BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 7 2013, pages 855–861
doi:10.1093/bioinformatics/btt058

Structural bioinformatics

Advance Access publication February 8, 2013

aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction

Douglas E. V. Pires[1,2,*], Raquel C. de Melo-Minardi[1,*], Carlos H. da Silveira[3], Frederico F. Campos[1] and Wagner Meira, Jr[1]

[1]Department of Computer Science, [2]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha Belo Horizonte - MG, 31270-901, and [3]Advanced Campus at Itabira, Universidade Federal de Itajubá, Rua Irmã Ivone Drumond, 200 - Distrito Industrial II Itabira - MG, 35903-087, Brazil

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Receptor-ligand interactions are a central phenomenon in most biological systems. They are characterized by molecular recognition, a complex process mainly driven by physicochemical and structural properties of both receptor and ligand. Understanding and predicting these interactions are major steps towards protein ligand prediction, target identification, lead discovery and drug design.

Results: We propose a novel graph-based–binding pocket signature called aCSM, which proved to be efficient and effective in handling large-scale protein ligand prediction tasks. We compare our results with those described in the literature and demonstrate that our algorithm overcomes the competitor's techniques. Finally, we predict novel ligands for proteins from *Trypanosoma cruzi*, the parasite responsible for Chagas disease, and validate them *in silico* via a docking protocol, showing the applicability of the method in suggesting ligands for pockets in a real-world scenario.

Availability and implementation: Datasets and the source code are available at http://www.dcc.ufmg.br/~dpires/acsm.

Contact: dpires@dcc.ufmg.br or raquelcm@dcc.ufmg.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 20, 2012; revised on January 17, 2013; accepted on February 3, 2013

1 INTRODUCTION

1.1 Background

Molecular recognition plays a fundamental role in most cellular processes. The conditions responsible for the binding and interaction of two or more molecules are a combination of conformational and physicochemical complementarity (Kahraman *et al.*, 2007). Understanding the receptor-binding pocket requirements for this recognition process is a major step towards protein ligand prediction, target identification, lead discovery and drug design.

It is assumed that similar ligands have similar binding sites in terms of shape and physicochemical properties. Several methods were proposed to describe, compare and predict ligands with binding pockets. However, despite the relevant contributions of the majority of the works, methods that rely on multiple structure alignments and pairwise pocket comparisons might be prohibitively expensive for large-scale experiments. As the availability of biological data has been growing in an exponential fashion in the past years, scalability has become a crucial characteristic for the execution of such tasks in real-world scenarios.

To overcome these challenges, we proposed a novel methodology for receptor-based protein ligand prediction, which is supported by a graph-based pocket signature. We extract distance patterns from protein pockets modeling them as atomic graphs and performing a noise and dimensionality reduction preprocessing step, which granted not only an improvement in efficacy in comparison with competitors works but also scalability to the methodology.

Atomic distance patterns perceive the structure arrangements of the protein and, therefore, reflect its function. This way, using this information to describe ligand-binding pockets is an appropriate strategy, given the close relationship between protein structure and function as well as the importance of shape complementarity in the molecular recognition process. Furthermore, considering the physicochemical properties in these patterns, also an important requirement for recognition, gives the description power needed to successfully describe, compare and predict protein–ligand interactions.

Receptor-binding pockets can be seen as graphs where nodes are the protein atoms and the edges are the chemical interactions established among them. Topological and chemical properties can be extracted from these graphs and summarized in a molecular recognition signature. These compact signatures can then be used in large-scale ligand prediction tasks. In this work, we derive a novel pocket signature from the Cutoff Scanning Matrix [CSM (Pires *et al.*, 2011)], which is essentially a graph-based signature successfully used for structural classification and function prediction tasks. We propose an atomic labeled version of the signature (henceforth called aCSM) that is independent of molecular orientation and does not require any ligand information in its calculation.

Given the complexity of the recognition process, these signatures are expected to be robust for ligand prediction. One of these difficulties, and an important source of noise in data, is ligand flexibility, which leads to a great conformational diversity. For instance, Figure 1a illustrates five representatives of nicotinamide–adenine dinucleotide (NAD), a highly flexible ligand, obtained from the Protein Data Bank (PDB). They were aligned

*To whom correspondence should be addressed.

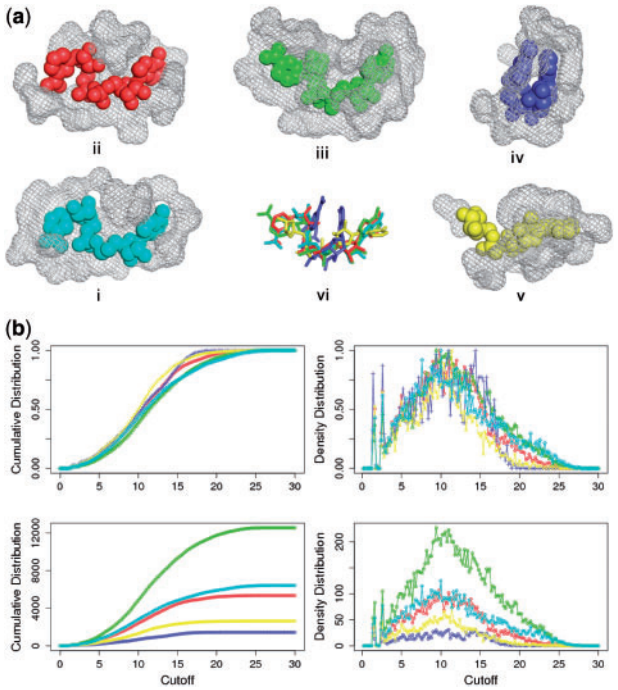

Fig. 1. Ligand conformational diversity. (**a**) NAD molecules presenting different conformations and the impact in its respective pockets (calculated using a distance threshold of 5 Å). The PDB IDs considered were (a.i) 3KSD:Q (ligand in cyan), (a.ii) 1A5Z:A (ligand in red), (a.iii) 1NAH:A (ligand in green), (a.iv) 1ZRQ:B (ligand in blue), (a.v) 2OOR:B (ligand in yellow). (a.vi) The ligands aligned by the program LigAlign. The cumulative and density distributions of aCSM signatures, fully explained in the next section, are presented in the same colors in (**b**), considering normalized (top) and absolute values (bottom)

and their pockets calculated by a distance criterion. We can see that the variability of conformations directly impacts on the binding pocket size and shape. Besides that, the induced fit mechanisms (Koshland, 1958) as well as allosteric regulations (Monod *et al.*, 1963) may promote expressive conformational changes in the protein target.

Other challenging factor is the several poses adopted by ligands in different pockets and its solvent accessibility when bound. Figure 2 presents an example where flavin–adenine dinucleotide (FAD) is bound to three pockets with different degrees of solvent accessibility. It is clear that these factors could dramatically affect pocket shape and size for the same ligand, which may impose severe limitations to methods that rely solely on structural alignments. In our case, it is also a considerable source of noise for the proposed signatures, which are based on atomic distance patterns.

To deal with these challenges and eliminate inherent noise, we apply singular value decomposition (SVD)-based noise and dimensionality reduction strategy. A detailed description about the technique as well as references can be found in Supplementary Material. Figure 1b presents the proposed signatures for the NAD-binding sites before noise reduction. We can see that, despite the similarity in the curves profile, we still see a considerable variability among them that is reduced by the data normalization achieved after SVD preprocessing. This preprocessing step is essential to extract from the original signatures,

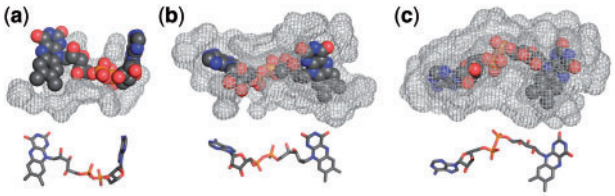

Fig. 2. Ligand solvent accessibility diversity. The figure shows FAD molecules with different degrees of solvent accessibility and its impact in the respective pocket (calculated using a distance threshold of 5 Å). The PDB IDs used were (**a**) 1O26:A, (**b**) 1AHV:A and (**c**) 1H83:C. Pymol CPK coloring scheme: Cs in gray, Ns in blue, Os in red and Ss in yellow

the components that are the most important to describe the pockets discarding redundancy and non-conserved dimensions.

1.2 Related works

To describe protein pockets to either compare them and/or predict their ligands, several methods have been proposed in the literature. Some of them are based on a paradigm of pocket similarity metrics. In the study conducted by Davies *et al.* (2007), the authors introduced a matching score for binding sites based on a probabilistic model and compared their metric with the Tanimoto index. Protein-binding pockets were compared by spherical harmonic decompositions (Morris *et al.*, 2005), technique also used in a study of their shape variation (Kahraman *et al.*, 2007). More recently, in a study conducted by (Hoffmann *et al.*, 2010), the authors proposed a method to quantify pocket similarity by representing them as clouds of atoms and comparing the resulting alignments with a convolution kernel. A measure of similarity was also derived in a recent work (Ueno *et al.*, 2012) from radial distribution functions of physicochemical properties of catalytic sites, information that was then used to cluster enzymes by function. In the study conducted by Gonçalves-Almeida *et al.* (2012), pockets are compared using hydrophobic patches represented by geometric centroids, and their conservation is detected despite sequence and structure dissimilarity.

Another set of methods attempts to compare ligand-binding sites based on multiple alignments. Similarity metrics were derived from the alignments of binding sites or cavity fingerprints represented by its physicochemical or topological properties in the studies conducted by Shulman-Peleg *et al.* (2008) and Schalon *et al.* (2008), whereas Spitzer *et al.* (2011) proposed a surface-based approach. There are also efforts that use multiple graph alignments and clique-based matching algorithms (Weskamp *et al.*, 2007; Najmanovich *et al.*, 2008) to perceive receptor-ligand interactions. Other alternative approaches in the study of binding mechanism include the use of docking and quantitative structure–activity relationships techniques (QSARs) (Sippl, 2000).

1.3 Summary of results

We showed the proposed signatures successfully deal with the challenging aspects of large-scale ligand prediction achieving an area under ROC curve (AUC) of 0.92 for a dataset composed by >35 000 enzyme pockets. As far as we are concerned, no other method was tested with a dataset of comparable volume. Despite the prominent variability of NAD, our methodology was able to retrieve their pockets with an AUC up to 0.96. We recovered as

well FAD sites presenting molecules with different solvent accessibilities achieving an AUC of 0.99. When compared with state-of-the-art methods, our approach achieves comparable or better results. Finally, we present a case study where we predict novel ligands for proteins from *Trypanosoma cruzi*, the parasite responsible for Chagas disease and validate them *in silico* via a docking protocol showing the applicability of the method in a real-world scenario.

2 MATERIALS AND METHODS

In this section, we explain the basis for defining our noise-free graph signatures, describe the datasets used in the experiments and explain the evaluation strategies. First, we describe the CSM method. Our strategy to reduce noise and dimensionality turning aCSM precise and robust as well as scalable to large datasets is explained in detail in Supplementary Material as well as how the classification algorithms used work and why they were chosen. Finally, we explain how the method was validated. Details about the used quality measures are also available in Supplementary Material. Figure 3 shows the aCSM-based ligand prediction workflow. It is divided into the following main steps: data collection, signature generation and noise/dimensionality reduction, supervised learning, ligand prediction and validation. A more detailed workflow can be found in Supplementary Figure S1.

2.1 aCSM-based signatures

CSM is a protein structural signature proposed by Pires *et al.* (2011) and successfully used in large-scale protein function prediction and structural classification tasks. The original CSM workflow generates, for each protein, a feature vector that represents distance patterns between protein residues represented by centroids, which are then used as evidence for the classification procedures. To reduce noise as well as data dimensionality, SVD (Demmel, 1997) was used as a preprocessing step.

Although other dimensionality reduction methods may be used, Supplementary Figure S2 shows that two well-established feature selection approaches were less effective than SVD in reducing noise and dimensionality of the data and, thus, improving classification performance.

Inter-residue distance patterns were also subject of study of our previous study (da Silveira *et al.*, 2009) and showed to be conserved across protein folds.

In the present work, we extend the inter-residue signature to an atomic level (atomic CSM or aCSM for short). The aCSM-based signatures are generated as follows: for each protein, we create a feature vector. First, we compute the Euclidean distance between all pairs of atoms and define a range of distances (cut-offs) to be considered and a distance step. We scan through these distances, computing the frequency of pairs of atoms that are close according to this distance threshold, i.e. the atoms in contact.

Furthermore, we propose in this work three new different types of aCSM-based signatures using atomic physicochemical properties.

- **aCSM:** generates one value per cut-off, corresponding to the number of atoms in contact according to this distance threshold.
- **aCSM-Hydrophobic patche (HP):** generates three values per cut-off, i.e. the frequency of hydrophobic–hydrophobic, hydrophobic–polar and polar–polar contacts.
- **aCSM-ALL:** considers eight categories: hydrophobic, positive, negative, acceptor, donor, aromatic, sulfur and neutral. The combination of these atoms labels generates 36 values per cut-off. The atoms classification was obtained by the program PMapper at pH 7. PMapper perceives pharmacophoric properties of atoms in a given molecular structures.

Algorithm 1 shows the function that calculates the atomic version of CSM. To compute a signature, one must supply the following input parameters: a set of proteins and the atomic categories to be considered, a cut-off range (D_{MIN} and D_{MAX}) and a cut-off step (D_{STEP}) in which each cut-off is discretized. In line 1, we define the prototype of the aCSM function. In line 2, we iterate through each i of the proteins of the input dataset. Line 3 shows the initialization of a variable used to index the signature array. In line 4, we call a function that computes the pairwise distances between all pairs of atoms of a protein and return and store these data in *distMatrix*. The loop in line 5 controls the iterations used to scan the *distMatrix* to compute the signature. In line 6, we iterate through the considered atom classes, and finally in lines 7–8, we call a function that computes the frequency of contacts for the current distance, between atoms of the given classes and store it in the corresponding signature array position. The aCSM generation runs in quadratic time, i.e. has time complexity of $O(n^2)$, where n is the number of atoms of the pocket, because of the pairwise distance computation. It is important to point out that the method is easily parallelizable, an important and desired characteristic for its efficient use in multi-core processor architectures.

In the experiments presented in this work, we vary the distance threshold from $D_{MIN} = 0.0$ Å to $D_{MAX} = 30.0$ Å, with a $D_{STEP} = 0.2$ Å, which generated for each type of signature a vector of 151, 453 and 5436 entries for each protein. In Supplementary Figure S3, one can see through the pocket diameter distribution that 30.0 Å accounts to ~95% of the pockets of the extensive enzyme dataset.

The aCSM might also be presented as a graph-based signature, as the information regarding each cut-off distance represents the number of edges of an atomic contact graph assembled considering this cut-off. Notice also the method generality, as it can be applied to predict both proteic and non-proteic ligands.

Algorithm 1. Atomic Cutoff Scanning Matrix calculation

```
1: function aCSM(ProteinSet, AtomClass, D_MIN, D_MAX, D_STEP)
2:   for all protein i ∈ (ProteinSet) do
3:     j = 0
4:     distMatrix ←calculateAtomicPairwiseDist(protein)
5:     for dist ← D_MIN; to D_MAX; step D_STEP do
6:       for all class ∈ (AtomClass) do
7:         aCSM[i][j] ← getFrequency(distMatrix, dist, class)
8:         j++
9:   return aCSM
```

2.2 Evaluation methodology

We evaluate the proposed method and compared it with other state-of-the-art algorithms using cross-validation computed metrics, especially AUC.

Cross-validation: is a traditional statistical analysis used to estimate the performance of predictive models. It consists of partitioning the dataset into two complementary subsets. The first is the training set used to build the model. The other is the test set used to measure the validity of the model. The dataset is partitioned, and the metrics are averaged over all the rounds. We use 10-fold cross-validation in all the experiments, but the comparative ones. For comparisons, we use leave-one-out cross-validation because this was the technique used by competitor's works.

AUC: is the area under ROC curve. ROC curves are explained in more detail in Supplementary Material. They provide a visual tool for examining the trade-off between the ability of a classifier to correctly identify positive cases and the number of negative cases that are incorrectly classified. An interesting feature of these curves is that the AUC can be used

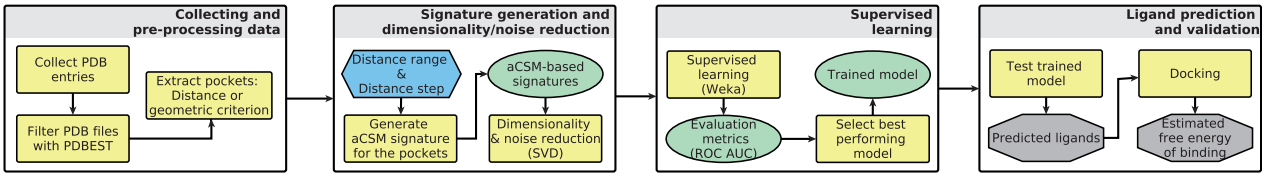

Fig. 3. aCSM-based ligand prediction workflow. The workflow is divided into four main steps: data collection, signature generation and noise/dimensionality reduction, supervised learning, ligand prediction and validation. Hexagonal blue boxes denote input files/parameters, ellipsoid green boxes are intermediate files generated, rectangular yellow boxes denote the intermediate steps and the octagonal gray boxes the outputs, i.e. the predicted ligands and its estimated binding free energy

as a measure of accuracy in many applications. The AUC ranges from 0 to 1, and a random classifier would have an AUC of 0.5.

2.3 Data

2.3.1 Datasets To support multiple types of experiments to validate our method's quality, generality and real-world applicability, we used four different databases with different purposes:

(a) **Large-scale enzyme dataset:** In a previous work (Pires *et al.*, 2011) we have proposed a dataset of the top 950 most-populated Enzyme Commission (EC) numbers, in terms of available structures, with at least nine representatives per class, concerning 55 474 chains. This dataset consists of reviewed enzymes from UniProt, i.e. the experimentally validated annotations from that database. Only ligands with seven or more atoms were considered, and also the pockets with <10 atoms were discarded. A total of 35 480 pockets were generated for 604 different ligands, with at least 10 representatives per ligand. This dataset is used to show the applicability of the method for very diverse real-world large enzyme database.

(b) **Kahraman dataset:** proposed by Kahraman *et al.* (2007), it comprises 100 protein-binding sites that are non-evolutionary related, X-ray–solved, complexed with 10 different ligands with various sizes and flexibility (namely, AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD). This dataset is used in comparisons of our method and its competitors.

(c) **Hoffmann Homogeneous dataset (HD):** proposed by Hoffmann *et al.* (2010), it is formed by 100 protein pockets complexed with 10 different ligands of similar size and was assembled by the authors to complement the *Kahraman dataset*, as it has ligands of different volumes. This dataset is also used to compare our method with its competitors.

(d) ***Trypanosoma cruzi* dataset:** composed of *T.cruzi* proteins. The criteria adopted for the protein selection was proteins solved by X-ray crystallography, with resolution <2.5 Å. A total of 104 PDB IDs, comprising 200 chains, were gathered. We used a 5 Å distance criteria to define the pockets. After removing crystallographic artifacts, 225 pockets were selected. This dataset is used to raise candidate ligands, using the aCSM signatures, and validate them via a docking protocol.

2.3.2 Data preprocessing All protein structures were collected from PDB, filtered and preprocessed using the PDBest toolkit. The proteins chains were split in separate files, and the binding pockets for each ligand were extracted.

2.3.3 Pocket computation Ligand pockets were extracted from protein structures in two ways:

- **Distance criterion:** considering a distance of 5 Å, i.e. only atoms within 5 Å from the ligand were selected. This criterion was used by the competitor's works.
- **Geometric criterion:** we compared the efficacy of two geometric methods in defining pockets, namely, the grid-based method that uses mathematical morphology, implemented by Ghecom (Kawabata, 2010), and the method based on α-shapes theory implemented by FPocket (Le Guilloux *et al.*, 2009). In both cases, we chose the pocket that has the closest atom from the ligand, i.e. the pocket that probably is more in contact with it.

3 RESULTS

To test and validate the ability of our signature to describe binding sites to support and aid in protein–ligand interaction prediction tasks, we designed an extensive set of experiments. First, we show our method can be used in large-scale ligand prediction and evaluate its precision in doing this task. Then, we compare the three proposed versions of aCSM signatures and evaluate which one presents the best descriptive power to ligand prediction. After that, we present the comparative results concerning state-of-the-art methods described in the literature and its respective datasets. Finally, we apply our methodology to pockets of *T.cruzi* proteins and predict ligands to them, comparing the binding free energies of the ligands docked with receptors with those of real complexes available at the PDB, via a redocking protocol, and with ligands randomly chosen.

3.1 Large-scale experiments

Figure 4 presents the AUC of our method for the large-scale enzyme dataset composed of reviewed enzymes from UniProt from which >35 000 pockets were extracted. We can see that the methods successfully predicted ligands in every type of experiment described with precisions going from 0.6 to 0.92. In the next sections, we explain the variations of the method that generated the results showed in the figure.

3.1.1 Signature types evaluation In the left-hand graph of Figure 4, we compare aCSM, aCSM-HP and aCSM-ALL in terms of their AUC achieved with different number of singular values used to approximate the original matrix.

In one hand, we can see that the more specific the signature is, in terms of physicochemical atom properties, the more precise it is in ligand prediction. With 100 singular values, for pockets extracted via a distance criterion (three upper curves), aCSM-ALL reaches an AUC of 0.92 as the basic aCSM presents an AUC of 0.75. On the other hand, with <20 singular values, the

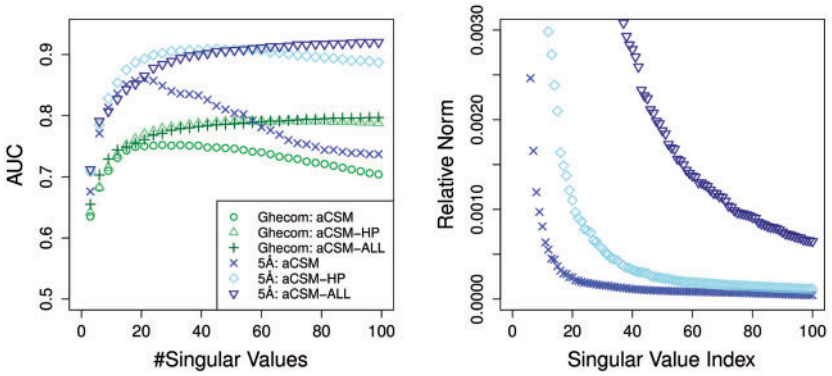

Fig. 4. Comparative prediction performance, in terms of the AUC metric, between two methods for defining the pocket: Ghecom (green shades/three lower curves) and threshold distance (blue shades/three upper curves). For each method, the performance of the three signatures types proposed is also compared. The large-scale enzyme dataset was used in this experiment, as well as the KNN classifier. The right-hand graph shows the relative norm for different dimensionality cut-offs as a metric to assess the quality of the data approximation provided by SVD. The low values of the relative norm at the maximum values of AUC indicate that the low-rank matrices approximate the original data well

difference between the different signatures is almost null reaching a high AUC score of 0.85.

It is interesting to notice that aCSM-ALL is the only one that seems to have benefited from the addition of many singular values. The first singular values respond to the higher data variability and are the most informative ones. As long as we add singular values to the signature, we add more and more noise to data, as well as we demand more computational time. These results show that aCSM presents an intrinsic limitation when 20 singular values are used, and a pick of AUC of ~0.86 is achieved. aCSM-HP reaches >0.90 with ~40 singular values. aCSM-ALL AUC is improved when we add successive singular values, and it does not converge until 100 singular values. It could indicate the absence of noise when we label atoms in such a specific way.

It is worth mentioning that the maximum value of AUC for the aCSM-ALL signatures occurs when 177 singular values are used (0.924). However, this improvement in AUC is marginal compared with the increase in computational cost.

To assess the quality of the data approximation provided by SVD, we measured the relative norm for each number of singular values being considered (right-hand graph of Fig. 4). The 2-norm can be used to measure the relative difference ρ_k between the reduced matrix A_k and the original matrix A. Considering the Theorem 1 from Supplementary Data and that the spectral norm $||A||_2$ is, simply, the first (largest) singular value of A, i.e. $||A||_2 = \sigma_1$, thus:

$$\rho_k = \frac{||A - A_k||_2}{||A||_2} \rightsquigarrow \rho_k = \frac{\sigma_{k+1}}{\sigma_1}. \quad (1)$$

Supplementary Table S2 shows the relative norm (1) at the maximum values of AUC for signatures 5 Å-aCSM, 5 Å-aCSM-HP and 5 Å-aCSM-ALL. These small values of ρ_k indicate that the low-rank matrix A_k approximates significantly the original data matrix A.

Supplementary Table S1 shows the comparison of the three proposed signatures, for the large-scale enzyme dataset, in terms of several quality measures, as well as in terms of mean execution time. The best results were achieved by aCSM-HP and aCSM-ALL after SVD preprocessing. aCSM-ALL is slightly better in terms of accuracy, even though it takes twice the time to run in comparison with aCSM-HP. In conclusion, aCSM-ALL is the better choice being the most accurate and having a non-prohibitive execution time.

3.1.2 Pocket detection method influence Figure 4 also shows the comparison of the aCSM signatures computed for pockets delimited using distance and geometric criterion (three lower curves). We can see that using the geometric method, the results are systematically ~12% worse than simply with distance criteria. This is probably because of loss of important molecular information when using Ghecom algorithm.

There is, in fact, a big difference in using a geometric criterion to detect pockets rather than using experimental data from complexes, as automatically detecting pockets is still challenging. Nevertheless, we believe that one of the main contributions of this work is the improvement in similarity assessment provided by aCSM when structures of complexes are available.

Even if we aggregate more atoms than the ones that were in fact accessible in the pocket with 5 Å cut-off, our method behaves robustly, being able to discard unnecessary or irrelevant information. Supplementary Figure S4 shows that for cut-off distances >5 Å, the predictive performance of our method increases.

We use 5 Å as a cut-off criterion because it was the same adopted by the competitors works. However, this value seemed to be defined arbitrarily and not necessarily reflects the best possible cut-off for every method. To evaluate this hypothesis, we contrast our signature performance according to pocket distance criterion for the large-scale enzyme dataset. In fact, in Supplementary Figure S5, we show that the best distance criterion for the aCSM signature, using the K-nearest neighbor algorithm (KNN) classifier, was actually 6.0 Å. This value is in agreement with other authors who also have investigated about the best atomic cut-off when the network of contacts is computed using a heavy atom proximity criterion (Zhang *et al.*, 1997; Kamagata and Kuwajima, 2006). It is important to stress that

with this cut-off, we have minimum noise in our signatures, as the best performing SVD cut-off is chosen.

3.2 Comparison with state-of-the-art methods

The experiments described later in the text were performed in two datasets (*Hoffmann HD* and *Kahraman*) already used by several related studies to compare them with our protein pocket signature *aCSM*. Leave-one-out cross-validation was used in all experiments regarding these two datasets; the same methodology was used in the related works.

Table 1 summarizes the results obtained. The aCSM signature achieved better results, considering the AUC score, in comparison with the other methods with low standard deviation. It is important to stress that leave-one-out cross-validation is computationally demanding and not suitable for a large-scale real-world scenario. Moreover, the two aforementioned datasets are small (only 100 pockets) and, in the case of the Kahraman dataset, divided into unbalanced classes, which makes the learning process of the classifiers difficult.

3.3 Case study: predicting ligands for *T.cruzi* proteins

Chagas disease is a tropical infection caused by the protozoan parasite *T.cruzi* that affects ~8 million people in Latin America (Rassi *et al.*, 2010) and is the leading etiology of non-ischemic heart disease worldwide. Unfortunately, the two available drugs for treatment (Nifurtimox and Benznidazole) have potential toxic side effects and variable efficacy (Canavaci *et al.*, 2010). The limitation of the current available treatment and interventions has been motivating several efforts towards the development of new drugs or a vaccine against *T.cruzi*. In fact, a recent study (Lee *et al.*, 2010) proposed a decision analytic Markov model that indicated that such vaccine could provide a substantial economic benefit. Some recent approaches to this problem described in the literature include screening efforts aiming inhibitors for *T.cruzi* known targets and development of high-throughput assays to validate anti-*T.cruzi* compounds (Canavaci *et al.*, 2010).

In this section, we used trained classification models to predict potential novel ligands to *T.cruzi* proteins with structures available at the PDB.

After an extensive analysis of the proposed signatures, we selected the best performing model trained in the biggest datasets considered (>35 000 pockets). The pockets obtained from the *T.cruzi* proteins were tested against this model, and a single ligand was predicted for each pocket. The KNN algorithm was used.

To validate the predictions, we performed the docking of the ligands in the *T.cruzi* pockets using AUTODOCK. We compare the energies of binding of the predicted ligands with the ones from real ligands from the crystallographic complexes via a redocking protocol. To assess the methods statistical significance, we compared our results with a null model. We selected for each pocket three independent random/null ligands from the pool of the training dataset. The docking workflow adopted in the present work is shown in Supplementary Figure S7.

In Figure 5, we can see that the energy distribution for aCSM predictions is more similar in shape to the redocking energy profile than the profile of the null models. Paired *t*-tests reveal a high *P*-value (0.26) between aCSM prediction and redocking means, but a low *P*-value ($1.2e^{-9}$) for aCSM prediction and null models. This strongly suggests that ligands found by aCSM may have the same binding free-energy profile of redocking ligands, but they may differ significantly from null ligands.

In summary, we showed that the binding free-energies for ligands predicted by aCSM are better (lower) in comparison with those predicted by the null models. Furthermore, the energies from a redocking protocol are indistinguishable from those obtained for aCSM prediction.

Table 1. Comparative results evaluated by the mean and standard deviation of the AUC score

Method	Dataset	AUC
Sequence	Kahraman	0.550 ± 0.08
MultiBind	Kahraman	0.715 ± 0.17
SHD	Kahraman	0.770
PSIM	Kahraman	0.790 ± 0.19
sup-PI	Kahraman	0.815 ± 0.13
sup-CK_L	Kahraman	0.861 ± 0.13
aCSM signature	**Kahraman**	$\mathbf{0.901 \pm 0.07}$
Sequence	Hoffmann HD	0.577 ± 0.09
MultiBind	Hoffmann HD	0.690 ± 0.14
sup-PI	Hoffmann HD	0.702 ± 0.19
sup-CK_L	Hoffmann HD	0.752 ± 0.16
PSIM	Hoffmann HD	0.760 ± 0.15
aCSM signature	**Hoffmann HD**	$\mathbf{0.804 \pm 0.13}$

The aCSM-ALL was the best performing signature for these experiments. AUC values were directly obtained from Hoffmann *et al.* (2010) and Spitzer *et al.* (2011), and the results for the aCSM signature were generated using multinomial logistic regression.

4 CONCLUSIONS

In the present work we proposed a novel, scalable, graph-based pocket signature called aCSM. It prospects distance patterns from the atoms that compose the binding pockets generating a feature vector that represents a cumulative edge count of contact graphs defined for different cut-off distances, which are used as evidence by supervised learning algorithms. SVD is also used as a preprocessing step to reduce dimensionality, lessen computational costs and grant scalability to the methodology and also reducing the inherent noise of the data, which increased the success rate of the predictions. Some of the most remarkable advantages of the aCSM signatures are that it does not require any ligand information in its calculation and also is independent of molecular orientation. Additionally, our algorithm presents a notable signature generality, as it may be applied to predict both proteic and non-proteic ligands to any type of biomolecular target (not only protein).

aCSM was successful when applied in ligand prediction tasks, presenting compatible or superior efficacy in comparison with state-of-the-art competitors. Besides, as a requirement and demand for its application to databases that are continuously growing, it proved scalable for large-scale scenarios and was able to perform well in a dataset composed by >35 000 pockets.

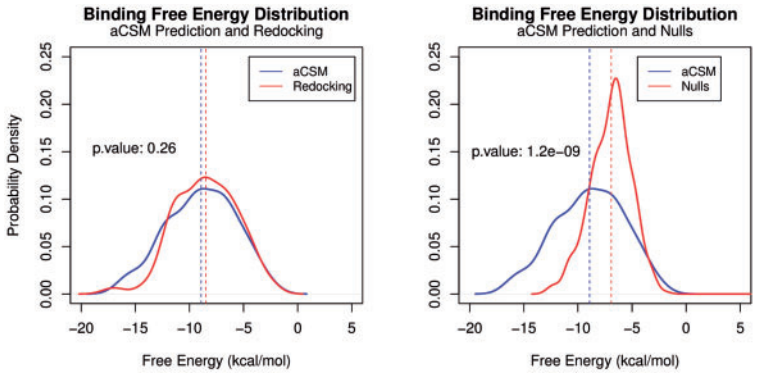

Fig. 5. Comparison of binding free-energies distribution of the docked complexes for aCSM prediction, redocking and null models. Dashed lines indicate the mean values, and the student *t*-test *P*-values for the significance of the means are also presented. Binding free energies for ligands predicted by aCSM are lower (better) in comparison with those predicted by the null models and indistinguishable from those obtained via a redocking protocol

On top of that, we applied the methodology to predict potential novel inhibitors to *T.cruzi* proteins. The validation of this step via docking confirmed that inhibitors predicted represent good candidates for further experimental validation.

We intend to predict inhibitors for proteins from other pathogenic organisms of interest, a study already in progress in our group. Finally, we plan to expand the signature to ligand-based lead discovery.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); Financiadora de Estudos e Projetos (FINEP); Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

Conflict of Interest: none declared.

REFERENCES

Canavaci,A.M.C. *et al.* (2010) In vitro and *in vivo* high-throughput assays for the testing of anti-trypanosoma cruzi compounds. *PLoS Negl. Trop. Dis.*, **4**, e740.

da Silveira,C.H. *et al.* (2009) Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, **74**, 727–743.

Davies,J.R. *et al.* (2007) The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics*, **23**, 3001–3008.

Demmel,J.W. (1997) *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Gonçalves-Almeida,V.M. *et al.* (2012) HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, **28**, 342–349.

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.

Kahraman,A. *et al.* (2007) Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.*, **368**, 283–301.

Kamagata,K. and Kuwajima,K. (2006) Surprisingly high correlation between early and late stages in non-two-state protein folding. *J. Mol. Biol.*, **357**, 1647–1654.

Kawabata,T. (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins*, **78**, 1195–1211.

Koshland,D., Jr (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA*, **44**, 98.

Le Guilloux,V. *et al.* (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.

Lee,B.Y. *et al.* (2010) The potential economic value of a trypanosoma cruzi (chagas disease) vaccine in latin america. *PLoS Negl. Trop. Dis.*, **4**, e916.

Monod,J. *et al.* (1963) Allosteric proteins and cellular control systems. *J. Mol. Biol.*, **6**, 306–329.

Morris,R.J. *et al.* (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, **21**, 2347–2355.

Najmanovich,R. *et al.* (2008) Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, **24**, i105–i111.

Pires,D.E.V. *et al.* (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, **12** (**Suppl 4**), S12.

Rassi,A., Jr *et al.* (2010) Chagas disease. *Lancet*, **375**, 1388–1402.

Schalon,C. *et al.* (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, **71**, 1755–1778.

Shulman-Peleg,A. *et al.* (2008) MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.*, **36** (web server issue), W260–W264.

Sippl,W. (2000) Receptor-based 3D QSAR analysis of estrogen receptor ligands - merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J. Comput. Aided Mol. Des.*, **14**, 559–572.

Spitzer,R. *et al.* (2011) Surface-based protein binding pocket similarity. *Proteins*, **79**, 2746–2763.

Ueno,K. *et al.* (2012) Exploring functionally related enzymes using radially distributed properties of active sites around the reacting points of bound ligands. *BMC Struct. Biol.*, **12**, 5.

Weskamp,N. *et al.* (2007) Multiple graph alignment for the structural analysis of protein active sites. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **4**, 310–320.

Zhang,C. *et al.* (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, **267**, 707–726.