

A new protein binding pocket similarity measure based on comparison of 3D atom clouds: application to ligand prediction

Brice Hoffmann*, Mikhail Zaslavskiy*, Jean-Philippe Vert and Véronique Stoven
Mines ParisTech, CBIO and CMM, France,
Institut Curie and INSERM U900, Paris F-75248 France.

* Contributed equally

July 9, 2009

Abstract

Motivation: Prediction of ligands for proteins of known 3D structure is important to understand structure-function relationship, predict molecular function, or design new drugs.

Results: We explore a new approach for ligand prediction in which binding pockets are represented by atom clouds. Each target pocket is compared to an ensemble of pockets of known ligands. Pockets are aligned in 3D space with further use of convolution kernels between clouds of points. Performance of the new method for ligand prediction is compared to those of other available measures and to docking programs. We discuss two criteria to compare the quality of similarity measures: area under ROC curve (AUC) and classification based scores. We show that the latter is better suited to evaluate the methods with respect to ligand prediction. Our results on existing and new benchmarks indicate that the new method outperforms other approaches, including docking.

Availability: The new method is available at <http://cbio.ensmp.fr/paris/>

Contact: mikhail.zaslavskiy@mines-paristech.fr

1 Introduction

One of the main goals of structural biology is to predict, from the 3D fold of a protein, its interacting partners, which in turn is related to its molecular function. However, understanding this structure-function relationship is still today an open question, and no reliable tool is available to permit such a prediction. Current efforts concentrate on local 3D approaches, focusing on identification and comparison of binding pockets, in order to predict the natural ligand for a protein, with the underlying idea that proteins sharing similar binding sites are expected to bind similar ligands. The same strategy also applies to the problem of identifying new drug precursors for a therapeutic target protein.

The comparison of 3D binding pockets is an active field of research, and during the last decade, many new methods were proposed. [MNKT05] considered a method based on using real spherical harmonic expansion coefficients, [GJ06] used a specialized geometric hashing procedure as the core of the SitesBase web server, [SPSNW08] used multiple common point set detection method. An approach proposed by [SSKR08] is based on a triangle-discretized sphere representation of binding pockets. [WHKK07] and [NKT08] considered graph-based representations of binding pockets and applied graph matching algorithms.

In this paper, we explore the potential of a new approach in which binding pockets are represented by clouds of atoms in 3D space potentially bearing additional labels such as partial charge or atom type. The new similarity measure is based on the alignment of protein pockets with further use of convolution kernel between 3D point clouds. We study how the proposed method may be used to predict a ligand for a given pocket by comparing it to a set of pockets with known ligand.

Here, we do not discuss the problem of pocket detection. In our experiments, we extracted pockets on the basis of known protein-ligand crystal structures as it was done by [KMLT07]. In cases where the binding site is unknown, various programs have been developed to locate depressions on protein surfaces and could be used to identify putative binding sites ([GMN⁺06]).

An important question in this paper is the evaluation of pocket similarity measures. We discuss two criteria to compare the quality of similarity measures on the basis of their ability to detect pockets binding the same ligand: area under ROC curve (AUC) and classification based scores. We compare our method with some existing state of the art algorithms on different benchmark datasets. Since we evaluate methods for binding pocket comparison according to their ability to predict ligands, we also report the performance of docking methods, on the same benchmark datasets. Finally, we also discuss possible extensions of the proposed method to other applications such as protein function prediction or ligand comparison.

2 Methods

2.1 Convolution kernel between atom clouds

In our model, a binding pocket is described by a set of atoms in 3D space. Our objective is to construct a similarity measure between pockets, which may be used to identify pockets binding the same ligand.

Let $P = (x_i, l_i)_{i=1}^N$ denote a binding pocket consisting of N atoms, where $x_i \in \mathcal{R}^3$ is a 3D vector representing atom coordinates, and l_i is a label (discrete or real valued) that may be used to bare additional information on the atoms (for example, atom type, atom partial charge, or amino acid type).

A classical approach for pocket comparison consists in iterative alignment of two pockets and further counting of overlapping atoms, usually within a tolerance of 1Å. Different implementations of this principle may be found in such methods as Tanimoto index [WWB86], the SitesBase algorithm (Poisson index), or the Multi-Bind algorithm [SPSNW08]. The alignment is made to maximize the number of overlapping atoms, which is generally a good indicator of pocket similarity.

However, atoms may have different positions but play equivalent roles in ligand binding, and the role of one atom in one pocket may be played by a group of atoms in another one. These observations lead us to the idea of an alternative smooth score which does not count the number of overlapping atoms, but rather uses a weighted number of atoms having closed positions. We first consider the case where labels are ignored, and only atom coordinates are used to measure the similarity between pockets, and then explain how the information on atom labels may be introduced in the new similarity measure.

Given two pockets P_1 and P_2 the similarity measure $K(P_1, P_2)$ is defined as follows

$$K(P_1, P_2) = \sum_{x_i \in P_1} \sum_{y_j \in P_2} e^{-\frac{\|x_i - y_j\|^2}{2\sigma^2}}. \quad (1)$$

This similarity measure defines in fact a positive definite kernel, i.e. it may be considered as a true scalar product on the set of atom clouds representing binding pockets [STV04]. Implicitly, it defines a distance between pockets: $D(P_1, P_2) = K(P_1, P_1) + K(P_2, P_2) - 2K(P_1, P_2)$ which has all standard properties of a true metric (non-negativity, identity of indiscernibles, symmetry, triangular inequality). The parameter σ characterizes the sensitivity of the similarity measure (1) to points relative displacements. When σ is small, only atoms of two pockets which are very close to each other significantly contribute to $K(P_1, P_2)$. On the contrary, when σ is large, almost all pairs of atoms contribute to $K(P_1, P_2)$.

The kernel (1) is an example of a convolution kernel [Hau99, GFKS02] between point sets. Alternative kernels may be constructed by substituting the Gaussian kernel $e^{-\frac{\|x_i - y_j\|^2}{2\sigma^2}}$ by any other kernel between 3D vectors x_i and y_j .

Interestingly, the kernel (1) may be seen as a particular case of kernel between point sets defined as a kernel between distribution function estimated from point sets [KJ03]. More precisely, let us represent each binding pocket P_i by a distribution of masses defined as the sum of Gaussian with bandwidth $\sigma/\sqrt{2}$ functions centered on the pocket atoms, namely:

$$f_{P_i}(x) = \sum_{x_i \in P_i} e^{-\frac{\|x - x_i\|^2}{\sigma^2}}.$$

Then kernel (1) between pockets P_1 and P_2 can be recovered, up to a scaling constant, as the scalar product in $L_2(\mathcal{R}^3)$ between the associated distributions because:

$$\begin{aligned} \langle f_{P_1}, f_{P_2} \rangle_{L_2(\mathcal{R}^3)} &= \int_{\mathcal{R}^3} \sum_{x_i \in P_1} e^{-\frac{\|x - x_i\|^2}{\sigma^2}} \sum_{y_j \in P_2} e^{-\frac{\|x - y_j\|^2}{\sigma^2}} dx \\ &= \sum_{\substack{x_i \in P_1 \\ y_j \in P_2}} \int_{\mathcal{R}^3} e^{-\frac{\|x - x_i\|^2}{\sigma^2}} e^{-\frac{\|x - y_j\|^2}{\sigma^2}} dx = C \sum_{\substack{x_i \in P_1 \\ y_j \in P_2}} e^{-\frac{\|x_i - y_j\|^2}{2\sigma^2}} = CK(P_1, P_2), \end{aligned}$$

where C is a positive constant.

However, formula (1) is not fully appropriate in practice, because the proposed measure is not invariant upon rotations and translations of the binding pockets. Therefore, we define a similarity measure *sup-CK* as the maximum of (1) over all possible rotations and translations of one of the two pockets:

$$\text{sup-CK}(P_1, P_2) = \max_{R, y_t} \sum_{x_i \in P_1, y_j \in P_2} e^{-\frac{\|x_i - (Ry_j + y_t)\|^2}{2\sigma^2}}, \quad (2)$$

where R is an orthonormal rotation matrix and y_t is a translation vector. *Sup-CK* is not a positive definite measure anymore, but it can still be used as a similarity score. Furthermore, to evaluate *sup-CK*, we now need

to maximize a non-concave function over the set of rotations and translations, which may have many local maxima. Exact maximization of this non-concave function is a hard optimization problem and we propose to estimate an approximate solution by running a gradient ascent algorithm, starting from many different initial points, and taking the best local maximum. The optimization algorithm may be significantly accelerated by choosing an initial point close to the global optimum. In the case of binding pockets, a good approximation of the optimal translation vector y_t is the vector which translates the geometric center of P_2 into the geometric center of P_1 , $y_t = \frac{1}{N_1} \sum_{x_i \in P_1} x_i - \frac{1}{N_2} \sum_{y_j \in P_2} y_j$. The approximated rotation matrix R superposes the first principal axis of P_2 with the first principal axis of P_1 , the second one with the second one, and the third one with the third one. Since principal vectors are defined up to a sign, the two signs for all principal vectors of one of the binding pockets have to be tested (there are 2^3 combinations). If some of the pocket axes have close lengths, then it may be also interesting to consider rotations which superpose the first principal axis of one pocket with the second principal axis of the other one.

Gradient ascent method requires to calculate the gradient of the function in (2) with respect to R and y_t . Calculation of the gradient components related to y_t is straightforward:

$$\nabla_{y_t} = \frac{1}{\sigma^2} \sum_{x_i \in P_1, y_j \in P_2} (x_i - (Ry_j + y_t)) e^{\frac{\|x_i - (Ry_j + y_t)\|^2}{2\sigma^2}}.$$

Since the set of rotation matrices is a 3D manifold embedded in 9D space, we cannot take derivatives with respect to each element of matrix R . Instead, we use the Euler representation of the rotation matrix:

$$R = R_X R_Y R_Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & \sin \phi \\ 0 & -\sin \phi & \cos \phi \end{bmatrix} \begin{bmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \psi & \sin \psi & 0 \\ -\sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (3)$$

where R is expressed as a function of $(\phi, \theta, \psi) \in [0; 2\pi)^3$. The derivatives of the maximand in (2) are now calculated with respect to (ϕ, θ, ψ) , for instance,

$$\nabla_{\theta} = \frac{1}{\sigma^2} \sum_{\substack{x_i \in P_1 \\ y_j \in P_2}} e^{\frac{\|x_i - (Ry_j + y_t)\|^2}{2\sigma^2}} (x_i - (Ry_j + y_t))^T (R_X \frac{\partial R_Y}{\partial \theta} R_Z y_j).$$

As mentioned above, it may be interesting to use additional information on binding pocket atoms (such as atom type or charge). Let us suppose that this information is represented by labels l_i (which may be discrete or real variables, or multidimensional vectors) with an associated similarity measure. For example, to measure the similarity between categorical labels like atom types, the Dirac function $1_{l_i=l_j}$ may be used. In our experiments, we use atom partial charges as atom labels, with a Gaussian kernel $K_{\mathbb{L}}(l_i, l_j) = e^{-\frac{(l_i - l_j)^2}{\lambda}}$. Of course, other similarity measures may be used as well.

Finally, atom labels are used to re-weight the contribution of two atoms x_i and y_j by $K_{\mathbb{L}}(l_i, l_j)$ in (2):

$$\text{sup-CK}_{\mathbb{L}}(P_1, P_2) = \max_{R, y_t} \sum_{\substack{x_i \in P_1 \\ y_j \in P_2}} e^{-\frac{(l_i - l_j)^2}{\lambda}} e^{\frac{\|x_i - (Ry_j + y_t)\|^2}{2\sigma^2}}, \quad (4)$$

where parameter λ controls the sensitivity of our measure to atom labels, for example to partial charges. When λ is large, impact of labels is negligible, which corresponds to a purely geometrical approach. When λ is close to zero, only pairs of atoms which have exactly the same partial charge contribute to our measure. In general, the smaller λ , the greater the contribution of the atom labels to the binding pocket similarity measure. Since the function $K_{\mathbb{L}}$ does not depend on R and y_t in (4), the same optimization procedure can be used to optimize (4) or (2).

Finally, it is important to notice that the *sup-CK* measure of similarity can be used to compare *any* set of atoms in 3D. While the primary goal of this research is to use it for comparison of binding pockets, we can also use it to compare, e.g., 3D conformations of ligands. This possibility is investigated in the experiments below.

2.2 Related methods

In this section we briefly review some of the existing methods for pocket comparison, which we compare to *sup-CK* in our experiments.

Spherical harmonic decomposition (SHD). [MNKT05] proposed to model pockets by star-shapes built using the SURFNET program. The star-shape representation is defined by a function $f(\theta, \phi)$, representing the distance from the pocket center to the pocket surface for a given (θ, ϕ) . To measure the similarity of binding pockets P_1 and P_2 , the corresponding functions f_1 and f_2 are first decomposed into spherical harmonics, and

the pocket similarity is then computed as the standard Euclidean metric between vectors of decomposition coefficients. [KMLT07] presented three different variants of *SHD*, using only the shapes of binding pockets, the sizes of the binding pockets (keeping only the zero-th order in the spherical harmonics expansion), and their combination. We only present the results of the latter in section 4, because it provided the best performance.

Poisson index (sup-PI). As we already mentioned in Section 2.1, many binding pockets similarity measures are based on pocket alignment with further counting of overlapping atoms. In particular this kind of approach is used in the *Poisson index* model [DJMT07]. More precisely the *Poisson index* model is based on normalized number of overlapping atoms $PI(P_1, P_2) = \frac{L}{\#P_1 + \#P_2 - L}$ where L is the number of overlapping atoms, and $\#P_1$ and $\#P_2$ are the respective numbers of atoms in the two pockets. The *PI* score may be computed for any pocket superposition method. While [DJMT07] used the geometric hashing algorithm, we use in our experiments the superposition made by *sup-CK* method, with further superposition refining to maximize the number of overlapping atoms.

Multibind. [SPSNW08] represent pockets by pseudo-atoms labeled with physico-chemical properties. Pockets are aligned using a geometric hashing technique. This algorithm was mainly designed for multiple alignment of binding sites, but it may be used for pairwise alignment of pockets, as was performed in this study.

Other simple methods. We also consider two simple methods based on the comparison of simple binding pockets characteristics. These methods represent each pocket by an ellipsoid constructed on the basis of pocket principal axis. The first method, referred to as *Vol*, estimates the similarity between pockets P_1 and P_2 by the absolute value of the difference between the volumes of their corresponding ellipsoids: $Vol(P_1, P_2) = |Vol(P_1) - Vol(P_2)|$. The second method, called *Princ-Axis*, estimates the similarity score between pockets by $\sum_{i=1}^3 (\lambda_i^{P_1} - \lambda_i^{P_2})^2$, where $\lambda_i^{P_1}$ and $\lambda_i^{P_2}$ are the lengths of the three principle axis of pockets P_1 and P_2 , respectively.

Combination of sup-CK and Vol. Since volume information was found to be important by [KMLT07], we also test a linear combination of the *sup-CK* and *Vol* methods, called *sup-CK-Vol*, where the coefficient of linear combination is learned as other model parameters in the double cross validation scheme. This linear combination takes advantage of the *Vol* method to separate very different pockets like PO4 and NAD, and of the *sup-CK* algorithm to allow finer discrimination.

2.3 Performance criteria

There are various ways to measure the similarity between binding pockets, some of them were discussed in the previous section. To evaluate the quality of a given similarity measure, one may compare it to some "ideal" similarity measure between binding pockets, but the problem is that such measure does not exist. As an example, given two alternative measures SM1 and SM2 applied to two pockets P1 and P2 such that SM1(P1,P2)= 0.3 and SM2(P1,P2)= 0.4, there is no way to decide which one is the best because we do not have any absolute reference. The choice of the optimal measure, thus, may depend on a particular problem of interest. In the context of ligand prediction, the quality of a similarity measure can be evaluated according to its ability to regroup together pockets binding the same ligand, which can be used to predict ligands for previously unseen binding pockets. To evaluate the regrouping quality of the similarity measures, we use two different scores.

AUC score. [KMLT07] use the AUC score which is computed as follows. Let us consider a set of pockets (P_1, \dots, P_N) and a similarity measure SM . To estimate the AUC score of a given pocket P_* , we rank all other pockets according to their similarity to P_* , $SM(P_i, P_*)$ (descending order), and we plot the ROC curve, i.e., the number of pockets binding the same ligand versus the number of pockets binding a different ligand among the top n pockets, when n varies from 0 to N . The ranking quality of SM is measured by the surface of area under the ROC curve, which defines the AUC score. An "ideal" SM function will rank all pockets binding the same ligand as P_* on the top of the list, leading to an AUC score equal to 1.0. On the contrary, if these pockets have random positions in the ranked list, the AUC score will be equal to 0.5 (worst possible case). Finally, to evaluate the overall AUC score of a method, we consider its mean value over all pockets.

While the AUC score represents an intuitive and natural way to evaluate the quality of similarities measures, in some situations it may fail. Consider the case of a dataset containing two types of pockets L_1 and L_2 (i.e. they bind two different ligands), and a similarity measure that correctly clusters pockets according to their type. If clusters are close to each other (see clusters A and C in Figure 1), the AUC score of pockets situated near the border (pockets p_1 and p_2 in Figure 1) will be low. The situation becomes even worse, if pockets binding ligand L_1 form several clusters, as shown in Figure 1, leading to low AUC scores for almost all pockets binding ligand L_1 . This similarity measure will have an overall poor AUC score, although it produces perfect separation of pocket types. This happens, for example, when the database contains proteins that underwent convergent evolution and bind the same ligand under highly different conformations. Therefore, a poor AUC score does not necessarily correspond to a poor pocket separation, and AUC scores may not be suited to evaluate the quality of similarity measures.

Classification error. These remarks lead us to employ another quality score based on classification error. To

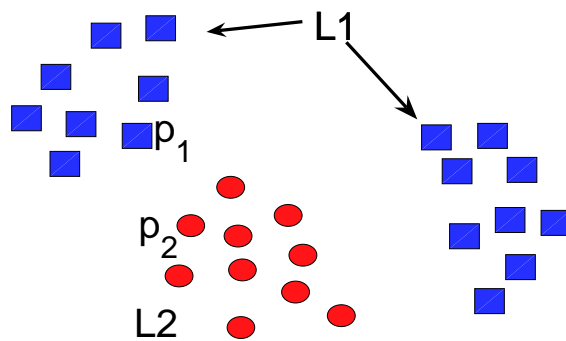


Figure 1: AUC score versus classification error as an evaluation of binding pocket similarity measure. Red circles represents pockets fixing ligand L_1 , blue squares represents pockets fixing ligand L_2 . The AUC score does not reflect the fact of good pocket clusterization, while the classification error does.

evaluate the quality of the similarity measure SM we try to predict a ligand (class) for each pocket from that of its neighbors. The smaller the classification error (proportion of bad predictions), the better the similarity measure.

In this work, we use a K nearest neighbors (KNN) classifier. To evaluate the classification error, we applied a leave-one-out double cross validation methodology. Namely, each pocket from the dataset is considered one by one, and all other pockets are used as the training set. Parameters of the model (k — number of neighbors, σ if we consider sup-SM method) are estimated on the training data via cross-validation technique, and the class (i.e. the ligand) of the pocket under consideration is predicted using the training data and the estimated parameters of the model.

2.4 Docking

Since docking programs may also predict ligands, we used the *Fred* [Nic05] and *FlexX* [RKLK96] programs. We chose these two programs because they are well referenced, and represent different strategies for ligand placement in the binding site. In all docking experiments, the active sites were the same as those used by the *sup-CK* methods. *Fred* performs rigid docking of molecules. Flexibility of ligands is taken into account by using pre-calculated conformers of a molecule. These conformers are ranked according to their estimated interaction energy with the protein, which defines the docking score (*chemgauss3* score) [MAN⁺03]. For each pocket, the predicted ligand was the most frequent molecule observed among the K first ranked molecules (K was optimized for each dataset).

FlexX performs flexible docking of molecules by fragmentation and incremental rebuilding inside the binding site. Therefore, only one ligand conformation is required as input, and the docking results are expected to be independent from that conformation. To predict a ligand for a given pocket, we choose the molecule of best docking score. In all cases, *FlexX* was run using standard parameters, with formal charges, and multiple conformations for rings were computed with Corina [GSS⁺96].

To evaluate the performance of docking programs we can use only classification error score. *Fred* and *Flex* may be used to predict binding ligands, but they do not measure similarity between binding pockets, so we can not compute the AUC score.

3 Datasets

For all protein structures, the binding pockets were extracted as follows: protein atoms situated at less than $R\text{\AA}$ of one of the ligand atoms were selected, where R is considered as a model parameter and is learned in the double cross-validation scheme. In our experiments, in most cases the optimal value of R was equal to 5.3\AA , this distance cutoff is in the range of that above which most interaction energy terms between a protein and a ligand usually become negligible. Finally, pockets are represented by 3D atom clouds with atom labeled by their partial charge, but other labels representing chemical properties such as amino-acid type could be included. Atom partial charges were attributed according to the GROMACS (FFG43a1) force field [STM⁺99].

We consider several benchmark datasets. The first one, referred to as the *Kahraman dataset*, comprises the crystal structures of 100 proteins in complex with one of ten ligands (AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD). It was proposed by [KMLT07] and is described in the Supplementary Materials. We

built an extended version of the Kahraman dataset (called *extended Kahraman Dataset* below), also described in the Supplementary Materials, in which we added protein structures in complex with one of the same ten ligands, leading to a total of 972 crystal structures. The added proteins present pairwise sequence identities less or equal to 30%, to avoid potential bias by inclusion of close homologs.

The Kahraman dataset contains only holo protein structures. However, apo structures may differ from holo structures when the latter undergo structural rearrangement upon ligand binding, a phenomenon called induced fit of the protein in order to adjust to the ligand [Bos01]. We tested a few examples of predictions for eight apo structures to evaluate the robustness of our method with respect to atom positions variability. We considered 8 apo structures corresponding to proteins able to bind one ligand from the Kahraman database: 1ADE for AMP, 1B8P for NAD, 1E4F for ATP, 1OMP for GLC, 1WS9 for FAD, 2RG7 for HEM, 1X56 for PO4 and 1N05 for FMN. These proteins share less than 30% sequence identity with any of the proteins of the extended Kahraman dataset, and had an holo structure available. The LigASite website ¹ was used for this selection. The holo and apo structures of these proteins were superposed, and the coordinates of the ligand in the holo structure were used to extract the pocket in the apo structure.

The Kahraman dataset comprises ligands of very different sizes and chemical natures. However, the real challenge is to test methods on pockets that bind ligands of similar size. Therefore, we created a third dataset comprising 100 structures of proteins in complex with ten ligands of similar size (ten pockets per ligand). This dataset will be referred to as the *Homogeneous Dataset* (HD), and is described in Supplementary Materials.

4 Results

The methods were tested on two datasets (Section 3 and Supplementary Materials). The performance of all methods is evaluated on the basis of the AUC score and the classification error (Section 2.3). The *sup-CK* method is compared to *sup-PI*, *SHD*, *Vol*, *Princ-Axis* and *MultiBind* algorithms (Section 2.2). Among the pocket extraction methods used in the *SHD* approach, we considered the results corresponding to the Interact Cleft Model, which is similar to our pocket extraction method. Results provided by the docking programs are called *Fred* and *FlexX*.

Pocket representation is subject to extraction noise. To estimate the method performance on unnoisy systems, algorithms for pockets comparison were also employed to compare ligands (except for the *MultiBind* method which is designed to be employed only on proteins).

4.1 Kahraman Dataset

Results of all methods on the Kahraman Dataset are presented in Table 1. According to the AUC score, simple

Table 1: Performances for all algorithms evaluated by the mean AUC scores and the mean classification errors (CE), over all pockets. We report only classification error for the Fred and Flex docking programs, because they can not be used to evaluate similarity between binding pockets. Column “Pockets” reports AUC and CE scores based on comparison of binding pockets. Column “Ligands” represents the same thing, but on the basis of ligands, for more explanations see text.

| Method | Pockets | | Ligands | |
|--------------------------|-------------|------|-------------|------|
| | AUC | CE | AUC | CE |
| sup-CK | 0.858±0.14 | 0.36 | 0.964±0.006 | 0.04 |
| sup-CK _L | 0.861±0.13 | 0.27 | — | — |
| sup-CK-Vol | 0.889±0.14 | 0.34 | 0.985±0.06 | 0.03 |
| sup-CK _L -Vol | 0.895±0.12 | 0.26 | — | — |
| Vol | 0.875±0.14 | 0.39 | 0.897±0.13 | 0.30 |
| Princ-Axis | 0.853±0.13 | 0.35 | 0.938±0.10 | 0.16 |
| sup-PI | 0.815±0.13 | 0.42 | 0.927±0.09 | 0.05 |
| SHD ^a | 0.770 | 0.39 | 0.920 | 0.07 |
| MultiBind | 0.715 ±0.17 | 0.42 | — | — |
| Fred | — | 0.47 | — | — |
| Flexx | — | 0.62 | — | — |

^aAUC scores are taken directly from [KMLT07], CE scores are estimated from data provided by authors

methods like *Vol* and *Princ-Axis* give surprisingly good results. The same effect was observed by [KMLT07]

¹<http://www.bigre.ulb.ac.be/Users/benoit/LigASite/>

when they used simple measure based on comparison of pocket sizes. The AUC scores of all the new methods (*sup-CK*, *sup-CK-Vol*, with or without use of partial charges) are higher than those of *ICM*, *MultiBind*, and *sup-PI*, and are in the same range than those of *Vol* and *Princ-Axis*. The best results are obtained by the *sup-CK-Vol* algorithm, which seems to benefit from the association of volume information and of more subtle geometric details provided by the *sup-CK* algorithm. Another observation, is that information on atom partial charges only leads to modest improvement of the *sup-CK* methods.

To evaluate the classification error, we tried to predict a ligand (a class) for each pocket using a K Nearest Neighbors classifier (see Section 2.3). Note that in a ten class (10 ligands) classification problem, a random classifier would have an error of 0.9, which represents baseline performance for all classifiers.

Table 1 shows that methods with higher AUC scores tend to have smaller classification errors, but this correlation is not strict. This indicates that the AUC score is not appropriate to compare similarity measures with respect to the problem of ligand identification, and underlines the interest of the classification approach.

The *sup-CK* and *sup-CK-Vol* algorithms have lower classification errors than other methods, which means that they are well suited to the problem of ligand prediction. Interestingly, atom partial charges information significantly reduces classification errors of both methods, which was not the case for AUC scores. Addition of more information for the description of pockets may improve the quality of ligand prediction. The *SHD* and *MultiBind* methods provide reasonable prediction quality, although they do not perform as well as *sup-CK*. The only difference between the *sup-PI* and *sup-CK* methods is the similarity measure used after superposition. The *sup-PI* method requires to determine the number of overlapping atoms. On the contrary, the *sup-CK* measure is based on a weighted number of atoms having close positions score taking into account, which probably leads to better results.

Docking is now widely used for ligand prediction [LSP06], and it is therefore interesting to compare its performances to those of pocket comparison methods. Table 1 shows that, on this benchmark, both docking programs do not perform as well as the *sup-CK* method, although *Fred* has better results than *FlexX*. Comparison of docking programs performances is beyond the scope of this paper, but it has been widely discussed that relative performances of docking programs strongly depend on the datasets [WAC⁺06]. They were here overall modest, but both docking programs better classified pockets associated to large ligands like FAD (flavin-adenine dinucleotide) or FMN (flavin mononucleotide), and poorly those that bind smaller ligands. These results are consistent with the fact that small ligands make few interactions, leading to low docking scores.

Since *sup-CK* method relies on 3D atom cloud representation of protein pockets, we applied it to compare ligands using their coordinates in the protein-ligand complex structures. We also recall the performances of the *SHD* algorithm for ligands of this dataset. No method reaches an AUC score of 1.0, or perfectly classifies the ligands (i.e. perfectly assign the correct ligand type). This indicates that ligands adopt different conformations in this dataset. However, performances of all algorithms are better for ligands than for pockets. Pockets have to be extracted from the protein structure, which introduces some noise that is absent in the case of ligands. This may explain better results, and represent the best expected performances for each method. In the case of ligand comparison, the best results are obtained with the *sup-CK* algorithms, although those of *SHD* and *sup-PI* are very good. The *Vol* and *Princ-Axis* methods have significantly lower results in terms of ligand classification than other methods, although their AUC scores were in the same range. Similarly, the *SHD* and *sup-PI* AUC scores are close to that of *Princ-Axis*, but they both perform much better in ligand classification than the latter.

Extension of Kahraman dataset.

To evaluate the ability of the *sup-CK* method to improve its performance when trained on a larger dataset, we consider an extension of Kahraman dataset consisting of 972 pockets that bind one of the 10 ligands of the original dataset (see Section 3). Pocket comparison and ligand prediction was performed with the *sup-CK* method including atom partial charges. The mean AUC score and classification error were equal to 0.87 and 0.18. In particular, 79% of the binding pockets of the original Kahraman dataset were correctly classified, compared to 73% on the original dataset (see Table 1). The results of the new method improve when trained on a larger dataset, which shows its ability to learn. The quality of predictions might again improve by including more structures available at the PDB.

It is also interesting to study the structure of the dataset according to the metric associated to the *sup-CK* method. We performed kernel principal component analysis [SSM99] on the pockets similarity matrix of the *sup-CK* method (this matrix is not positive definite, but we can extract principal components associated to the largest positive eigenvalues). Figure 2(a) represents the projection of 972 binding pockets on the first two principal components. Overall, we observe a clustering of binding pockets according to their ligands, which illustrates the good performances of this method for ligand prediction. Looking into more details, we notice that the clusters of pockets that bind ATP, AMP or PO₄ overlap. Indeed, proteins that binds ATP usually also bind AMP or PO₄, although with different affinities. Furthermore, some pockets (for example pockets that bind glucose GLC or FAD) are found far from their main cluster, or form secondary clusters, which illustrates that pockets having different geometrical characteristics may bind the same ligand. In the classification approach

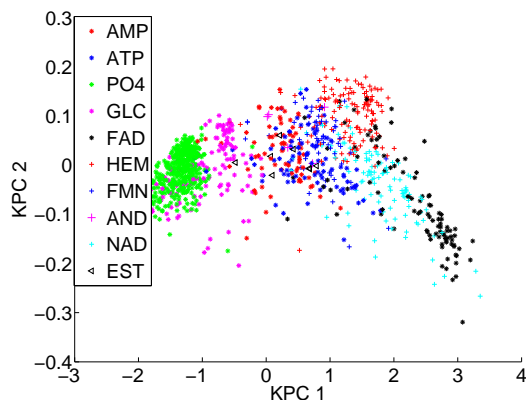


Figure 2: Projection of ext-KD on the two first kernel principal components.

employed here, prediction of a ligand for a given pocket uses the classes of its neighbors, which allows to handle the case of pockets belonging to secondary clusters.

Prediction on apo structures.

The Kahraman dataset includes protein structures in complex with a ligand which was removed, and then predicted in a "leave-one-out" procedure. However, in practice, the relevant problem will be to predict ligand for apo structures. Apo structures may differ from holo structures due to the induced fit phenomenon. Therefore we tested the performance of our method on eight apo structures (Section 3). The ligands for the eight considered apo pockets were predicted by the *sup-CK* algorithm, and the only misclassified pocket was that of 2RG7, a protein which binds HEM.

4.2 Homogeneous dataset (HD)

The Kahraman dataset contains ligands of very different sizes. It is important to test methods on a benchmark containing pockets binding ligands of similar sizes. For this reason, we built the Homogeneous dataset. Table 2 presents the performances of different algorithm on this dataset.

Table 2: Performances for all algorithms evaluated by the mean AUC scores and the mean classification errors, over all pockets.

| Method | Pockets | | Ligands | |
|--------------------------|------------|------|------------|------|
| | AUC | CE | AUC | CE |
| sup-CK | 0.710±0.19 | 0.47 | 0.892±0.14 | 0.12 |
| sup-CK _L | 0.752±0.16 | 0.38 | — | — |
| sup-CK-Vol | 0.722±0.18 | 0.46 | 0.909±0.17 | 0.12 |
| sup-CK _L -Vol | 0.766±0.17 | 0.38 | — | — |
| Vol | 0.648±0.15 | 0.89 | 0.812±0.15 | 0.54 |
| Princ-Axis | 0.650±0.18 | 0.71 | 0.830±0.20 | 0.28 |
| sup-PI | 0.702±0.19 | 0.47 | 0.880±0.14 | 0.12 |
| MultiBind | 0.69±0.14 | 0.48 | — | — |
| Fred | — | 0.54 | — | — |
| Flex | — | 0.85 | — | — |

Table 2 shows that the performance of all algorithms are lower than on the Kahraman dataset, which illustrates that the Homogeneous dataset is a more difficult benchmark. The *Vol* and *Princ-Axis* display stronger degradation of performances, with AUC scores equal to 0.65, and classification errors of 89% and 71%, respectively. This is due to the fact that the size information is less discriminative on this dataset. In terms of AUC scores, the best performance is obtained by the *sup-CK* and *sup-CK-Vol* algorithms, but volume information only provides a slight improvement of 1%, compared to 3% on the Kahraman dataset. On the contrary, partial charges information leads to a significant improvement of 4% for the *sup-CK* and *sup-CK-Vol* algorithms. This shows that addition of physico-chemical information is critical for discriminating pockets of similar sizes. In terms of classification error, volume information is useless, but the use of information on partial charge leads to significant improvement of 9%.

The same conclusions also hold for ligands comparison: performances are lower than on the Kahraman

dataset, for all methods, and degradation of the classification errors is much stronger for the *Vol* and *Princ-Axis* methods. On this dataset, the docking programs did not perform as well as methods based on pocket comparison in terms of classification errors.

5 Discussion

An important characteristic of the *sup-CK* algorithm is its ability to adapt to the pocket variability. Parameter σ of the *sup-CK* method controls the sensitivity of the similarity measure to atom relative displacements. The larger the variability of pockets binding the same ligand, the greater should be the value of σ . Figure 3a shows how the AUC score and classification error vary with σ on the Homogeneous dataset. In both cases, the optimum is reached when σ is equal to one. Note that, in our experiments (section 4), we did not use the same value of σ estimated from all pockets. For each pocket, the optimal value was estimated on the basis of the 99 training pockets to avoid overfitting to the data. However, we observed that, in most cases (90%), $\sigma = 1$ was chosen. Similarly, when information on atom partial charges is used, parameter λ (4) conditions the sensitivity of the method to relative values of atom charges. Figures 3b and 3c present the variation of AUC scores and classification error as functions of σ and λ . We observe that for the AUC score, the optimum is reached when σ equals to 2 and λ equals to 0.25, while for the classification error optimal σ is equal to 4.

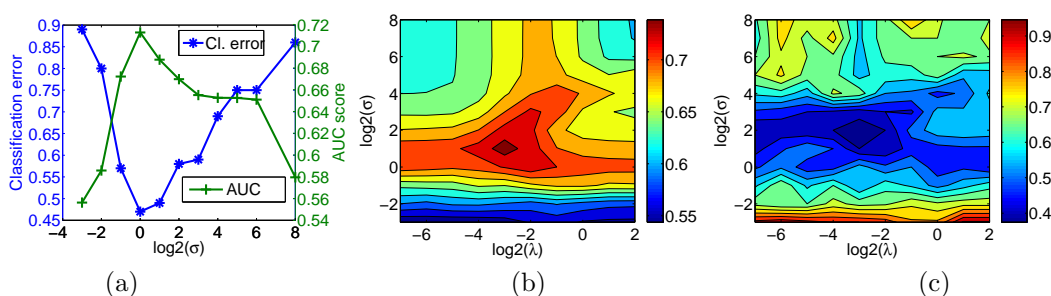


Figure 3: Homogeneous database. (a) AUC score and prediction error as functions of σ in the *sup-CK* method (pure geometrical version, $\lambda = \infty$), (b) AUC score and (c) classification error as functions of σ and λ when information on atom partial charges is used.

Figures 4b and 4c illustrate the optimal alignment found for two ATP binding pockets. While this alignment was estimated on the basis of pocket atom coordinates, the bound ligands are found well aligned, which suggests a good quality of pocket alignment. Note, that *sup-CK* does not try to superpose individual atoms, but rather superposes atom sets.

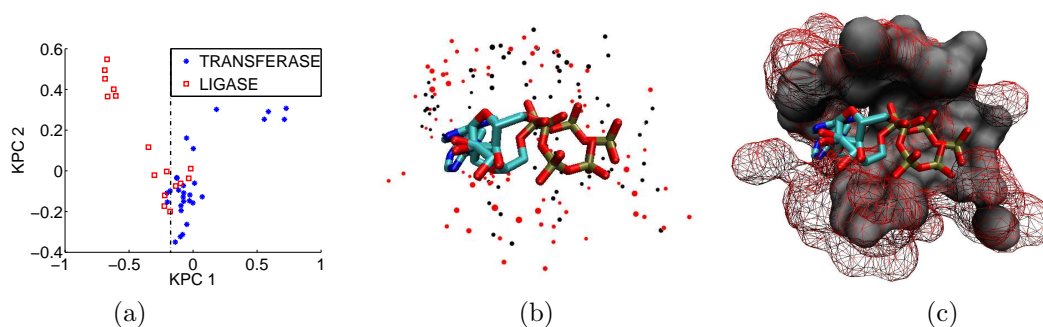


Figure 4: (a) Projection of ATP binding pockets on the two first kernel principal components of *sup-CK*. (b,c) Alignment of two ATP pockets made by *sup-CK*, atoms of different pockets are represented by black and red points in (b) and by black and red surfaces in (c), two ATP ligands are traced in licorice.

The running time of the *sup-CK* method depends on the value of stopping criterion used in the gradient ascent method and on the number of atoms. In our experiments, the algorithm running time varied between 0.2 and 1.3 seconds (2.5 GHz CPU) per pocket pair. This running time is already quite reasonable to process large protein databanks, however a pre-filtering on the basis of simple pocket descriptors (like volume or size) may be quite useful in the further acceleration of the *sup-CK* method. We defined pockets as the set of all protein atoms within 5Å of a bound ligand. Similar approaches were used by [KMLT07] (Interacted Cleft Model), and

similar pockets may also be retrieved by methods like *Q-SiteFinder* [LJ05] without any information on ligand coordinates.

In our experiments, docking programs (*FlexX* and *Fred*) did not perform as well for ligand prediction as most methods based on pockets similarity measure. Docking programs have many parameters that can be tuned to particular protein-ligand systems [ATL⁺07]. Fine preparation of the active site, such as assignment of amino acid protonation states, is also critical. Such tuning for each pocket is hardly automatized in large scale datasets (up to almost 1000 proteins in this study), and therefore, the performance of docking programs is underestimated.

An important topic is the relation between methods for binding pockets comparison and algorithms in field of computer vision for comparison of 3D shapes. A complete review of 3D shape comparison methods is out of scope of this article, and interested readers may consult [IJL⁺05] for a detailed review. Interestingly, most of existing methods for binding pocket comparison have an analogue in the domain of computer vision. For example, methods based on real spherical harmonic expansion used in [MNKT05] for binding pocket comparison are also discussed by [PPPT07, SV01] in the context of general 3D shape matching. Principles used in another popular method for matching and comparison of 3D forms, called Iterative Closest Point algorithm [ZHA92], and its variants are used in *Poisson index* and *MultiBind* algorithms. Examples of approaches based on graph representation of 3D forms and graph matching methods may be found in [WHKK07] for binding pockets comparison, as well as in [BMM⁺04] for 3D shapes comparison. Nevertheless, binding pockets are not continuous shapes but discrete clouds of points. They can be transformed into 3D shapes [MNKT05, KMLT07], but this transformation may be a source of noise. Moreover, a similarity measure between binding pockets should be rotationally and translationally invariant, which is not always the case in computer vision methods. However, we believe that the adaptation of appropriate methods may be very fruitful for the recognition of binding pockets.

The prediction of protein ligands is related to the problem of predicting the protein molecular function. We analyzed the repartition of the ATP binding pockets generated by this similarity measure on the extended Kahraman dataset. Figure 4a presents the projection of ATP pockets annotated as transferases or ligases, on the first two principal components of the *sup-CK* similarity matrix. We observed that these two families of enzymes are essentially separated. Although these are very preliminary results, they show that *sup-CK* method may be useful in the prediction of protein molecular functions.

The *sup-CK* algorithm showed a good performance in ligand prediction for apo structures. This is an important preliminary result, in order to apply the method to real case studies, or to proteins with no known experimental structure but for which a homology model can be constructed [LS08].

References

- [ATL⁺07] C. David Andersson, Elin Thysell, Anton Lindström, Max Bylesjö, Florian Raubacher, and Anna Linusson. A multivariate approach to investigate docking parameters' effects on docking performance. *J Chem Inf Model*, 47(4):1673–1687, 2007.
- [BMM⁺04] Silvia Biasotti, Simone Marini, Michela Mortara, Giuseppe Patane, Michela Spagnuolo, and Bianca Falcidieno. 3d shape matching through topological structures. In *Discrete Geometry for Computer Imagery*, pages 194–203. Springer Berlin / Heidelberg, 2004.
- [Bos01] H. R. Bosshard. Molecular recognition by induced fit: how fit is the concept? *News Physiol Sci*, 16:171–173, Aug 2001.
- [DJMT07] J.R. Davies, R.M. Jackson, K.V. Mardia, and C.C. Taylor. The poisson index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics*, 23(22):3001–3008, Nov 2007.
- [GFKS02] T. Gärtner, P.A. Flach, A. Kowalczyk, and A.J. Smola. Multi-Instance Kernels. In C. Sammut and A. Hoffmann, editors, *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 2002.
- [GJ06] N.D. Gold and R.M. Jackson. Sitesbase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res*, 34:D231–D234, Jan 2006.
- [GMN⁺06] F. Glaser, R. J. Morris, R. J. Najmanovich, R. A. Laskowski, and J. M. Thornton. A method for localizing ligand binding pockets in protein structures. *Proteins*, 62(2):479–488, February 2006.
- [GSS⁺96] Johann Gasteiger, Jens Sadowski, Jan Schuur, Paul Selzer, Larissa Steinhauer, and Valentin Steinhauer. Chemical information in 3d space. *Journal of Chemical Information and Computer Sciences*, 36(5):1030–1037, 1996.

- [Hau99] D. Haussler. Convolution Kernels on Discrete Structures. Technical Report UCSC-CRL-99-10, UC Santa Cruz, 1999.
- [IJL⁺05] Natraj Iyer, Subramaniam Jayanti, Kuiyang Lou, Yagnanarayanan Kalyanaraman, and Karthik Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-Aided Design*, 37(5):509–530, April 2005.
- [KJ03] Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *In International Conference on Machine Learning (ICML)*, 2003.
- [KMLT07] Abdullah Kahraman, Richard J Morris, Roman A Laskowski, and Janet M Thornton. Shape variation in protein binding pockets and their ligands. *J Mol Biol*, 368(1):283–301, Apr 2007.
- [LJ05] Alasdair T. R. Laurie and Richard M. Jackson. Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.
- [LS08] Guillaume Launay and Thomas Simonson. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics*, 9:427, 2008.
- [LSP06] Andrew R Leach, Brian K Shoichet, and Catherine E Peishoff. Prediction of protein-ligand interactions. docking and scoring: successes and gaps. *J Med Chem*, 49(20):5851–5855, Oct 2006.
- [MAN⁺03] Mark R. McGann, Harold R. Almond, Anthony Nicholls, Andrew J. Grant, and Frank K. Brown. Gaussian docking functions. *Biopolymers*, 68(1):76–90, 2003.
- [MNKT05] R. J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355, May 2005.
- [Nic05] A. Nicholls. Oechem, version 1.3.4, openeye scientific software. website, 2005.
- [NKT08] R. Najmanovich, N. Kurbatova, and J. Thornton. Detection of 3d atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 24(16):i105–i111, Aug 2008.
- [PPPT07] Panagiotis Papadakis, Ioannis Pratikakis, Stavros Perantonis, and Theoharis Theoharis. Efficient 3d shape matching and retrieval using a concrete radialized spherical projection representation. *Pattern Recogn.*, 40(9):2437–2452, 2007.
- [RKLK96] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, Aug 1996.
- [SPSNW08] Alexandra Shulman-Peleg, Maxim Shatsky, Ruth Nussinov, and Haim J. J. Wolfson. Multibind and mappis: webservers for multiple alignment of protein 3d-binding sites and their interactions. *Nucleic Acids Res*, 36:260–264, May 2008.
- [SSKR08] Claire Schalon, Jean-Sbastien Surgand, Esther Kellenberger, and Didier Rognan. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, 71(4):1755–1778, Jun 2008.
- [SSM99] B. Schölkopf, A.J. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 327–352. MIT Press, 1999.
- [STM⁺99] Walter R. P. Scott, Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, and Peter Kruger. The gromos biomolecular simulation program package. *J. Phys. Chem. A*, 103:3596–3607, 1999.
- [STV04] B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, The MIT Press, Cambridge, Massachusetts, 2004.
- [SV01] Dietmar Saupe and Dejan V. Vranic. 3d model retrieval with spherical harmonics and moments. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, pages 392–397, London, UK, 2001. Springer-Verlag.

- [WAC⁺06] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, and M.S. Head. A critical assessment of docking programs and scoring functions. *J Med Chem*, 49(20):5912–5931, Oct 2006.
- [WHKK07] Nils Weskamp, Eyke Hullermeier, Daniel Kuhn, and Gerhard Klebe. Multiple graph alignment for the structural analysis of protein active sites. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 4(2):310–320, 2007.
- [WWB86] Peter Willett, Vivienne Winterman, and David Bawden. Implementation of nearest-neighbor searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences*, 26(1):36–41, 1986.
- [ZHA92] Zhengyou ZHANG. Iterative point matching for registration of free-form curves. Technical report, Institut National de Recherche en Informatique et en Automatique (INRIA), 1992.