Predicting structural and functional sites in proteins by searching for maximum-weight cliques

Franco Mascia and **Elisa Cilia** and **Mauro Brunato** and **Andrea Passerini**
Information Engineering and Computer Science Department
University of Trento - Via Sommarive, 14 I-38100 Trento - Italy
{mascia,cilia,brunato,passerini@disi.unitn.it}

Abstract

Fully characterizing structural and functional sites in proteins is a fundamental step in understanding their roles in the cell. This extremely challenging combinatorial problem requires determining the number of sites in the protein and the set of residues involved in each of them. We formulate it as a distance-based supervised clustering task, where training proteins are employed to learn a proper distance function between residues. A partial clustering is then returned by searching for maximum-weight cliques in the resulting weighted graph representation of proteins. A novel stochastic local search algorithm is proposed to efficiently generate approximate solutions. Our method achieves substantial improvements over a previous structured-output approach for metal binding site prediction. Significant improvements over the current state-of-the-art are also achieved in predicting catalytic sites from 3D structure in enzymes.

Introduction

In order to accomplish their biological function, proteins often interact with different types of external molecules such as metal ions, prosthetic groups and various organic compounds. Metalloproteins (Bertini, Sigel, and Sigel 2001) bind metal ions in order to stabilize their three-dimensional structure, induce conformational changes or assist protein function, such as electron transfer in cytochromes. Metal binding sites are characterized by the set of protein atoms directly involved in binding the ion, called ligands, and the overall geometry of the site. Furthermore, the same protein often binds multiple ions, with typical numbers ranging from one to four. Enzymes are a fundamental type of proteins which accelerate chemical processes within a cell, by complexing with the substrate and thus lowering the activation energy of the reaction. Functional residues play various roles in the catalytic process, such as donating electrons or polarizing cofactor bonds (Bartlett et al. 2002). Solely binding substrates, cofactors or metals, which are often involved in enzymatic reactions, does not characterize a residue as catalytic according to the Catalytic Site Atlas (CSA) (Porter, Bartlett, and Thornton 2004).

Being able to predict metal binding sites as well as enzyme active sites in novel proteins is a fundamental step in understanding their functioning. Both problems have been mostly addressed as a binary classification task at the residue level: given a protein sequence, predict for each residue whether it is involved in a metal binding site (Passerini et al. 2006), (Shu, Zhou, and Hovmoller 2008) or an active site (Tong et al. 2009), (Cilia and Passerini 2010) respectively. Most existing approaches for modeling the full metal binding geometry assume knowledge of the 3D structure of the protein (Ebert and Altman 2008; Babor et al. 2008) and focus on detecting apo-proteins, i.e. proteins solved without the ion. A recent attempt (Frasconi and Passerini 2008) to predict metal binding geometry from sequence formulates the problem as a structured-output task. The proposed solution is a search algorithm greedily assigning residues to ions (or a default *nil* ion if predicted as free) guided by a scoring function trained to rank correct moves higher than incorrect ones. The algorithm is guaranteed to find the solution maximizing the overall score, given the matroid structure of the problem. However, the scoring function is learned from examples and there is no guarantee that it correctly approximates the true underlying function.

We take here a different viewpoint and formalize the problem as a distance-based supervised clustering task (Basu 2005). Given a set of training instances, we first learn a similarity function predicting whether two residues jointly participate in a certain metal or active site. The learned similarity measure is subsequently fed to a maximum-weight clique algorithm collecting sets of residues maximizing their pairwise similarities. The algorithm has a number of desirable features including automatic selection of the number of clusters, natural handling of overlapping clusters, and scalability to large datasets. Experimental results show a substantial improvement over the structured-output approach for metal binding geometry prediction. Significant improvements over the state-of-the-art are also obtained for active site prediction from protein 3D structure, where both node and edge weights are employed in order to exploit both local predictions and spatial constraints.

Problem description and formalization

Given a protein sequence as a string of characters in the alphabet of 20 amino acids, the problem consists of: detecting the number of binding or catalytic sites; collecting for each site the set of protein residues involved. Metal binding

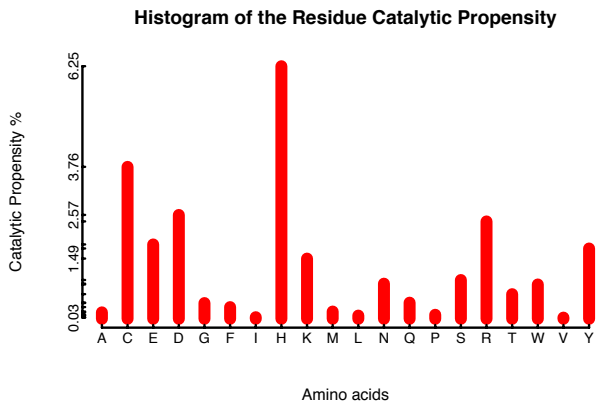

Figure 1: Histogram of the catalytic propensities of the residues in the experimental dataset *HA superfamily* (see experimental section for details).

sites tend to be rather specific in terms of possible ligands with cysteine (C), histidine (H), aspartic (D) and glutamic (E) acids being by far the most common ligands in transition metals. Cysteines and histidines are the vast majority of ligands in structural sites, while aspartic and glutamic acids are quite common in proteins and their relative binding frequency is thus very limited (Passerini et al. 2006). A more complex situation can be observed with alkali and alkaline-earth metals, which often bind proteins through the oxygen in backbone carbonyl groups. Catalytic propensity is even less specific, given the number of different roles that a residue can play within the active site. Figure 1 reports the catalytic propensity of the whole set of amino acids, showing that only few of them can be safely discarded. Previous results (Cilia and Passerini 2010) on the simpler binary classification task actually indicate that keeping all candidates produces slightly better results on average: the predictor occasionally manages to correctly predict rare amino acids as catalytic without significantly affecting precision.

Concerning the number of sites, metalloproteins usually contain between one and three sites, sometimes four and occasionally more. The coordination number of a bound ion, i.e. the total number of its ligands, varies from one to about eight depending on the metal. Values between two and four are the most frequent for transition metals. Figure 2 shows the metal binding geometry of the equine herpes virus-1 (PDB code 1CHC), where candidate ligands in $\mathcal{L} = \{C, H\}$ not binding any ion are marked in grey. Contrarily to metal binding sites, enzymes tend to have a single catalytic site involving a larger number of residues, ranging from 1 to 9 in the experimental dataset we used. Multiple active sites can actually be found in some multimeric proteins, such as the 3-isopropylmalate dehydrogenase (PDB code 1A05). Figure 3 shows the active site of cloroperoxidase T (PDB code 1A7U and UniProtKB entry O31168) with seven residues corresponding to seven different amino acids involved. Note that proximity in sequence only partially relates to involvement in the same site, as the three-dimensional arrangement of

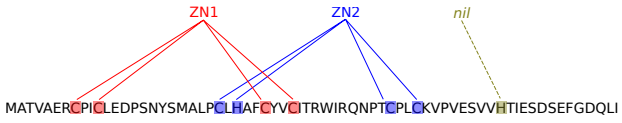

Figure 2: Sequence of the equine herpes virus-1 (PDB code 1CHC). Residues composing the metal binding sites are highlighted in different colors.

```
MPFITVGQEN STSIDLYYED HGAGQPVVLI HGFPLSGHSW  40
ERQSAALLDA GYRVITYDRR GFGQSSQPTT GYDYDTFAAD  80
LNTVLETLDL QDAVLVGFSM GTGEVARYVS SYGTARIAKV 120
AFLASLEPFL LKTDDNPDGA APKEFFDGIV AAVKADRYAF 160
YTGFFNDFYN LDENLGTRIS EEAVRNSWNT AASGGFFAAA 200
AAPTTWYTDF RADIPRIDVP ALILHGTGDR TLPIENTARV 240
FHKALPSAEY VEVEGAPHGL LWTHAEEVNT ALLAFLAK
```

Figure 3: Sequence of the cloroperoxidase T (PDB code 1A7U and UniProtKB entry O31168). Residues composing the active site are highlighted in red.

the protein can bring quite distant residues closer. However, additional features contribute to characterize target residues, such as conservation profile and residue neighborhood.

Given these premises, we formulate the problem as a supervised clustering task. We provide a common formulation for both metal binding site and active site prediction. Slightly abusing terminology, we refer to residues involved in either type of site as *ligands*. While the two problems are treated as separate tasks in the experiments, they are indeed highly correlated as metal binding sites are often part of a larger active site. We are planning to extend our work to predict a structured set of sites in order to jointly address these problems.

A protein sequence is represented as the set x of its candidate ligands, that is residues belonging to $\mathcal{L}$. The output y for the sequence is a subset of the powerset of x, i.e. $y \subseteq \mathcal{P}(x)$. Outputs for proteins in Figures 2 and 3, for instance, would be represented as $\{\{c_1, c_2, c_4, c_5\}, \{c_3, h_1, c_6, c_7\}\}$ and $\{f_2, s_8, m_2, a_{14}, p_7, d_{18}, h_6\}$ respectively, assuming $\mathcal{L}$ is equal to $\{C, H\}$ for metal binding sites and the whole set of amino acids for catalytic sites. The desired output is thus a partial clustering of residues, where only predicted ligands are reported. Furthermore, at least for metal binding sites, clusters can overlap, as the same residue can simultaneously bind two ions, as happens for glutamic and aspartic acids with their two side-chain oxygen atoms. For comparison with previous approaches, experiments only deal with non-overlapping clusters, but our approach can naturally handle overlaps, as described in the next section.

Distance-based supervised clustering with maximum-weight cliques

A training set of labelled proteins can be easily obtained from experimentally solved protein structures and catalytic annotations, and a supervised clustering approach can thus be pursued. We opt for a distance-based supervised ap-

proach (Basu 2005), where training instances are used to learn an appropriate distance (or similarity) measure to be later used in the clustering. The learning stage simply consists of training a pairwise classification function $F(x^i, x^j)$ predicting for each pair of residues x^i and x^j in x whether they belong to the same site. We employ a pairwise support vector machine (SVM) as the underlying classification function. More complex alternatives can be pursued, as will be detailed in the Discussion.

Given a learned similarity function F, we represent a set x as a weighted graph, removing edges whose weight is below a certain threshold ϕ and rescaling remaining weights to be positive. A maximum-weight clique algorithm is then run on the graph in order to return a set of maximal cliques, which correspond to the predicted sites. The rationale for the approach is that given a reasonable pairwise similarity measure, the algorithm should isolate few densely connected components which correspond to the desired solution while discarding most of the nodes in the graph. The algorithm can be asked to return a single large cluster, as typical of the active site prediction task, or a set of possibly overlapping maximal cliques, as for the metal binding site case, where the number of clusters cannot be specified *a priori*.

The maximum-weight clique clustering algorithm

Maximum Clique is a paradigmatic NP-hard problem with relevant applications in many areas; its weighted versions originate from fields such as computer vision, pattern recognition and robotics (Ballard and Brown 1982). A survey on recent literature on Weighted Maximum Clique algorithms can be found in (Pullan 2008).

In the following we introduce our heuristic algorithm. We describe it for weighted edges only. Its extension for dealing with weights on both nodes and edges, as well as the case where weights are averaged on the number of nodes, is straightforward.

Given a set of residues R, in the previous section we defined a learned symmetric similarity function F that maps each pair of residues onto a measure of likelihood that they belong to the same cluster. Given a positive threshold value ϕ, we define a weighted undirected graph as a triplet $G_\phi \equiv (R, E_\phi, F)$ where the vertex set R is composed by the residues, the edge set E_ϕ is defined by vertex pairs whose similarity function F is above the threshold ϕ

$$E_\phi = \big\{\{u, v\} \subset R : u \neq v \wedge F(u, v) \geq \phi\big\},$$

and the weight of every edge $e \in E_\phi$ is given by $F(e)$. From now on, subscript ϕ shall be removed for clarity.

A *clique* in graph G is defined as a completely connected subgraph of G, i.e., any subset $R' \subseteq R$ such that for every pair of nodes $u, v \in R'$ the pair $\{u, v\}$ belongs to E. The *Edge-Weighted Maximum Clique Problem* requires to find the clique in R that maximizes the sum of weights:

$$R'_{\max} = \operatorname*{arg\,max}_{\substack{R' \subseteq R \\ R' \text{ clique in } G}} \sum_{u,v \in R'} F(u, v).$$

Input	Meaning
R, E, F_E	Edge-weighted undirected graph

Variable	Meaning
t	Current iteration index
T	Prohibition period
L_v	Last iteration when $v \in R$ was added/removed
$\bar{R}$	Current configuration
P	List of nodes that can be added to $\bar{R}$
w	Clique weight
v	Chosen node
a	Action to be taken (`Add` or `Drop`)

```
1  function WMC(R, E, F_E)
2     L_v ← −∞ for v ∈ R
3     t ← 0; R̄ ← ∅; P ← R; w ← 0
4     repeat
5        UPDATEPROHIBITION(R̄, T)
6        (v, a) ← CHOOSENODE(L, R̄, P, T, t, E, F_E)
7        if a = Add
8           R̄ ← R̄ ∪ {v}
9        else
10          R̄ ← R̄ \ {v}
11       recompute P and w incrementally
12       L_v ← t
13       if too many iterations without improvements
14          RESTART()
15       t ← t + 1
16    until termination condition is met
17    return best R̄ found
```

Figure 4: The main section of WMC: the local search step is repeated and the best clique is returned (bookkeeping operations such as best configuration maintenance are not shown).

Being a generalization of the Maximum Clique Problem, the edge-weighted version is also NP-hard. In this paper, we introduce the Reactive Local Search optimization heuristic for Weighted Maximum Clique finding (RLS-WMC, in the following WMC for short), based on the RLS-MC heuristic for Maximum Clique finding (Battiti and Protasi 2001), with a novel dynamic behavior adapted from (Battiti and Mascia 2010).

The reaction technique of the WMC heuristic, described below, offers an effective diversification mechanism that provides a thorough exploration of the search space, and is therefore capable of dealing with problem instances for which exhaustive enumeration is infeasible.

The WMC heuristic, whose main section is shown in Fig. 4, is a stochastic local search (SLS) algorithm. In SLS algorithms for the MC problem, a "current" configuration (subset of vertices) $\bar{R} \subseteq R$ is maintained throughout the search, being initially the empty set (line 4), and is modified by incremental moves consisting in the addition or in the removal of a node (lines 10–13). At every step the "current" configuration is required to be a clique in the original graph (the system generally moves only within feasible solutions),

1 **function** CHOOSENODE($L, \bar{R}, P, T, t, E, F_E$)
2 $S \leftarrow \left\{ w \in P : \begin{array}{c} L_w > t - T \wedge \\ \wedge\, w \text{ maximizes future expectations} \end{array} \right\}$
3 $a \leftarrow$ `Add`
4 **if** $S = \emptyset$
5 $S \leftarrow \left\{ w \in \bar{R} : \begin{array}{c} L_i > t - T \wedge \\ \wedge\, w \text{ maximizes future expectations} \end{array} \right\}$
6 $a \leftarrow$ `Drop`
7 Pick $v \in S$
8 **return** (v, a)

Figure 5: The CHOOSENODE procedure: choose the non-prohibited node having the best chance to lead to better cliques in the future; if no nodes can be added, pick one for removal.

therefore the addition move will only consider nodes that maintain the clique property, i.e., that are connected to all nodes in $\bar{R}$. Such set of eligible nodes is called P in Fig. 4, and is maintained incrementally during the search.

The WMC heuristic completes the generic SLS framework by defining the criteria by which the incremental moves are selected. In particular, a parameter T, called *prohibition period*, is set and a vector $(L_v)_{v \in R}$, storing the last iteration at which node v was added or removed to the current clique $\bar{R}$, is initialized (line 3) and maintained (line 15). Nodes that have been used in the last T iterations, called *prohibited*, are not considered for addition or removal. This mechanism, known as *Tabu Search*, prevents the system from getting stuck in local optima and encourages diversification.

The move selection routine CHOOSENODE, whose purpose is the choice of the next node to be added or removed, is outlined in Fig. 5. Array (L_v) is used to check prohibitions. Since more than one non-prohibited node is usually eligible for addition to $\bar{R}$, other selection criteria intervene in order to maximize the chance that a large clique will be obtained, for instance by choosing the node that maximizes the average edge weight (line 2), with ties broken randomly (line 7). If no nodes are eligible for insertion in the current configuration $\bar{R}$ (either because there are no more nodes connected to all nodes in $\bar{R}$, or all of them are prohibited), then a non-prohibited node chosen within $\bar{R}$ is selected for removal (lines 4–6).

The value of the prohibition period T is critical for the good behavior of the algorithm. Small values of T tend to be insufficient for the system to efficiently escape local optima, while high values highly reduce the flexibility of the search procedure by reducing the number of eligible nodes. Rather than relying on an ideal value of T as a function of the graph size and of its density, WMC determines it dynamically (line 8 of Fig 4) by calling a function, UPDATEPROHIBITION, that detects anomalous situations where a change would benefit the search. To achieve this, recent configurations are stored in a hash table; if a configuration is visited (i.e., becomes the current one) too often, then the T parameter is increased in order to improve the differentiation capabilities of the algorithm. If, on the other hand, no configuration is revisited for a given time, T is reduced. Further details on the dynamic adaptation of T are available in (Battiti and Mascia 2010).

Finally, a RESTART mechanism is provided (lines 16–17): if the best solution is not improved in a while, then the algorithm is restarted, so that new regions of the search space are visited. The RLS-WMC algorithm maintains the weight of the current configuration $\bar{R}$ by incrementally updating it at every move.

For the purposes of this paper, cliques within the expected size are stored along with their weight, and are post-processed in order to determine which ones represent the correct clusters. Bookkeeping operations such as the computation of the clique weight, storage of the visited cliques and of the best clique are not detailed here.

Experimental results

Predicting geometry of metal binding sites

We tested our method on the task of predicting metal binding sites in metalloproteins. We used the same setting described in (Frasconi and Passerini 2008), with 30 random 80/20 train/test splits. We encoded pairs of residues by concatenating their features vectors, thus comparing residues according to their order in the sequence. This option was shown (Frasconi and Passerini 2008) to provide better results with respect to alternative approaches such as averaged pairwise comparisons, possibly because sequential ordering is relevant in characterizing sites. Pairs were labeled positive if both residues bind to the same metal ion and negative otherwise, and an SVM was used as the pairwise classifier.

All parameters concerning the SVM and the maximum weighted clique algorithm described below were selected by an inner-fold cross-validation on the training set of the first split and kept fixed for all remaining folds. As a result of this model selection phase, we employed a second degree polynomial kernel and a cost factor $j = 3$ outweighing error on positive with respect to negative examples. In building the weighted graph, we discarded edges having weight smaller than -0.9, and rescaled remaining weights to have positive values. The weight of each clique was averaged over the number of its nodes. The algorithm returned the set of non-overlapping solutions with at most four residues. We made no further selection of the returned solutions, except for limiting the number of solutions to 4.

We present here a set of measures including those reported in (Frasconi and Passerini 2008). Note that we are not trying to predict the identity of an ion (e.g. the "first" zinc, the "second" iron or so), but only the subset of residues which jointly bind the same one. Thus, when evaluating the quality of a certain clustering, we assign each ion to the cluster containing the highest number of its true ligands (if any). An equivalent approach was employed in (Frasconi and Passerini 2008). P_E, R_E, and F_E are the precision, recall, and F_1 of the correct assignment between a ligand and a metal ion. P_S, R_S, and F_S are the precision, recall, and F_1 of the correct prediction of binding sites, i.e., how many sites are entirely correctly predicted over the total number of sites in the chain. P_B, R_B, and F_B are the precision, recall,

and F_1 of the correct prediction of the bonding state of the residues in the chain, i.e. regardless of which ion they actually bind. Tables 1 and 2 report the mean and standard deviation of these performance measures averaged over the 30 splits. The breakdown of these measures for proteins binding different numbers of metal ions (i.e. from 1 to 4) is also reported.

# sites	SVM + WMC			(Frasconi and Passerini 2008)		
	P_E	R_E	F_E	P_E	R_E	F_E
any	$79 \pm 3\bullet$	$59 \pm 5\bullet$	$62 \pm 5\bullet$	66 ± 5	52 ± 4	53 ± 4
1	84 ± 4	73 ± 7	73 ± 6	66 ± 7	58 ± 6	57 ± 6
2	70 ± 8	33 ± 5	42 ± 6	67 ± 7	44 ± 9	48 ± 9
3	70 ± 15	22 ± 8	32 ± 11	69 ± 19	24 ± 13	32 ± 12
4	42 ± 30	16 ± 13	23 ± 18	42 ± 31	20 ± 19	26 ± 22
	P_S	R_S	F_S	P_S	R_S	F_S
any	$42 \pm 7\bullet$	$30 \pm 7\bullet$	$31 \pm 7\bullet$	20 ± 7	17 ± 6	16 ± 6
1	50 ± 8	41 ± 9	41 ± 9	25 ± 10	22 ± 8	22 ± 8
2	25 ± 14	8 ± 7	11 ± 9	15 ± 9	7 ± 7	7 ± 7
3	23 ± 32	4 ± 7	5 ± 11	0 ± 2	0 ± 1	0 ± 2
4	9 ± 21	3 ± 6	5 ± 9	2 ± 7	1 ± 5	1 ± 5
	P_B	R_B	F_B	P_B	R_B	F_B
any	$88 \pm 3\bullet$	63 ± 5	$67 \pm 4\bullet$	79 ± 4	64 ± 6	64 ± 4
1	84 ± 4	73 ± 7	73 ± 6	74 ± 5	68 ± 7	65 ± 6
2	92 ± 8	45 ± 6	58 ± 7	88 ± 5	60 ± 11	66 ± 10
3	100 ± 0	34 ± 12	49 ± 15	98 ± 5	38 ± 22	50 ± 20
4	67 ± 45	25 ± 18	36 ± 25	65 ± 44	32 ± 28	40 ± 31

Table 1: Comparison on the metalloproteins dataset. The means and standard deviations are computed on the 30 random splits. A bullet indicates that the performance differences are statistically significant ($p < 0.05$).

	# sites				
	any	1	2	3	4
SVM + WMC	$27 \pm 6\bullet$	40 ± 9	1 ± 4	0 ± 0	0 ± 0
(Frasconi and Passerini 2008)	14 ± 6	20 ± 8	3 ± 7	0 ± 0	0 ± 0

Table 2: Experimental results on the metalloproteins dataset. A_G is the accuracy at a chain level, i.e., the number of entire configurations correctly predicted. A bullet indicates that the performance differences are statistically significant ($p < 0.05$).

Our SVM+WMC approach achieves significant improvements over the previous structured-output approach in edge, site and bonding state prediction, as measured by paired Wilcoxon tests ($p < 0.05$).

The most significative improvement over (Frasconi and Passerini 2008) lies in the number of sites entirely correctly predicted. The overall P_S, R_S, and F_S, is consistently better for any number of metal ions in the protein.

Active sites prediction

We applied our approach to the prediction of active sites in enzymes. We focused on the *HA superfamily* dataset (Chea and Livesay 2007), the largest dataset employed as benchmark in the literature. Prediction of catalytic residues was previously addressed starting from either sequence or structural information. We considered both settings, relying on previous state-of-the-art results by a simple support vector machine exploiting residue structural neighborhood (Cilia and Passerini 2010). The detailed description of the features employed for both sequence-based and structure-based predictions can be found in this previous work. Given that most proteins contain a single active site, and the labeling found in the CSA (Porter, Bartlett, and Thornton 2004) does not include information on different sites, we considered a single site prediction setting. Common examples of multiple active sites are those of polymeric proteins in which a pair of specular sites is found at the interface of two identical chains. We plan to extract this additional information from known 3D structures in order to fully characterize overall geometry in an extended version of the work.

For sequence-based prediction, we employed a setting analogous to the metal binding site case, with pairs of residues represented as ordered pairs of feature vectors from (Cilia and Passerini 2010). Following (Cilia and Passerini 2010), we employed a linear kernel and a 6 to 1 subsampling of negative (i.e. non-catalytic) residues, resulting in a 61/1 proportion of negative vs positive residue pairs. Following the site size distribution in training instances, we fixed the maximum size of cliques to six.

For structure-based prediction, we took a slightly different approach, since we could also exploit the spatial information provided by the protein structure. We modified the maximum-weight clique algorithm in order to consider both edge and node weights. Edge weights were in this case inverse Euclidean distances between corresponding residues, pruned for distances over 14 Å. This threshold was chosen according to the distribution of distances between catalytic residues in the training set. The idea of constraining candidate solutions based on their pairwise 3D distances was actually used in the MBG prediction approach by Babor et al. (Babor et al. 2008) as an initial filtering stage. However the 3D constraint is much less stringent in catalytic sites, as shown by the quite large threshold (14 Å) we derived from data. Node weights encoded catalytic propensity as predicted by the state-of-the-art support vector machine predictor described in (Cilia and Passerini 2010). Node and edge weights were normalized in order to fall within the same range of values.

Experimental comparisons with the local approach in (Cilia and Passerini 2010) are shown in Table 3, where the protein-level precision, recall and F_1 measures averaged across folds are reported.

	(Cilia and Passerini 2010)			SVM+WMC		
	P	R	F_1	P	R	F_1
seq	20 ± 4	59 ± 7	25 ± 4	22 ± 2	41 ± 4	$27 \pm 3\bullet$
struct	23 ± 3	65 ± 6	28 ± 3	35 ± 7	43 ± 7	$34 \pm 6\bullet$

Table 3: Comparison of the results (performance $\pm$ st.d.) obtained in active site prediction. A bullet indicates that the performance differences are statistically significant ($p < 0.05$).

The SVM+WMC approach achieves significant improvements at $p < 0.05$ in both sequence-based and structure-based predictions according to a paired Wilcoxon test. Note that the average protein-level F_1 of the local predictor is quite lower than the F_1 computed from average protein-level precision and recall. This happens because the local SVM produces rather unbalanced predictions, either maximizing

recall with low precision or (more rarely) vice versa, and for a number of proteins it outputs completely wrong predictions. The SVM+WMC approach is much more stable and balanced in its predictions. Note also that the improvement in F_1 is not simply due to a better choice of the decision threshold with respect to the standard local approach. The best F_1 value which could be obtained with local sequence-based predictions by optimizing the threshold (on the test set) is just 0.256. Results from the structured-based prediction significantly improve the current state-of-the-art thanks to an effective use of the spatial geometry information. In particular, the algorithm finds cliques that discard many of the classifier false positives.

Discussion

We address the problem of predicting geometry of structural and functional sites in proteins by casting it into a supervised clustering task. We propose a novel distance-based supervised clustering approach in which the learned pairwise distance is employed to turn instances into weighted graphs. A maximum-weight clique algorithm is executed on the graph to return a small set of densely connected components corresponding to candidate sites. Supervised clustering is an active area of research and a number of different approaches have been proposed in the literature (Basu 2005). We use a very simple distance learning approach based on pairwise classification of instances. The maximum-weight clique clustering algorithm is however independent of this stage, and can be easily integrated in more complex supervised clustering approaches such as the structured-output formulation proposed in (Finley and Joachims 2005).

The algorithm substantially improves over the only existing approach in predicting geometry of metal binding sites from sequence alone. Focusing on small components with large overall weights, our algorithm is more robust to a possibly incorrect bonding state prediction. On the other hand, the structured-output approach in (Frasconi and Passerini 2008) is capable of exploiting the full relational structure of partial solutions in order to evaluate them, instead of being limited to networks of pairwise interactions. Indeed, such approach is superior when bonding state information is assumed to be known. We are planning to extend our algorithm in order to deal with clique-based weights, thus combining some of the advantages of the two formulations: the ability of a structured-output approach to better model the quality of candidate solutions, and the robustness of stochastic local search strategies in dealing with a scoring function which only approximates conditions guaranteeing greedy optimality (Frasconi and Passerini 2008).

Significant improvements over the state-of-the-art are also obtained in predicting active sites from 3D structure. The algorithm naturally handles the lack of knowledge in the number of clusters, partial clusterings with many outliers and overlapping clusters. We are planning to extend it to return a structured set of solutions, such as metal binding sites as parts of wider active sites, a quite common situation in enzymes.

Acknowledgements

Many thanks to Michèle Sebag for very fruitful discussions.

References

Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; and Edelman, M. 2008. Prediction of transition metal-binding sites from apo protein structures. *Proteins* 70(1):208–217.

Ballard, D., and Brown, C. 1982. *Computer Vision*. Englewood Cliffs: Prentice-Hall.

Bartlett, G.; Porter, C.; Borkakoti, N.; and Thornton, J. 2002. Analysis of catalytic residues in enzyme active sites. *J Mol Bio 2002* 324(1):105–121.

Basu, S. 2005. *Semi-supervised clustering: probabilistic models, algorithms and experiments*. Ph.D. Dissertation, University of Texas at Austin.

Battiti, R., and Mascia, F. 2010. Reactive and dynamic local search for max-clique: Engineering effective building blocks. *Computers & Operations Research* 37(3):534–542.

Battiti, R., and Protasi, M. 2001. Reactive local search for the maximum clique problem. *Algorithmica* 29(4):610–637.

Bertini, I.; Sigel, A.; and Sigel, H., eds. 2001. *Handbook on Metalloproteins*. Marcel Dekker, New York, 1 edition.

Chea, E., and Livesay, D. R. 2007. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics* 8:153+.

Cilia, E., and Passerini, A. 2010. Automatic prediction of catalytic residues by modeling residue structural neighborhood. *BMC Bioinformatics* 11(1):115.

Ebert, J. C., and Altman, R. B. 2008. Robust recognition of zinc binding sites in proteins. *Protein Sci* 17(1):54–65.

Finley, T., and Joachims, T. 2005. Supervised clustering with support vector machines. In *ICML*.

Frasconi, P., and Passerini, A. 2008. Predicting the geometry of metal binding sites from protein sequence. In *NIPS*, 465–472.

Passerini, A.; Punta, M.; Ceroni, A.; Rost, B.; and Frasconi, P. 2006. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins* 65(2):305–316.

Porter, C. T.; Bartlett, G. J.; and Thornton, J. M. 2004. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32(Database issue).

Pullan, W. 2008. Approximating the maximum vertex/edge weighted clique using local search. *Journal of Heuristics* 14:117–134.

Shu, N.; Zhou, T.; and Hovmoller, S. 2008. Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* 24(6):775–782.

Tong, W.; Wei, Y.; Murga, L.; Ondrechen, M.; and Williams, R. 2009. Partial order optimum likelihood (POOL): Maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. *PLoS Computational Biology* 5(1):e1000266+.