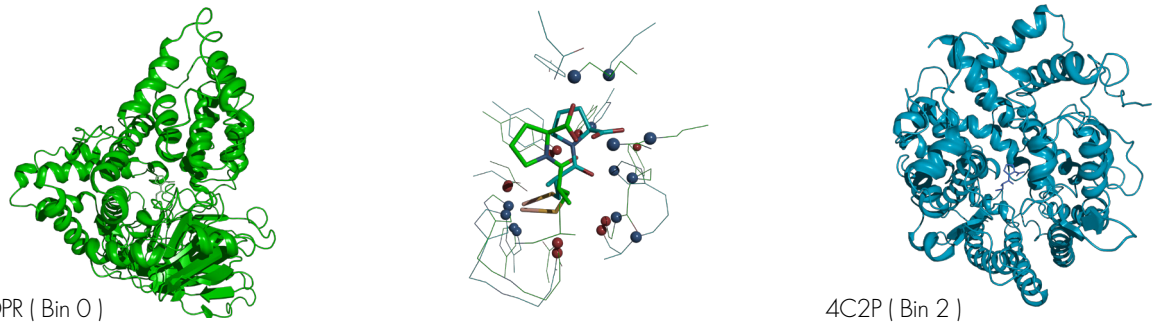
Fig. 1



Bin classification according to eye-scored site similarity. The overall fold of the structure of two captopril binding proteins from test case are clearly dissimilar (cartoon representation). Nevertheless, when sites of both structures are superimposed (residues in line and ligands in stick, in the middle) atoms exhibit some similarity (materialized with spheres : nitrogen and oxygen) which denotes a Bin 2 according to our classification.

FIT FOR FITTING: BIONEXT-DSS, A DATASET OF SIMILARS SITES

SUMMARY

Predicting novel binding pockets for a given medicine is a challenging task with numerous pharmaceutical applications. Based on the principle that "similar binding sites share similar ligands"[4], many binding pocket prediction algorithms are based on the evaluation and retrieval of pockets being similar to a known binding site. Two distinct yet connected problematics can hence be described : (i) the ability to retrieve known sites for a given ligand in distinct proteins, and (ii) the retrieval of similar sites. Nowadays, many protocols and datasets exist for the evaluation of ligand pocket prediction, mainly based on pioneering works from Kahraman[1] and Hoffmann[2]. However, to our knowledge, no dataset exist for the specific evaluation of similarity based algorithm probably because the very notion of "site similarity" eludes a straight definition. For these reasons we compiled "BionextDSS", a dataset of similar binding sites which contains 35 test cases connected to a pharmaceutical application. Each case includes a reference site, as well as several positive sites further classified in 6 bins according to their similarity level to the reference. We moreover provide a list of noise structures carefully selected so as to be used as negative controls in algorithms evaluation. Our dataset would be used to bench our approach (BioBind[3]) to several others. *The BionextDSS dataset is freely available on our servers (https://goo.gl/n5rAOp).*

BIONEXT-DSS CONSTRUCTION

Four distinct elements are compiled for each test case:

- The reference site (bin 0) belongs to a protein of pharmaceutical interest and is systematically selected as the binding site of a ligand of pharmaceutical interest according to the Linpisky rules5. Many of the small ions, sugars, solvent or bulky HEME cofactors that can be found in Kahraman1 and Hoffmann2 datasets were for instance ruled out.
- The positive sites were selected as sites containing the same ligand or a ligand very similar to that of the reference.
- Similarity assessment (Binning) : each positive site was manually aligned with its reference site using PyMol6. Chemical groups of positive sites matching those of the reference site were materialized by spheres (see Fig.1) and a visual assessment of the structural similarity of each positive site to its reference was then performed based on the number and relevance of such spheres. Although this classification is very subjective, we tried to conform to certain rules: Bin 1 contains sites very similar to the reference (up to the overall fold); Bin 2 includes sites with enough fluctuation whilst the global pattern is still retained; Bin 3 contains sites where the similarity is still visible yet getting blurred. In higher bins, few local similarities are still observed whilst the global pattern can be considered dissimilar.
- Noise : we believe that control of negative structures is as important as control over positive structures when it comes to algorithms evaluation. Therefore, compared to previous works, we provide "noise" structures instead of using positives from other test cases. We achieved this by clustering structures in the Protein Data Bank (PDB) to 30% similarity. Structures quality was moreover taken into account (only structures with a resolution under 3 Å and chains containing at least 80 amino acids were retained). Construction of a dataset of true negative sites is difficult. Therefore we provide instead a list of "noise" structures built so as to be as exempt as possible of sites similar to a positive site. We ensured this by removing structures sharing Interpro annotation7 with any of the positive structures.

« **DSS: 35 innovative testcases to define site similarity in a pharmaceutical context.**

Dataset

Benchmark

Site similarity

POSSIBLE USES OF BIONEXT-DSS

DSS can be used for the two aforementioned problematics :
- In order to evaluate the effectiveness of similarity retrieval, one can use the reference site as query and probe the ability to retrieve positive sites versus noise through the measurement of the area under the ROC curve (AUC).
- In order to evaluate the algorithms ability to retrieve all known pockets of a given ligand, one can use alternatively all positive sites of a test case as a query and calculate an average AUC.

As a complement, the structural alignment of positive sites with the reference provided in DSS can be used to discriminate valid prediction of positive sites.

Dataset	Bionext-DSS	Kahraman	Hoffman
Positives Sites	222	64	97
Test cases	35	7	10

Table 1: Number of test cases and positive sites for the studied datasets.

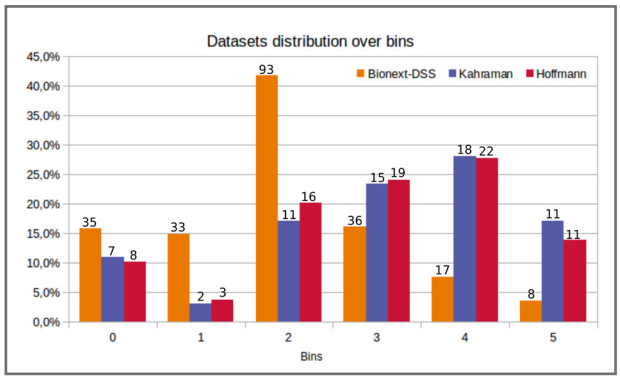


Fig. 2: Percentage of positive sites per bin for each dataset. Numbers on the top of each bar refer to the numbers of positive sites.

FACTS AND COMPARISONS

For comparisons we applied our method for the evaluation of site similarity to the sites of the Kahraman and Hoffmann datasets.

As can be seen in Table 1, DSS contains a larger amount of test cases as well as a larger amount of positive sites. Fig. 2 moreover reveals that DSS contains a bigger proportion of trivial similarities (bins 1 and 2), making our dataset more suited to evaluate algorithms on their ability to detect similarity. Moreover, the amount of difficult cases (bins 3 to 5) in DSS is comparable to that in Kahraman and Hoffmann, which makes our dataset also fit for the assessment of the more classical problematics of ligand binding pocket retrieval.

With this dataset, we propose an experimental definition of similarity compliant to the intuition of a pharmaceutically trained bioinformatician. To our knowledge, this is the first effort in that direction, and we believe this work is necessary in order to tune similarity-based algorithms. Although our classification in similarity bins is necessarily subjective, experiments conducted with our in-house algorithm BioBind as well as with ProBis8 show that lower bins were retrieved first, thus confirming an agreement between our intuitive definition of similarity and these algorithms.

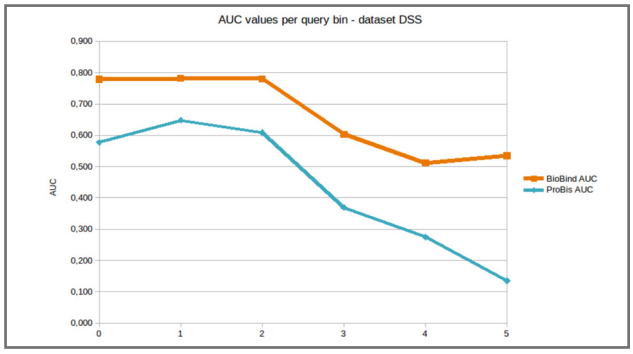


Fig.3: Application of DSS to BioBind and ProBis algorithms.

REFERENCES

1. Kahraman, A. et al. (2007). Shape Variation in Protein Binding Pockets and their Ligands. Journal of Molecular Biology, 368(1), 283–301.

2. Hoffmann, B. et al. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC bioinformatics, 11, 99.

3. BioBind: algorithm patented by Bionext to compare and classify physico-chemical surfaces. Cf other Bionext communications for more information.

4. Klabunde, T. (2007). Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br J Pharmacol, 152(1), 5–7

5. Lipinski, C. A (2004). Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today, Dec(4), 337-41

6. DeLano W. (2002). The PyMOL molecular graphics system. San Carlos, CA, USA: DeLano Scientific LLC. http://pymol.sourceforge.net .

7. Robert D.Finn. (2016) InterPro in 2017 - beyond protein family and domain annotations. Nucleic Acids Res (2017), 45 (D1), D190-D199.

8. Konc, J. et al. (2015) ProBiS-CHARMMing: Web Interface for Prediction and Optimization of Ligands in Protein Binding Sites. J. Chem. Inf. Model., 55, 2308-2314.

ABOUT BIONEXT

BIONEXT is a French bioinformatics company founded in 2009 and dedicated to the better understanding and treatment of various diseases. As such, BIONEXT focuses on pharmaceutical problematics with a major interest in guiding the prediction of biochemical compounds for pharmaceuticals targets. BIONEXT aims to accelerate and improve the effectiveness of drug development and biological research by developing programs and software-based tools to tackle problematics of prime importance in the pharmaceutical field, such as drug profile assessment or suggestion of drug repurposing. Our first application BioBind is based on the now accepted principle that similar receptors bind similar ligands.

BIONEXT
Delivering the *in silico* promise
WWW.BIONEXT.COM