Practice on Virtual Screening

www.lephar.com
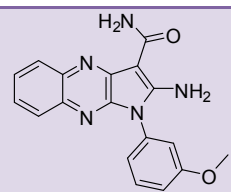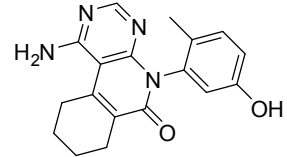
Abstract

Here we describe a virtual screening tutorial using the professional LeDock, which supports the SDF format and is 30% faster with improved accuracy. The pose-analysis tool LePose allows for an easy definition of expected interactions for removing undesired poses/compounds. In addition, it provides a scaffold clustering function by maximum common substructure. We further discussed why docking score only is not sufficient to discriminate bioactives from inactive ones.

Introduction

Here we are going to rediscover the two compounds (Table 1), which were initially identified via docking-based virtual screening. To mimic a real virtual screening campaign, about 50 compounds from the ChemBridge library were added. The tutorial uses command-line software, which may sound scaring to many Windows users, but actually they are not. As you can see, it is rather easy to use and more efficient than a graphic interface. Of course, you are expected to know a little bit about very basic Linux commands.

Table 1. EphB4 kinase inhibitors discovered by virtual screening.

	IC_{50} (nM)	PDB entry	Reference
	300	4P4C	PLoS One 2011, 6, e19923 J Med Chem 2014, 57, 6834-44
	160	4G2F	ACS Med Chem Lett 2012, 3, 834-8

Experimental

Protein Preparation. The X-ray structure of the EphB4 kinase domain in complex with a small molecule (PDB entry 2VWX) is included in the directory. We firstly need to remove water and co-crystal ligands, and then add hydrogen to the protein at the physiological pH using the following command

$ bin/lepro_linux_x86 2vwx.pdb

This will output two files. The file 'pro.pdb' is the protein file used for docking. The second file 'dock.in' defines docking parameters. The binding site is determined to include any atom within 4 Å of the co-crystal ligand.

Ligand Preparation. We will skip the step of generating 3D structures of ligands. Now we simply make a list of compounds to be docked

$ ls all.sdf > ligands

The file 'all.sdf' contains the 52 compounds to be docked.

Molecular docking. The professional LeDock will be used to perform molecular docking. It supports the SDF format, and is about 30% faster with improved accuracy.

$ bin/ledock_go dock.in

Each compound in the library will be sequentially docked, and all docking poses will be written in the file 'all_dock.sdf'.

Post processing. The docking poses can be processed according to user-defined interactions. The details of defined interactions e.g., H-bonding and/or lipophilic, will be discussed later on.

$ bin/lepose all_dock.sdf poses.txt

Visual Analysis. We recommend the free ICM Browser for visual analysis.

Results and Discussions

Let's open the all_dock.sdf using the ICM browser. What's shown is a chemical table, which is just like an Excel spread sheet (Figure 1). Compounds/poses will be ranked upon clicking the header of column 'Score'. We then make a histogram plot of Score by clicking the bottom icon on the far right pane.
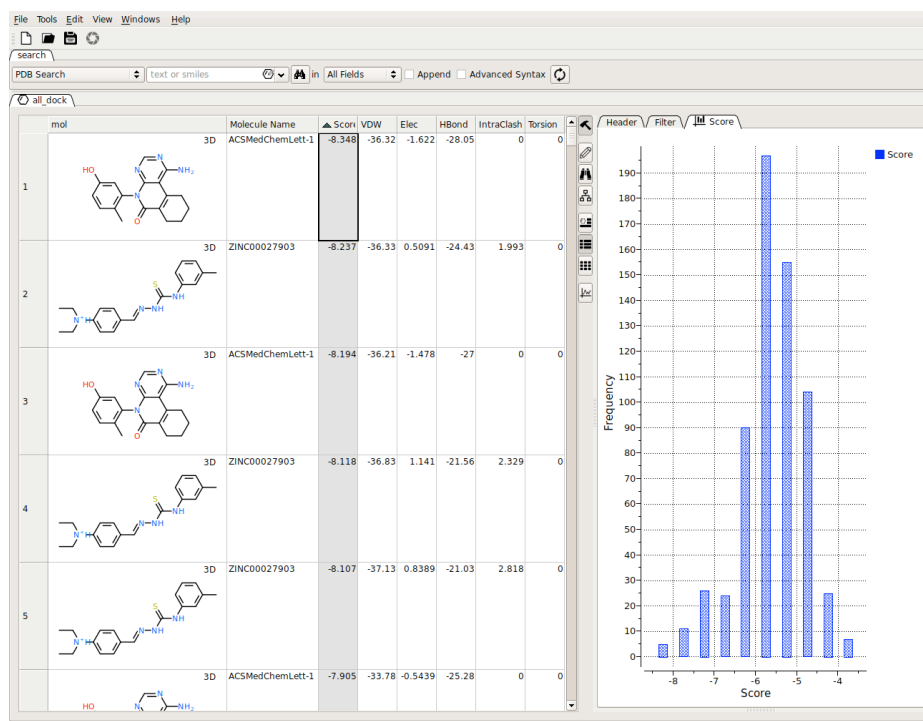

Figure 1. Screenshot showing the rank of docking poses by docking score and the histogram analysis.

3

The docking scores range from -3.5 to -8.5 kcal/mol. Although one of the two compounds in Table 1 is top ranked, the second one isn't. Docking scores from a high-throughput virtual screening often have a narrow distribution with a standard deviation of about 2 kcal/mol. On the other hand, even the most accurate affinity prediction method, i.e., free energy perturbation, has an error of 1 kcal/mol. It is a true miracle to see that ranking by docking score only could lead to discovery of novel active compounds. As we cannot expect miracles to happen all the time, we need additional tricks.

The reason why docking score only isn't enough to discriminate active compounds from false positives is because of so many approximations in the scoring function, which often adopts a soft-energy scheme in molecular docking. It is a good practice to remove poses which are unlikely to be true based on our knowledge. As we aim to find novel hinge binders, which often form at least one hydrogen bond with the hinge. We could require at least one H-bond with the hinge as an acceptor

[O,n&X2]   2.6   15.920 5.552   12.530

The first field defines an acceptor, which could be oxygen or deprotonated aromatic nitrogen. The second field is the distance. The cutoff of H-bonding interaction between two heavy atoms is 3.6 Å, but is 2.6 Å between the acceptor and the donor hydrogen. The last three fields are Cartesian coordinates of the hinge donor hydrogen (Met696 backbone HN). We simply copy the coordinate from the 'pro.pdb' file. The above criteria require a ligand acceptor atom (either oxygen or aromatic nitrogen) within 2.6 Å of the hinge Met696 HN.

The typical H-bonding patters in hinge binders are DAD. One may define more H-bonding interactions if necessary. The EphB4 kinase has a small gatekeeper and occupation of the back-site by a ligand could increase the binding affinity dramatically. We define a second interaction

[a]   3.0   7.765   1.203   13.027

The first field defines an aromatic atom, as the moiety in the back-site is often an aromatic ring. The coordinate is taken from the phenyl of the co-crystal ligand, which helps define the location of the back-site.

Score   -7
LE   -0.3
WithinTopScore   0
ClusterByMCS   0.8

Finally, we filter poses by docking score and ligand efficiency, and we will keep only the pose with the highest score. Often from virtual screening, one scaffold/chemotype may have tens or even hundreds of analogues. It makes difficult to spot on novel scaffolds. Here we will do a scaffold clustering by maximum common substructure. The output of LePose is shown in Figure 2.


```
                                                          LePose | http://lephar.com
============================================================================
--Docking Poses Processing with:
  WithinTopScore                          0.0
  Score                                  -7.0
  LE                                     -0.3
----------------------------------------------------------------------------
..Interactions..
  [O,n&X2]                                      within 2.6 A of (  15.920    5.552   12.530 )
  a                                             within 3.0 A of (   7.765    1.203   13.027 )
============================================================================

  Clustering 4 Compounds by Maximum Common Substructure with a coefficient of 0.8
  ... 2 Clusters ...

  Written to: all_dock_pass.sdf
```

Figure 2. Screenshot of output by LePose.

Now, let's open all_dock_pass.sdf and the protein structure 2vwx.pdb. As shown in Figure 3, the added 50 compounds were filtered out and, the 4 active ones were kept and clustered into two chemotypes. It is recommended to take a diversity set for experimental validation, instead of taking all compounds from one or couple of chemotypes.
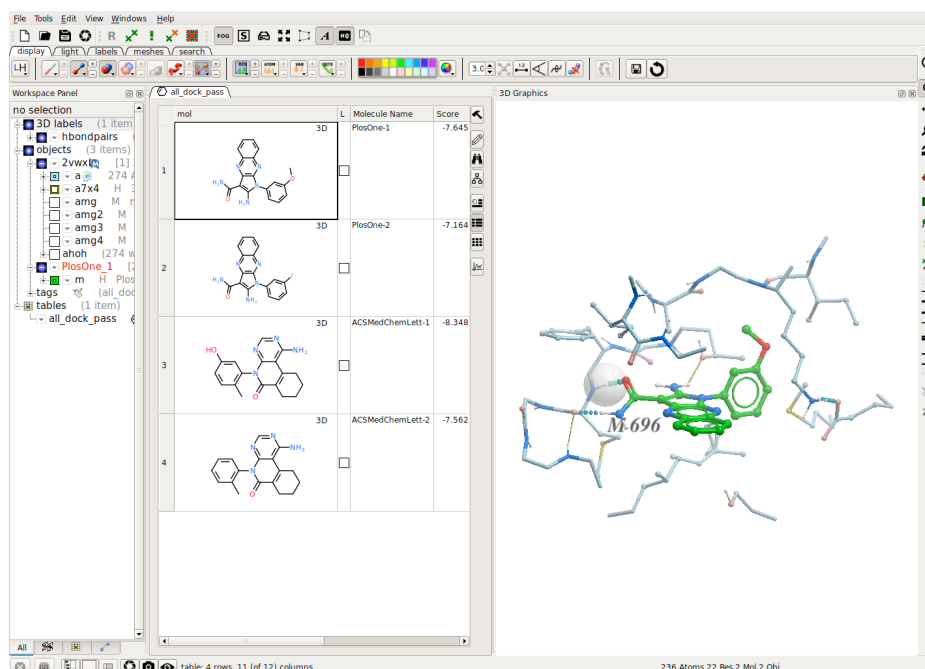

Figure 3. Screenshot showing passed compounds and their docking poses in the binding site.