J Comput Aided Mol Des (2009) 23:171–184
DOI 10.1007/s10822-008-9249-9

VSDMIP: virtual screening data management on an integrated platform

Rubén Gil-Redondo · Jorge Estrada · Antonio Morreale · Fernando Herranz · Javier Sancho · Ángel R. Ortiz

Received: 29 May 2008 / Accepted: 28 September 2008 / Published online: 22 October 2008

Abstract A novel software (VSDMIP) for the virtual screening (VS) of chemical libraries integrated within a MySQL relational database is presented. Two main features make VSDMIP clearly distinguishable from other existing computational tools: (i) its database, which stores not only ligand information but also the results from every step in the VS process, and (ii) its modular and pluggable architecture, which allows customization of the VS stages (such as the programs used for conformer generation or docking), through the definition of a detailed workflow employing user-configurable XML files. VSDMIP, therefore, facilitates the storage and retrieval of VS results, easily adapts to the specific requirements of each method and tool used in the experiments, and allows the comparison of different VS methodologies. To validate the usefulness of VSDMIP as an automated tool for carrying out VS several experiments were run on six protein targets (acetylcholinesterase, cyclin-dependent kinase 2, coagulation factor Xa, estrogen receptor alpha, p38 MAP kinase, and neuraminidase) using nine binary (actives/inactive) test sets. The performance of several VS configurations was evaluated by means of enrichment factors and receiver operating characteristic plots.

Keywords Docking · Virtual screening · Drug design · Database · Platform

Abbreviations

AChE	Acetylcholinesterase
fXa	Coagulation factor Xa
CDK2	Cyclic dependant kinase 2
Era	Estrogen receptor a
p38MAP	MAP Kinase P38
VS	Virtual Screening
EF	Enrichment Factor
ROC	Receiver Operating Characteristic

Introduction

Launching a new molecule to the market requires tremendous effort in research, development and money investment. Recent studies estimate in 15 years and around 800 million dollars the average time and cost per approved molecule [1]. After more than 30 years of using combinatorial chemistry and high-throughput screening (the two techniques that researchers thought would be the solution to the drug discovery bottleneck), the ratio between the number of new drugs obtained and the funds invested in their generation is well below the initial expectations [2, 3]. In some sense this has fuelled the development of theoretical techniques that attempt to accelerate the initial steps in the drug design cycle. Theoretical methods, if correctly derived, allow pinpointing the more promising candidates (hits) out of pools of thousands or even millions of molecules (chemical libraries). This reduced set can then be subjected to experimental analysis, and any promising compound can be subsequently optimized to attain the desired pharmacological profile and become a lead.

Ángel R. Ortiz deceased on May 5, 2008.

R. Gil-Redondo · A. Morreale (✉) · F. Herranz · Á. R. Ortiz
Unidad De Bioinformática, Centro De Biología Molecular
Severo Ochoa (CSIC-UAM), C/Nicolás Cabrera 1,
Campus De Cantoblanco, Madrid 28049, Spain
e-mail: amorreale@cbm.uam.es

J. Estrada · J. Sancho
Departamento de Bioquímica y Biología Molecular y Celular,
Facultad de Ciencias and BIFI –Instituto de Biocomputación y
Física de Sistemas Complejos, c/Pedro Cerbuna 12,
Universidad de Zaragoza, Zaragoza 50009, Spain

From a theoretical perspective different scenarios can be envisaged depending on the structural data available [4, 5]. The most favourable one is when the structures of both the target and the ligand(s) are known. In this case docking and VS techniques are the methods of choice [6]. More elaborate approaches based on molecular dynamics [7] or free energy perturbation coupled to thermodynamic integration can also be used [8]. The problem here is that the amount of time required to perform the calculations becomes prohibitive when a large collection of molecules is used.

The goal of docking is to identify, among the large number of possible orientations of a ligand within the binding site of the target, the one closest to the experimental structure of the complex [4]. This is done by using a mathematical function that accounts for the goodness of the coupling between ligand and target. Consequently, two key elements of the docking problem are: (a) a good sampling method, and (b) an accurate scoring function. VS is the extrapolation of docking to the case in which a large database of molecules is going to be processed [9]. Here the primary goal is somewhat different from that in docking because promising hit candidates are picked out from a database of mostly non-binders and no attempt is made to correctly classify all the molecules in the library or to identify all the actives.

Generating the experimental pose for a ligand is feasible with today's sampling techniques. However, positioning this pose at the top of a prioritized list of candidates is more problematic [10]. The main reason for this limitation is that the physical effects that describe the binding process are incompletely represented in the scoring function despite the fact that the underlying principles are reasonably well understood. In particular, the influence of the solvent, entropic effects and target flexibility are difficult to implement without compromising speed. Although many advances are being made to overcome these deficiencies, we are still at an early developmental stage.

Besides the problems outlined above, filtering millions of molecules using molecular descriptors and properties, docking them with one or more programs endowed with different accuracies, dealing with complex effects such as desolvation and/or protein flexibility at post-docking stages, etc. generates huge amounts of data that cannot be easily stored or utilized. It would then seem that the introduction of relational databases and potent database managing tools could be of aid in this regard.

Here we propose a flexible, fully automated computational platform to perform VS experiments that combines all the necessary steps to generate a short list of candidates starting from a database of 2D molecular structures. VSDMIP is intended to fill an existing gap in the docking and VS fields in relation to the storage and handling of the data. We are committed to making this platform available to interested parties so that the scientific community at large can benefit from it.

Methods

VSDMIP software and architecture

The VSDMIP architecture (Fig. 1) consists of (1) a database (in a multithreaded multi-user Structured Query Language [MySQL] DataBase Manager System), (2) a library of service interfaces and plugins, and (3) a set of workflows and its implementing commands. All small molecule data and VS results are stored in the VSDMIP database. The user controls the platform through different command line utilities and configures it using XML files. VSDMIP currently runs on Linux/x86 platforms (in our

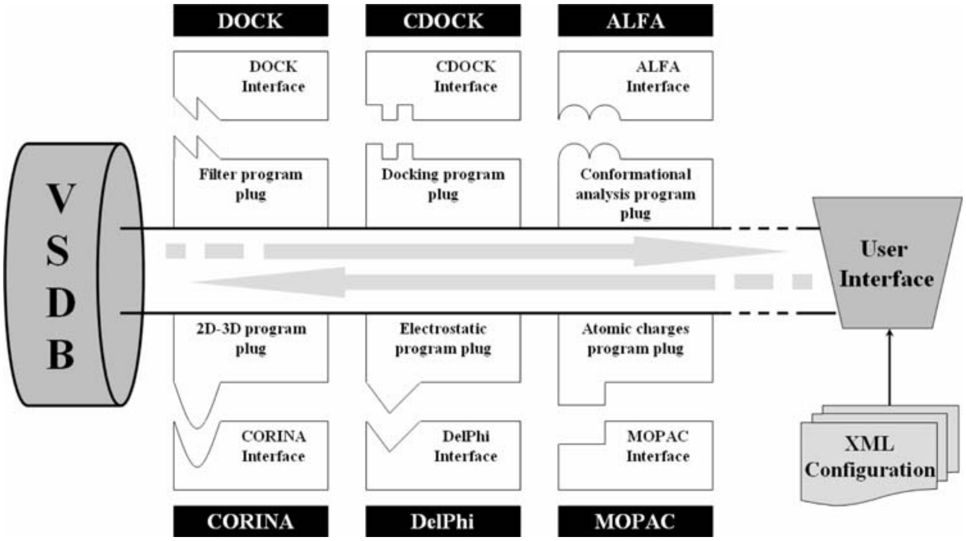

Fig. 1 Pictorial representation of the VSDMIP's architecture. VSDB refers to the Virtual Screening Data Base where data is stored

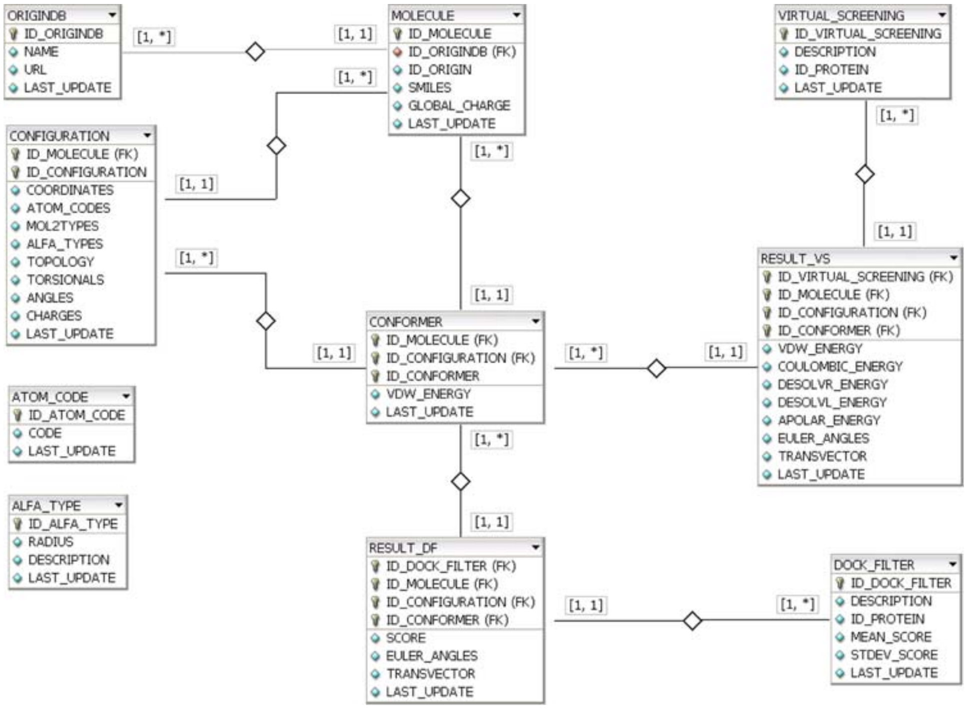

Fig. 2 Entity-Relationship schema of the database used by VSDMIP

case, a cluster of ten 2.4 and thirty-two 3.0 GHz Xeon Biprocessor CPUs with 2 GB RAM each).

Relational database

The database (Fig. 2) contains tables for the compound libraries (ORIGINDB, MOLECULE, CONFIGURATION, CONFORMER), results from filtering experiments (DOCK_FILTER, RESULT_DF), and results from VS experiments (VIRTUAL_SCREENING, RESULT_VS). Each compound library has its origin entered in an ORIGINDB entry. Each molecule has a MOLECULE entry with its global charge and a SMILES string representing its topology. The different stereoisomers of each molecule in combination with the different ring conformations are stored, together with a 3D structure, in a CONFIGURATION entry; discrete conformational changes on each of these entries are stored in the CONFORMER table, with a score resembling its internal energy (VDW_ENERGY). Filtering experiments are identified by a DOCK_FILTER entry. The best docking pose for each of the selected conformers included in the experiment is stored in a RESULT_DF entry, together with its score. These poses are stored as a translation and rotation of the conformer with respect to the original ligand coordinates stored in the database. More accurate docking experiments are identified by a VIRTUAL_SCREENING entry, and each individual conformer solution has an entry in RESULT_VS, with the same translation and rotation information as explained for RESULT_DF. Besides, the docking score can be stored with the details of its interaction energy terms.

VSDMIP software library

This C/C++ library manages the database, interconnects the different applications used, and offers biochemical and geometrical utilities. For each external application it is possible to add functionalities to the platform by creating an interface class and a storage class, under a common framework. The interface class provides methods for managing the application configuration attributes, preparing the execution of the application, performing the execution, managing errors and storing the results in the database. Six service interfaces currently exist: 3D structure generation, conformational analysis, atomic charge calculation, filtering, VS, and electrostatic calculations. Several plugins have been developed to interface with CORINA 3.0.5 [11] (3D generation), ALFA [12] (conformational analysis), MOPAC 7 [13] (atomic charges), DOCK 3.5 [14] and FRED 2.2 [15] (filter interface), CDOCK [16] and Autodock 3.0.5 [17] (VS interface), and for DelPhi 4 [18] and ISM [19] (electrostatic calculations).

Process workflows

The currently included commands allow inserting molecules into the database, docking a set of molecules within a protein binding site (for filtering or VS), rescoring the results of a docking experiment, and retrieving the results from a filtering or docking step. Each command uses one or more of the services plugged into VSDMIP and task execution is allowed to take place in a computer cluster.

Validation tests

Eleven different VS protocols (VSP) were designed and applied to nine different ligand datasets involving six protein targets. The protocols differ in several aspects: (i) the use of a filter and the number of molecules and conformers allowed to pass the filter, (ii) the docking engine used, and (iii) the scoring function(s) employed for ranking (Table 1). The active compounds for the validation tests (Table 2) were obtained from the binary (active/inactive) datasets available from CHEMINFORMATICS.ORG. Inactive compounds were randomly selected from a 9862 subset of the Maybridge Hit-Finder collection. Most of these molecules follow Lipinski's rule of 5 [20]. Table 3 shows the general properties of the database. The results were evaluated using receiver operating characteristic (ROC) plots [21], which represent the sensitivity (y-axis, true positives rate, see Eq. 1) versus (1—specificity) (x-axis, false positives rate, see Eq. 2), as well as areas under the ROC curves (AUC), enrichment factors (EF), and computing time.

Table 2 Relevant information for each of the datasets used in the virtual screening experiments

VS experiment	# of actives	# of inactives	Total
fXa (Fontaine)	432	500	932
fXa (Jacobsson)	127	500	627
fXa (Jorissen-Gilson)	50	500	550
AChE (Jacobsson)	54	1000	1054
CDK2 (Jorissen-Gilson)	50	1000	1050
ERa (Jacobsson)	142	1000	1142
ERa (Stahl)	50	1000	1050
Neuraminidase (Stahl)	17	1000	1017
p38MAP (Stahl)	22	1000	1022

$$\text{Sensitivity} = \frac{\text{TruePositives}}{\text{True Positives} + \text{False Negatives}} \quad (1)$$

$$\text{Specificity} = \frac{\text{TrueNegatives}}{\text{True Negatives} + \text{False Positives}} \quad (2)$$

EF represents the ratio of active compounds detected in a fixed percentage of the scanned database to the total number of actives in the whole database (Eq. 3),

Table 1 Virtual screening protocols used in the validation of VSDMIP

Virtual screening tests		
VSP id.	Description	Description
1	DOCK_100_1_CS	DOCK is performed on the 100 lower-energy conformers per molecule. The best conformer is selected using contact scoring.
2	DOCK_100_1_FF	DOCK is performed on the 100 lower-energy conformers per molecule. The best conformer per molecule is selected using force-field scoring.
3	DOCK_100_1_CS-XS	The same as in #1 but a final re-scoring is performed with XSCORE.
4	CDOCK_50_VDW_COUL	CDOCK is performed on the 50 lower-energy conformers per molecule. The best conformer per molecule is selected using the sum of van der Waals and coulombic interaction energies.
5	CDOCK_50_VDW	The same as in #1 but the final rank is done only with van der Waals interaction energies
6	CDOCK_50_ALL500	#4 is followed, and then for the 500 top-ranking molecules the solvation correction term is calculated using DelPhi. The final score is obtained by summing up the van der Waals and desolvation energies.
7	DOCK_100_1_CS + CDOCK_ZS3.0	#1 is followed, and then those compounds with ZScore $\geq$ 3.0 are submitted to CDOCK.
8	DOCK_100_10_CS + CDOCK_ZS3.0	The same as in #7 but now the best 10 conformer for each molecule are passed on to CDOCK.
9	DOCK_100_1_CS + CDOCK_ZS1.5	The same as in #7 but with ZScore $\geq$ 1.5.
10	DOCK_100_10_CS + CDOCK_ZS1.5	The same as in #8 but with ZScore $\geq$ 1.5.
11	DOCK_100_1_CS + CDOCK_ZS3.0_ALL500	#7 is followed, and then solvation correction using DelPhi is calculated for the first 500 molecules. The final score is obtained by summing up the van der Waals and desolvation energies.

Table 3 Mean and standard deviation of Lipinsky's properties for the Maybridge database ligands extended with the number of rotatable bonds

	MW[a]	HBA[b]	HBD[c]	RB[d]	logP[e]
Mean	311	4.82	1.12	4.58	2.65
Standard deviation	75.7	1.87	1.06	2.39	1.65

As calculated with OpenEye's Filter 2.0. The rules are MOL_WT [0, 500], ROT_BONDS [0, 20], LIPINSKI_DONORS [0, 5], LIPINSKI_ACCEPTORS [0, 10] and XLOGP [−5.0, 5.0]. 743 molecules failed Lipinski's test

[a] Molecular weight (in Da); [b] Number of hydrogen bond acceptors; [c] Number of hydrogen bond donors; [d] Number of rotatable bonds; [e] log of octanol/water partition coefficient

$$\mathrm{EF} = \frac{\{\mathrm{NAc}_{\mathrm{subset}}/\mathrm{NT}_{\mathrm{subset}}\}}{\{\mathrm{NAc}_{\mathrm{total}}/\mathrm{NT}\}} \tag{3}$$

where subset is the fixed percentage of the database, $\mathrm{NAc}_{\mathrm{subset}}$ is the number of active molecules found in the subset, $\mathrm{NT}_{\mathrm{subset}}$ is the total number of molecules in the subset, $\mathrm{NAc}_{\mathrm{total}}$ is the total number of active molecules in the entire database, and NT is the total number of molecules in the database. For comparative purposes we have computed the best EF found ($\mathrm{EF}_{\mathrm{best}}$), the maximum EF ($\mathrm{EF}_{\max}$) possible for each experiment, and the percentage of the database at which $\mathrm{EF}_{\mathrm{best}}$ is obtained. To calculate $\mathrm{EF}_{\mathrm{best}}$, only subsets with an $\mathrm{NT}_{\mathrm{subset}}$ multiple of b are considered, where b is max{10, 0.01 × NT}. This avoids artefacts in $\mathrm{EF}_{\mathrm{best}}$ at the start of database scanning.

Setup of small molecule databases

All molecules were first converted to isomeric Simplified Molecular Input Line Entry Specification (SMILES) [22] format, using the OpenEye OEChem 1.4 library. A 10,000 subset of the Maybridge HitFinder database was obtained in 2D Structure Data Format (SDF) [23]. The stereo information was not considered, and only those molecules bearing 3 or less stereogenic centres were selected. An additional set of 12 molecules had improbable connectivities and were also discarded by the VSDMIP insertion application, leaving a total of 9,862 molecules. The active datasets were prepared from SDF or SMILES input files. Duplicates (most likely arising from different stereoisomers devoid of the stereo information) were removed. When known, stereo and protonation information was preserved. Otherwise, protonation was assigned following OpenEye's Filter 2.0 rules for pH 7.4. Changes in bond order and number of hydrogens were done in some molecules to obtain standard valences. Using the final SMILES strings, the totally automatic process, ***insertVSDB***, inserted the molecules into the database after carrying out the following steps: (i) conversion from SMILES to 3D MOL2 using CORINA: up to 6 stereogenic centres were considered, ring conformations were generated, hydrogen atoms were added, and salt ions were removed; (ii) atomic charge calculations with MOPAC: single point calculations were performed with the MNDO semiempirical method [24] obtaining atom centred charges via electrostatic potential fitting techniques on each single structure provided from CORINA (one or more for each SMILE string depending on the number of stereogenic centres and ring conformations); (iii) atom type assignment and conformational analysis using ALFA.

Protein set up

All protein structures were obtained from the PDB and correspond to X-ray experiments with resolutions of 2.6 Å or better. The structures were chosen based on previous published VS and docking experiments: 1f0r (chain A) for coagulation factor Xa (fXa) [25, 26]; 1eve for acetylcholinesterase (AChE) [27, 28], 1e1x for cyclin-dependent kinase 2 (CDK2) [29, 30], 3ert for estrogen receptor alpha (Era) [31, 32], 1nsc (chain B) for neuraminidase [33, 34], and 1p38 for p38 MAP kinase [35, 36]. All HETATM records, including water molecules and ions, were removed from the PDB files. Side chains with missing atoms were rebuilt using SCWRL3 software[37]. In 1e1x, MODELLER v6.2 [38] was employed to reconstruct a missing loop. Then, the AMBER 8 [39] ff99 force field [40] was used to assign atom types and partial charges to each atom in the proteins, and hydrogen atoms were added assuming standard protonation states of titratable groups.

The H++ web server [41] was employed to study and assign protonation states for key interacting residues in the binding site (see below). The Poisson-Boltzmann (PB) method [42, 43] was used, at pH 7.4, 0.15 M salt concentration, and using internal and external dielectric constants of 4 and 80, respectively. Based on the information provided by H++, the following residues were protonated: HIS57 and ASP189 in 1f0r; HIS440, GLU278, and GLU443 in 1eve; HIS125 in CDK2; and ASH323 in 1nsc. The modified proteins were subjected to 10,000 steps of energy minimization (500 initial steps of steepest descent, followed by conjugate gradient), in vacuum, and only the hydrogen atoms were allowed to move. A further 10,000 steps of energy minimization were done with a Generalized-Born (GB) implicit solvent model [44–46] during which all atoms were allowed to move but heavy atoms were positionally restrained with a harmonic potential (100 kcal/molÅ^2). To check for consistency, the optimized structures were submitted again to the H++ web server. No significant changes were observed. The final minimized structures differed less than 0.05 Å (for Cα atoms) compared to the initial X-ray structure.

Binding site definition and characterization

For each protein structure, the initial binding site was defined as the space delimited by the axis-parallel box containing the co-crystallized ligand, augmented by 5 Å in each axis direction. Structure 1p38 was structurally aligned to PDB 1ywr [47], and the co-crystallized ligand found in 1ywr was used to define the binding site in the p38MAP studies. Protein interaction grids covering the binding site (1.0 Å spacing in all directions) were calculated for atom probes C, N, O, S, P, H, F, Cl, Br, and I. Each grid point represents the interaction between the protein and the probe atom as the sum of a van der Waals Lennard-Jones 12–6 potential and an electrostatic term modelled with a sigmoidal dielectric screening function [48]. For each binding site, a set of about 20 interaction points were defined to guide docking studies. These points were selected from the GAGA [49] centers of the gaussian sphere functions that best captured the interaction maps between the protein and benzene, water, and methanol molecules, as calculated by docking experiments using CDOCK. The set of points thus selected represent hydrophobic, hydrophilic and hydrogen bonding interactions. For docking using DOCK, the binding site was defined as the space delimited by the axis-parallel box containing the selected interaction points, augmented by 7.5 Å in each axis direction. The interaction grids used by DOCK and covering the binding site had a spacing of 0.3 Å. DOCK grids (see DOCK documentation) represent electrostatic, van der Waals and contact scores. For docking studies using CDOCK, the binding site was the same as that defined for DOCK, adding an extra 2.5 Å (for a total of 10 Å) in each axis direction. CDOCK grids had a spacing of 0.5 Å, and considered both protein interactions with different atom probes and electrostatic interactions, as explained above.

Filter plugin

For the tests shown in this article, DOCK 3.5 was used as a fast initial filter of the chemical libraries to be screened. DOCK uses a sphere-matching algorithm to fit ligand atoms to spheres in the binding pocket. We defined such points using GAGA as explained before. We chose to evaluate docking poses using the DOCK contact score in most cases and the DOCK force-field score in one case (VSP 2). For each conformer the single best DOCK solution obtained was retained. For VSP 1, 2, 3, 7, 9, and 11 DOCK scores for the best conformer of each molecule were stored; for VSP 8 and 10 the scores for the 10 best conformers of each molecule were kept. For VSP 3 the best solution per molecule was rescored using XSCORE [50] v. 1.2.1. but no improved performance was found and therefore, in the following, no further reference to this program will be made. Stored scores were normalized to ZScore (Eq. 4),

$$\mathrm{ZScore}_i = \frac{\mathrm{score}_i - \overline{\mathrm{score}}}{\sigma} \tag{4}$$

where $\overline{\mathrm{score}}$ is the mean and σ is the standard deviation values for the scores.

The VSDMIP application ***runDOCKFilter*** extracts information for all selected molecules from the database, performs the docking and stores the results in the database. Then ***getResultsFromDOCK*** extracts the docking results in a single coordinate file in MOL2 format for all the conformers with a ZScore higher than that provided by the user.

Docking plugin

We have used our in-house program CDOCK for the detailed docking phase of our protocols. Using the interaction energy grids calculated with CGRID, CDOCK exhaustively docks each molecule within the binding site. The centres of mass of the molecules are positioned on grid points equally spaced 1 Å, and discrete rotations of 27° on each axis are performed. A molecular mechanics force-field scoring function is used to score each pose (Eq. 5),

$$E_{\mathrm{MM}} = \sum_{i}^{\mathrm{prot}} \sum_{j}^{\mathrm{lig}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{D(r_{ij}) r_{ij}} \right] \tag{5}$$

where A_{ij} and B_{ij} are the van der Waals parameters of the atom types to which atom i and j belong, r_{ij} is the distance between the ith atom in the protein and the jth from the ligand, q_i and q_j are the partial charges of atom i and j, respectively. $D(r_{ij})$ is a sigmoidal dielectric function that accounts for solvent screening (Eq. 6),

$$D(r_{ij}) = \frac{\varepsilon + 1}{1 + \mathrm{k}e^{-\lambda(\varepsilon+1)r_{ij}}} \tag{6}$$

where ε is the dielectric constant for water (78.39), k is $(\varepsilon - 1)/2$, λ is $\alpha/(\varepsilon + 1)$ and α is 1.0367. The docking score for each pose (van der Waals plus electrostatic) is then calculated using a trilinear interpolation method.

The VSDMIP application ***runCDOCK*** extracts information for all selected molecules from the database, performs the docking and stores the results in the database. Then ***getResultsFromVS*** extracts the docking results in a single coordinate file in MOL2 format [51] containing the number of molecules defined by the user, and creates a text file with the score for each result.

Rescoring plugin

CDOCK scores were corrected with solvation energies (Eq. 7) obtained by solving the Poisson equation (electrostatic part, hereafter PB).

$$\Delta G_{\mathrm{desolv}} = \Delta G_{\mathrm{ele}} + \Delta G_{\mathrm{np}} \tag{7}$$

The electrostatic part of the energy (Eq. 8) is the sum of the electrostatic interactions between the ligand and protein in the complex (E_{ele}^{LR}), the change in solvation energy of the ligand upon binding ($\Delta G_{\mathrm{desolv}}^{L}$) and the change in solvation energy of the receptor upon binding ($\Delta G_{\mathrm{desolv}}^{R}$).

$$\Delta G_{\mathrm{ele}} = E_{\mathrm{ele}}^{LR} + (\Delta G_{\mathrm{desolv}}^{L} + \Delta G_{\mathrm{desolv}}^{R}) \tag{8}$$

The first term in Eq. 8 was computed as the product of ligand charges times the electrostatic potential created by the protein at each charge. The ligand desolvation energy was computed as the difference in energy between the solvated ligand and the energy of the ligand complexed to the uncharged receptor. The receptor desolvation term was computed analogously. All calculations were performed by numerically solving the linear Poisson equation using the finite difference method as implemented in DelPhi, PARSE atomic radii [52], AMBER ff99 partial charges for protein atoms, and MOPAC-calculated charges for ligand atoms, as described above. Each complex was immersed in a cubic box occupying 65% of the total volume with a grid spacing of 0.5 Å. The solute dielectric constant was set to 4 while that of the solvent was set to 80. The dielectric boundary was calculated using a solvent probe radius of 1.4 Å and a minimum separation of 11 Å was allowed between any solute atom and the box walls. The potentials at the grid points delimiting the box were calculated analytically by treating each charge atom as a Debye-Hückel sphere. The non-polar part of the desolvation (Eq. 9) was modelled as a linear relationship to the change of solvent-accessible surface area (SASA),

$$\Delta G_{\mathrm{np}} = a + b\Delta \mathrm{SASA} \tag{9}$$

where a is 0.092 kcal/mol, b is 0.00542 kcal/molÅ^2, and the change in SASA refers to the complex SASA minus the sum of that of the protein and the ligand alone. SASAs were calculated using the analytical method implemented in TINKER [53].

The VSDMIP application ***runDelPhiAndApolar*** extracts information for all selected molecules from the database, performs the calculation and stores the results as new entries into the database.

Visualization

Visualization of results is accomplished using the well-documented, free, open-source program Pymol [54].

Results

VSDMIP architecture and relational database

The architecture of VSDMIP is depicted in Fig. 1, and has been commented on, to some extent, in the Methods section. The main role played by the VSDMIP database is to act as a common origin and destination for all stages of the VS process. The XML configuration files and the different plug-in interfaces allow the user to configure different custom-made VS protocols, as exemplified below.

The Entity-Relationship schema of the database used by VSDMIP is shown in Fig. 2. The table primary keys are depicted with a key symbol, and foreign keys are marked with FK in brackets. The schema shows a compact way of storing and organizing chemical libraries and VS results through the MOLECULE, CONFIGURATION, and CONFORMER tables, and their relationships with RESULT_DF and RESULT_VS. Many different VS protocols can be composed by using as many DOCK_FILTERs (and their related RESULT_DF) and VIRTUAL_SCREENINGs (and their related RESULT_VS) as desired, and allow the results to be easily reused. Only the results from the last step are needed to run the following step in the workflow.

Performance of the individual docking tools (Tables 4–6)

DOCK performance is really poor for all but one system (ERa Jacobsson set), where it is marginally better than random selection. The results obtained using the force-field scoring function (VSP 2) are always better than those obtained with the contact scoring function (VSP 1), but the difference is not significant, the greater being around 0.3 U of AUC. This is also the range of variation in AUC values for different proteins with both scoring schemes. The EFs are very low and far from EF$_{\mathrm{max}}$ in all the cases and no significant differences are found when the force-field scoring function is used instead of the contact one. The best EFs are obtained for the ERa Jacobsson set (EF$_{\mathrm{best}}$ = 3.35). In most cases, the CDOCK AUCs are above 0.6, with the exception of the ERa receptor, for which results are below random (Jacobsson set) and slightly above random (Stahl set). In all the cases, except for the three fXa test sets, compound selection based only on van der Waals interaction energies (VSP 5) leads to smaller AUC values but this effect is not important. A noteworthy exception is the neuraminidase example where the electrostatic energy term improves the AUC by 0.47. Variations in AUC values across protein systems are significant (around 0.6 U) but independent from the scoring function (van der Waals plus electrostatic or van der Waals alone). For complete

CDOCK scoring (van der Waals plus electrostatic, VSP 4) EF are always better than when using DOCK (with both scoring functions, VSP 1 and 2) and above 10% of EF_{max}, reaching this value in one case (fXa Fontaine set), 80% (AChE), and 55% (ERa, Sthal set). If only the van der Waals term is used (VSP 5), the EF is greatly reduced, and this decrease is most dramatic in the AChE (30%), ERa (Sthal set, 20%), and neuraminidase (16%) examples.

Performance of mixed protocols: DOCK as a filter for CDOCK (Tables 4–6)

(a) Considering different ZScores. Filtering using DOCK, selecting molecules with a ZScore above 3.0, and then employing CDOCK (VSP 7) yields AUC values that are very close to those obtained from a random screen. No appreciable differences are observed when a more restrictive ZScore is used (1.5 vs. 3.0, VSP 9). The two exceptions in this case are AChE and neuraminidase. In the former, from an almost random performance (0.49) to a respectable value of 0.71, a difference of 0.22 units of AUC, while in the latter the difference is 0.14, thus reaching an AUC value close to 0.70. The differences across target proteins are of the same order (0.2 for VSP 7 and 0.3 for VSP 9). In only two cases the EF is above 50% of EF_{max} (ERa Sthal set 64% and fXa Fontaine set 70%) while the rest range between 4% and 35%. Small variations in EFs are observed for most of the sets when the lower ZScore value is used. These variations are more important in AChE (18% increase) and ERa (Sthal set, 28% decrease).

(b) Selecting 10 (instead of 1) conformations for each ligand from DOCK. When the ZScore is 3.0 (VSP 8) the AUC values are always above 0.5, the exception being the ERa example (Jacobsson set). Variations across targets are of the order of 0.3 AUC units. For a ZScore value of 1.5 (ID VSP 10) better results were obtained. In general, values for AUC above 0.6 are common, except for ERa (Jacobsson set) once again. Here, variations across the targets are as high as 0.5 AUC units. Three cases present EFs values above 50% of EF_{max} while the others range between 7% and 27%, when ZScore is 3.0. With the lower value, i.e. 1.5, the major differences are observed in AChE (18% increase), neuraminidase (14% increase), and ERa (Sthal set, 32% decrease).

Including solvent effect: PBSA as a correction term (Tables 4–6)

Molecular mechanics interaction energies (CDOCK scoring function) are corrected for desolvation effects using the PBSA method [55] on results directly obtained from CDOCK (VSP 6) or after a combined protocol encompassing DOCK and CDOCK (VSP 11). In the first case, the AUCs obtained are somewhat above random in many of the tests. The best AUC are for AChE and neuraminidase targets, while ERa (Jacobsson test) is the only case with AUC below random. In this case, the introduction of solvent effects via PBSA does not lead to an improvement over plain CDOCK results. Five sets show EFs above 30%, with fXa (Fontaine set) achieving EF_{max}. Except for AChE, the inclusion of desolvation always reduces the EF values. The reduction range goes from negligible (ERa, Jacobsson set, 2% or p38, 5%) to notorious (fXa Fontaine set, 30% or ERa Sthal set, 28%). In the second case, the AUC values

Table 4 Area Under the Curve (AUC), EF_{best}, EF_{max} for the three fXa sets obtained for eachVSP used

VSP id.	fXa								
	Fontaine set			Jacobsson set			Jorissen-Gilson set		
	AUC	EF_{best} (EF_{max} = 2.16)	log *t*	AUC	EF_{best} (EF_{max} = 4.94)	log *t*	AUC	EF_{best} (EF_{max} = 11)	log *t*
1	0.39	1.00 (0.99)	1.30	0.41	1.00 (0.99)	0.99	0.36	1.00 (1.00)	0.93
2	0.37	1.00 (1.00)	1.32	0.41	1.01 (0.96)	0.91	0.39	1.02 (0.98)	0.93
3	0.33	1.00 (1.00)	1.34	0.33	1.00 (1.00)	1.05	0.31	1.00 (1.00)	0.99
4	0.67	2.16 (0.01)	3.62	0.61	1.65 (0.05)	3.42	0.68	3.30 (0.04)	3.38
5	0.70	1.94 (0.01)	3.62	0.70	1.97 (0.05)	3.42	0.68	3.30 (0.02)	3.38
6	0.57	1.51 (0.14)	3.72	0.52	1.23 (0.22)	3.55	0.60	1.80 (0.20)	3.51
7	0.54	1.51 (0.01)	1.43	0.55	1.73 (0.06)	1.28	0.58	3.30 (0.02)	1.26
8	0.59	1.44 (0.03)	2.19	0.55	1.32 (0.05)	NA	0.65	2.57 (0.05)	1.99
9	0.58	1.58 (0.03)	1.76	0.56	1.38 (0.32)	1.53	0.61	2.93 (0.05)	1.46
10	0.64	1.43 (0.09)	2.58	0.57	1.29 (0.33)	2.37	0.65	2.48 (0.07)	2.31
11	0.54	1.25 (0.16)	2.51	0.54	1.48 (0.02)	2.29	0.58	1.83 (0.05)	2.29

Values in brackets refer to the percent (in decimal form) of the database scanned at which EF_{best} is found

Table 5 Area Under the Curve (AUC), EF_{best}, EF_{max} for the AChE and ERa sets obtained for each VSP used

VSP id.	AChE			ERa					
	Jacobsson set			Jacobsson set			Stahl set		
	AUC	EF_{best} (EF_{max} = 19.52)	log *t*	AUC	EF_{best} (EF_{max} = 8.04)	log *t*	AUC	$EF_{bes}t$ (EF_{max} = 21)	log *t*
1	0.30	1.08 (0.86)	0.84	0.56	2.01 (0.01)	0.90	0.28	1.00 (1.00)	1.24
2	0.32	1.02 (0.98)	0.85	0.60	3.35 (0.04)	0.91	0.57	1.73 (0.22)	1.23
3	0.26	1.02 (0.98)	1.17	0.36	1.03 (0.97)	0.99	0.17	1.00 (1.00)	1.29
4	0.97	15.97 (0.01)	3.09	0.35	1.00 (1.00)	3.02	0.55	11.45 (0.01)	3.13
5	0.94	9.58 (0.05)	3.09	0.33	1.00 (1.00)	3.02	0.52	7.64 (0.01)	3.13
6	0.93	15.97 (0.01)	3.46	0.47	1.00 (1.00)	NA	0.64	5.73 (0.01)	3.32
7	0.49	5.32 (0.01)	1.08	0.43	1.00 (1.00)	1.15	0.63	13.36 (0.01)	1.35
8	0.69	12.42 (0.01)	NA	0.42	1.00 (1.00)	1.74	0.66	17.18 (0.01)	1.80
9	0.71	14.20 (0.01)	1.25	0.43	1.00 (1.00)	NA	0.65	7.64 (0.01)	1.46
10	0.93	15.97 (0.01)	NA	0.45	1.00 (1.00)	NA	0.71	10.5 (0.02)	2.09
11	0.49	1.77 (0.01)	2.75	0.43	1.00 (1.00)	2.44	0.63	11.45 (0.01)	2.47

Values in brackets refer to the percent (in decimal form) of the database scanned at which EF_{best} is found

Table 6 Area Under the Curve (AUC), EF_{best}, EF_{max} for the CDK2, neuraminidase and p38MAP sets obtained for each VSP used

VSP id.	CDK2			Neuraminidase			p38MAP		
	Jorissen-Gilson set			Stahl set			Stahl set		
	AUC	EF_{best} (EF_{max} = 21)	log *t*	AUC	EF_{best} (EF_{max} = 59.82)	log *t*	AUC	EF_{best} (EF_{max} = 46.45)	log *t*
1	0.36	1.00 (1.00)	1.20	0.35	1.09 (0.87)	1.26	0.23	1.00 (1.00)	0.95
2	0.41	1.91 (0.01)	1.18	0.56	1.36 (0.39)	0.37	0.42	1.17 (0.31)	1.00
3	0.41	1.02 (0.98)	1.27	0.20	1.00 (1.00)	1.35	0.26	1.03 (0.97)	1.10
4	0.67	2.45 (0.07)	3.02	0.89	10.88 (0.01)	3.25	0.64	4.22 (0.02)	NA
5	0.63	2.12 (0.09)	3.02	0.42	1.06 (0.78)	3.25	0.60	4.22 (0.01)	NA
6	0.57	2.06 (0.14)	3.33	0.72	6.22 (0.08)	3.24	0.55	1.92 (0.12)	NA
7	0.55	1.91 (0.13)	1.28	0.55	10.88 (0.01)	1.40	0.65	4.75 (0.09)	1.36
8	0.63	3.82 (0.01)	1.67	0.63	5.44 (0.01)	1.96	0.73	4.22 (0.03)	2.02
9	0.59	2.67 (0.05)	1.40	0.69	10.88 (0.01)	1.56	0.72	4.22 (0.01)	1.57
10	0.61	2.10 (0.10)	1.95	0.82	13.6 (0.02)	NA	0.75	4.22 (0.01)	NA
11	0.55	1.91 (0.02)	2.54	0.53	1.81 (0.03)	2.56	0.64	3.75 (0.10)	2.60

Values in brackets refer to the percent (in decimal form) of the database scanned at which EF_{best} is found

are almost unaltered compared with the situation in which no solvent effects are introduced. Again, performance on the ERa target (Jacobsson test) was worse than random. In terms of variations across the targets, 0.46 AUC units are found when PBSA is applied to CDOCK without previous DOCK filter and 0.21 with the DOCK filter. The EFs here are similar to those commented on before, but only four sets are above 30% (instead of five) and none reaches EF_{max}. Again, a general reduction in these values is observed when desolvation is taken into account, the range being between 2% (p38) and 18% (AChE).

Finally, and related to the percentage of the database at which the best EF is found, almost the entire database has to be explored when DOCK is used. There are, however, some exceptions (6 out of 99): VSP 2 for ERa (Jacobsson and Sthal sets), CDK2, neuraminidase and p38MAP, and VSP 1 for ERa (Jacobsson test). For the rest of the protocols, the best EF is achieved very early for most of the cases, except in those for which over 25% of the database has to be screened: ERa (Jacobsson set, VSP 9 and 10), neuraminidase (VSP 5), and fXa (Jacobsson set, VSP 9 and 10).

Discussion

A plethora of methods are available to propose new drug candidates starting simply from 2D sketches of virtual chemical libraries. It should then be possible to integrate all the needed software elements to create customized workflows for any desired project. But the data flow between the different steps (input/output connections) seems to be an important problem mainly due to the great variety of formats that can be used to describe molecular structures. Although some advances have been done (e.g., SMILES and InChI [IUPAC International Chemical Identifier]), a consensus format is far from having been adopted. On the other hand, life sciences, in general, and computer-aided drug design, in particular, witness a data deluge coming from the target side, i.e., 3D protein structures available from structural genomics projects, as well as from the ligand side, i.e., huge chemical libraries with millions of molecules to be screened. Finally, the amount of data to be processed, stored and managed requires potent database engines. These three aspects have motivated us to develop VSDMIP as an integrative platform for handling these data. The main advantages of VSDMIP are: (1) the possibility to perform automated VS experiments; (2) easy comparison of different protocols; (3) total flexibility to design VS protocols; (4) the implementation of an XML mechanism to plug in new software pieces to customize protocols at will; and (5) the generation of a coupled relational database to have all the data organized and ready to use. VSDMIP presently lacks a graphical user interface (GUI), but it may be added in the future. When this is done, it will be able to additionally provide information regarding both the receptor and the ligand-binding site interactions. Small changes in the database schema will make it possible to store the results from docking engines that generate new ligand conformations as solutions, thus extending the current capability that basically works with docking engines that rely on pre-computed sets of conformers.

As similar approaches have been published recently, a brief discussion of some of them in comparison with VSDMIP is in order. Probably the most similar approach is that reported by SciTegic, Inc. named Pipeline Pilot [56, 57]. This commercial software uses the technology known as data pipeline to construct and execute customizable workflows using components that encapsulate mainly cheminformatics-based algorithms (although docking can also be performed using programs GOLD and FLEXX). Hassan et al. have shown the usefulness of this platform in a recent review [57], where they also show results from virtual screening experiments using Bayesian learning technique on several targets. It has an underlying database in common with VSDMIP, while performing many different types of calculations and methods.

Along the the same line, Astex Therapeutics [58] reported a proprietary web-based platform that integrates an ORACLE relational database to store molecules (from the Astex Technology Library of Available Substances [ATLAS] database), properties, target data, binding interactions, and results. VS experiments can be performed using compounds obtained directly from the database or created virtually thorough stored chemical reactions. Molecules in SMILES 2D strings are converted into 3D structures using CORINA and are docked with GOLD. It also contains a GUI to set up the experiments and visualize the results. VSDMIP has some added advantages such as its modularity, the possibility for users to generate their own chemical libraries, and the fact that it will be available to academic users upon request. On the down side, to the present day VSDMIP works through a command line interface.

The idea of employing user-configurable XML files to perform customized drug design-related workflows (as we do in VSDMIP) has been implemented by Lehtovuori and Nyrönen [59]. In their SOMA approach, the user decides which protocol to use by selecting, from a web browser interface, the steps and adequate parameters (even though in the reported implementation only molecular structure-based properties calculation and/or docking experiments with GOLD can be performed). Then program *Grape* manages the workflow by joining the needed applications in the order established by the user through the execution nodes (to run and manage the applications). Finally, the same web browser interface is used to retrieve, visualize and analyze the results. The output produced by each executable node is encoded, labelled (to keep track of the process at any time), and updated with the output information generated by the subsequent steps (depending on the selected protocol). Another important component is the toolkit that contains utility programs to perform intermediate tasks (file format conversions and generation of execution files, among others). The entire SOMA protocol is encoded within a unique XML file, which means that, after the required input has been provided, the user does not need to interact with the system until the entire workflow has been completed. This is very convenient for already established protocols. On the contrary, VSDMIP is more focused on decision making after every step. This is so because if more than one docking program is going to be used, or different filters are available, it is necessary to stop at each step to check the results before carrying on. A fine tuning at this level is not easily tractable nor is it feasible to implement in an automatic protocol. Another aspect that deserves to be commented on in relation to VSDMIP is the lack of a database to manage the results. SOMA stores the results of the entire workflow in a large single file that is displayed as a table within the interface. It is unclear whether the molecules and results already obtained could be used in another set of experiments. VSDMIP is totally flexible in this respect (i.e. once a molecule has been inserted into VSDMIP it can be reutilised as many times as desired). The availability of the SOMA XML schema and the web-implemented GUI makes this approach very attractive to drug designers with low programming skills who want to focus more on the chemical/biochemical aspects of the problem in hand rather than on the technical details. A final aspect in common with VSDMIP is the possibility to incorporate new applications in a relatively easy way using XML scripts.

Although not strictly comparable with VSDMIP, some other related tools also merit some comments. For example, the Autogrid/Autodock suite, which has attracted a lot of interest in recent years and is probably called to become one of the most widely used programs for docking purposes. Besides the automatic tools available at the developers' web site, two other applications for VS experiments have to be noted: BDT [60] and DOVIS [61]. The former allows the user to interact with the program code thorough a graphical front-end application that automatically performs grids preparation and their combination (to allow for receptor flexibility), docking computation and analysis of the results. The latter also incorporates an additional step for ligand preparation. The main advantage of DOVIS over BDT is that DOVIS allows docking to be performed in parallel using Linux clusters (with or without a queuing system). In both cases, however, the user is restricted to just one docking program and no database exists to manage different projects. Nonetheless, the free distribution of the programs and the easy-to-use graphical interfaces make them ideal tools for researchers who are more interested in getting answers to their particular problems than in the docking process itself.

To test the performance of our platform we used six different targets, and two of these with different sets of active compounds. The compilation of 11 VSP allows us to discuss some important effects in VS experiments, although more detailed studies will have to be conducted to assess the performance of other more sophisticated protocols. Three main questions are particularly addressed here: (1) the possible advantage of a combined docking protocol (using two programs, a first one as a filter, and a second more exhaustive one as a final docking tool) over a single one; (2) the effect of the number of molecules and conformers per molecule that are passed from the filter on to the final docking tool; and (3) the impact of incorporating desolvation using a continuum method as a rescoring function. For reference, the AUCs obtained for all the protocols applied to the fXa and neuraminidase test sets are depicted in Fig. 3 and the time employed per database molecule in these same cases is graphically shown in Fig. 4. As expected, better results are obtained when the VS is performed directly with CDOCK (compare VSP4–VSP6 with VSP1–VSP3) due to its superior scoring function and the exhaustive search within the binding site. More interesting, however, is the fact that when DOCK is used as a filter preceding CDOCK (VSP7–VSP11) reasonable results are obtained, which attests to its suitability to remove undesirable ligand structures. On the other hand, as clearly shown in Fig. 4, the best results are obtained for VSP4, which is also the most time-consuming protocol. On the contrary, VSP10 displays similar results compared to VSP4 but the computer time increases between 1 and 2 orders of magnitude. In other words, the use of a filter saves computational time while retaining most of the AUC values. The representative ROC plots depicted in Fig. 5 illustrate the aforementioned effect.

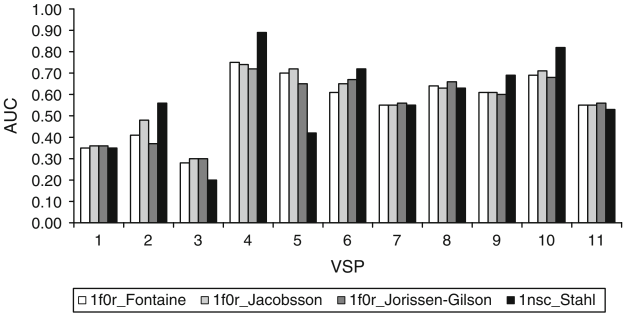

Fig. 3 Area Under the Curve (AUC) versus virtual screening protocol (VSP) identification number for the three sets of fXa and neuraminidase

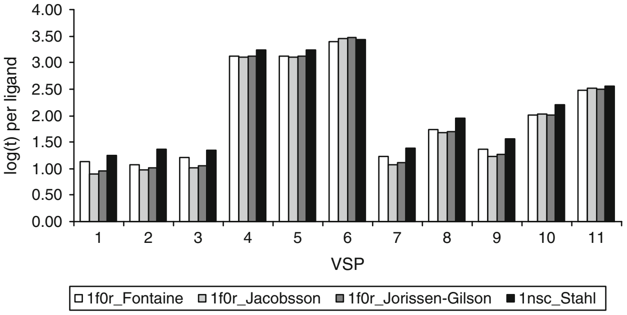

Fig. 4 CPU time (seconds, in logarithmic units) per ligand required for each virtual screening protocol (VSP) for the three sets of fXa and neuraminidase

Given the amount of molecules present in standard databases, it is not yet computationally feasible to conduct VS experiments using extremely accurate docking programs. Instead, a common approach is to use some sort of concatenated filters to reduce the number of molecules to be docked. In other studies, after initial filtering and docking, different scoring functions are employed and then candidates are selected on the basis of a consensus criterion [62]. More promising alternatives make use of more than one docking program in increasing order of accuracy [63, 64]. We chose DOCK as an initial docking program because it is fast enough to screen a molecule in a few seconds, the total time depending on the number of conformers and the number of spheres used to describe the

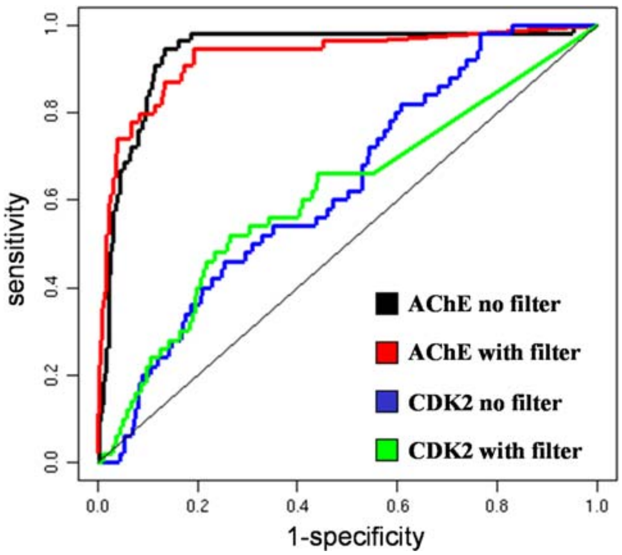

Fig. 5 ROC plots for AChE and CDK2 validation tests using VSP4 (without DOCK filter) or VSP10 (with DOCK filter)

binding site. In order to compare the relative performance of DOCK when used alone or in conjunction with another more exhaustive docking tool, an initial study was done considering DOCK alone. The performance of our configuration for DOCK is not better than random either with a contact- or a force-field-based scoring function, as can be seen from the AUC and EF_{best} values. We subjected our in-house docking program CDOCK to the same test and the results show that CDOCK clearly outperforms this particular DOCK configuration both in terms of AUC and EF_{best} values. As a trade-off for this increased accuracy, however, CDOCK employs more computer time. Another observation is that the elimination of the coulombic component does not really produce any variation in AUC or EF_{best} values, a result that is presently being scrutinized.

As stated before, a good approach consists of using a sequential combination of docking programs in increasing order of accuracy. Here we tested different protocols using first DOCK and then CDOCK. Variations among them relate to the number of compounds and conformations per compound to be passed from DOCK to CDOCK. In all cases, 100 conformers are used in DOCK. First, two cut-off values for the numbers of molecules passed to CDOCK are set using ZScore (3.0 and 1.5), DOCK contact scores and only one conformer per molecule. The AUC values do not seem to be dependent on ZScore, whereas the EF_{best} values show some degree of variation depending on the type of target. This means that although most of the molecules that survive after applying a low ZScore value are in fact inactive, they can be recognized and discarded by CDOCK. In general, the combination of both docking programs shows an intermediate degree of performance, as could be expected, between DOCK alone and CDOCK alone. Secondly, better results are obtained when instead of a single conformation 10 are passed on to CDOCK. This is a commonly observed effect, provided that conformational sampling has been performed adequately, and simply states that with more conformations per molecule the docking algorithm stands a better chance of detecting the correct pose for a binder [65]. Finally, use of the same ZScore but a different number of conformers per molecule (1 or 10) does not lead to any appreciable changes in AUC or EF_{best} values, the variation in the latter case depending on the type of target.

The third point addressed here is related to the introduction of solvent effects via PBSA as a post-scoring function. Together with flexibility issues [66], an adequate representation of the desolvation process that accompanies ligand-receptor binding is a major hurdle in VS studies. It is also a problem in traditional docking but, due to the fact that a small number of molecules are going to be studied, more elaborated solvation methods can be applied in this case, even when time consumption for such a calculation is often seen as a shortcoming. Methods based on PB or GB approximations are common but become impractical at the large scales a VS experiment requires although some promising approaches have already been published [19, 67, 68]. We have observed a small influence of solvent effects on AUC values using PB after CDOCK and no effect at all when used after the combination of DOCK and CDOCK. Again, the influence on EF_{best} is target-dependent, but in all cases a decrease in these values has been found.

VSDMIP has already proved successful in some recent scientific applications, such as one devoted to the discovery of new inhibitors of the DNA repair protein O6-alkylguanine DNA alkyltransferase [69]. Four compounds selected out of 3.5 million molecules from the ZINC database [70] showed acceptable in vivo and in vitro activities. In another example using in vivo screenings of a chemical combinatorial library and VSDMIP, we were able to develop small molecules that compete with ubiquitin E2 variant (UEV) for its interactions with ubiquitin-conjugating enzyme UBC13 and inhibit its enzymatic activity. The UEV–UBC13 complex is also implicated in mechanisms of DNA repair (unpublished results).

Conclusion

An integrated computational platform to perform VS experiments has been developed that includes an associated relational database which stores (i) molecules and molecular properties (energies, conformations, charges, etc.), (ii) results from docking filters, and (iii) final VS results. This procedure allows the inserted molecules to be reused in as many VS experiments as desired as well as the continued incorporation of new molecules. Also, it is easy and fast to create and allows a battery of analyses to be performed in order to test a particular VS protocol. The modular idea underlying its design is one of the stronger points, as the user is able to replace existing modules with new ones to create customizable protocols. Finally, and under development, is the idea to include protein set-up in an automatic fashion within the database, allowing the storage of geometrical and energetic characteristics of the binding site, which should serve as a classification tool for binding sites. The platform has been prepared as a bundled package to be distributed to the scientific community upon request from the authors [71]. In brief, all the programs implemented in the platform (except those that need to be purchased, by a modest prize, such as CORINA or DelPhi) are either free (MOPAC, DOCK, FRED, AutoDock) or will be released under a scientific/academic non-profit and non-commercial license as is the case for ALFA, CGRID, CDOCK, and ISM. The scripts to create the database structure as well as the XML configuration files will be also provided.

Acknowledgements Work at the CBM-SO was partially supported by a grant from "Comunidad de Madrid" thorough BIPEDD project (SBIO-0214–2006) and from "Ministerio de Educación y Ciencia" (BIO2005–0576). J.S. and J.E. were funded by grants BFU2007–61476/BMC (MEC, Spain) and PM076/2006 (DGA, Spain). J.E.'s research stage at CBM "Severo Ochoa" was funded by grants DGA (CONSI + D)/CAI (Spain) and FPU (MEC, Spain). J.E. is recepient of an FPU grant (MEC, Spain). J.E. thanks Alejandra Leo-Macías for help in using the MODELLER software. A.M. and R.G.-R. thank David Abia and Rubén Muñoz for technical support. We also acknowledge the generous allocation of computer time at the Barcelona Supercomputing Center. This work would not have been possible without the encouraging help of Ángel R. Ortiz, to whose memory this article is dedicated.

References

1. Smith A (2002) Nature 418:453
2. Lahana R (1999) Drug Discov Today 4:447. doi:10.1016/S1359-6446(99)01393-8
3. Ramesha CS (2000) Drug Discov Today 5:43. doi:10.1016/S1359-6446(99)01444-0
4. Perola E, Walters WP, Charifson PS (2004) Proteins 56:235. doi:10.1002/prot.20088
5. Warren GL, Andrews CW, Capelli AM et al (2006) J Med Chem 49:5912. doi:10.1021/jm050362n
6. Kitchen DB, Decornez H, Furr JR et al (2004) Nat Rev Drug Discov 3:935. doi:10.1038/nrd1549
7. Adcock SA, McCammon JA (2006) Chem Rev 106:1589. doi:10.1021/cr040426m
8. Brandsdal BO, Osterberg F, Almlof M et al (2003) Adv Protein Chem 66:123. doi:10.1016/S0065-3233(03)66004-3
9. Shoichet BK (2004) Nature 432:862. doi:10.1038/nature03197
10. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49:5851. doi:10.1021/jm060999m
11. Corina Molecular Networks (2000). GmbH Computerchemie Langemarckplatz 1, Erlangen, Germany. http://www.molecular-networks.com/software/corina/index.html. Accessed 24 Sept 2008
12. Gil-Redondo R (2006) Master Thesis: Implementación de una plataforma para el cribado virtual de quimiotecas. UNED, Madrid
13. Stewart JJ (1990) J Comput Aided Mol Des 4:1. doi:10.1007/BF00128336
14. Kuntz ID, Blaney JM, Oatley SJ et al (1982) J Mol Biol 161:269. doi:10.1016/0022-2836(82)90153-X
15. McGann MR, Almond HR, Nicholls A et al (2003) Biopolymers 68:76. doi:10.1002/bip.10207
16. Perez C, Ortiz AR (2001) J Med Chem 44:3768. doi:10.1021/jm010141r
17. Morris GM, Goodsell DS, Halliday RS et al (1998) J Comput Chem 19:1639. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B
18. Rocchia W, Sridharan S, Nicholls A et al (2002) J Comput Chem 23:128. doi:10.1002/jcc.1161
19. Morreale A, Gil-Redondo R, Ortiz AR (2007) Proteins 67:606. doi:10.1002/prot.21269
20. Lipinski CA, Lombardo F, Dominy BW et al (2001) Adv Drug Deliv Rev 46:3. doi:10.1016/S0169-409X(00)00129-0
21. Triballeau N, Acher F, Brabet I et al (2005) J Med Chem 48:2534. doi:10.1021/jm049092j
22. Weininger D (1988) J Chem Inf Comput Sci 28:31. doi:10.1021/ci00057a005
23. Ctfile Formats MDL (2007). Symyx, California. http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp. Accessed 24 Sept 2008
24. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:2338. doi:10.1021/ja00449a053
25. Maignan S, Guilloteau JP, Pouzieux S et al (2000) J Med Chem 43:3226. doi:10.1021/jm000940u
26. Murcia M, Ortiz AR (2004) J Med Chem 47:805. doi:10.1021/jm030137a
27. Jacobsson M, Liden P, Stjernschantz E et al (2003) J Med Chem 46:5781. doi:10.1021/jm030896t
28. Kryger G, Silman I, Sussman JL (1999) Structure 7:297. doi:10.1016/S0969-2126(99)80040-9
29. Arris CE, Boyle FT, Calvert AH et al (2000) J Med Chem 43:2797. doi:10.1021/jm990628o
30. Thomas MP, McInnes C, Fischer PM (2006) J Med Chem 49:92. doi:10.1021/jm050554i
31. Bissantz C, Folkers G, Rognan D (2000) J Med Chem 43:4759. doi:10.1021/jm001044l
32. Shiau AK, Barstad D, Loria PM et al (1998) Cell 95:927. doi:10.1016/S0092-8674(00)81717-1
33. Burmeister WP, Henrissat B, Bosso C et al (1993) Structure 1:19. doi:10.1016/0969-2126(93)90005-2
34. Murray CW, Baxter CA, Frenkel AD (1999) J Comput Aided Mol Des 13:547. doi:10.1023/A:1008015827877
35. Cavasotto CN, Abagyan RA (2004) J Mol Biol 337:209. doi:10.1016/j.jmb.2004.01.003
36. Wang Z, Harkins PC, Ulevitch RJ et al (1997) Proc Natl Acad Sci USA 94:2327. doi:10.1073/pnas.94.6.2327
37. Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) Protein Sci 12:2001. doi:10.1110/ps.03154503
38. Fiser A, Sali A (2003) Methods Enzymol 374:461. doi:10.1016/S0076-6879(03)74020-8
39. Case DA, Darden TA, Cheatham TE et al (2004) AMBER 8. University of California, San Francisco
40. Wang J, Cieplak P, Kollman PA (2000) J Comput Chem 21:1049. doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F
41. Gordon JC, Myers JB, Folta T et al (2005) Nucleic Acids Res 33:W368. doi:10.1093/nar/gki464
42. Honig B, Nicholls A (1995) Science 268:1144. doi:10.1126/science.7761829

43. Tishmack PA, Bashford D, Harms E et al (1997) Biochemistry 36:11984. doi:10.1021/bi9712448
44. Hawkins GD, Cramer CJ, Truhlar DG (1995) Chem Phys Lett 246:122. doi:10.1016/0009-2614(95)01082-K
45. Hawkins GD, Cramer CJ, Truhlar DG (1996) J Phys Chem 100:19824. doi:10.1021/jp961710n
46. Tsui V, Case DA (2000) Biopolymers 56:275. doi:10.1002/1097-0282(2000)56:4<275::AID-BIP10024>3.0.CO;2-E
47. Golebiowski A, Townes JA, Laufersweiler MJ et al (2005) Bioorg Med Chem Lett 15:2285. doi:10.1016/j.bmcl.2005.03.007
48. Mehler EL, Solmajer T (1991) Protein Eng 4:903. doi:10.1093/protein/4.8.903
49. Wang K, Murcia M, Constans P et al (2004) J Comput Aided Mol Des 18:101. doi:10.1023/B:jcam.0000030033.26053.40
50. Wang R, Lai L, Wang S (2002) J Comput Aided Mol Des 16:11. doi:10.1023/A:1016357811882
51. Tripos Mol2 File Format (2007). Tripos LP, Missouri. http://www.tripos.com/tripos_resources/fileroot/mol2_format_Dec07.pdf . Accessed 24 Sept 2008
52. Sitkoff D, Sharp KA, Honig B (1994) J Phys Chem 98:1978. doi:10.1021/j100058a043
53. Molecular Modeling Package TINKER (2004). http://dasher.wustl.edu/tinker. Accessed 24 Sept 2008
54. DeLano WL (2002). The PyMOL Molecular Graphics System DeLano Scientific, Palo Alto, CA. http://pymol.sourceforge.net. Accessed 24 Sept 2008
55. Kollman PA, Massova I, Reyes C et al (2000) Acc Chem Res 33:889. doi:10.1021/ar000033j
56. SciTegic, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA, http://accelrys.com/products/scitegic. Accessed 24 Sept 2008
57. Hassan M, Brown RD, Varma-O'brien S (2006) Mol Divers 10:283. doi:10.1007/s11030-006-9041-5
58. Watson P, Verdonk M, Hartshorn MJ (2003) J Mol Graph Model 22:71. doi:10.1016/S1093-3263(03)00137-2
59. Lehtovuori PT, Nyronen TH (2006) J Chem Inf Model 46:620. doi:10.1021/ci050388n
60. Vaque M, Arola A, Aliagas C et al (2006) Bioinformatics 22:1803. doi:10.1093/bioinformatics/btl197
61. Zhang S, Kumar K, Jiang X et al (2008) BMC Bioinformatics 9:126. doi:10.1186/1471-2105-9-126
62. Yang JM, Chen YF, Shen TW et al (2005) J Chem Inf Model 45:1134. doi:10.1021/ci050034w
63. Maiorov V, Sheridan RP (2005) J Chem Inf Model 45:1017. doi:10.1021/ci050089y
64. Miteva MA, Lee WH, Montes MO et al (2005) J Med Chem 48:6012. doi:10.1021/jm050262h
65. Knox AJ, Meegan MJ, Carta G et al (2005) J Chem Inf Model 45:1908. doi:10.1021/ci050185z
66. Teague SJ (2003) Nat Rev Drug Discov 2:527. doi:10.1038/nrd1129
67. Huang N, Kalyanaraman C, Irwin JJ (2006) J Chem Inf Model 46:243. doi:10.1021/ci0502855
68. Kuhn B, Gerber P, Schulz-Gasch T (2005) J Med Chem 48:4040. doi:10.1021/jm049081q
69. Ruiz FM, Gil-Redondo R, Morreale A (2008) J Chem Inf Model 48:844. doi:10.1021/ci700447r
70. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45:177. doi:10.1021/ci049714+
71. Gil-Redondo R, Estrada J, Morreale A, et al. (2008). VSDMIP. CBM "Severo Ochoa" (CSIC-UAM) and Universidad de Zaragoza, Spain. http://ub.cbm.uam.es/VSDMIP.htm. Accessed 24 Sept 2008