DUD Subset for Ligand-Based Virtual Screening (DUD LIB VS 1.0)

Andreas Jahn[1], Georg Hinselmann[1], Nikolas Fechner[1] and Andreas Zell[1]

Center of Bioinformatics Tübingen (ZBIT), University of Tübingen, Germany

1 Introduction

The text describes the compound preparation of the Directory of Useful Decoys (DUD) release 2 [1, 2] used in the work "Optimal assignment methods for ligand-based virtual screening" [3]. Only 13 data sets were used for the evaluation of the optimal assignment methods. Nevertheless, the preparation protocol of the data sets was applied on all 40 DUD targets. The complete compilation of the data sets is contained in this archive file. We use the same abbreviations of the target names as in the work of Huang et al. [1] If you use these data sets please cite the work of Huang et al. [1], Good and Oprea [4], and Jahn et al. [3]

2 Active Compounds

The starting point of the preparation protocol of the active structures was the clustered and filtered compounds provided by Good and Oprea [4]. The SD files are directly obtainable from the DUD site[1]. These files have only 2D coordinates. Therefore, CORINA3D was used to generate initial seed coordinates of all compounds. We used MarcoModel 9.6 [5] with the following settings to further optimize the 3D coordinates:

- Force field: OPLS force field 2005
- Optimization method: limited Broyden-Fletcher-Goldfarb-Shanno
- Maximum iterations: 5000
- Convergence parameter: 0.0001 RMSD (Avg. atomic movement between two iterations)

An overview of the number of compounds and clusters can be seen in Table 1.

3 Decoy Compounds

The original DUD decoy data set release 2 for each target was obtained from the DUD site[2]. To remove the bias of an artificial enrichment, we performed the same filter approach as Good and Oprea [4, 6]. The first step of the lead-like filter is

[1] http://dud.docking.org/clusters/
[2] http://dud.docking.org/r2/

Table 1. Overview of the number of active compounds and clusters for each target. MacroModel removed one compound of the ACHE and Thrombin data set.

Target	Ligands	Clusters
ACE[a]	46	19[d]
ACHE[a b]	99	18
ADA	23	8
ALR2	26	14
AMPC	21	6
AR	68	10
CDK2[a]	47	32
COMT	11	2
COX-1	23	11
COX-2[a]	212	44
DHFR	190	14
EGFR[a]	365	40
ER agonist	63	10
ER antagonist	18	8
FGFR1	71	12
FXA[a]	64	19
GART	8	5
GPB	52	10
GR	32	9
HIVPR	4	3
HIVRT[a]	34	17
HMGA	25	4
HSP90	23	4
INHA[a]	57	23
MR	13	2
NA	49	7
P38[a]	137	20
PARP	31	7
PDE5[a]	26	22
PDGFRB[a]	124	22
PNP	25	4
PPAR γ	6	6
PR	22	4
RXR α	18	3
SAHH	33	2
SRC[a]	98	21
Thrombin[c]	23	14
TK	22	7
Trypsin	9	7
VEGFR2[a]	48	31

[a] Data sets used in the work of Jahn et al. [3].
[b] The compound with the ZINC id 01903720 was removed by MacroModel.
[c] The compound with the ZINC id 03834162 was removed by MacroModel.
[d] Data set contains one compound that does not have any ring systems. Therefore, the reduced graph algorithm used by Good and Oprea was not be able to process this molecule. This structure is treated as one additional cluster.

an AlogP filter with a cutoff value of 4.5 (5.5 for nuclear hormone receptors: AR, ER agonist, ER antagonist, GR, MR, PPAR γ, PR, and RXR α). The AlogP values were calculated using dragonX 1.4 [7]. The removed structures can be found in the AlogP-fail folder. The second filter step applies a molecular weight filter removing all structures with a molecular weight (MW) >= 450 $\frac{\text{g}}{\text{mol}}$. The compounds that did not pass the filter can be found in the MW-fail folder. The two filters have the same setup as the filters used by Good and Oprea. Therefore, the bias of an artificial enrichment as a result of filtering the active structures only is removed. The remaining compounds were optimized using the same setup of MacroModel. An overview of the final composition of the data sets and the number of compounds that were removed by the AlogP and MW filter can be seen in Table 2.

References

1. Huang N, Shoichet BK, Irwin JJ: **Benchmarking Sets for Molecular Docking**. *J. Med. Chem.* 2006, **49**(23):6789–6801.
2. **DUD - A Directory of Useful Decoys**[http://dud.docking.org].
3. Jahn A, Hinselmann G, Fechner N, Zell A: **Optimal assignment methods for ligand-based virtual screening**. *Journal of Cheminformatics* 2009, **1**:14.
4. Good AC, Oprea TI: **Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: a Help or Hindrance in Tool Selection?** *J. Comput.-Aided Mol. Des.* 2008, **22**(3–4):169–178.
5. Schrödinger, LLC: *MacroModel.* version 9.6, New York, NY 2008.
6. Oprea TI, Davis AM, Teague SJ, Leeson PD: **Is There a Difference between Leads and Drugs? A Historical Perspective**. *J. Chem. Inf. Comput. Sci.* 2001, **41**(5):1308–1315, [http://pubs.acs.org/doi/abs/10.1021/ci010366a].
7. Talete srl, Milano, Italy: **dragonX 1.4 for Linux (Molecular Descriptor Calculation Software)**.

Table 2. Overview of the number of final decoy compounds and compounds that do not pass the AlogP and MW filter. MacroModel removed several compounds of the CDK2, DHFR, FXA, and GPB data sets.

Target	Decoys	AlogP fail	MW fail
ACE[a]	1796	1	0
ACHE[a]	3859	33	0
ADA	927	0	0
ALR2	986	9	0
AMPC	786	0	0
AR	2848	6	0
CDK2[a b]	2070	3	0
COMT	468	0	0
COX-1	910	1	0
COX-2[a]	12606	645	38
DHFR[c]	8350	13	1
EGFR[a]	15560	432	4
ER agonist	2568	2	0
ER antagonist	1058	306	84
FGFR1	3462	274	814
FXA[a d]	2092	572	3079
GART	155	2	722
GPB[e]	2135	0	0
GR	2585	359	3
HIVPR	9	560	1469
HIVRT[a]	1494	25	0
HMGA	1423	44	13
HSP90	975	4	0
INHA[a]	2707	556	3
MR	636	0	0
NA	1713	0	0
P38[a]	6779	2360	2
PARP	1350	1	0
PDE5[a]	1698	79	201
PDGFRB[a]	5603	377	0
PNP	1036	0	0
PPAR γ	40	1079	2008
PR	920	121	0
RXR α	575	173	2
SAHH	1346	0	0
SRC[a]	5679	317	323
Thrombin	1148	6	1302
TK	891	0	0
Trypsin	718	1	945
VEGFR2[a]	2712	191	3

[a]Data sets used in the work of Jahn et al. [3].
[b]The compound with the ZINC id 03997306 was removed by MacroModel.
[c]The compounds with the ZINC ids 03997305, 03997306, and 04475324 were removed by MacroModel. [d]The compounds with the ZINC ids 03983347, and 03983409 were removed by MacroModel. [e]The compounds with the ZINC ids 01583034, 01583034, 01660425, 04097362, and 04293792 were removed by MacroModel.