Mining PDB Subpockets to Rebuild Ligand Binding Conformations

Jean-Yves Trosset[1], Maud Vieillevoye[2], Stewart Adcock[2], François Delfaud[2,3]

[1]BIRL Sup'Biotech, 66 rue Guy Môquet, 94800 Villejuif, France
[2]Felix Concordia SARL, 400 av de Roumanille, Bat 7, 06906 Sophia Antipolis, France
[3]Medit SA, 2 rue du Belvédère, 91120 Palaiseau, France

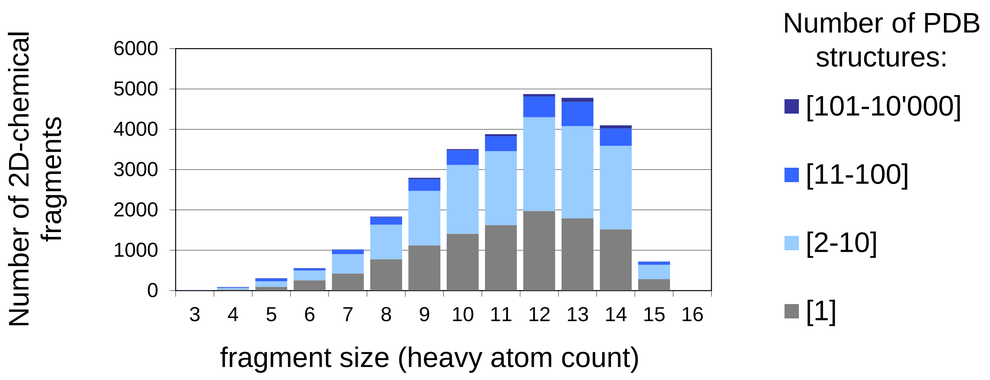
Contact : fdelfaud@felixc.eu
http://www.felixc.eu/

ABSTRACT

We constituted a database of subpocket-fragment complexes (Fragmentor) through deconvolution of each PDB (Protein Data Bank) ligand in all possible fragments that match one of the chemical molecule contained in the Pubchem database. After application of a Matriochka filter, we obtained a set of 28.482 2D-fragments and 398.236 3D-fragments (PDB conformations). Subpockets were defined as protein surfaces located at 4,5 Å around fragments.

The goal of this work was to determine if there is enough information in the PDB to successfully rebuilt the binding mode of ligands, starting from the target protein structure. To test this hypothesis, we selected ligands in unique PDB entry to build a test set of 2292 protein-ligand complexes and tried to recover the position and conformation of the ligands using our Fragmentor database. We were able to predict at least 80% of the 3D-structure from 1091 ligands (48%). In conclusion, this study highlights the quality of the information contained in the PDB and supports the use of its structural information for docking tools or fragment-based drug design.

FRAGMENTOR: A DATABASE CONTAINING 398.236 SUBPOCKET-FRAGMENT COMPLEXES



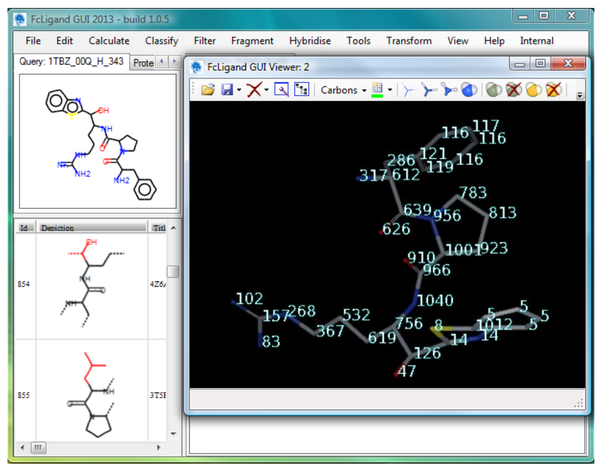
We constituted a database of subpocket-fragment complexes derived from PDB data. 40% of the 28.482 2D-fragments are present only in one PDB file (grey columns) whereas the other 60% can be retrieved from multiple PDB files (blue columns).

WORKFLOW

1 Compare the binding site of the target protein with *Fragmentor* database

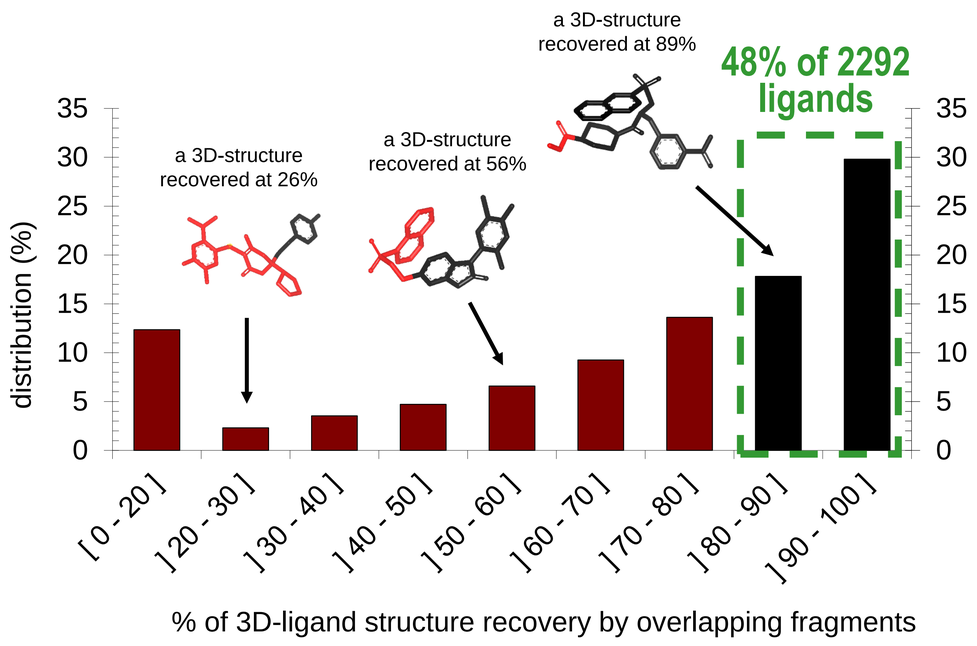
2 Retrieve subpockets and associated 3D-fragments that superpose onto the binding site of the target

3 Calculate how many times the atoms of the reference ligand are recovered at a deviation less than 1Å with a common substructure of collected 3D-fragments



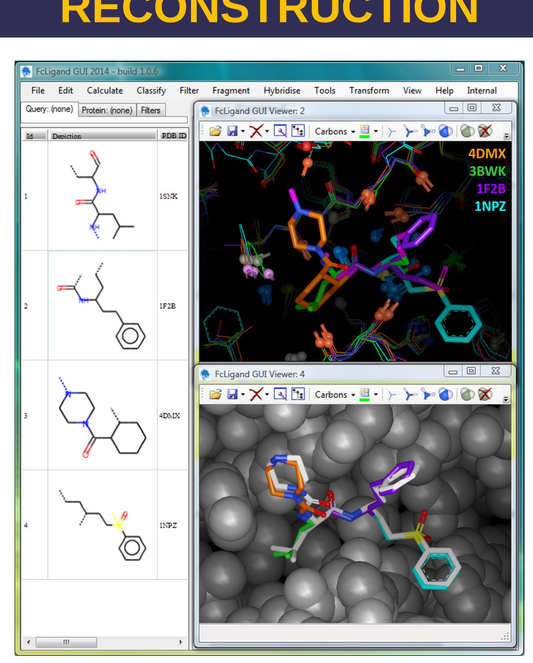
Pairwise comparisons of subpockets were carried out with the **MEDP-SiteClassifier** software, an evolution of **MED-SuMo** software. The left column shows the overlapping chemical moieties (in black) between selected 3D-fragments and the reference ligand (on top); a MCS (Maximum Common Substructure), coupled to a 3D deviation detection, turns non overlapping atoms in red. A javascript in **FcLigand** viewer illustrates the number of times the atoms of the reference ligand were recovered with overlapping 3D-chemical fragments.

PERCENTAGE OF 3D-STRUCTURE RECOVERY OF 2292 LIGANDS : 48% CAN BE ALMOST RECONSTRUCTED



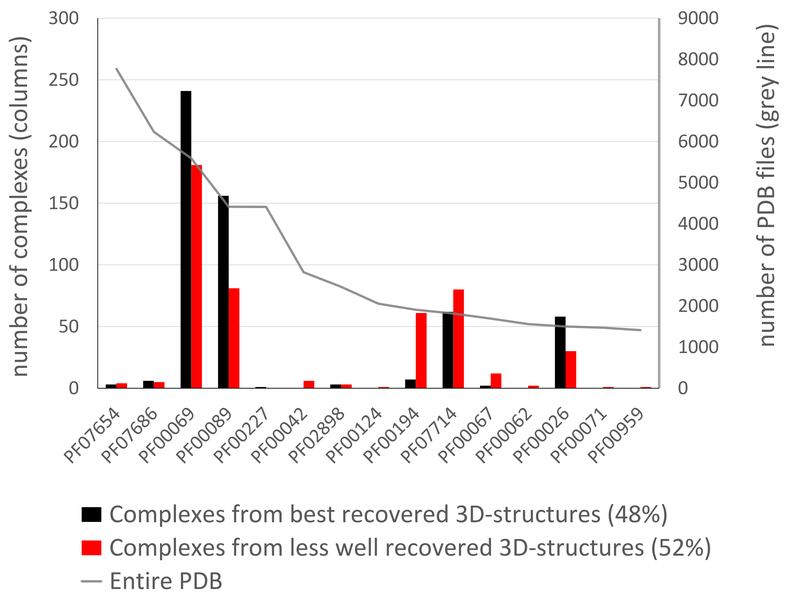
The analysis of 2292 protein-ligand complexes reveals that the PDB contains sufficient structural information to rebuilt 48% of ligands with at least 80% of their 3D-structure. Three ligands are depicted to illustrate the graph (in red: atoms which were not predicted).

EXAMPLE OF SUCCESFUL LIGAND RECONSTRUCTION



MED-SuMo in combination with **FcLigand** software were able to successfully rebuilt vinyl sulfone inhibitor, a ligand of cathepsin K (ID: 1MEM) by hybridization of retrieved 3D-fragments including a substructure constraint to the inhibitor. Four fragments which are associated with different proteins in the PDB (cathepsin K, falcipain-3, cruzain protein, cathepsin S) were necessary for this reconstruction. Top panel: the four fragments superposed onto their targets before hybridization. Lower panel: The hybrid molecule (in colour) and vinyl sulfone inhibitor (in grey) in the binding site of cathepsin K.

INFLUENCE OF PFAM AFFILIATION ON LIGAND RECOVERY



The horizontal axis lists the 15 most frequent Pfam families present in the PDB. The grey curve (right axis) indicates the total number of structures that belong to the considered Pfam family in the PDB. The complexes for which the ligand was well predicted are represented by red columns whereas the complexes for which the ligand was predicted with less than 80% of its 3D-structure are represented by black columns (left axis).

It seems that the quantity of structural data available for the Pfam family of the target does not significantly influence the percentage of recovery of the ligand (success rate: 54% for ligands stemming from the 15 most frequent families versus 48% for the whole dataset)

PFAM	DESCRIPTION	BLACK COLUMNS	RED COLUMNS
PF07654	Immunoglobulin C1-set domain	0,43	0,57
PF07686	Immunoglobulin V-set domain	0,55	0,45
PF00069	Protein kinase domain	0,57	0,43
PF00089	Trypsin	0,66	0,34
PF00227	Proteasome subunit	1,00	0,00
PF00042	Globin	0,00	1,00
PF02898	Nitric oxide synthase, oxygenase domain	0,50	0,50
PF00124	Photosynthetic reaction centre protein	0,00	1,00
PF00194	Eukaryotic-type carbonic anhydrase	0,10	0,90
PF07714	Protein tyrosine kinase	0,44	0,56
PF00067	Cytochrome P450	0,14	0,86
PF00062	C-type lysozyme/alpha-lactalbumin family	0,00	1,00
PF00026	Eukaryotic aspartyl protease	0,66	0,34
PF00071	Ras family	0,00	1,00
PF00959	Phage lysozyme	0,00	1,00
	WEIGHTED MEAN	0,54	0,46
	TOTAL MEAN (2292 complexes)	0,48	0,52

CONCLUSION

- We generated of a rich subpocket-fragment database from PDB data which was used for the present proof of concept
- 48% of ligands could be recovered by 80% of their 3D-structure using a protocol based on **MEDP-SiteClassifier** software
- The percentage of ligand recovery does not depend on the pfam family of the target
- We were able to successfully reconstruct a vinyl sulfone inhibitor in **FcLigand** software after hybridization of fragments collected by **MED-SuMo**.