J Comput Aided Mol Des (2014) 28:455–462
DOI 10.1007/s10822-014-9707-5

Virtual screening of the SAMPL4 blinded HIV integrase inhibitors dataset

Claire Colas · Bogdan I. Iorga

Received: 18 November 2013 / Accepted: 10 January 2014 / Published online: 24 January 2014

Abstract Several combinations of docking software and scoring functions were evaluated for their ability to predict the binding of a dataset of potential HIV integrase inhibitors. We found that different docking software were appropriate for each one of the three binding sites considered (LEDGF, Y3 and fragment sites), and the most suitable two docking protocols, involving Glide SP and Gold ChemScore, were selected using a training set of compounds identified from the structural data available. These protocols could successfully predict respectively 20.0 and 23.6 % of the HIV integrase binders, all of them being present in the LEDGF site. When a different analysis of the results was carried out by removing all alternate isomers of binders from the set, our predictions were dramatically improved, with an overall ROC AUC of 0.73 and enrichment factor at 10 % of 2.89 for the prediction obtained using Gold ChemScore. This study highlighted the ability of the selected docking protocols to correctly position in most cases the *ortho*-alkoxy-carboxylate core functional group of the ligands in the corresponding binding site, but also their difficulties to correctly rank the docking poses.

Keywords Virtual screening · Docking · Scoring function · HIV integrase inhibitors · SAMPL4 blind challenge

Introduction

The Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenges represent unique opportunities for the molecular modeling community to evaluate, in "blind" conditions, the performance of the computational chemistry tools and methods currently available. Two of the previous editions featured a protein-ligand binding prediction component: SAMPL1 (2008) was focused on the pose prediction on kinases, and SAMPL3 [1, 2] (2011) involved the virtual screening of a fragment-like dataset of bovine trypsin inhibitors.

The data set provided for the virtual screening SAMPL4 challenge [3] in 2013 consisted of 321 compounds that are potential binders to the HIV-1 integrase (for which we will use the term "HIV integrase" throughout the paper) [4]. This protein features several binding sites, and the SAMPL4 challenge was focused on three of them, the so called LEDGF/p75 (for which we will use the term LEDGF throughout the paper), fragment and Y3 sites (Fig. 1) [5–7]. After the removal of some problematic and duplicate structures, the SAMPL4-VS dataset was reduced to 305 compounds, which were used for the evaluation of submissions. This dataset proved to be particularly challenging, the main difficulties being related to the high structural

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9707-5) contains supplementary material, which is available to authorized users.

C. Colas · B. I. Iorga (✉)
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Centre de Recherche de Gif-sur-Yvette, Labex LERMIT, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France
e-mail: bogdan.iorga@cnrs.fr

Present Address:
C. Colas
Department of Pharmacology and Systems Therapeutics, Tisch Cancer Institute, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

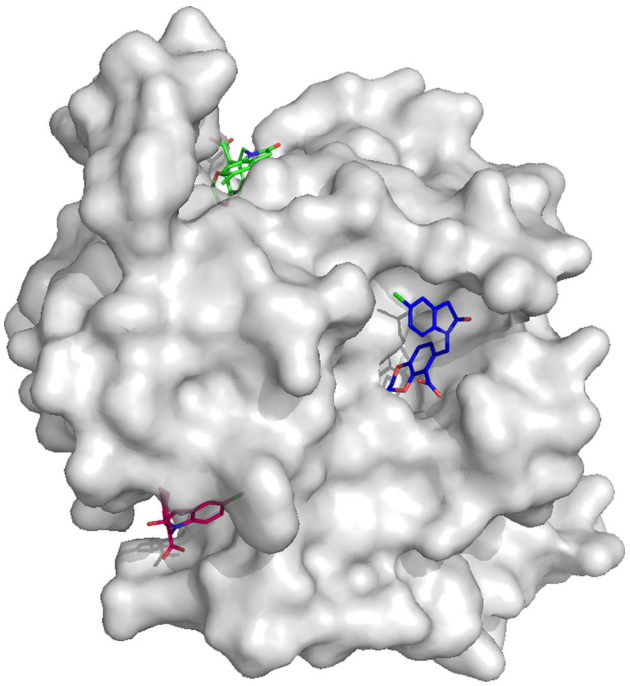

Fig. 1 Surface representation of the X-ray structure of HIV integrase protein (PDB code 3NF8) with the CDQ ligand present in the three binding sites considered in this study: LEDGF (*magenta*), fragment (*blue*) and Y3 (*green*)

similarity between active and inactive compounds, but also to the presence of three different binding sites [3].

Methods

In this work, we followed a protocol similar to those reported for our SAMPL3 virtual screening study of bovine trypsin inhibitors [8]: (1) generation of a training set of HIV integrase ligands from the structural data available in the Protein Data Bank (PDB) [9, 10]; (2) evaluation of the performance of different docking software and scoring functions on HIV integrase using this training set; (3) use of the best docking parameters identified in the previous step to predict the binding properties of the SAMPL4-VS dataset.

Generation of a training set of HIV integrase ligands from the structural data available

In the first step, the available structures for the protein-ligand complexes of HIV integrase were downloaded from the PDB. The complete list of these structures is presented in Table S1 (Electronic Supplementary Material). The PDB codes for the ligands present within these structures were identified, together with their distribution on the different HIV integrase binding sites (Table S1). The SMILES strings corresponding to the structures of these ligands were downloaded from the PDB Ligand Expo (http://ligand-expo.rcsb.org/) and were converted into 3D using CORINA version 3.44 (http://www.molecular-networks.com). Their protonation state at pH 7.0 was assigned using LigPrep module from the Schrödinger Suite (http://www.schrodinger.com/). This procedure that we used to generate the training set of HIV integrase ligands, although longer and more complex, was preferred over the direct extraction of ligand coordinates from the PDB structures, in order to avoid any bias related to the initial position of the ligand in the docking process.

Choice of docking software and scoring parameters

The training set of HIV integrase ligands generated previously was then used to identify the docking software, scoring functions and docking parameters that are the most appropriate for dealing with this SAMPL4-VS challenge. In addition to the correct positioning of ligands in the binding site, we were interested to evaluate the ability of the docking software to discriminate between the LEDGF, Y3 and fragment sites of HIV integrase.

The ligands from the training set were docked in each of these three binding sites using the HIV integrase three-dimensional structure provided by the SAMPL4 organizers (PDB code 3NF8 [11]) and six different combinations of docking software and scoring function: (1) Autodock 4.2 [12] with the default scoring function, (2) Glide 5.7 (http://www.schrodinger.com/) with the SP scoring function, and Gold 5.2 [13] with the (3) ASP, (4) ChemPLP, (5) ChemScore and (6) GoldScore scoring functions. It should be noted that normal docking (not virtual screening) parameters were used, as our previous studies [8] showed that these parameters provided better results, due to improved conformational sampling within the binding site. Of course, the computational cost was more important using these conditions, but it was still affordable for the number of compounds evaluated for this challenge.

The analysis of docking poses and ROC curves, for each individual binding site and on the ensemble of the three sites, allowed us to choose two docking software/scoring function combinations that were the most efficient on HIV integrase: Glide with the SP scoring function and Gold with the ChemScore scoring function.

Virtual screening of SAMPL4-VS compounds

In the next step, the 321 ligands from the SAMPL4-VS dataset [4] (MOL2 structures provided by the SAMPL4 organizers which were used directly, without any additional preparation) were docked in all three sites using the

two docking software/scoring functions selected above. In both cases, the docking results for the three sites were fusioned and the best score over the three sites was retained for each compound. The data was formatted according to the template file provided.

Graphics

Figures were generated using PyMol (http://www.pymol.org/) and the ROC plots using the XmGrace package (http://plasma-gate.weizmann.ac.il/Grace/). The chemical structures were drawn with ChemDraw version 13 (http://www.cambridgesoft.com/).

Results and discussion

Choice of docking software and scoring parameters

Empirical ROC AUCs were calculated for all combinations of the six docking software/scoring functions with the three binding sites, which showed that a different docking software is more appropriate for each binding site. Gold with ChemScore, GoldScore and ChemPLP behaves very well for the LEDGF site, with ROC AUCs around 0.75 (Figure S1). The binding into the fragment site was very difficult to predict for most of docking software/scoring functions (ROC AUC less than 0.5) with the exception of Glide SP (ROC AUC 0.64) and Gold ASP (ROC AUC 0.58) (Figure S2). Autodock behaves particularly well (ROC AUC 0.87) for the Y3 site, the others providing acceptable results (ROC AUC ranging from 0.66 to 0.81) (Figure S3). However, the submission format of the SAMPL4-VS challenge required a global prediction over all the three sites, which cannot be obtained using a different tool for each site. In these conditions, empirical ROC AUC were also calculated for all docking software/scoring functions on the ensemble of the three sites (Fig. 2). Glide SP gave the best global prediction (ROC AUC 0.72), followed by Gold ChemScore (ROC AUC 0.60). It should be noted that with Gold ChemScore all the first 30 % of ranked compounds were correctly identified as actives. The other docking software/scoring functions provided less interesting predictions (ROC AUC around 0.57–0.59, but especially with a poor prediction profile for the first 30 % of ranked compounds).

Virtual screening of SAMPL4-VS compounds

Considering the results obtained in the previous step, only two protocols (Glide SP and Gold ChemScore) were chosen to perform the virtual screening of the SAMPL4-VS dataset, using the same docking parameters as for the training set, and therefore two predictions were submitted for this challenge.

Soon after the submission deadline of the SAMPL4-VS challenge, the organizers revealed that there were 55 binders, out of the 305 compounds present in the final, cleaned dataset. The chemical structures of these binders are presented in Fig. 3 and their distribution over the three sites, as well as our predictions for these compounds, are presented in Table S2. It can be seen that most of compounds (48) bind exclusively in the LEDGF site, two bind in both LEDGF and Y3 sites and one exclusively in the Y3 site. There were also four compounds that bind in the fragment site.

Using the two protocols mentioned above, we were able to predict correctly 11 and 13 binders out of 55 (the total number of binders), which means 20.0 and 23.6 % of correct predictions, respectively (Fig. 3, Table S2). Only 5 of them were predicted by both protocols, and this fact denotes some specificity of each docking software for different subfamilies from the SAMPL4-VS dataset, whereas some other subfamilies are not recognized by any of them. All of the correctly predicted compounds were found to bind in the LEDGF site, which is in agreement with the results obtained with the training set. However, it should be noted that a random selection of 55 compounds from the SAMPL4-VS dataset would return on average 10 binders, which is very close to the 11 and 13 binders detected by the two selected docking-scoring protocols. This was also confirmed by the negligible enrichment factors that we obtained for our submissions (see below).

The overall ROC AUCs obtained for the ensemble of the three sites (Fig. 4) are rather disappointing, with values around 0.56, which are much lower than those obtained for the training set. This fact might be explained by the important structural differences between the compounds that were present in our training set and in the SAMPL4-VS dataset, as well as by the limited number of binders on Y3 and fragment sites within the training set. It is noteworthy that the early detection of binders was also lower for the SAMPL4-VS dataset (enrichment factors at 10 % of 1.25) compared to the training set (enrichment factors at 10 % of 3.5–4.0).

However, when the analysis of the SAMPL4 virtual screening challenge was carried out by removing all alternate isomers of binders from the set (the size of the set to be analyzed is therefore reduced from 305 molecules to 189, see [3] for the reasoning behind this alternate analysis), our results were dramatically improved, with an overall empirical ROC AUC of 0.73 and enrichment factor at 10 % of 2.89 for the prediction obtained using Gold ChemScore (Fig. 5). These values are closer to those

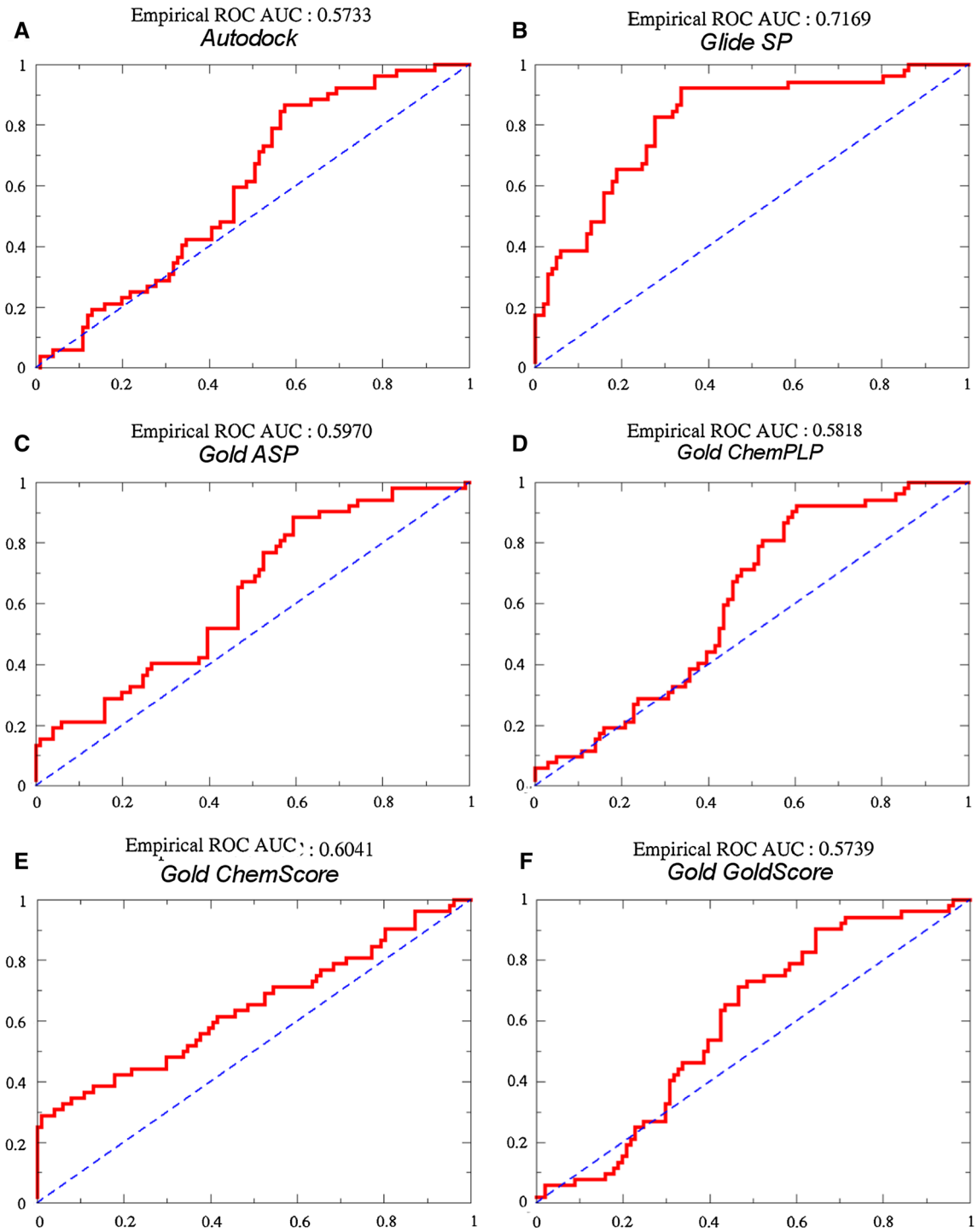

Fig. 2 ROC curves for the training set docking on the ensemble of the three sites, using different docking software and scoring parameters: Autodock (**a**), Glide with the SP protocol (**b**), and Gold with ASP (**c**), ChemPLP (**d**), ChemScore (**e**) and GoldScore (**f**) as scoring functions

obtained for the training set and rank our prediction with Gold ChemScore on the second position of the SAMPL4-VS challenge with this alternate analysis [3].

The docking conformations of the correctly predicted compounds are shown in Fig. 6, whereas the binders that were not correctly predicted are depicted in Figure S4. It

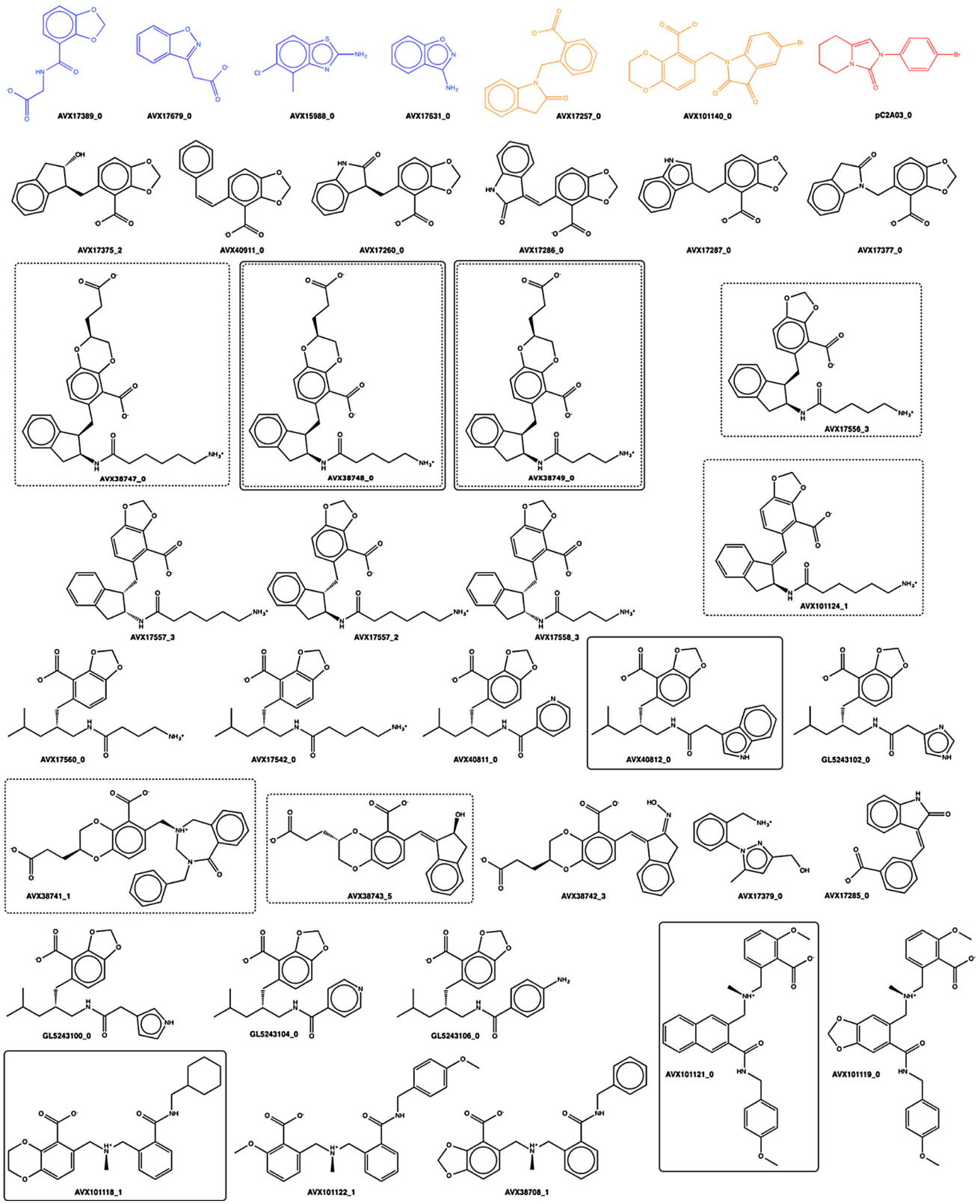

Fig. 3 Chemical structures of the 55 HIV integrase binders from the SAMPL4 dataset. Compounds *colored in blue* bind in the fragment site, the one *colored in red* binds in the Y3 site, and the compounds *colored in orange* bind in both LEDGF and Y3 sites. All the other compounds (*colored in black*) bind exclusively in the LEDGF site. Structures that were correctly predicted using the Glide SP and Gold ChemScore protocols are represented with *dotted* and *plain contour lines*, respectively

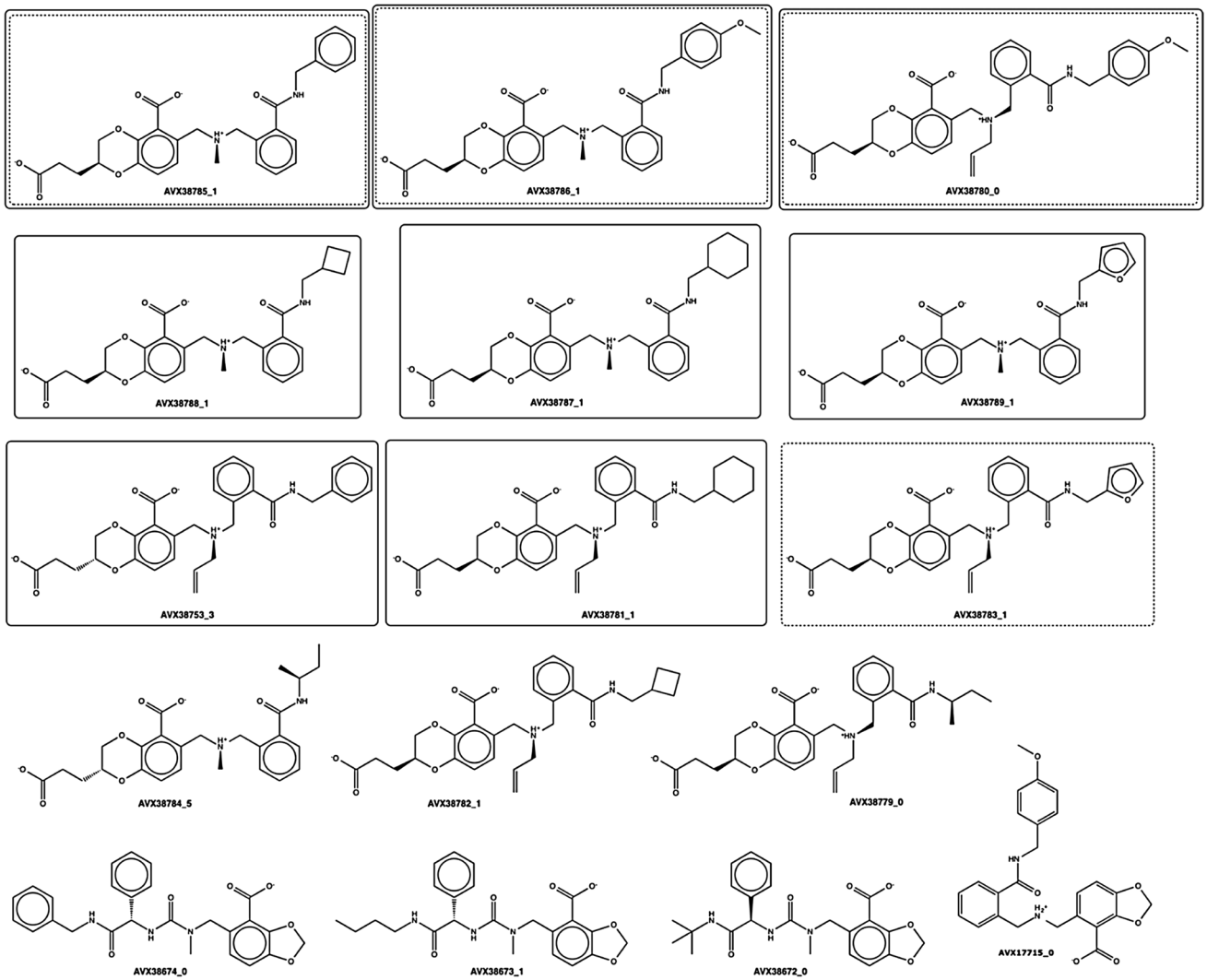

Fig. 3 continued

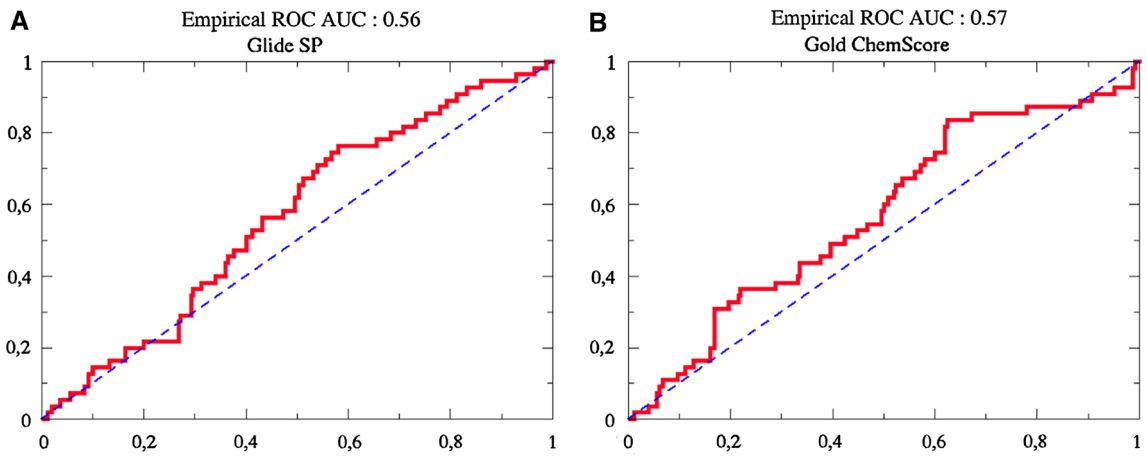

Fig. 4 ROC curves for the virtual screening of SAMPL4-VS dataset using Glide SP and Gold ChemScore

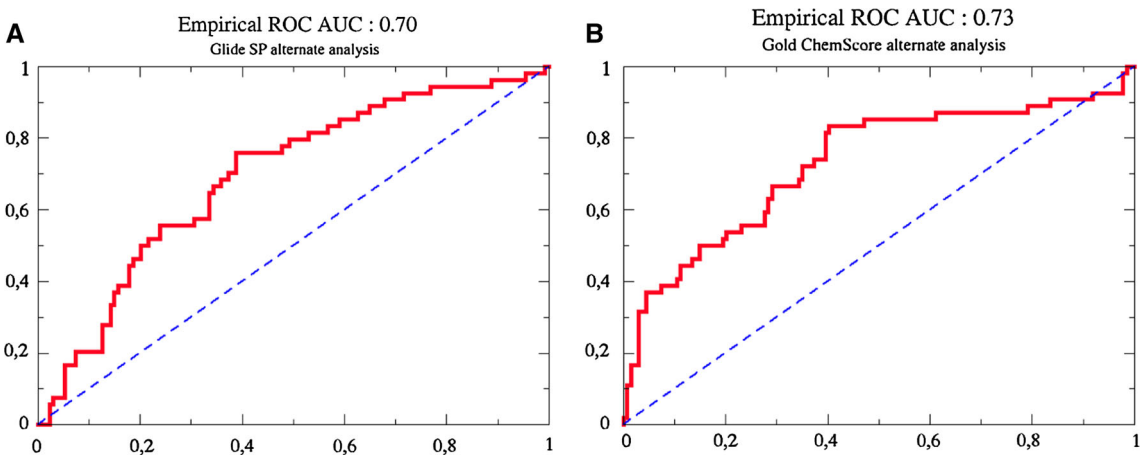

Fig. 5 ROC curves for the virtual screening of SAMPL4-VS dataset using Glide SP and Gold ChemScore with the alternate analysis

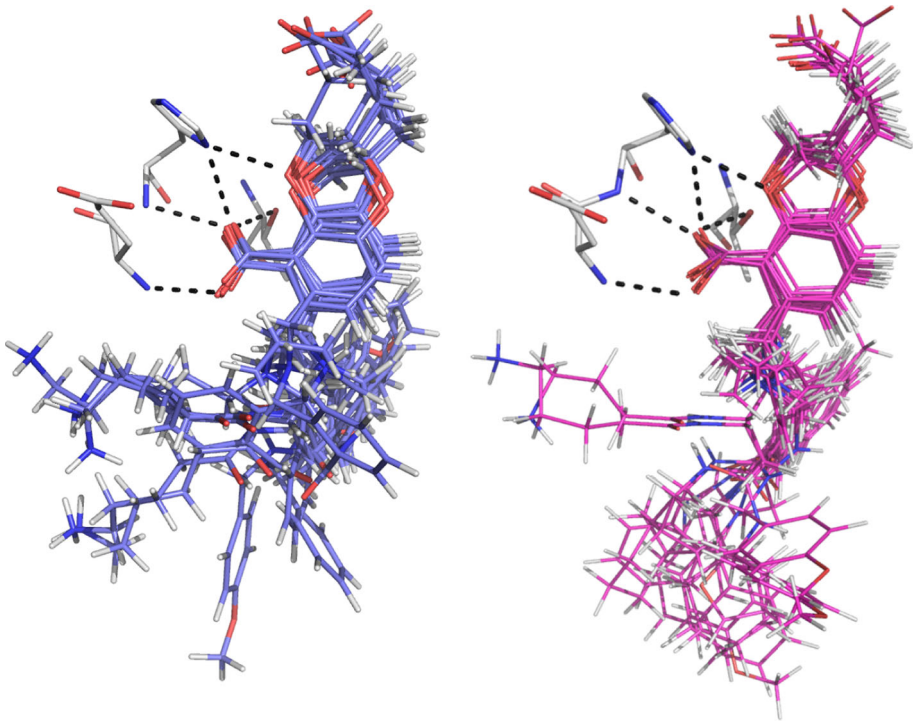

Fig. 6 Docking conformations of the compounds correctly identified to bind on HIV integrase (all in the LEDGF site), using Glide SP (left, 11 compounds) and Gold ChemScore (right, 13 compounds). Hydrogen bond interactions with the residues Glu170, His171 and Thr174 are highlighted

can be observed that in both cases the *ortho*-alkoxy-carboxylate fragment, that is common to all binders of the LEDGF site, is correctly positioned to interact with the residues Glu170, His171 and Thr174. Therefore, our protocols are able to position correctly this core fragment of the ligands in the LEDGF site, but seem to be not very efficient in scoring the right compounds on higher ranking positions. We also observe a much higher variability in the positioning of the flexible regions of these molecules (Fig. 6), which is in agreement with the lower density observed in the corresponding crystallographic structures for these regions [3].

Conclusion

In this study, we have "blindly" assessed the ability of several combinations of docking software and scoring functions to predict the binding of a dataset of potential HIV integrase inhibitors. We found that different docking software were appropriate for each one of the three binding sites considered in this work (LEDGF, Y3 and fragment sites), and the most suitable two docking protocols, involving Glide SP and Gold ChemScore, were selected using a training set of compounds identified from the structural data available. These protocols could successfully predict

respectively 20.0 and 23.6 % of the HIV integrase binders, all of them being present in the LEDGF site. When a different analysis of the results was carried out by removing all alternate isomers of binders from the set, our predictions were dramatically improved, with an overall ROC AUC of 0.73 and enrichment factor at 10 % of 2.89 for the prediction obtained using Gold ChemScore. This study highlighted the ability of the selected docking protocols to correctly position in most cases the *ortho*-alkoxy-carboxylate core functional group of the ligands in the corresponding binding site, but also their difficulties to correctly rank the docking poses.

Acknowledgments Our laboratory is a member of the Laboratory of Excellence in Research on Medication and Innovative Therapeutics (LERMIT) supported by a grant from the French National Research Agency (ANR-10-LABX-33). We would like to thank the SAMPL4 organizers, with a special mention to David L. Mobley, for providing the experimental data required for the evaluation of our predictions, as well as for the alternate analysis of the virtual screening results. The pertinent comments and suggestions of the manuscript reviewers are also kindly acknowledged.

References

1. Skillman AG (2012) SAMPL3: blinded prediction of host-guest binding affinities, hydration free energies, and trypsin inhibitors. J Comput Aided Mol Des 26(5):473–474. doi:10.1007/s10822-012-9580-z
2. Newman J, Dolezal O, Fazio V, Caradoc-Davies T, Peat TS (2012) The DINGO dataset: a comprehensive set of data for the SAMPL challenge. J Comput Aided Mol Des 26(5):497–503. doi:10.1007/s10822-011-9521-2
3. Mobley DL, Liu S, Lim NM, Wymer KL, Perryman AL, Forli S, Deng N, Su J, Branson K, Olson AJ (2014) Blind prediction of HIV integrase binding from the SAMPL4 challenge. J Comput Aided Mol Des
4. Peat TS, Dolezal O, Newman J, Mobley D, Deadman JJ (2014) Interrogating HIV integrase for compounds that bind—a SAMPL challenge
5. Mtifiot M, Marchand C, Pommier Y (2013) HIV integrase inhibitors: 20-year landmark and challenges. Adv Pharmacol 67:75–105. doi:10.1016/B978-0-12-405880-4.00003-2
6. Hazuda DJ (2012) HIV integrase as a target for antiretroviral therapy. Curr Opin HIV AIDS 7(5):383–389. doi:10.1097/COH.0b013e3283567309
7. Quashie PK, Sloan RD, Wainberg MA (2012) Novel therapeutic strategies targeting HIV integrase. BMC Med 10:34. doi:10.1186/1741-7015-10-34
8. Surpateanu G, Iorga BI (2012) Evaluation of docking performance in a blinded virtual screening of fragment-like trypsin inhibitors. J Comput Aided Mol Des 26(5):595–601. doi:10.1007/s10822-011-9526-x
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res Suppl 28(1):235–242
10. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980. doi:10.1038/nsb1203-980
11. Rhodes DI, Peat TS, Vandegraaff N, Jeevarajah D, Le G, Jones ED, Smith JA, Coates JAV, Winfield LJ, Thienthong N, Newman J, Lucent D, Ryan JH, Savage GP, Francis CL, Deadman JJ (2011) Structural basis for a new mechanism of inhibition of HIV-1 integrase identified by fragment screening and structure-based design. Antivir Chem Chemother 21(4):155–168. doi:10.3851/IMP1716
12. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30(16):2785–2791. doi:10.1002/jcc.21256
13. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. Proteins 52(4):609–623. doi:10.1002/prot.10465