small variations, growing asymptotically. This guarantees that, even for large populations, the number of elements to be studied is both manageable and representative of the entire population.

Sampling method

A stratified sampling technique (SST) [57] was used in this study. This sampling method allows the creation of S_n strata from an entire population. It constitutes a major advantage over other sampling methods since data might be grouped based on similar characteristics or descriptors (*e.g.*, molecu-

lar weight, composition, presence of particular amino acids, etc.). As added value, the procedure for creating strata is less sensitive to a poor representation of the drawn sample and enables more accurate inference in each stratum, which could be otherwise lost in a sample chosen classical simple random sample (SRS) method.

Once all strata were defined, each independent stratum was sampled using SRS method. The sample size for each stratum $n_i, i = 1, 2 \ldots, S_n$ satisfied that $n = \sum_{i=1}^{S_n} n_i$. Usually the assignment for the sample size n_i is determined proportional to the corresponding size of the stratum in consideration:

$$n_i = n\frac{N_i}{N}, \tag{4}$$

where $N = \sum_{i=1}^{S_n} N_i$ and $N_i, i = 1, 2 \cdots, S_n$, are the size of each stratum.

The combination of these two sampling methods is not meant only to draw a sample from the population, but also for both of them to play a role as classifiers. This approach creates a convenient and representative sample. It is unbiased and less prone to casuistic sampling errors, as opposed to SRS method alone, or to any of the knowledge-based and incremental construction methods previously mentioned [43, 44].

Classification of data based on chemical characteristics

All peptides were initially placed in non-overlapping strata, hence the sampling of possible receptors was handled in a more comprehensive way. The creation of those strata was carried out using the amino acid composition of peptides as a descriptor and classifier for populating each stratum. Specifically for this study, five strata were created attending to the predominance of particular types of amino acids (Aliphatic, Positively charged, Negatively charged, Polar and Aromatic) in peptide sequences, disregarding their positions. A sixth stratum contained peptides with no predominant types of amino acids. This approach allowed the partitioning of the population according to similarities in peptide sequences, providing an unbiased, representative way of exploring sample space.

The size of the six strata created ($S_n = 6$) was: $N_1 = 2254$, $N_2 = 486$, $N_3 = 224$, $N_4 = 1250$, $N_5 = 486$ and $N_6 = 3300$ (Table 2). Group sizes were quite different; however, this fact is not determinant in the methodology.

Algorithm description

Analysis of data and strata creation The first step of the SP algorithm uses the nature of data to classify all population elements (i.e., tripeptides) and create the initial non-overlapping S_n strata. In the present work, these considerations were of chemical nature, regarding the type of amino acids present in peptide sequences. However, the criterion or descriptor used to allocate elements to each stratum is flexible and depends on the system studied, the information available, and goals pursued. This is important since the creation of strata is a requirement to calculate the *strength index*.

Next, all the N elements in the initial population are allocated to each stratum. This is possible since given a peptide length, the total number of peptides constituting the population can be easily determined, their sequences and tridimensional structures can be readily generated, and their binding affinity for a particular ligand can be computed whenever necessary.

In order to extract the sample from the population a classical approach is used. Sample size n is computed according to Eq. (1) and the size of each stratum $n_i, i = 1, 2 \ldots, S_n$ according to Eq. (4). To simplify notations, in this document an association has been made between $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1\mathcal{A}_2, \cdots, \mathcal{A}_{20}\}$ and the 20 amino acids $\{AGILMPVRHKDECNQSTFWY\}$, $(i.e., \mathcal{A}_1 = A\,(alanine), \mathcal{A}_5 = P\,(proline), \mathcal{A}_{19} = W\,(tryptophan)$, etc.).

Evaluation of binding affinity: computation of scores for the selected sample As previously mentioned, the SP algorithm is not dependent on the method used to estimate binding affinities. In fact, two different approaches were used in this work. Disregarding the method used, the evaluation of binding affinities must be carried out, computing the n score values for the selected samples, henceforth denoted by $(X_i), i = 1, 2, \ldots, n$, since this is the response variable that allows future classification. Scores in this case are related to the binding free energy, thus lower scores indicate lower energies, higher affinities and in consequence, more stable complexes.

Determination of the two target populations attending to scores ranking In this step, the SP algorithm tries to separate the good receptors from the bad ones, and assign them to their corresponding subpopulations. The sample of n peptides is sorted in decreasing affinity order according to their scores. Two cutoff values are introduced to discriminate, based on scores, whether a peptide produces a stable complex or not. This is done by assuming that the first 100β % of the entire population are the peptides with higher affinities and the last 100γ % the ones with the lower affinities (where β and γ are arbitrary values in the [0, 1] range). Considering that scores $(X_{(i)}), i = 1, 2, \ldots, n$ can be sorted, where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, the cutoff values are established as follows:

$$C_\mathrm{b} = X_{([\beta n])} \tag{5a}$$

$$C_\mathrm{w} = X_{(n-[\gamma n])}, \tag{5b}$$

where subscripted [...] stands for the integral part of X and C_b, C_w represent the cutoff values for discriminating the best and worst complexes formed, respectively. It is straightforward that in fact three subpopulations can be differentiated:

- G_b: Subpopulation formed by the $n_b = [\beta n]$ elements having the best scores.

$$G_\mathrm{b} = \left\{ X_i^B = X_{(i)}, \forall i : X_{(i)} \leq C_\mathrm{b} \right\}$$

- G_w : Subpopulation formed by the $n_w = \left[\gamma n\right]$ elements with worst scores.

$$G_\mathrm{w} = \left\{ X_i^\mathrm{w} = X_{(i)}, \forall i : X_{(i)} \geq C_\mathrm{w} \right\}$$

- G_r: Subpopulation formed with remaining $n - n_\mathrm{b} - n_\mathrm{w}$ elements, discarded in next step.

This procedure always guarantees non-empty subpopulations G_b and G_w. In an initial stage, taking $\beta = \gamma$ is a simple, logical assumption that considers identical behavior in both subpopulations, equally weighed and does not affect the next steps. These cutoff values are recalculated in each iteration of the SP algorithm, refining the classification criteria based on the information accumulated from previous iterations.

Computation of the Strength Index according to the strata information This is the core of the SP algorithm. Here, selected peptides are grouped in strata and placed in either subpopulations G_b, G_w or G_r. For each subpopulation, the frequency in which an amino acid appears in any of the positions of the peptides sequences is counted and stored in the following matrix:

$$\mathcal{C}^{\mathrm{b},p} = \left(\mathcal{C}_{ij}^{\mathrm{b},p} \right) \in M_{20\times p}, \tag{6a}$$

where $\mathcal{C}_{ij}^{\mathrm{b},p}$ is the number of times the amino acid i appears in position j of the peptide sequence, in the subpopulation G_b; $M_{20\times p}$ is a matrix with 20 columns (one for each amino acid) and 3 rows (one for each possible sequence position). In an analogous way, another matrix can be defined for subpopulation G_w:

$$\mathcal{C}^{\mathrm{w},p} = \left(\mathcal{C}_{ij}^{\mathrm{w},p} \right) \in M_{20\times p}, \tag{6b}$$

Non-zeros values of $\mathcal{C}^{\mathrm{b},p}$ and their equivalents in $\mathcal{C}^{\mathrm{w},p}$ are fundamental. Supported by the assumption that some amino acids in specific positions (or PAPs) are key in stabilizing the supramolecular complex, it would be logical to presume that those PAPs will not appear in the G_w subpopulation, or at most, they will have a minor presence. The opposite is also valid.

In practice, however, equivalent positions in both matrices can have significant or similar values. One possible cause for this is that in the previous assumption, the chemical environment (other amino acids, covalently bonded in the sequence) is not considered.

Summarizing, there are two different situations:

- $\mathcal{C}_{ij}^{\mathrm{w},p} = 0$

The corresponding amino acid i only appears in subpopulation G_b. With this information alone the conjecture is that PAP is relevant to the formation of the association complex. To evaluate how good it actually is, strata information is used to compute the *SI*, in others words, to measure the quality of this amino acid. *SI* is defined as follows:

$$SI_{S_n}^b(i,j) = \sum_{k=1}^{S_n} \mathcal{J}_{\mathcal{G}_k}^j(\mathcal{A}_i) \tag{7}$$

where $\mathcal{G}_k$ is the subset of amino acids in the sample that belongs to the stratum k, k=1, 2, ..., S_n, and

$$\mathcal{J}_{\mathcal{G}_k}^j(\mathcal{A}_i) = \begin{cases} 1, & \text{If } \mathcal{A}_i \text{ appears in } \mathcal{G}_\mathrm{k} \text{ in position } j \\ 0, & \text{otherwise} \end{cases}$$

If $SI_{S_n}^\mathrm{b} \in (1, S_n)$ where $SI_{S_n}^b(i,j) = S_n$, then that amino acid i appears in all the strata at the same position j and that gives an idea of its strength.

- $\mathcal{C}_{ij}^{\mathrm{w},p} > 0$

In this case, the following hypothesis tests are used:

$$H_0 : p_{ij}^\mathrm{b} = p_{ij}^w \tag{8a}$$

$$H_A : p_{ij}^\mathrm{b} > p_{ij}^w \tag{8b}$$

where p_{ij}^b and p_{ij}^w are the probabilities associated to the presence of amino acid i in position j of the peptide sequence, in the subpopulations with better and worst scores, respectively; H_0 and H_A are the null and alternative hypotheses, respectively. Preferred amino acids are those in which previous hypothesis test is rejected. The critical region with signification level α is defined by:

$$Z = \frac{\hat{p}_{ij}^b - \hat{p}_{ij}^\mathrm{w}}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_\mathrm{b}} + \frac{1}{n_\mathrm{w}}\right)}} > Z_{1-\alpha} \tag{9}$$

where,

$$\hat{p}_{ij}^\mathrm{b} = \frac{\mathcal{C}_{ij}^{\mathrm{b},p}}{n_\mathrm{b}}, \tag{10a}$$

Table 1 List of the 26 groups initially formed with the 159 ligands according to their functional groups

Main groups	Group	Compound type	Quantity	Training	Test
10 Pure 105 (31)	1	Alcohols	11	5 (1)	6 (1)
	2	Aldehydes	14	7 (1)	7 (2)
	3	Ketones	12	6 (3)	6 (3)
	4	Esters	22	11 (4)	11 (4)
	5	Halogenated compounds	4	2 (1)	2 (1)
	6	Aromatic	14	7 (2)	7 (5)
	7	Cyclic	6	3 (1)	3 (1)
	8	Thiolated compounds	9	3	6
	9	Nitrogenated compounds	10	7	3
	10	Others	3	3 (1)	0
13 Doubles 44 (18)	11	Alcohol + Ketone (Hydroxypyranones/Furanones)	4	2	2
	12	Alcohol + Aromatic (Phenols)	6	4	2
	13	Alcohol + Cyclic (Substituted Hexanols)	2	0	2
	14	Aldehyde + Aromatic (Benzaldehydes)	7	5 (1)	2
	15	Ketone + Aromatic (Benzofuranone)	1	0	1 (1)
	16	Ketone + Cyclic	6	3 (3)	3 (2)
	17	Ketone + Nitrogenated	2	2	0
	18	Substituted benzene dicarboxylates	4	1 (1)	3 (3)
	19	Halogenated + Aromatic (Halogenated Benzenes)	6	3 (1)	3 (2)
	20	Halogenated + Cyclic	1	0	1 (1)
	21	Halogenated + Nitrogenated (Halogenated Aniline)	2	1	1 (1)
	22	Aromatic + Nitrogenated (Nitrobenzenes)	2	1	1 (1)
	23	Cyclic + Thiolated	1	0	1 (1)
3 Triples 10	24	Halogenated Phenols	6	4	2
	25	Alcohol + Aromatic + Nitrogenated (Nitrophenols)	3	1	2
	26	Alcohol + Aromatic + Ketone	1	0	1
		Total	159	81 (20)	78 (29)

Column 4 shows the number of ligands in each group. Columns 5 and 6 indicate the number of compounds in the training and test sets, respectively. In parentheses, the number of discarded elements

$$\hat{p}_{ij}^{\mathrm{w}} = \frac{\mathcal{C}_{ij}^{\mathrm{w},p}}{n_{\mathrm{w}}}, \tag{10b}$$

$$\hat{p} = \frac{\mathcal{C}_{ij}^{\mathrm{b},p} + \mathcal{C}_{ij}^{\mathrm{w},p}}{n_{\mathrm{b}} + n_{\mathrm{w}}}, \tag{10c}$$

In the cases were hypothesis test is rejected, the *SI* is computed, else, it can be assumed that the presence of that amino acid in both population (according to their probabilities) is similar, or at least, there is no evidence to think otherwise. In those cases, the amino acid is not representative in any subpopulation, hence its contribution to identify good and bad receptors is not conclusive.

The procedure described here is meant to calculate the *SI* for G_{b}; it works *mutandis mutatis* for G_{w}. The combination of these two indexes allows the overall identification of preferred and forbidden positions for specific amino acids in peptides chains to form reliable association complexes with a particular ligand.

Purge and resampling of remaining data, based on information provided by the SI This is the last step of a single iteration of the SP algorithm. In this step, once the *SI* has been calculated, this information is used as a feedback in the next iteration of the algorithm to purge unlikely solutions from the remaining population. This elimination means that these peptides are not considered in the next iteration. Since the *SI* is dynamic and because of the stochastic nature of the algorithm, a PAP marked as relevant, can be either confirmed or refuted in subsequent iterations of the algorithm as it converges to the optimal solutions.

In order to do this, another sample of size n_{r}, is chosen according to Eq. (1), using $N - n$ instead of N:

$$n_{\mathrm{r}} = \frac{Z_{1-\alpha}^2 \, (N-n) \, pq}{\varepsilon^2 \, (N-n-1) + Z_{1-\alpha}^2 \, pq}. \tag{11}$$

Two new groups are then defined: the first one, contains all peptides having in any position at least one of the amino acids

that satisfies $SI^b_{S_n} \geq \left[\frac{S_n}{2}\right]$ and $SI^{\mathrm{w}}_{S_n} < \left[\frac{S_n}{2}\right]$; the second is formed by the remaining peptides not satisfying this condition. The number $[S_n/2]$ was chosen arbitrarily in a way that a PAP having a $SI^b_{S_n}$ greater than this number, has a greater contribution to the best scores.

This assumption was made, in accordance to the characteristics and goals pursued in this work. For example if $S_n = 6$, then $[S_n/2] = 3$, therefore, if $SI^b_{S_n} \geq 3$ it means that this particular PAP has been found relevant in peptides present in at least a half plus one of all strata. However, it is not convenient to have a large number of strata; in the case it cannot be avoided, another condition for $SI^b_{S_n}$ should be chosen. A similar condition could be defined for $SI^w_{S_n}$. In order to create these two subpopulations, their sample sizes are defined as follows:

$$n_{\mathrm{r}}^{\mathrm{b}} = \frac{\beta}{\beta+\gamma} n_r \tag{12a}$$

$$\text{and } n_{\mathrm{r}}^{\mathrm{w}} = \frac{\gamma}{\beta+\gamma} n_{\mathrm{r}} \tag{12b}$$

Attending to this last equation, samples are drawn from each subpopulation.

Screening process: application of the SP algorithm

The ligands studied were classified and grouped in training and test sets. These ligands were the target molecules for the tripeptide receptors. Chemical characteristics were of particular importance, in consequence, 26 disjointed groups containing from 1 to 22 compounds were formed as shown in Table 1. In these groups, 10 contained compounds with only one functional group of interest; 13 grouped compounds with double functional group; and the rest in the remaining 3 groups.

The training and test sets for the algorithm were prepared so that all groups were represented whenever possible, but in some cases the training/test ratio was altered on purpose. This aimed to analyze whether the predictive accuracy of the method was the same, even with types of structures not included in the set of ligands used for training the algorithm. In particular, when forming groups, to incorporate perturbations in the training of the SP algorithm and check its robustness, two actions were taken. First, none of the ligands in the special group "Others" was considered in the testing stage. This group was formed by ligands with functional groups poorly represented, or with high variability. Second, the distribution for Nitrogenated and Thiolated compounds was altered to favor the training and test sets, respectively. It is important to notice that some ligands, even though classified in one of the main 9 groups, presented multiple functional groups (see Supporting Information).

Application of SP scheme and comparison with other methods

For the purpose of the present work, the following parameters $\varepsilon = 0.05, \alpha = 0.01, p = q = 1/2$ were used to determine the value of n and $\beta = \gamma = 0.05$ for the third step of the method. These values are the classical ones in the former theory of random sampling: ε is the desired error associated to the method and α is associated with the probability to obtain a good estimation in our sampling process, specifically, this probability is $1 - \alpha$.

Parameters β and γ are both related with the number of good and bad peptides in the database, respectively. They establish the percentage of the population that can be classified as good complexes or bad ones. They were arbitrarily chosen, with value 0.05 as starting point; however, in the final stage of this work, values 0.02, 0.03, 0.04, 0.075, and 0.1 were also tested.

For the training and testing of the algorithm, the affinity of each of the 8000 tripeptides vs. each of the 159 ligands was evaluated. This was done in each corresponding stage in order to have a complete description of the system, including maximum and minimum values, global best, and worst receptors, respectively. With each ligand, 1000 simulations with a truncated SP algorithm (2 iterations) were conducted, computing the best and worst positions in each case. Probabilities were estimated in each iteration and the results were compared with the ones obtained in an analogous way using a classical SRS method.

Different metrics were used to assess the accuracy of the method and its predictive ability for the systems studied (i.e., ligands). First, an analysis of the Receiver Operating Characteristic curves (ROC curve) was carried out. This analysis included ROC shape, the area under the ROC curve (*AUC*), variability for different types of ligands, reproducibility in the course of several simulations for the same ligand, the influence of parameterization, and sample size. The implementation of these metrics was extracted from the work of Truchon and Bayly [58].

Also, another measure of the effectiveness of the method is provided: the enrichment factor (*EF*). The *EF* is another metric commonly used to compare virtual screening methods. It is calculated according to the expression:

$$EF^{X\,\%} = \frac{A_S N_T}{A_T N_S}; \tag{13}$$

where A_S is the number of Actives sampled, A_T is the total of possible Actives (PP group), N_T is the size of the population, and N_S is the sample size. For a method in which the proportion of Actives in the population corresponds to the one found in the sample (i.e., methods such as SRS), the $EF = 1$. An $EF = 5$, for example, indicates that the sam-

ple has 5 times more Actives than expected according to the proportion in the population. In this work, a variant of this metric was also used, namely the population enrichment factor (denoted EF_p in this work). The difference with the traditional EF is that EF_p accounts for how favorable is the "Actives Ratio" in the list of sampled elements with respect to the entire population. This was necessary since in traditional virtual screening studies the chemical library analyzed is considered as the population. That is not the case in this study since the chemical library analyzed (sampled peptides) is just a fraction of a larger population.

Results and discussion

Distribution of association energies (scores) in populations

The scores computed over all complexes formed between the 8000 peptides and each of the 159 ligands were sorted in ascending order. The curves obtained had similar distributions (Fig. 2). Score values comprised within the range of 6 to −6 kcal/mol in all simulations. In a single simulation using a given ligand versus the 8000 peptides, the average variability of scores was 3.4 ± 0.8 kcal/mol. To allow an unbiased comparison between the different systems under analysis, cutoff values were relative and expressed in terms of percentage of the population, instead of a unique, absolute value. Initially, 5 % was selected as cutoff because in all simulations this value delimited the zones in both ends of the curves in which the steeper slope changes were observed.

These cutoffs separated the 5 % of the complexes with higher scores and the 5 % with the worst scores from the rest of the population. These two subpopulations were arbitrarily labeled "prospective positives" (PP) and "prospective negatives" (PN), respectively. PP was defined as the group of complexes in which all members had a negative score under the lower cutoff value (VMin) established for that population. Similarly, PN was defined as the group with scores higher than the upper cutoff value (VMax). In PN, the scores were either positive or negative, depending on the ligand analyzed. However, since scores represent the association energies in the complexes, and positive values indicate that the association is unlikely, only the complexes with scores less than or equal to 0.5 kcal/mol were assigned to PN for VMax = 5 %. This concession was made because the rigid-body docking software used is not configured to work with such small receptors, and besides, the function used for the assessment of association energies is based on molecular mechanics, which is not very accurate by definition.

Ligands studied

Positive scores were obtained in more than 50 % of the hypothetical complexes formed with ligands having molecular weights (M) over 150 Da (with the exception of phenolic derivatives and compounds having hydroxyl or carbonyl groups). As result, 49 ligands were discarded. These ligands yielded positive scores in more than 20 % of the complexes formed, even after considering the tolerance margin of 0.5 kcal/mol previously mentioned.

Preliminary results indicate that discarded ligands had molecular weights comparable to those of the receptors. In particular for these ligands, the M(receptor)/M(ligand) ratio was always greater than 0.5. The sequence pattern Pro-X-Ala (X stands for any amino acid) showed a high tendency to yield positive scores. This behavior was observed, in general,

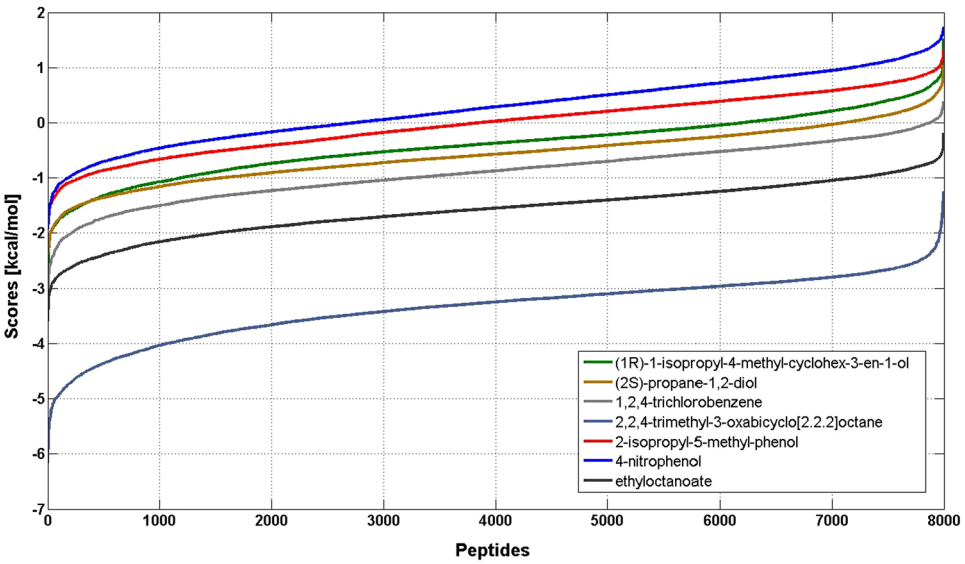


Fig. 2 Typical distributions of scores obtained in the simulations. The *graph* shows the behavior of scores in the association complexes formed between seven randomly selected ligands versus the 8000 tripeptides. The data are sorted in ascending order of score, thus not necessarily a correspondence must exist between the positions of the peptides in each curve

for most of the peptides with proline in first position. Conversely, aromatic amino acids phenylalanine, tyrosine, and tryptophan, tended to improve the affinity of the complexes formed. The most obvious case was that of tryptophan, where its presence, even in complexes formed with ligands among the 49 discarded, yielded negative scores.

These results indicate that using a higher number of ligands to test and validate the methodology can be misleading. Very similar ligands may result in overfitting of the algorithm. On the other hand, the larger the number of ligands, the higher the probability of including ligands for which the energy assessment method is not well parameterized, resulting in unreliable results and/or misleading predictions of association energies. In consequence, rather than increasing the number of ligands or functional groups, a large number of trials should be performed (repetitions of the simulations for each ligand included in the study) to demonstrate the stability of the SP algorithm performance and its superiority over an SRS alternative.

Classifier algorithm: strength index (SI)

As explained in the *Methods* section of this paper, the strata grouped peptides according to the predominance of an amino acid type in their sequences (Table 2). The *SI* was calculated for each PAP depending on its presence in other strata.

In the scope of present work, the maximum *SI* value for a given PAP was 5 because it counted how many times that particular PAP was in other peptides classified as PP but in different strata. Thus, the Total Strength Index (SI_T) for a given peptide in the systems studied was restricted to the range [−15,15], with $SI_T = \sum_{i=1}^{n} SI_i\,(PAP_{PP}) - \sum_{i=1}^{n} SI_i\,(PAP_{PN})$, $n = 3$ represents peptide length; $SI_i\,(PAP_{PP})$ and $SI_i\,(PAP_{PN})$ are the Strength Indexes corresponding to the PAP identified in the PP and PN groups, respectively.

Table 2 Features and composition of strata

Group	Condition	Amino acids	Type	Quantity
G1	At least two amino acids of the same type	AGILMPV	Aliphatic	2254
G2		RHK	Positive	486
G3		DE	Negative	224
G4		CNQST	Polar	1250
G5		FWY	Aromatic	486
G6	Do not meet the above condition	All	Indefinite	3300

In this work, 6 strata representing the predominance of a type of amino acid were used. The quantity of ligands included in each stratum is shown in the last column, totalizing 8000 tripeptides

The SI_T parameter was selected as classifier in the algorithm because it did not depend on prior knowledge of scores for the entire population and the scores are not entirely reliable. This approach also has the advantage of favoring or penalizing sequence patterns in a manner not correlated with score values although maintaining the same trend.

Behavior of the algorithm "Select and Purge" (SP): algorithm SP versus SRS

The simple random sampling (SRS) method was used as a control and reference to benchmark the effectiveness of the SP algorithm. The SRS was modified so it could calculate the *SI* for the PAPs, in order to make the results comparable to those obtained with the SP.

The ligands were separated into different chemical families (Table 1) to evaluate whether the behavior of the SP algorithm implementation in its current form (using OEDocking for rigid-body molecular docking) was applicable to those families of ligands. Covering all possible functional groups and their combinations was not an objective; so, ligands in groups 10–26 with double or triple functional groups were combined into groups named Doubles and Triples, respectively.

The ability of each method to extract a significant amount of complexes from the population and assign them into any of the PP or PN groups was compared. VMin and VMax values were the criteria used to select the members of each group. The *SI* was used only for internal calculations in both algorithms, not to rank or refine the results.

Tables 3 and 4 show the performance of the two methods for both the training and test sets. Based on these results, the SP algorithm has a probability 2–3 times higher than SRS to sample the best complex from the population. The difference is even greater when analyzing the top 20 results sampled. The probability of finding at least 5 complexes, belonging to the PP group, among the top 20 sampled results is 6 to 8 times higher in the SP when compared to the SRS. Something similar occurs in the analysis for the worse results, although in this case the probabilities for the SP are almost halved. This happens because the algorithm favors the search for "good" complexes rather than the "bad" ones, which only serves as feedback for the algorithm to eliminate false positives. The higher variability in SP results can be observed for compounds in Thiolated and Nitrogenated groups, the later showing a predictive capability above expectations, based on training set results.

Figure 3 shows the probability of identifying at least 5 of the best receptors with both methods, in the course of 100 simulations for a randomly selected ligand. Even though the SP has a more disperse behavior than the SRS, it is at all times higher. The SRS instead shows little variation due to the random nature inherent to the method.

Table 3 Performance of SP and SRS algorithms for the training set

Groups	Best complexes				Worst complexes			
	P (Best)		*P* (PP_{20})		*P* (Worst)		P (PN_{20})	
	SP	SRS	SP	SRS	SP	SRS	SP	SRS
Alcohols	0.22	0.12	0.35	0.09	0.25	0.13	0.51	0.09
Aldehydes	0.31	0.13	0.63	0.09	0.27	0.12	0.45	0.08
Ketones	0.35	0.12	0.66	0.10	0.18	0.13	0.36	0.09
Esters	0.37	0.12	0.64	0.09	0.22	0.12	0.36	0.09
Halogenated	0.33	0.12	0.67	0.09	0.25	0.11	0.30	0.09
Aromatic	0.28	0.12	0.73	0.09	0.18	0.13	0.27	0.08
Cyclic	0.37	0.13	0.71	0.10	0.16	0.12	0.19	0.09
Thiolated	0.24	0.12	0.47	0.08	0.19	0.13	0.26	0.10
Nitrogenated	0.35	0.12	0.62	0.09	0.24	0.13	0.52	0.09
Doubles	0.32	0.12	0.64	0.09	0.21	0.13	0.35	0.09
Triples	0.38	0.12	0.70	0.10	0.21	0.13	0.40	0.09

The table shows the probabilities over 1000 simulations. *P* (Best) and *P* (worst) are the probabilities of sampling the best/worst complexes according to their scores, from the entire population. PP (PP_{20}) and PN (PN_{20}) account for the probability of finding at least 5 of the complexes among the top 20 and the worst 20 results, respectively, for each algorithm. Groups in Table 1 were condensed from 10–23 and 24–26 in the groups Doubles and Triples, respectively

Table 4 Performance of SP and SRS algorithms for test set

Groups	Best complexes				Worst complexes			
	P (Best)		*P* (PP_{20})		P (Worst)		*P* (PN_{20})	
	SP	SRS	SP	SRS	SP	SRS	SP	SRS
Alcohols	0.22	0.12	0.35	0.09	0.13	0.12	0.55	0.09
Aldehydes	0.31	0.12	0.60	0.09	0.21	0.12	0.43	0.09
Ketones	0.32	0.13	0.69	0.09	0.17	0.11	0.33	0.09
Esters	0.37	0.12	0.63	0.09	0.26	0.12	0.37	0.09
Halogenated	0.34	0.13	0.67	0.09	0.14	0.12	0.22	0.09
Aromatic	0.29	0.13	0.78	0.09	0.15	0.12	0.24	0.09
Cyclic	0.35	0.11	0.58	0.09	0.16	0.12	0.24	0.10
Thiolated	0.37	0.13	0.59	0.09	0.18	0.12	0.38	0.09
Nitrogenated	0.23	0.12	0.49	0.10	0.18	0.12	0.32	0.09
Doubles	0.33	0.12	0.68	0.09	0.18	0.12	0.31	0.09
Triples	0.15	0.12	0.62	0.09	0.17	0.13	0.35	0.09

Table caption reported in Table 3

Depending on the ligand studied, a slight variability can be observed in the SP algorithm performance. This is an interesting result and may be useful for future refinement of the method because it implies that the algorithm is sensitive to the chemical nature of the interactions. This is mainly a result derived from molecular docking methodology; however, creation of strata, criteria used for classification, and how the *SI* is calculated, may also influence the accuracy of the results.

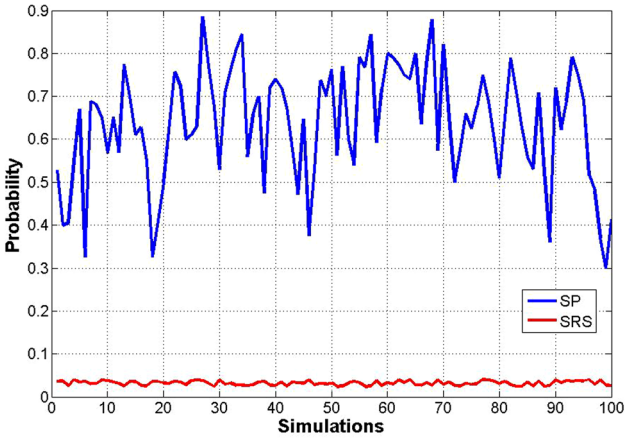


Fig. 3 Comparison between the curves representing the probability of finding at least 5 peptides of the PP group among the top 20 complexes sampled by SP (*blue*) and SRS (*red*) methods

However, the strength of the SP partially lies in its ability to sample a higher proportion of Actives when compared to the SRS. Figure 4 shows how frequently each method samples multiple Actives in the course of 1000 simulations versus 10 randomly selected ligands. The figure shows the behavior for a sample of approximately 11 % of the peptides available (two iterations of the SP method samples approximately 950 peptides) with both methods, using a 5 % cutoff value. Frequencies indicate that the SP method identifies a larger number of peptides with scores better than the VMin cutoff distance. The SP method has greater variability when it comes to the number of "good" receptors it is capable of identifying. However, for samples of the same size, the SP identifies an amount of prospective "good" receptors equal or greater than the amount identified by the SRS. For this graph, ligands were randomly selected, each one belonging to groups presented in Table 1.

The algorithm largely depends on the availability of time and computational resources. It deals with sample size selection and incorporates a sampling criterion in which sample size in a single iteration of the algorithm is always both manageable and representative of the population. The way samples are selected, combined with the purge of unlikely combinations of amino acids-positions offer a screening alternative in which the exploratory analysis of sample space is conducted toward the regions in which best results can be found, as shown schematically in Fig. 5.

In the iterative process of the SP algorithm (top 3 tiles from left to right), the samples are chosen according to strata populations. In each successive iteration, specific zones of sample space are marked as irrelevant (dark gray regions) and discarded in the sampling process. Sample space is thus reduced in each iteration increasing the probability of picking the correct elements (regions dotted in light gray) in the fol-

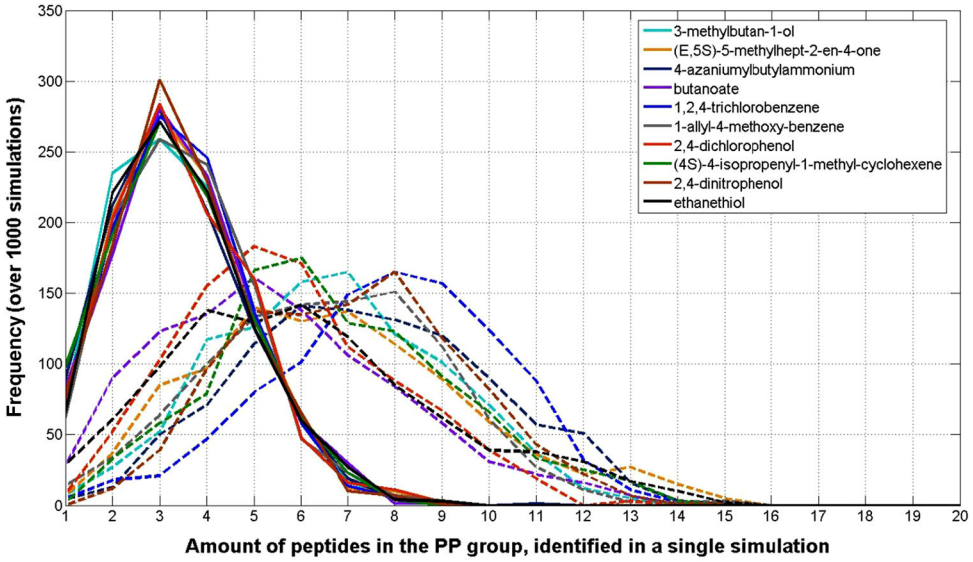


Fig. 4 Comparison of distributions of the top-ranked peptides among the 20 best-sampled results, identified by the SP method (*dashed lines*) and SRS (*solid lines*) over 1000 simulations for 10 randomly selected ligands. The behavior of SRS is much more homogeneous than the SP with a lower detection threshold (mode of 3 and 6 for SRS and SP, respectively)

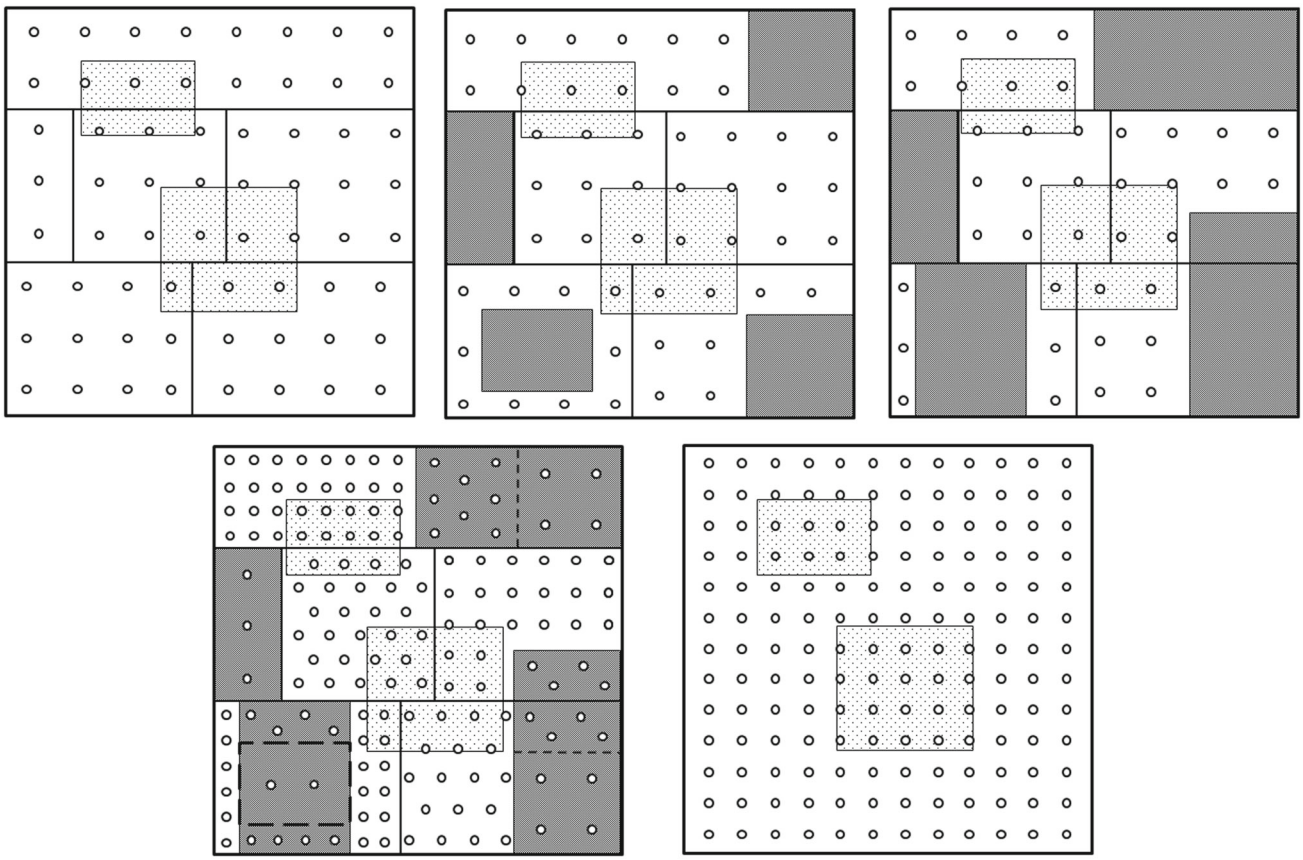


Fig. 5 Schematic comparison between SP and SRS exploration of sample space (*bottom*). The process represented in the figure corresponds to two iterations of the SP algorithm (*top*). The distribution of points is merely indicative

lowing iteration. The final distribution of sampled elements with the SP algorithm (bottom left tile), compared with an equivalent sample size using the SRS approach (bottom right tile), shows that sampling is more focused to regions with higher propensity to yield good results.

Measuring the performance of the SP algorithm: ROC curves and EF

Previous section establishes the ability of SP algorithm to select an actual subset of possible "good" receptors. How-

ever, this information is determined by a priori knowledge about the behavior of scores for the entire population. Such information in a prospective application of the method is not available. Instead, the only information available would be the scores calculated for each of the sampled complexes. In practice, scores alone cannot be used as the sole criterion to establish whether a receptor is appropriated or not, and an analysis with a higher level of theory for all possible candidates is not possible, given the number of systems to be analyzed and the complexity thereof. Because of this, a classifier parameter, calculated by the method itself was required. As mentioned before, in the scope of this algorithm, this classifier was the SI_T.

The ability of both algorithms (SP and SRS) to differentiate the best results from the rest was evaluated in each simulation. The classification was based on the scores obtained for each peptide in the corresponding association complex with a particular ligand. According to this, two classes were defined:

TP: True Positives (equivalent to Actives in traditional screenings), all complexes sampled belonging to the PP group, or with a score lower than Vmin for the initial population vs. the ligand analyzed.

NP: Non-Positives (equivalent to Inactives), the rest of the sample.

Figure 6 shows examples of ROC curves average behavior for both methods. In this case, 20 simulations were performed vs. ethanol, one of the ligands showing the higher deviations from global trends. Two iterations of the SP algorithm were performed in each case for Vmin = Vmax = 5 %. For the SRS, an equivalent sample size to that of the SP was used.

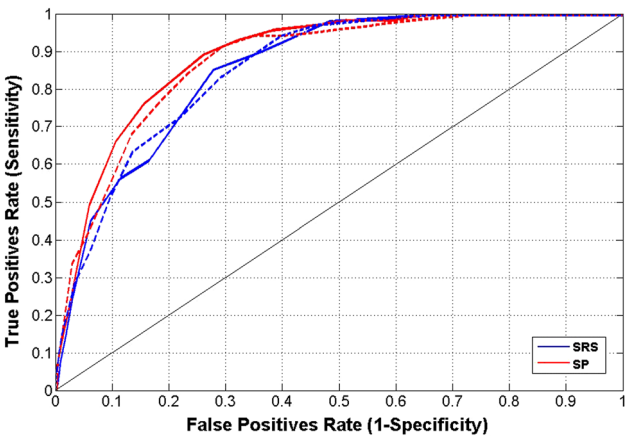


Fig. 6 Average ROC curves for 20 simulations with ethanol. The simulations were performed with both algorithms, the SP (*red*) and SRS (*blue*) using OEDocking (*continuous line*) and AutoDock Vina (*discontinuous line*). In both cases, the curves shape showed minor changes and are very similar, with AUC values comprised between 0.91 and 0.94

Simulations were conducted with OEDocking (rigid-body docking + molecular mechanics) and with AutoDock Vina (flexible docking + empiric scoring function).The first aspect to note in this graph is that both methods yield very similar curves. Furthermore, the shape of the curves is close to ideal behavior and experience minor variations when repeating the simulations under the same conditions even though the approaches used for energy estimation were changed.

All these factors indicate that both implementations of the SP achieve a high identification rate of TPs, maintaining a low rate of false negatives. However, a sensitivity greater than or equal to 0.8 is attained (averaged for all ligands studied) for values near 6 and 3 for SP and SRS methods, respectively (see Table 6). This indicates that although the shape of the curves is very similar, the SP method is more specific than the SRS, and complexes identified with this method have better scores than those identified by the SRS.

In terms of the areas under the curves, the average *AUC* for the system shown in Fig. 6, for example, is 0.907 and 0.895 for the SP and SRS, respectively, when using OEDocking, each with a standard deviation of 0.015. With AutoDock Vina, *AUCs* values were 0.900 and 0.887 for SP and SRS methods, respectively, and 0.02 standard deviation. The *AUC* experienced some variation when analyzing different ligands, but as can be seen in Table 5, it was not significant.

This shows that the performance of the SP algorithm does not depend on which method is used to assess the binding affinity, nor the accuracy of the calculations, or the practical application of simulated results. If the energy assessment method is able to correlate the amino acids types, their positions in the sequence, and the affinity of the resulting peptide for a given ligand, then the algorithm is capable of finding some of the best simulated results. The performance of the methods and the results obtained are not the same thing. The former refers to the likelihood of the methods to yield similar TP/FN ratio meanwhile the later are the actual sequence patterns, or PAP found to be favorable/unfavorable in the molecular association process. In this study, the results obtained for AutoDock Vina were poorer as expected due to that software parameterization. However, the algorithm performance barely varied when compared to the implementation with OEDocking as the docking software.

In Table 5, three important results can be observed: average values of *AUCs*, good reproducibility, and stability of results for both methods. The reproducibility of results is evident in the small deviations shown by the *AUCs* when repeating the simulations under the same conditions. Moreover, the methods seem to be very stable and behave similarly with different types of ligands. In this case, for 10 ligands (on behalf of their respective groups, reported in Tables 3 and 4), the variation in the corresponding *AUCs* is on the 3rd significant figure compared to the global average values for *AUC* for the SP and SRS methods, which are 0.907 and

Table 5 AUCs for 10 simulations with 10 ligands, representative of their corresponding groups

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Simulations	SP									
1	0.95	0.92	0.92	0.88	0.91	0.90	0.88	0.91	0.93	0.89
2	0.94	0.93	0.92	0.90	0.90	0.90	0.87	0.92	0.93	0.89
3	0.93	0.90	0.93	0.87	0.91	0.91	0.90	0.89	0.90	0.89
4	0.93	0.92	0.93	0.89	0.89	0.91	0.90	0.89	0.90	0.92
5	0.94	0.88	0.92	0.92	0.89	0.89	0.91	0.91	0.93	0.88
6	0.92	0.94	0.95	0.91	0.90	0.90	0.87	0.88	0.92	0.90
7	0.92	0.93	0.91	0.90	0.90	0.91	0.89	0.93	0.92	0.91
8	0.92	0.90	0.91	0.91	0.91	0.92	0.89	0.91	0.91	0.91
9	0.93	0.91	0.94	0.91	0.90	0.90	0.89	0.91	0.91	0.91
10	0.95	0.91	0.91	0.89	0.91	0.92	0.89	0.91	0.92	0.93
Average	0.93	0.92	0.92	0.90	0.90	0.90	0.89	0.91	0.92	0.90
SD	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02
Simulations	SRS									
1	0.91	0.93	0.90	0.90	0.88	0.89	0.88	0.91	0.88	0.91
2	0.92	0.90	0.93	0.91	0.87	0.88	0.90	0.93	0.92	0.91
3	0.96	0.88	0.92	0.90	0.88	0.89	0.89	0.92	0.90	0.91
4	0.89	0.91	0.91	0.88	0.88	0.87	0.89	0.91	0.90	0.90
5	0.92	0.90	0.89	0.90	0.86	0.90	0.90	0.91	0.89	0.91
6	0.92	0.88	0.93	0.89	0.88	0.88	0.87	0.92	0.88	0.92
7	0.92	0.92	0.91	0.90	0.87	0.90	0.91	0.92	0.90	0.87
8	0.95	0.90	0.91	0.87	0.86	0.88	0.88	0.93	0.93	0.91
9	0.93	0.89	0.90	0.88	0.87	0.90	0.89	0.93	0.92	0.91
10	0.95	0.91	0.93	0.84	0.85	0.90	0.86	0.93	0.90	0.92
Average	0.93	0.90	0.91	0.89	0.87	0.89	0.89	0.92	0.90	0.91
SD	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01

The mean AUC values for each ligand used and its standard deviation are reported. The data were calculated for Vmin = Vmax = 5 %, two iterations of the SP (or ≈10% of the population with the SRS). The ligands, L1… L10, correspond to the same ligands used in Fig. 4, in the same order

0.900, respectively. In all cases the values are above 0.85 which ensures its effectiveness as a classification algorithm. This does not necessarily imply that both algorithms perform equally. As mentioned before, their curves are similar, as well as their *AUCs* just because the threshold they use to obtain such results is the optimum for each method, that is, a SI_T of 6 and 3 for the SP and SRS, respectively, as presented in Table 6 and in Figure 4. However, if both methods were to run under the same threshold conditions, the curves will be completely different, as well as their *AUCs* and *EFs*.

Table 6 shows the averaged values over 100 simulations performed with the SP algorithm for a selection of 25 ligands with differences in terms of sizes and functional groups. The analysis of these results shows that, in general, the method behaves with small variations when applied to different systems. The value in column 5 indicates that a value of $SI_T = 6$ ensures that at least 80 % of the complexes belonging to PP group and present in the sample, will be correctly classified by the method as TP. The false positive rate (*i.e.*, NP misclassified as TP) is lower than 30 % in the worst case.

It can be easily verified by analyzing the values in Table 6 that the SP algorithm improves the ratio of TPs compared to the SRS. In fact, the first step in the stratified SP uses an SRS method for selecting the initial sample. The EF_p of this first step is close to one (Table 6, column 7). In the next iteration of the algorithm the information from previous iteration and *SI* (columns 8 and 9) are used, increasing to more than double the proportion of TPs found. The method has a total EF_p average of 1,745 (column 6). These data indicate that the SP algorithm is not only able to sample a greater number of TPs but is also able to classify them correctly with an acceptable margin of error. The EF_p represents how much the TP/NP ratio is improved when compared to the population. As discussed in the *Methods* section, this is not exactly the way this metric is defined because the population includes elements not even sampled by the algorithm. The traditional *EF* is plotted in Fig. 7.

Table 6 Summary of SP algorithm behavior

Gp.	Ligand	TPR	FPR	SI_T	$EF_p^{17\%}$	EF_{p1}	EF_{p2}	EF_{p3}
1	(3S)-3,7-dimethyloct-6-en-1-ol	0.90	0.21	6	1.63	1.03	2.30	1.78
1	Ethanol	0.86	0.29	4	1.50	0.97	1.73	1.38
2	Furan-2-carbaldehyde	0.92	0.15	5	2.08	0.98	4.16	1.39
2	Octanal	0.92	0.16	6	1.67	1.02	2.45	1.78
3	Acetone	0.90	0.20	5	1.47	0.94	1.88	1.76
3	Butane-2,3-dione	0.90	0.21	5	1.59	0.96	2.53	1.45
4	Ethyl acetate	0.90	0.22	5	1.62	0.99	2.49	1.61
4	Methyl propanoate	0.90	0.21	5	1.62	1.01	2.49	1.53
6	Naphthalene	0.92	0.17	6	1.93	1.02	3.51	1.53
6	2-Methoxy naphthalene	0.91	0.18	6	1.95	1.02	3.50	1.58
7	(4S)-1,4-dimethylcyclohexene	0.91	0.18	6	1.95	1.00	3.53	1.63
8	Methylsulphanyl methane	0.90	0.20	6	1.63	1.02	2.54	1.54
9	Pyridine	0.93	0.11	6	2.36	1.06	4.98	1.42
9	2-Nitroaniline	0.87	0.24	5	1.51	0.98	2.22	1.51
10	Hexane	0.89	0.22	6	1.54	1.01	2.30	1.49
11	2-Ethyl-3-hydroxy-pyran-4-one	0.89	0.22	5	1.51	0.99	2.28	1.42
12	o-Cresol	0.89	0.21	5	1.59	0.99	2.50	1.43
12	Phenol	0.88	0.24	5	1.43	0.97	1.97	1.45
14	Benzaldehyde	0.90	0.21	6	1.81	0.96	3.06	1.67
17	1-(2,3,4,5-Tetrahydropyridin-6-yl) ethanone	0.92	0.16	6	2.20	0.92	4.60	1.34
19	1,2,4-Trichlorobenzene	0.91	0.18	6	2.01	1.01	3.95	1.35
21	4-Chloroaniline	0.89	0.22	6	1.60	0.94	2.24	1.86
22	Nitrobenzene	0.89	0.22	5	1.87	0.99	3.42	1.45
24	2,4,6-Trichlorophenol	0.91	0.17	6	2.08	0.95	3.89	1.76
25	2-Nitrophenol	0.90	0.20	5	1.80	1.04	3.14	1.48
	Average	0.90	0.20	6*	1.76	0.99	2.95	1.54
	SD	0.02	0.03		0.25	0.03	0.86	0.15

The results correspond to the calculated 100 simulations for each of the 25 ligands average. This analysis was conducted for a Vmin = 5 % and three iterations of the algorithm.

* The reported value is the mode, not the average

Gp Group; *TPR* ratio of true positives; *FPR* ratio of false negatives; SI_T= Total Strength Index recommended setting; $EF_p^{17\%} = EF_p$ for the final result of the simulation, sampling 17% of the population; $EF_{p1-3} = EF_p$ for steps 1–3 of the algorithm

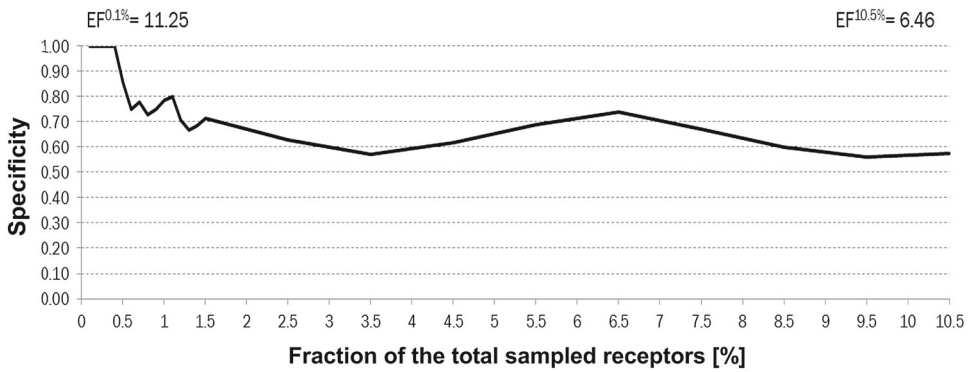


Fig. 7 Enrichment factor (EF) for 2,3,4,5,6-pentachlorophenol, with $EF_p^{17.7\%} = 1.78$ and TPR = 0.830

This figure represents the behavior of the TPs retrieved for a representative ligand that was chosen because of the similarity of its *EFp* and TPR with the overall behavior observed for all ligands analyzed and reported in Table 6. The values here are reported versus the actual sample size selected from the original 8000 peptides. The ratio of TPs is stable and

Table 7 Comparison between the effects of cutoff values and sample sizes for SP and SRS methods, in the AUCs for phenol

Cutoffs (%)	SP						SRS					
	2 Iter		3 Iter		5 Iter		≈10 %		≈15 %		≈25 %	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
2.0	0.90	0.03	0.90	0.02	0.88	0.02	0.87	0.04	0.87	0.03	0.89	0.02
3.0	0.88	0.03	0.88	0.02	0.87	0.02	0.86	0.04	0.87	0.02	0.87	0.02
4.0	0.88	0.03	0.88	0.02	0.87	0.02	0.86	0.03	0.87	0.02	0.86	0.02
5.0	0.87	0.03	0.87	0.02	0.87	0.01	0.85	0.02	0.85	0.02	0.85	0.02
7.5	0.87	0.02	0.87	0.01	0.86	0.01	0.85	0.02	0.84	0.02	0.83	0.01
10.0	0.85	0.02	0.86	0.01	0.85	0.01	0.83	0.02	0.83	0.02	0.82	0.01

over 0.5 even though 10 % of the total peptides selected in the sample are considered. The *EF* is also high, and close to the maximum *EF*. For example, at 10.5 % of the sorted list, for this particular ligand, the maximum *EF* is 9.35.

Effect of sample size and cutoff values

The values calculated in Table 6 were obtained using three iterations of the algorithm instead of the two used in the above calculations, sampling approximately 17 % of the population instead of the 10–11 % previously used. To further extend this analysis, the SP and SRS algorithms were tested by sampling 10, 15, and 25 % of the entire population of peptides. Also, as the initial classification of peptides in PP and PN groups can have a direct impact on the predictive ability of the method, and since the initial selection of Vmin and Vmax was arbitrary, some tests were performed with different cutoff values for Vmin and Vmax, at 2, 3, 4, 7.5, and 10 % of the total population. These results are exemplified in Table 7 for Phenol, selected as a worst case scenario since with this ligand both methods tend to classify a higher proportion of false positives when compared to overall trend.

In absolute terms, the larger the sample, the greater the number of TPs identified and vice versa. However, there is a risk of identifying higher number of false positives as well. This is true and is reflected in Table 7 in the *AUC* values, but the variation is so small that it is not statistically significant. The small differences observed in *AUCs* averages, the low standard deviations, and shape similarities in the curves indicate that both methods (SP and SRS) have a very stable behavior regardless of the type of ligand used, cutoff values, and sample sizes.

But despite this, the SP algorithm proved to be quite robust when considering that even when the cutoff criterion is fivefold increased, the *AUC* decrement is under 0.05. There is, however, a correlation between the AUCs and the cutoff values. For the selected ligands it was possible to determine that in 95 % of cases the relationship is given by the expression $AUC\,(vc) = aln\,(vc)+b$; where vc is the cutoff in percent; a and b are two parameters in the range $a = [-0.001, -0.003]$, $b = [0.850, 0.950]$ for the samples analyzed. This relationship must be subject of in-depth future research because the number of complexes analyzed must be even larger to make this result extensive to other possible systems. Nevertheless, our results show a regular and confident correlation and even though the tests were performed with tripeptides, peptide sizes should not pose a handicap to the application of the algorithm.

Conclusions

The proposed SP algorithm is more efficient than the SRS implementation as a sampling method for VS techniques. The stratified sampling approach along with the inclusion of *SI* doubled the probability of retrieving the highest-ranked peptide when compared to the SRS method. In addition, the overall probability of finding multiple top-ranked peptides in a single simulation using the SP algorithm is approximately seven times that of the SRS. Results obtained for training and test sets are consistent using both approaches.

The SP, however, showed small deviations from overall expected values throughout the diversity of ligands analyzed. This was expected because of the stochastic nature of the algorithm, but the trends observed indicate that chemical considerations somehow influence the outcome of the algorithm. Nevertheless, in all simulations performed, SP performance was always better than SRS, even though in a theoretical worst case scenario, they would yield similar results.

Using this new approach a more thorough search is conducted in some zones of the sample space at the expense of other zones discarded by the algorithm. Under these conditions, even the entire strata might be discarded. Purging the population in such a way reduces the number of remaining elements to be tested and increases the likelihood of retrieving relevant elements from the population. Results for longer sequences will most likely deviate from results currently obtained, in particular for peptide sequences with 7

or more amino acids because of the appearance of structural motifs. These effects were not considered in the initial model of the system. Also, the number of iterations (and in consequence, the volume of calculations) required to obtain similar results to the ones presented in this work may experience substantial variations. Despite all this, the algorithm presents a much more viable and efficient alternative to a traditional SRS approach for virtual screening, specifically in the computational design of peptide-based receptors from scratch when no information at all is available for the complementarity of interactions with the ligand studied.

References

1. Bajorath J (2002) Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1:882–894. doi:10.1038/nrd941
2. Bleicher KH, Böhm H-J, Müller K, Alanine AI (2003) A guide to drug discovery: hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov 2:369–378. doi:10.1038/nrd1086
3. Yang W, MD M, Hameed A, Hamza A, Zhan C-G, (2012) New inhibitor of 3-phosphoinositide dependent protein kinase-1 identified from virtual screening. Bioorg Med Chem Lett 22:1629–1632. doi:10.1016/j.bmcl.2011.12.121
4. Hamza A, Zhao X, Tong M, Tai H-H, Zhan C-G (2011) Novel human mPGES-1 inhibitors identified through structure-based virtual screening. Bioorg Med Chem 19:6077–6086. doi:10.1016/j.bmc.2011.08.040
5. Perez-Pineiro R, Burgos A, Jones DC, Andrew LC, Rodriguez H, Suarez M, Fairlamb AH, Wishart DS (2009) Development of a novel virtual screening cascade protocol to identify potential trypanothione reductase inhibitors. J Med Chem 52:1670–1680. doi:10.1021/jm801306g
6. Waszkowycz B (2008) Towards improving compound selection in structure-based virtual screening. Drug Discov Today 13:219–226. doi:10.1016/j.drudis.2007.12.002
7. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-based virtual screening for drug discovery: a problem-centric review. AAPS J 14:133–141. doi:10.1208/s12248-012-9322-0
8. Kuntz ID (1992) Structure-based strategies for drug design and discovery. Science 257:1078–1082. doi:10.1126/science.257.5073.1078
9. Jain AN (2004) Ligand-based structural hypotheses for virtual screening. J Med Chem 47:947–961. doi:10.1021/jm030520f
10. Ripphausen P, Nisius B, Bajorath J (2011) State-of-the-art in ligand-based virtual screening. Drug Discov Today 16:372–376. doi:10.1016/j.drudis.2011.02.011
11. Ebalunode JO, Dong X, Ouyang Z, Liang J, Eckenhoff RG, Zheng W (2009) Structure-based shape pharmacophore modeling for the discovery of novel anesthetic compounds. Bioorg Med Chem 17:5133–5138. doi:10.1016/j.bmc.2009.05.060
12. Yang SY (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov Today 15:444–450. doi:10.1016/j.drudis.2010.03.013
13. Braga RC, Andrade CH (2013) Assessing the performance of 3D pharmacophore models in virtual screening: how good are they? Curr Top Med Chem 13:1127–1138. doi:10.2174/1568026611313090010
14. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 47:409–443. doi:10.1002/prot.10115
15. Schneider G, Bohm HJ (2002) Virtual screening and fast automated docking methods. Drug Discov Today 7:64–70. doi:10.1016/s1359-6446(01)02091-8
16. Houston DR, Walkinshaw MD (2013) Consensus docking: improving the reliability of docking in a virtual screening context. J Chem Inf Model 53:384–390. doi:10.1021/ci300399w
17. Yuriev E, Ramsland PA (2013) Latest developments in molecular docking: 2010–2011 in review. J Mol Recogn 26:215–239. doi:10.1002/jmr.2266
18. Lee HS, Choi J, Kufareva I, Abagyan R, Filikov A, Yang Y, Yoon S (2008) Optimization of high throughput virtual screening by combining shape-matching and docking methods. J Chem Inf Model 48:489–497. doi:10.1021/ci700376c
19. Choi J, He N, Kim N, Yoon S (2012) Enrichment of virtual hits by progressive shape-matching and docking. J Mol Graph Model 32:82–88. doi:10.1016/j.jmgm.2011.10.002
20. Vyas VK, Ghate M, Goel A (2013) Pharmacophore modeling, virtual screening, docking and in silico ADMET analysis of protein kinase B (PKB beta) inhibitors. J Mol Graph Model 42:17–25. doi:10.1016/j.jmgm.2013.01.010
21. Kantardjiev AA (2012) Quantum. Ligand. Dock: protein-ligand docking with quantum entanglement refinement on a GPU system. Nucleic Acids Res 40:W415–W422. doi:10.1093/nar/gks515
22. Vilar S, Ferino G, Phatak SS, Berk B, Cavasotto CN, Costanzi S (2011) Docking-based virtual screening for ligands of G protein-coupled receptors: not only crystal structures but also in silico models. J Mol Graph Model 29:614–623. doi:10.1016/j.jmgm.2010.11.005
23. Umamaheswari A, Kumar MM, Pradhan D, Marisetty H (2011) Docking studies towards exploring antiviral compounds against envelope protein of yellow fever virus. Interdiscip Sci 3:64–77. doi:10.1007/s12539-011-0064-y
24. Chen H, Dong X, Zhou M, Shi H, Luo X (2011) Docking-based virtual screening of potential human P2Y12 receptor antagonists. Acta Biochim Biophys Sin (Shanghai) 43:400–408. doi:10.1093/abbs/gmr023
25. Baba N, Akaho E (2011) VSDK: virtual screening of small molecules using AutoDock Vina on Windows platform. Bioinformation 6:387–388
26. Zhang Q, Yu C, Min J, Wang Y, He J, Yu Z (2011) Rational questing for potential novel inhibitors of FabK from *Streptococcus pneumoniae* by combining FMO calculation, CoMFA 3D-QSAR modeling and virtual screening. J Mol Model 17:1483–1492. doi:10.1007/s00894-010-0847-9
27. Gharaghani S, Khayamian T, Keshavarz F (2011) A structure-based QSAR and docking study on imidazo[1,5-a][1,2,4]-triazolo[1,5-d][1,4,]benzodiazepines as Selective GABA(A) alpha5 inverse agonists. Chem Biol Drug Des 78:612–621. doi:10.1111/j.1747-0285.2011.01183.x
28. Lan P, Huang ZJ, Sun JR, Chen WM (2010) 3D-QSAR and molecular docking studies on fused pyrazoles as p38alpha mitogen-activated protein kinase inhibitors. Int J Mol Sci 11:3357–3374. doi:10.3390/ijms11093357
29. Wang F, Li Y, Ma Z, Wang X, Wang Y (2012) Structural determinants of benzodiazepinedione/peptide-based p53-HDM2 inhibitors using 3D-QSAR, docking and molecular dynamics. J Mol Model 18:295–306. doi:10.1007/s00894-011-1041-4
30. Dong B-L, Liao Q-H, Wei J (2011) Docking and molecular dynamics study on the inhibitory activity of N, N-disubstituted-trifluoro-3-amino-2-propanols-based inhibitors of cholesteryl ester transfer protein. J Mol Model 17:1727–1734. doi:10.1007/s00894-010-0881-7
31. Yui T, Shiiba H, Tsutsumi Y, Hayashi S, Miyata T, Hirata F (2010) Systematic docking study of the carbohydrate binding module protein of Cel7A with the cellulose 1alpha crystal model. J Phys Chem B 114:49–58. doi:10.1021/jp908249r

32. Wang J-C, Chu P-Y, Chen C-M, Lin J-H (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res 40:W393–W399. doi:10.1093/nar/gks496
33. Lauro G, Masullo M, Piacente S, Riccio R, Bifulco G (2012) Inverse virtual screening allows the discovery of the biological activity of natural compounds. Bioorg Med Chem Lett 20:3596–3602. doi:10.1016/j.bmc.2012.03.072
34. Chen YZ, Zhi DG (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 43:217–226
35. Hui-Fang L, Qing S, Jian Z, Wei F (2010) Evaluation of various inverse docking schemes in multiple targets identification. J Mol Graph Model 29:326–330. doi:10.1016/j.jmgm.2010.09.004
36. Kamper A, Apostolakis J, Rarey M, Marian CM, Lengauer T (2006) Fully automated flexible docking of ligands into flexible synthetic receptors using forward and inverse docking strategies. J Chem Inf Model 46:903–911. doi:10.1021/ci050467z
37. Grinter SZ, Lianga Y, Huanga S-Y, Hydera SM, Zou X (2011) An inverse docking approach for identifying new potential anti-cancer targets. J Mol Graph Model 29:795–799. doi:10.1016/j.jmgm.2011.01.002
38. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, Wang X, Jiang H (2006) TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 34(Suppl 2):W219–W224. doi:10.1093/nar/gkl114
39. Abdulhameed MDM, Chaudhury S, Singh N, Sun H, Wallqvist A, Tawa GJ (2012) Exploring polypharmacology using a ROCS-based target fishing approach. J Chem Inf Model 52:492–505. doi:10.1021/ci2003544
40. Padhy BM, Gupta YK (2011) Drug repositioning: re-investigating existing drugs for new therapeutic indications. J Postgrad Med 57:153–160. doi:10.4103/0022-3859.81870
41. Pavan S, Berti F (2012) Short peptides as biosensor transducers. Anal Bioanal Chem 402:3055–3070. doi:10.1007/s00216-011-5589-8
42. Hussain M, Wackerlig J, Lieberzeit P (2013) Biomimetic strategies for sensing biological species. Biosensors 3:89–107. doi:10.3390/bios3010089
43. Perez G, Mascini M, Sergi M, Del Carlo M, Roberta C, Montero-Cabrera LA, Compagnone D (2013) Peptides binding cocaine: a strategy to design biomimetic receptors. J Proteomics Bioinform 6:15–22. doi:10.4172/jpb.1000255
44. Perez G, Mascini M, Lanzone V, Sergi M, Del Carlo M, Esposito M, Compagnone D (2013) Peptides trapping dioxins: a docking-based inverse screening approach. J Chem 2013:1–8. doi:10.1155/2013/491827
45. Mascini M, Montesano C, Sergi M, Perez G, De Cicco M, Curini R, Compagnone D (2013) Peptides trapping cocaine: docking simulation and experimental screening by solid phase extraction followed by liquid chromatography mass spectrometry in plasma samples. Anal Chim Acta 772:40–46. doi:10.1016/j.aca.2013.02.027
46. Mascini M, Del Carlo M, Compagnone D, Perez G, Montero-Cabrera LA, Gonzalez S, Yamanaka H (2011) Multiple minima hypersurfaces procedures for biomimetic ligands screening. Sensors Microsyst 91:403–407
47. Chianella I, Karim K, Piletska EV, Preston C, Piletsky SA (2006) Computational design and synthesis of molecularly imprinted polymers with high binding capacity for pharmaceutical applications-model case: Adsorbent for abacavir. Anal Chim Acta 559:73–78. doi:10.1016/j.aca.2005.11.068
48. Chianella I, Lotierzo M, Piletsky SA, Tothill IE, Chen B, Karim K, Turner APF (2002) Rational design of a polymer specific for microcystin-LR using a computational approach. Anal Chem 74:1288–1293. doi:10.1021/ac010840b
49. Heurich M, Altintas Z, Tothill IE (2013) Computational design of peptide ligands for ochratoxin A. Toxins (Basel) 5:1202–1218. doi:10.3390/toxins5061202
50. Cannon EO (2012) New benchmark for chemical nomenclature software. J Chem Inf Model 52:1124–1131. doi:10.1021/ci3000419
51. Wlodek S, Skillman A, Nicholls A (2010) Ligand entropy in gas-phase, upon solvation and protein complexation. Fast estimation with Quasi-Newton Hessian. J Chem Theory Comput 6:2140–2152. doi:10.1021/ct100095p
52. Pedretti A, Villa L, Vistoli G (2004) VEGA-an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. J Comput Aided Mol Des 18:167–173. doi:10.1023/b:jcam.0000035186.90683.f2
53. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model 50:572–584. doi:10.1021/ci100031x
54. Mcgann MR (2011) FRED pose prediction and virtual screening accuracy. J Chem Inf Model 51:578–596. doi:10.1021/ci100436p
55. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem 31:455–461. doi:10.1002/jcc.21334
56. Cochran WG (2007) Sampling techniques. John Wiley & Sons, New York
57. Fuller WA (2011) Sampling statistics. John Wiley & Sons, New York
58. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47:488–508. doi:10.1021/ci600426e