BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 3 2012, pages 342–349
doi:10.1093/bioinformatics/btr680

Structural bioinformatics

Advance Access publication December 9, 2011

HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids

V. M. Gonçalves-Almeida[1,2,*], D. E. V. Pires[1,2], R. C. de Melo-Minardi[1,*], C. H. da Silveira[3], W. Meira[1] and M. M. Santoro[2]

[1]Department of Computer Science, [2]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte and [3]Advanced Campus at Itabira, Universidade Federal de Itajubá, Itajubá, Brazil

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Protein–protein interfaces contain important information about molecular recognition. The discovery of conserved patterns is essential for understanding how substrates and inhibitors are bound and for predicting molecular binding. When an inhibitor binds to different enzymes (e.g. dissimilar sequences, structures or mechanisms what we call cross-inhibition), identification of invariants is a difficult task for which traditional methods may fail.

Results: To clarify how cross-inhibition happens, we model the problem, propose and evaluate a methodology called HydroPaCe to detect conserved patterns. Interfaces are modeled as graphs of atomic apolar interactions and hydrophobic patches are computed and summarized by centroids (HP-centroids), and their conservation is detected. Despite sequence and structure dissimilarity, our method achieves an appropriate level of abstraction to obtain invariant properties in cross-inhibition. We show examples in which HP-centroids successfully predicted enzymes that could be inhibited by the studied inhibitors according to BRENDA database.

Availability: www.dcc.ufmg.br/∼raquelcm/hydropace

Contact: valdetemg@ufmg.br; raquelcm@dcc.ufmg.br; santoro@icb.ufmg.br

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

Received on July 25, 2011; revised on November 17, 2011; accepted on December 3, 2011

1 INTRODUCTION

Enzyme inhibition occurs when a molecule binds to an enzyme, thus decreasing its activity. Inhibitors may be proteic or non-proteic; they can decrease the enzyme's ability to bind substrates or can lower the enzyme's catalytic activity or a combination of both. Inhibition is an important biochemical mechanism that is involved in metabolism regulation. It controls many intra- and extracellular pathways, inflammatory and immunological processes, virus replication and many other biological functions (Barrett *et al.*, 2004) Furthermore, once this natural phenomenon is understood, it might be used for biotechnological purposes including the development of drugs, insecticides, pesticides and disinfectants.

A particular case is the inhibition of peptidases; on this subject, the MEROPS database is currently one of the most important peptidase repositories (Rawlings *et al.*, 2008). The MEROPS database groups both proteases and inhibitors hierarchically into families (sequence-related entities) and clans (structure-related entities). A careful MEROPS search highlighted a well-known but intriguing phenomenon: some protease inhibitors lack specificity and involve different 3D structures and catalytic mechanisms. For instance, Turkey Ovomucoid and Englin C act in different serine peptidase clans such as PA(S) (all β Trypsin-like folds) and SB (α/β Subtilisin-like folds) and soybean Kunitz trypsin inhibitor decays proteolytic activity as much in serine peptidases as in metallopeptidases (which have very different enzymatic mechanisms). We call this lack of specificity *cross-inhibition*. Our main challenge in this article is to create a methodology that helps to understand and predict this phenomenon.

Protease–inhibitor recognition and binding are determined by a complex orchestration of interactions and entropic factors that involve the entire protease–inhibitor–solvent system. Fortunately, the experimental binding energetics of many protease–inhibitor complexes have already been thermodynamically determined. It is known, for example, that the binding of Turkey Ovomucoid with Elastase at 25°C is characterized by a negative Gibbs free energy in which enthalpy change is almost negligible but entropy change is largely positive (Baker and Murphy, 1997). Furthermore, we spot a clear trend of higher apolar/polar accessible surface area ratio toward interface (Supplementary Fig. S1), which is an evidence of the importance of the hydrophobic interactions in protease–inhibitor complex formation. That said, we particularly focus our attention on the search for conserved hydrophobic interaction patterns. We define these patterns as invariant hydrophobic regions (or patches) that are in contact with the same apolar complementary parts of the inhibitor (Supplementary Figs 3 and 4). We show (Supplementary Fig. S2) a strong linear relationship (Pearson's correlation coefficient of 0.98) between the inferred solvation entropy change and the extension of hydrophobic patches, measured in terms of the number of hydrophobic atoms inside them.

Although there are many biochemical studies that analyze diversity in inhibition processes [e.g. (Bode *et al.*, 1986; Chakrabarti and Janin, 2002a; Laskowski and Qasim, 2000; Qasim *et al.*, 1997)], experimental characterization of inhibition is a labor-intensive process. The large amount of possible inhibitors for a given enzyme can make tests costly; hence, *in silico* methods can contribute to predicting inhibitor–enzyme recognition.

Despite its evident importance, there are few models and algorithms that identify recognition and interaction patterns that

*To whom correspondence should be addressed.

could help to clarify how cross-inhibition occurs. In this context, a pattern is a conserved set of interface attributes that is used to explain or predict binding.

Traditionally, sequence comparison and/or structural alignment methods have been used in conservation detection (Melo-Minardi *et al.*, 2007; Ribeiro *et al.*, 2010; Zhang *et al.*, 2011). According to Tuncbag *et al.* (2011), structures are more conserved than sequences, and interface-forming residues (IRFs) are even more conserved than the whole structure. However, these classical methods are inappropriate because in cross-inhibition we may deal with very dissimilar sequences and even completely distant folds.

Indeed, in cross-inhibition pattern detection with traditional methods, we identify essentially known conserved residues that directly participate in the catalysis process, such as the catalytic triad, the specificity pocket and oxyanion-binding sites. We note that to correctly assess the eventual hydrophobic contribution of the entire protease–inhibitor interface, we should abstract the residue semantics and should assess patches at the atomic level. A similar approach has been used to characterize the core of protein domains with similar folds but very divergent sequence compositions (Soundararajan *et al.*, 2010). The atomic level is more appropriate because all residues have apolar portions. Lysine, for example, is considered a positively charged residue (at neutral pH), but there are also several hydrophobic methyl groups.

Enzyme–inhibitor recognition is determined by a network of interactions between atoms; hence, graph modeling is a straightforward approach. We model hydrophobic atoms as nodes of a graph and the contacts between them as the edges. We use the graph to obtain conserved hydrophobic patches or, in other words, connected components.

Supposing that the most important property of a hydrophobic patch is where it is positioned to interact with the ligand, we abstract from its composition volume, shape and density, and we represent the patch as a geometric centroid that we call the HP-centroid(hydrophobic patch centroid). In this work, we propose a novel model and algorithms to detect conserved HP-centroids in cross-inhibition.

Finally, we present a qualitative case study that consists of two examples of cross-inhibition, Trypsin-like and Subtilisin-like enzymes, both of which belong to the serine proteases family. They present completely different 3D structures and the sequence identity is as low as 20% (Wallace *et al.*, 1996). However, they possess exactly the same Ser-His-Asp triad on their active sites. In the first case, we have complexes of Trypsin-like and Subtilisin-like enzymes inhibited by Eglin C (Betzel *et al.*, 1993), and in the second case, we have complexes of the same families with Turkey Ovomucoid (Papamokos *et al.*, 1982). We verify that the HP-centroids obtained from the complexes are present in a set of sequence-diverse apo structures that are conserved throughout the family.

2 MATERIALS AND METHODS

Each step of the proposed methodology, called HydroPaCe, is described below. A complete workflow of the methodology is presented in Figure 1.

2.1 Data selection and preparation

As explained previously, we have chosen serine proteases to test our algorithm. We chose them because there are few other examples of cross-inhibition structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000). Moreover, this is a well-studied family that presents some peculiarities and similarities in catalytic sites (Page and Di Cera, 2008). Although Trypsin-like and Subtilisin-like have very different 3D structures, they hydrolyze their substrates by the same mechanism (Ekici *et al.*, 2008; Lesk and Fordham, 1996; Siezen and Leunissen, 1997).

Enzyme–inhibitor complexes: we found five non-redundant complexes involving the Eglin C inhibitor: four bond to Subtilisin-like (PDB IDs: 1TEC, 1CSE, 1MEE and 1SBN) and one to Trypsin-like (PDB ID: 1ACB) enzymes. Likewise, we found four complexes involving the Ovomucoid inhibitor: three complexed with Trypsin-like (PDB IDs: 1CHO, 1PPF and 3SGB) and one with Subtilisin-like (PDB ID: 1R0R) enzymes. Despite the large amount of information on enzymatic complexes involving these two families, there is much redundant information regarding the sequence identities, and this leaves only a small number of non-redundant complexes to be analyzed.

Apo enzymes: we selected a set of non-redundant apo enzymes from the two families by removing enzymes that presented >50% of sequence identity. Hence, we use 9 samples from Subtilisin-like and 35 from Trypsin-like families. The complete list of PDB ids is presented in the Supplementary Material.

All the structures were submitted to standardization processes using the PDB Enhanced Structures Toolkit (PDBest) (Pires *et al.*, 2007).

2.2 IFRs

The current analysis is restricted to regions of the molecular interface of the enzyme and its inhibitor. The IFRs can be determined by three different methods. The first defines the interface simply by using a cut-off distance between the residues of the interacting molecules (Chothia and Janin, 1975; Conte *et al.*, 1999). The second approach computes the interactions based on differences in solvent-accessible surface area (ASA) when the monomers are separated (Chakrabarti and Janin, 2002b; Janin *et al.*, 1990). Finally, the last approach defines interfaces through computational geometry using Voronoi diagrams and the alpha shapes theory (Pontius *et al.*, 1996). We used the ASA method because it is the most used method and is therefore more consolidated.

Enzyme–inhibitor IFRs: we computed the IFRs in the cross-inhibition complexes using the ASA approach with the STING Millennium Suite platform (SMS) (Neshich *et al.*, 2003).

Projection of IFRs from complexes into apo enzymes: for the apo proteins, the projection was derived by structural alignment using an enzyme–inhibitor complex and the computed IFR. Moreover, the structures were solvated using Gromacs. After applying the treatment to PDB files, all structures, including the complex model, were superimposed using the program MultiProt. Finally, the residues that aligned with the interface of the complex model were considered the interfaces of the apo proteins. This process was performed for analysis of both sets (Trypsin-like and Subtilisin-like) of single-chain proteins in our database.

2.3 Problem modeling

The proposed method is based on the search for conserved hydrophobic patches (HP-centroids). In what follows, we detail each step of our model:

Graph construction: the first step of our model consists of the representation of IFRs as graphs. The nodes are atoms from the IFR residues, and the edges are the presumed contacts. According to our previous work (da Silveira *et al.*, 2009), there are two main approaches to identify contacts in proteins: the first is cut-off dependent (CD), and the other is independent (CI). Although in the above-mentioned study we found that, at the residue level, the CD approach was a simpler, more complete and more reliable technique than some CI techniques, here we chose to use a CI methodology because we did not find a reliable cut-off value at the atomic level. This paradigm uses classical computational geometry algorithms to compute a Voronoi diagram (VD) (Poupon, 2004) and its dual problem, the Delaunay

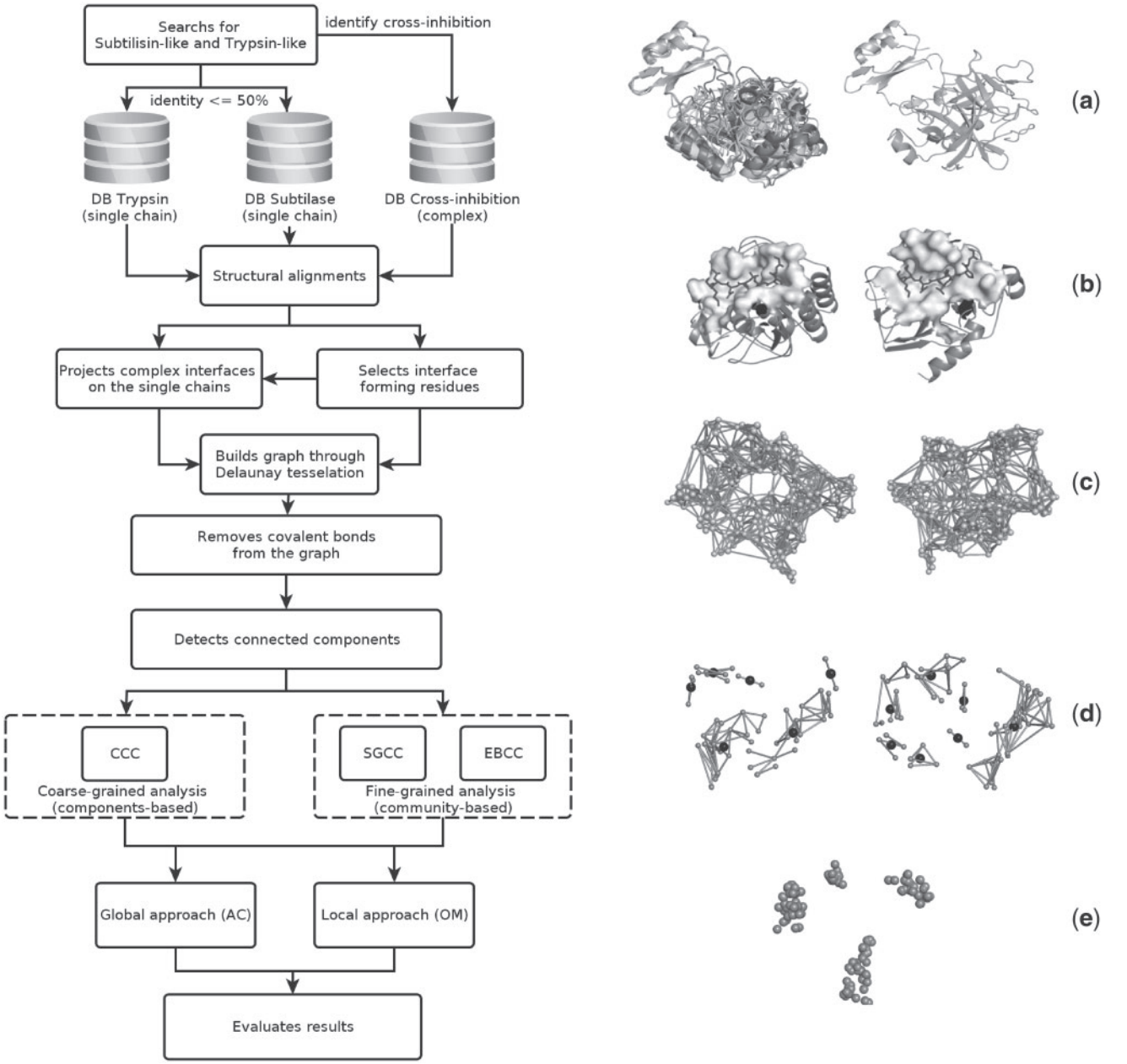

Fig. 1. HydroPaCe workflow: We searched for 3D structures of subtilisin-like and trypsin-like families in the PDB database. The PDB ids were separated into protein–inhibitor complexes and apo proteins. The structures with sequence identities that were >50% identical to other selected sequences were discarded. The cross-inhibition complexes were aligned by the inhibitor's chain (**a**), and the interfaces of contact (also called IFRs) were identified using ASA methodology (**b**). The apo proteins were aligned by their single chains by using an enzyme–inhibitor complex to project the interface. Using DT, possible interatomic contacts were computed, resulting in the edges of a graph where nodes are atoms (**c**). We considered only the hydrophobic interactions between atoms and removed edges that represented covalent bonds. We then identified the connected components that represent the hydrophobic patches in these graphs (**d**). We propose two levels of abstractions to represent the hydrophobic patches, both of which are based on geometric centroids (HP-centroids). The first is a coarse-grained analysis that consists of computing a centroid for each connected component, and the second is a fine-grained analysis that searches for dense subregions using two different community detection algorithms and calculating the HP-centroids on communities. The obtained HP-centroids were clustered using the OM and AC methods (**e**). Finally, the HP-centroids were evaluated using PRM, which accounts positively for coverage in terms of enzymes and negatively for enzyme redundancy. In (a–d), the left-hand structure is Subtilisin-like and the right-hand structure is Trypsin-like.

tessellation (DT) (Dupuis *et al.*, 2005). In the 3D view, the VD decomposes the volume by associating a polyhedron with each site (which is called a Voronoi cell). Each face of these polyhedrons is composed of a plane that bisects the line and links each site to each of its near sites, thus mapping a neighborhood with the closest (not occluded) contacts (da Silveira *et al.*, 2009).

Deletion of covalent edges: we are interested only in non-covalent interactions; hence, we remove covalent bond edges in a post-processing step.

Deletion of polar edges: once we have a geometrical inference of non-occluded interactions, we classify them into hydrophobic and polar interactions based on the classification rules proposed in Sobolev *et al.* (1999). The complete table with the classifications of all the atoms can be found in the Supplementary Material. As discussed previously, we restrict our analysis to hydrophobic interactions type by removing polar contact edges. Nevertheless, the analysis can be extended to deal with polar areas.

Computation of hydrophobic patches: we use a depth-first search to efficiently detect the connected components, which are natural representations of the hydrophobic patches.

Abstraction of hydrophobic patches through centroids: hydrophobic patches may occur in different shapes and volumes; our model considers two levels of abstractions to represent them, both of which are based on geometric centroids (HP-centroid). The first, which we call the coarse-grained analysis, consists of computing a centroid for each connected component. The second is a fine-grained analysis that divides the original connected components into dense subgraphs or communities. A community is a subgraph in which the nodes are much more connected with the other nodes in the community than with the external nodes. In this approach, the HP-centroids are computed based on communities.

In conclusion, our method is based on the computation of hydrophobic patches and their abstraction through geometric centroids (HP-centroids) that can represent the entire patch (coarse-grained) or communities of these patches (fine-grained). Considering the HP-centroids of a set of cross-inhibition complexes, we propose algorithms to cluster the centroids and to detect those that are conserved across all of them. We describe the algorithms in the next section and then explain how to evaluate the clusters obtained.

2.4 Algorithms

Here, we describe in more detail the different approaches (coarse- and fine-grained) that we propose to abstract from the hydrophobic patches. We briefly describe the paradigms for community detection used in fine-grained decomposition of hydrophobic patches. Finally, we explain the algorithms that we use to cluster the HP-centroids: one attempts to globally match similar centroids and the other locally clustered centroids in an agglomerative manner.

CCC: *Connected Component Centroids* is the name we give to the coarse-grained approach.

EBCC: the *Edge Betweenness Community Centroid (EBCC)* (Newman and Girvan, 2004) is a divisive approach in which the most central edges are broken one after another until the modularity of the graph is maximized. The edge centrality is computed through the edge betweenness, which counts the number of shortest paths that traverse through that edge. The higher the value of edge betweenness, the more the edge is used or the more central it is. In other words, this value indicates when there are no redundant edges to cross between different communities and when the edge joins two different communities.

SGCC: the *Spin Glass Community Centroid (SGCC)* (Reichardt and Bornholdt, 2006) tries to find communities in graphs via a spin-glass model and simulated annealing. That is, it uses simulated annealing to maximize graph modularity. The modularity of a possible division of a graph into communities is defined as the fraction of edges that falls within a given community minus the expected value of this fraction if edges were randomly distributed. Commonly, the randomization of the edges is done in such a way as to preserve the degree of each vertex.

OM: we have developed a linear programming *Optimization Model (OM)* that is based on the transport problem and that attempts to match points by globally minimizing the differences between the edge sizes between all possible pairs of points. The optimization functions that we want to minimize, as well as the associated restrictions, are explained in detail in the Supplementary Material.

AC: this method is a local strategy based on *Agglomerative Clustering (AC)*. It matches the closest HP-centroids through an iterative bottom-up agglomerative process. In this case, there is an important decision about when to stop the process to ensure that we have high-quality clusters. The strategy for determining this stopping point, and a detailed explanation of the algorithm, are presented in the Supplementary Material.

2.5 Evaluation

To perform a quantitative evaluation of the clusters formed by the matches, we propose a metric based on the concept of recall that is penalized when different HP-centroids of the same protein (redundant centroids) are grouped together. We have it called the penalized recall metric (PRM) and is formalized below:

$$\mathrm{PRM} = \frac{\mathbb{C}\binom{D}{2}}{\mathbb{C}\binom{P}{2}} - \frac{\mathbb{C}\binom{E}{2}}{\mathbb{C}\binom{P}{2}} \quad (1)$$

where $\mathbb{C}\binom{D}{2}$ is the number of pairs of centroids from different enzymes in the same cluster, $\mathbb{C}\binom{E}{2}$ is the number of pairs of HP-centroids from the same protein in the same cluster, $\mathbb{C}\binom{P}{2}$ is the total number of pairs of HP-centroids in the cluster and the values of D and E are limited to P.

The metric produces values in the range of $[-1; 1]$ where -1 is the worst case, with minimum recall and maximum redundancy. It will result in 1 when we have maximum recall and minimum redundancy. When we have similar values for recall and redundancy, the metric approaches 0.

The average of the PRM of the clusters was used to evaluate the three different approaches (CCC, EBCC and SGCC). It cannot be used to compare between the OM and AC. In the OM, clusters are formed with total variability by definition; in other words, there are no pair of centroids of the same enzyme in a cluster.

In this case, we use traditional intra- and intercluster average distances. *A priori*, a high-quality cluster must have low intracluster and high intercluster distances. That is because, in an ideal clustering, similar elements must be grouped together and dissimilar ones must be separated.

In conclusion, we compare the proposed approaches in the light of the PRM (the closer to 1, the better) and the average intra- and intercluster distances (it is better to have low intracluster and high intercluster distances).

3 RESULTS

In this section, we present and discuss the results of the case study of serine peptidases (Trypsin-like and Subtilisin-like) that are cross-inhibited by Eglin C and Turkey Ovomucoid. We also compare the quality of the conserved HP-centroids that are obtained by the different proposed methods.

3.1 The Eglin C Inhibitor

Eglin C is a small monomeric protein (70 residues) that belongs to the Potato Chymotrypsin Inhibitor I family of serine protease inhibitors that occurs naturally in the Leech *Hirudo medicinalis*. Functionally, Eglin C can inhibit more than one proteinase family with non-homologous structures (Hyberts *et al.*, 1992). In the BRENDA database (Scheer *et al.*, 2011), we found 12 different EC numbers that are known to be inhibited by this molecule. In this section, we present the analysis with the five non-redundant existing experimental complexes, four of which are Subtilisin-like and one of which is Trypsin-like.

As explained previously, we use different approaches to find the HP-centroids. The OM has no parameters and it clusters all of the centroids. With the AC, we must supply the number of clusters as an input parameter. Figure 2a shows the distributions of mean PRM and intracluster distances. We observe that PRM is maximized and intradistance values are stable with 12 clusters. With this configuration, we obtain five high-quality clusters according to the

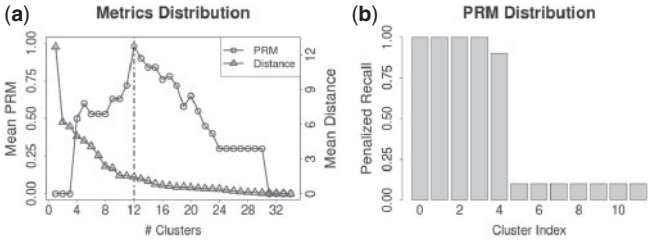

Fig. 2. The CCC approach. In (**a**), we present the distribution used to maximize the mean PRM metric as well as the respective mean intracluster distance distribution. (**b**) Shows the PRM distribution for the best configuration achieved, with 12 clusters.

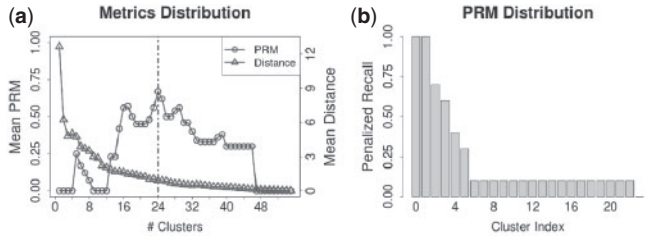

Fig. 3. The SGCC approach. In (**a**), we present the distribution used to maximize mean PRM metric as well as the respective mean intracluster distance distribution. (**b**) Shows the PRM distribution for the best configuration achieved, with 24 clusters.

Table 1. Quantitative comparison of the proposed algorithms for Eglin C cross-inhibition.

		Mean intra (Å)	Mean inter (Å)	Mean PRM
CCC	AC	3.435	13.835	**0.98**
	OM	5.460	9.294	–
SGCC	AC	2.450	13.138	0.82
	OM	5.093	11.231	–
EBCC	AC	2.679	12.670	0.90
	OM	5.339	9.986	–

The best mean PRM value is in bold.

PRM (Fig. 2b). This set of conserved HP-centroids presents a very high recall value (i.e. they are present in almost all the cross-inhibition complexes) and furthermore, there is only one case where two points in a cluster come from the same complex.

The same experiment was performed using the fine-grained approach, as presented in Figure 3. At this level of abstraction, we could not identify a threshold that clearly distinguishes high-quality clusters from poor-quality ones. Since we aim to find as many conserved HP-centroids as possible, the coarse-grained approach systematically presents better results. This might indicate that the cross-inhibition pattern depends on the inhibitor-relative positions of the conserved HP-centroids regardless of their density.

Table 1 shows the complete set of results. AC performs better, especially in the coarse-grained analysis, achieving low intra- and high intercluster distances combined with a very high PRM value (0.98).

The semantics of the five hydrophobic patches represented by the conserved HP-centroids is presented in Figure 4. We can see

Table 2. Quantitative comparison of the proposed algorithms for Turkey Ovomucoid cross-inhibition.

		Mean intra (Å)	Mean inter (Å)	Mean PRM
CCC	AC	4.803	13.239	**0.94**
	OM	8.009	10.402	–
SGCC	AC	2.901	10.999	0.75
	OM	9.303	11.014	–
EBCC	AC	3.419	14.045	0.75
	OM	6.459	11.997	–

The best mean PRM value is in bold.

why the proposed method reaches an abstraction level that is useful for identifying relevant cross-inhibition patterns. When we compare the residues that compose cluster IV, we can see for a Trypsin-like enzyme the presence of LEU-143, THR-151, ALA-149, TYR-146, CYS-220, CYS-191 and MET-192. At the counterpart cluster in a Subtilisin-like enzyme, we find PHE-193, ASN-163 and THR-224. Despite the very dissimilar residue compositions, patch volumes and densities, the method selects HP-centroids that are spatially conserved according to the inhibitor. Additional graphs for the other three samples are presented in the Supplementary Material.

3.2 The Turkey Ovomucoid inhibitor

Ovomucoids are the glycoprotein protease–inhibitors of avian egg whites. There are several protease inhibitors in egg white. The Turkey Ovomucoid is from a Kazal-type inhibitor family of serine protease inhibitors, which occurs naturally in Meleagris gallopavo. It is a significant contaminant of crude Ovomucoid preparations, and it acts on Bovine Trypsin and Chymotrypsin as well as on Porcine Elastase and Fungal Proteinase (Fujinaga *et al.*, 1987; Robertson *et al.*, 1988).

We analyze the four non-redundant existing complexes, of which three have Trypsin-like enzymes and one has Subtilisin-like enzymes. By conducting similar experiments to those presented in the previous section, and by varying the number of clusters, we observe that the mean intradistance stabilizes from four clusters on. We obtain three high-quality clusters according to the PRM (Supplementary Material).

Table 2 shows the results for the algorithm comparisons. As in the previous analysis, AC presents a combination of low intracluster distances, high intercluster distances and the highest PRM value (0.94) indicating a consistent match of the patterns.

According to these results, the coarse-grained approach once more achieved better results than the fine-grained approach.

The three hydrophobic patches that were conserved in the Ovomucoid complexes are presented in Figure 5. Again, we can see a very dissimilar cluster composition and an interesting conservation of position according to the common inhibitor. We present additional graphs for Ovomucoid cross-inhibition in the Supplementary Material.

According to Baker and Murphy (1997), hydrophobic interactions are essential for explaining how inhibition happens in proteases. Our results are in agreement with this hypothesis. Searching for conserved abstractions of hydrophobic patches at the atomic level (HP-centroids) in protease–inhibitor interfaces, we proposed and evaluated a global and a local algorithm to cluster centroids. We aimed to find conserved centroids at coarse- and fine-grained levels. We conclude that the coarse-grained AC local algorithm was able to

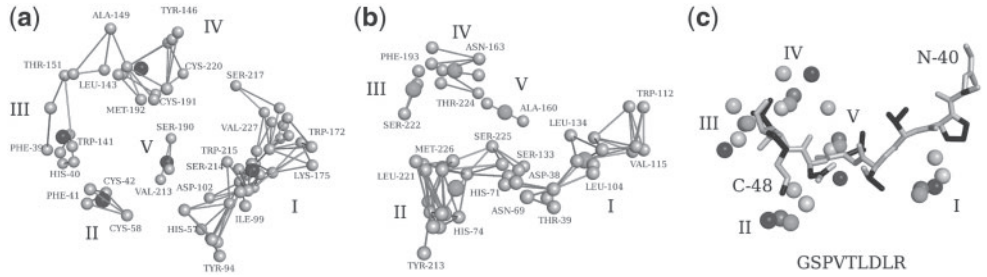

Fig. 4. Hydrophobic patches for cross-inhibition by Eglin C. In (**a**) PDB id 1ACB:E, we can see a sample with Trypsin-like enzyme and in (**b**) PDB id 1TEC:E, with Subtilisin-like enzyme (the hydrophobic patches for the five complexes are in the Supplementary Material). We show an atomic graph in which the residue types and numbers are presented and the red (a) and green (b) spheres are the HP-centroids that represent each of the patches. In the last part of the figure (**c**), we present the inhibitor (residues from 40 to 48) as gray sticks (in black, the apolar portions), and the five centroids are superposed in colors. The green shades are the Subtilisin-like HP-centroids and the red ones are Trypsin-like.

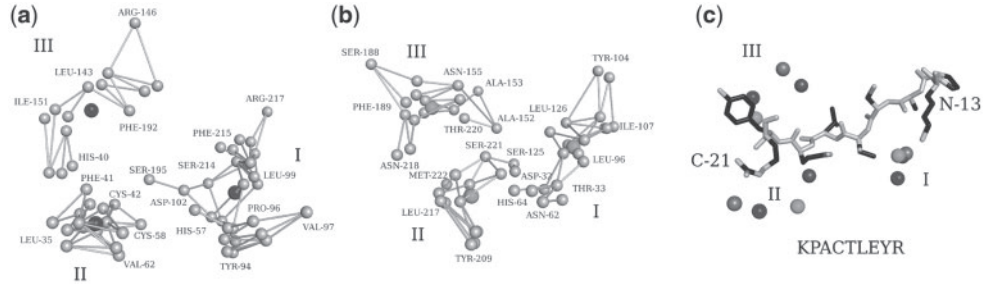

Fig. 5. Hydrophobic patches for cross-inhibition by Turkey Ovomucoid. In (**a**) PDB id 1R0R:E, we can see a sample with Subtilisin-like enzyme and in (**b**) PDB id 1PPF:E, a sample with Trypsin-like enzyme (the hydrophobic patches for the four complexes are shown in the Supplementary Material). We show an atomic graph in which the residue types and numbers are presented and the red (a) and green (b) spheres show the HP-centroids that represent each patch. In (**c**), we present the inhibitor (residues from 13 to 21) as gray sticks (apolar portions in black), and the five HP-centroids are superposed in colors. The green shades are the Subtilisin-like centroids and the red ones are the Trypsin-like centroids.

identify the more complete set of invariant HP-centroids across the protease–inhibitor cross-inhibition examples.

Certainly, the contribution of polar interactions must be studied in more detail in future work; interestingly, however, we have found a minimum of three invariant centroids in all cross-inhibition cases. As proteins are 3D objects, we conjecture that for a molecule to bind and to hold another one, there must exist at least three non-collinear contact points. It is possible that the conserved hydrophobic patches obtained are responsible for binding and holding inhibitors at the enzyme binding sites.

3.3 The use of HP-centroids for inhibition prediction

Once we have the problem of scarcity of experimental complexes representing cross-inhibition examples, it is intriguing to ask whether we can generalize the conserved HP-centroids to binding sites of apo enzymes of the studied families. We extended the analysis to a set of non-redundant apo structures of serine proteases (a list of proteins is in the Supplementary Material). We project the IFR obtained from the cross-inhibition complexes to the apo enzymes by using structural alignments, and we verify a strong conservation of the HP-centroids found through complex analysis.

Due to the low conservation of residues, it is not possible to understand how inhibition occurs by examining only sequence-level conservation (even when sequence alignments are done by structural alignments as shown in Fig. 6). Notice that we can find some conserved residues (marked with *) that are known to participate in the catalysis (catalytic triad, oxyanion role) or in the specificity binding sites. Apart from these residues, no other interest conservation can be easily identified in these logos.

However, our hypothesis is that for inhibition to occur, we must have very conserved hydrophobic patches in specific positions to accommodate each of the inhibitors. For example, PHE-215 in the Trypsin-like enzymes in Figures 6b and 5b is a voluminous hydrophobic residue that is equivalent to the hydrophobic portions of residues in positions LEU-96, ILE-107 and LEU-126 in the Subtilisin-like enzymes in Figures 6a and 5a. This is an example in which conserved patterns cannot be inferred from the sequence or structure but are clearly identified in our conserved HP-centroid I.

Going further, we believe that these patterns could be used to predict inhibition for other enzymes for which structures are available but no experimental evidence of inhibition is known. For instance, we used eight samples of non-redundant Subtilisin-like apo enzymes (listed in the Supplementary Material) belonging to five different EC numbers (3.4.21.62 / 64 / 66 / 75 / 97). We considered only those enzymes for which the ECs are complete with the four levels of annotation. According to the BRENDA database, three of these are inhibited by Eglin C (3.4.21.62 / 66 / 75), and we can say that this constitutes successful predictions. As far as we are concerned, the other two enzymes (3.4.21.64 / 97), which represent Proteinase K and Assemblin Protease, are

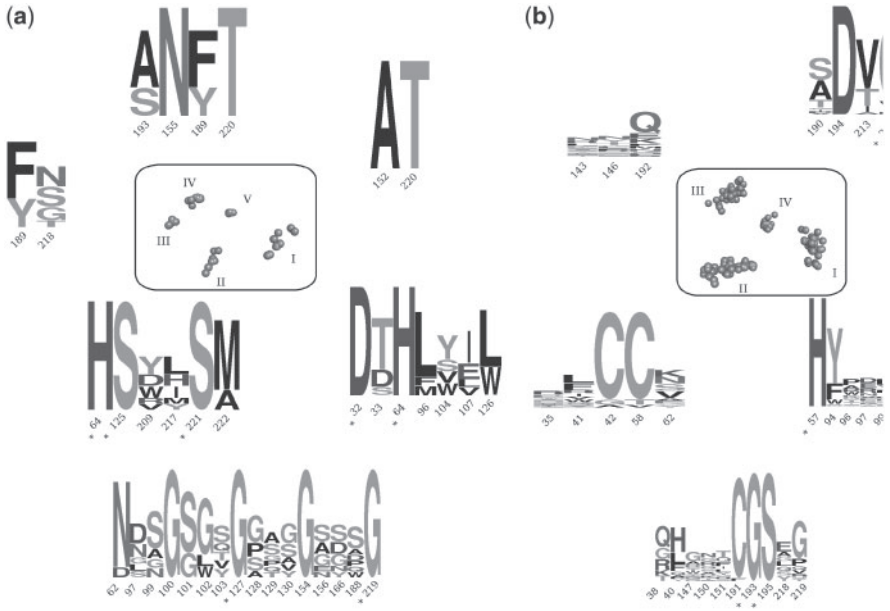

Fig. 6. IFR projections of HP-centroids found in serine proteases that are cross-inhibited by Ovomucoid. In (**a**), we show results for nine non-redundant superposed Subtilisin-like enzymes (residue numbers according to PDB id 1R0R:E) and in (**b**) we show results for 35 non-redundant superposed Trypsin-like enzymes (residue numbers according to PDB id 1PPF:E). On both sides, the bottom logos show the residues that are in the IFR but that are not part of a conserved cluster.

not mentioned in the literature but present the same pattern as do the other Subtilisin-like enzymes. It would be very interesting to verify experimentally whether they can be inhibited by Eglin C, as they present the same HP-centroids as do other complexes with this inhibitor. Similar analyses for Ovomucoid are presented in the Supplementary Material, and we can also verify successful predictions and several unknown inhibition possibilities.

4 CONCLUSIONS

In this work, we model the problem of understanding and predicting enzyme cross-inhibition. We propose and evaluate algorithms to detect conserved hydrophobic patch centroids (HP-centroids) to clarify how these centroids occur in proteases. Our model is based on the importance of apolar interactions to inhibition in this family and on the fact that these hydrophobic portions should be studied at an atomic level. We model the interfaces between enzymes and inhibitors as graphs of atomic apolar interactions, detect connected components to represent hydrophobic patches, summarize them using centroids and show how to obtain as complete a set of conserved centroids as possible. One of the strengths of the method is that it achieves the appropriate level of abstraction to detect the invariant properties involved in cross-inhibition. One of the main difficulties in the study and understanding of this complex phenomenon through classical methods is that dissimilar sequences and structures might be inhibited by the same inhibitor. Despite the lack of conservation at the sequence and structure levels, the proposed HP-centroids appear to be promising, as they are very conserved across the studied cases of cross-inhibition.

As we have few non-redundant experimental complexes available, we test the generality of HP-centroids with a set of non-redundant apo enzymes representing entire families. By comparing with experimental data available in the BRENDA database, we also show some successful examples of how HP-centroids can be used to predict enzymes that could be inhibited by the studied inhibitors. Finally, we raise some questions about possible enzymes that might be inhibited by Eglic C and/or Turkey Ovomucoid and expose them to further experimental validation.

We believe that this work should be extended to other enzyme families for which entropic changes are known to be important factors in inhibition processes. It would also be interesting to verify whether this method should be used in other problems of protein–protein interaction pattern detection.

Funding: Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG); Financiadora de Estudos e Projetos (FINEP).

Conflict of Interest: none declared.

REFERENCES

Baker,B.M. and Murphy,K.P. (1997) Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. *J. Mol. Biol.*, **268**, 557–569.

Barrett,A.J. *et al.* (eds) (2004) *Handbook of Proteolytic Enzymes*, vol. 1–2, 2 edn. Elsevier, London.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Betzel,C. *et al.* (1993) Structure of the proteinase inhibitor eglin c with hydrolysed reactive centre at 2.0 a resolution. *FEBS Lett.*, **317**, 185–188.

Bode,W. *et al.* (1986) X-ray crystal structure of the complex of human leukocyte elastase (pmn elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO J.*, **5**, 2453–2458.

Chakrabarti,P. and Janin,J. (2002a) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.

Chakrabarti,P. and Janin,J. (2002b) Dissecting protein-protein recognition sites. *Proteins Struct. Funct. Genet.*, **47**, 334–343.

Chothia,C. and Janin,J. (1975) Principles of protein-protein recognition. *Nature*, **256**, 705–708.

Conte,L.L. *et al.* (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

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.

Dupuis,F. *et al.* (2005) Voro3d: 3d voronoi tessellations applied to protein structures. *Bioinformatics*, **21**, 1715–1716.

Ekici,O.D. *et al.* (2008) Unconventional serine proteases: variations on the catalytic ser/his/asp triad configuration. *Protein Sci.*, **17**, 2023–2037.

Fujinaga,M. *et al.* (1987) Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 a resolution. *J. Mol. Biol.*, **195**, 397–418.

Hyberts,S.G. *et al.* (1992) The solution structure of eglin c based on measurements of many noes and coupling constants and its comparison with x-ray structures. *Protein Sci.*, **1**, 736–751.

Janin,J. *et al.* (1990) The structure of protein-protein recognition sites. *Structure*, **265**, 16027–16030.

Laskowski,M. and Qasim,M.A. (2000) What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim. Biophys. Acta*, **1477**, 324–337.

Lesk,A.M. and Fordham,W.D. (1996) Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *J. Mol. Biol.*, **258**, 501–537.

Melo-Minardi,R.C. *et al.* (2007) Finding protein-protein interaction patterns by contact map matching. *Genet. Mol. Res.*, **6**, 946–963.

Neshich,G. *et al.* (2003) Sting millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, **31**, 3386.

Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.

Page,M.J. and Di Cera,E. (2008) Serine peptidases: classification, structure and function. *Cell. Mol. Life Sci.*, **65**, 1220–1236.

Papamokos,E. *et al.* (1982) Crystallographic refinement of japanese quail ovomucoid, a kazal-type inhibitor, and model building studies of complexes with serine proteases. *J. Mol. Biol.*, **158**, 515–537.

Pires,D.E.V. *et al.* (2007) Pdbest: Pdb enhanced structures toolkit. In *Proceedings of the 3rd International Conference of Brazil Association for Bioinformatics.* AB3C Publishing, São Paulo, p. 39.

Pontius,J. *et al.* (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, **264**, 121–136.

Poupon,A. (2004) Voronoi and voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.*, **14**, 233–241.

Qasim,M.A. *et al.* (1997) Interscaffolding additivity. association of p1 variants of eglin c and of turkey ovomucoid third domain with serine proteinases. *Biochemistry*, **36**, 1598–1607.

Rawlings,N.D. *et al.* (2008) Merops: the peptidase database. *Nucleic Acids Res.*, **36**, D320–D325.

Reichardt,J. and Bornholdt,S. (2006) Statistical mechanics of community detection. *Phys. Rev. E*, **74**, 016110.

Ribeiro,C. *et al.* (2010) Analysis of binding properties and specificity through identification of the interface forming residues (ifr) for serine proteases in silico docked to different inhibitors. *BMC Struct. Biol.*, **10**, 36.

Robertson,A.D. *et al.* (1988) Two-dimensional NMR studies of kazal proteinase inhibitors. 1. sequence-specific assignments and secondary structure of turkey ovomucoid third domain. *Biochemistry*, **27**, 2519–2529.

Scheer,M. *et al.* (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, **39**, 670–676.

Siezen,R.J. and Leunissen,J.A. (1997) Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Sci.*, **6**, 501–523.

Sobolev,V. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.

Soundararajan,V. *et al.* (2010) Atomic interaction networks in the core of protein domains and their native folds. *PLoS One*, **5**, e9391.

Tuncbag,N. *et al.* (2011) Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Phys. Biol.*, **8**, 035006.

Wallace,A.C. *et al.* (1996) Derivation of 3D coordinate templates for searching structural databases: application to ser-his-asp catalytic triads in the serine proteinases and lipases. *Protein Sci.*, **5**, 1001–1013.

Zhang,Z. *et al.* (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.