BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 3029–3035
doi:10.1093/bioinformatics/btt519

Structural bioinformatics

Advance Access publication September 18, 2013

CLIPS-4D: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3D data

Jan-Oliver Janda[1,†], Andreas Meier[2,†] and Rainer Merkl[1,*]

[1]Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany and [2]Faculty of Mathematics and Computer Science, University of Hagen, D-58084 Hagen, Germany

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The precise identification of functionally and structurally important residues of a protein is still an open problem, and state-of-the-art classifiers predict only one or at most two different categories.

Result: We have implemented the classifier CLIPS-4D, which predicts in a mutually exclusively manner a role in catalysis, ligand-binding or protein stability for each residue-position of a protein. Each prediction is assigned a *P*-value, which enables the statistical assessment and the selection of predictions with similar quality. CLIPS-4D requires as input a multiple sequence alignment and a 3D structure of one protein in PDB format. A comparison with existing methods confirmed state-of-the-art prediction quality, even though CLIPS-4D classifies more specifically than other methods. CLIPS-4D was implemented as a multiclass support vector machine, which exploits seven sequence-based and two structure-based features, each of which was shown to contribute to classification quality. The classification of ligand-binding sites profited most from the 3D features, which were the assessment of the solvent accessible surface area and the identification of surface pockets. In contrast, five additionally tested 3D features did not increase the classification performance achieved with evolutionary signals deduced from the multiple sequence alignment.

Availability: CLIPS-4D is available as a web-service at http://www-bioinf.uni-regensburg.de.

Contact: rainer.merkl@ur.de

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

Received on April 3, 2013; revised on August 1, 2013; accepted on August 31, 2013

1 INTRODUCTION

An important goal of computational biology is the comprehensive annotation of proteins, which requires to predict a function and to identify all crucial residues. To assign function to a query sequence, BLAST (Altschul *et al.*, 1997) and other, more sensitive, algorithms of sequence comparison (Söding, 2005) are of utmost value. However, these methods are *per se* not capable to identify residues, which are critical for function or stability; thus, alternative approaches are needed. One technique is the mapping of known functional sites based on a sequence alignment (Lopez *et al.*, 2011), which is named homology transfer. Alternatively, if the 3D structure of a query is available, algorithms can exploit structural correspondences with annotated active sites to identify functional residues (Goyal *et al.*, 2007; Stark and Russell, 2003). More generally applicable are methods that do not require annotated proteins for comparison but assess each individual residue-position by means of a knowledge-based scoring system. Owing to the relevance of this task, a large number of such *in silico* approaches have been introduced, which are sequence-based (Berezin *et al.*, 2004; Capra and Singh, 2007; Casari *et al.*, 1995; Fischer *et al.*, 2008; Gutman *et al.*, 2005; Huang and Brutlag, 2001; Lichtarge *et al.*, 1996; Overington *et al.*, 1990; Sankararaman *et al.*, 2009; Tang *et al.*, 2009; Teppa *et al.*, 2012) or combine information from sequence and structure of a protein (Ashkenazy *et al.*, 2010; Capra *et al.*, 2009; Kalinina *et al.*, 2009; Laskowski *et al.*, 2005a; Panchenko *et al.*, 2004; Sankararaman *et al.*, 2010; Yahalom *et al.*, 2011; Yao *et al.*, 2003) to predict one or two functional categories.

As we were interested to classify more specifically, we have recently introduced a multiclass support vector machine (MC-SVM), which we named CLIPS-1D (Janda *et al.*, 2012). In contrast to other approaches, CLIPS-1D predicts in a mutually exclusively manner a role in catalysis, ligand-binding or protein structure by analyzing a multiple sequence alignment (MSA). Interestingly, not more than seven carefully selected features related to the conservation and the abundance of residues at individual sites and their local sequence neighborhood were sufficient to attain state-of-the-art performance.

Many of the inferred 3D features are orthogonal to the sequence-based features exploited by CLIPS-1D, and therefore we expected an increase of classification quality for a combination of both. This is why we have systematically determined the classification performance for combinations of 1D and 3D features and selected an optimal combination for a novel classifier, which we named CLIPS-4D. This program uses the 3D structure of a single protein chain to deduce the local environment of each residue and does not use the position of ligands. A comparison with CLIPS-1D made clear that the prediction of ligand-binding sites profited most from the integration of 3D features. Our approach compares favorably with state-of-the-art algorithms, although this MC-SVM distinguishes catalytic, ligand-binding and structurally relevant residue-positions.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 METHODS

2.1 1D features

2.1.1 Conservation of a residue-position The conservation measure $JSD(k)$ (Fischer *et al.*, 2008) has performed better than other conservation measures (see Capra and Singh, 2007) and was computed for a residue-position k according to

$$JSD(k) = H(\frac{f_K^{obs} - f^{backgr}}{2}) - \frac{1}{2}H(f_K^{obs}) - \frac{1}{2}H(f^{backgr}) \quad (1)$$

f_K^{obs} is the probability mass function for site k approximated as $f_K^{obs}(aa_i) = f_k(aa_i)$ by the amino acid frequencies observed in the respective column k of the MSA; the mean amino acid frequencies as found in the SwissProt database (Bairoch and Apweiler, 2000) were taken as background frequencies f^{backgr}. $H(.)$ is Shannon's entropy (Shannon, 1948). For classification, we used the z-score $cons_{JSD}(k)$:

$$cons_{JSD}(k) = \frac{JSD(k) - \mu_{JSD}}{\sigma_{JSD}} \quad (2)$$

Mean μ_{JSD} and standard deviation σ_{JSD} values were determined individually for each MSA.

2.1.2 Conservation of a sequence neighborhood To characterize the conservation of a sequence neighborhood, $cons_{neib}(k)$ was computed in analogy to Fischer *et al.* (2008):

$$cons_{neib}(k) = \frac{1}{|Neib|}\sum_{l \in Neib} w_l \, cons_{JSD}(k+l) \quad (3)$$

$Neib = \{-3, -2, -1, +1, +2, +3\}$ determined the set of neighboring positions. The weights were $w_{-1} = w_{+1} = 3$, $w_{-2} = w_{+2} = 2$, $w_{-3} = w_{+3} = 1$. The conservation of position k did not contribute to $cons_{neib}(k)$.

2.1.3 Propensities of catalytic sites, ligand-binding sites and positions important for structure Inspired by Bartlett *et al.* (2002), three scores named $abund(k, CLASS)$ were computed:

$$abund(k, CLASS) = \sum_{i=1}^{20} f_k(aa_i) \log \frac{f^{CLASS}(aa_i)}{f^{backgr}(aa_i)} \quad (4)$$

$f^{backgr}(aa_i)$ were the above background frequencies, and $f^{CLASS}(aa_i)$ were the frequencies of residues from one set $CLASS \in \{CAT\_sites, LIG\_sites, STRUC\_sites\}$. For the analysis of a single sequence with CLIPS-3D (see later in the text), $f_k(aa_i)$ was 1.0 for aa_s^k and zero for all other residues.

2.1.4 Scoring propensities of a neighborhood To assess the class-specific neighborhood of a site k, we introduced

$$abund_{neib}(aa_s^k, CLASS) = \frac{1}{|Neib|}\sum_{l \in Neib}\sum_{i=1}^{20} f_{k+l}(aa_i) \log \frac{f_{k+l}^{CLASS}(aa_i|aa_s)}{f^{backgr}(aa_i)} \quad (5)$$

Here, aa_s^k is the amino acid aa_s occurring at site k under consideration, $f_{k+l}(aa_i)$ is the frequency of aa_i at position l relative to k, and $f_{k+l}^{CLASS}(aa_i|aa_s)$ is the conditional frequency of aa_i at the same positional offset deduced from the neighborhood of all residues aa_s of a set $CLASS$. $Neib$ is the ± 3 neighborhood.

2.2 3D Features

2.2.1 Conservation of a 3D neighborhood To characterize the conservation of a 3D neighborhood of a residue aa_k in a protein, $3D - cons_{neib}(k)$ was computed in analogy to Formula (3):

$$3D - cons_{neib}(k) = \frac{1}{|3D - Neib|}\sum_{l \in 3D-Neib} cons_{JSD}(l) \quad (6)$$

$3D\text{-}Neib$ is the set of all residues aa_l in the vicinity of k possessing an $atom_s$ such that the distance between van der Waals spheres of at least one pair of sidechain heavy atoms ($atom_r$, $atom_s$) with $atom_r$ from aa_k is at most 0.5 Å. $cons_{JSD}(l)$ is a normalized Jensen–Shannon divergence; see Formula (2).

2.2.2 Scoring propensities of a 3D neighborhood To assess the class-specific 3D neighborhood, we introduced

$$3D - abund_{neib}(aa_s^k, CLASS) = \frac{1}{|3D - Neib|}\sum_{l \in 3D-Neib}\sum_{i=1}^{20} f_l(aa_i) \log \frac{f_l^{CLASS}(aa_i|aa_s)}{f^{backgr}(aa_i)} \quad (7)$$

Here, aa_s^k is the amino acid aa_s occurring at site k under consideration, $f_l(aa_i)$ is the frequency of aa_i at position l, and $f_l^{CLASS}(aa_i|aa_s)$ is the conditional frequency of aa_i deduced from the neighborhood of all residues aa_s of a set $CLASS$. The set $3D\text{-}Neib$ was determined as previously mentioned.

2.2.3 Mutual information score of a 3D neighborhood As proposed (Buslje *et al.*, 2010), we determined a proximity score pMI, which assesses the mutual information of pairs of residue-positions in the vicinity of k.

$$pMI(k) = \frac{1}{|3D - Neib|}\sum_{l \in 3D-Neib} cMI(l) \quad (8)$$

$cMI(l)$ is a cumulative mutual information value (see Buslje *et al.*, 2010), and $3D\text{-}Neib$ was determined as previously mentioned.

2.2.4 Assessing the B-factor of a residue In analogy to Petrova and Wu (2006), a normalized B-factor $BF(k)$ was computed.

$$BF(k) = \frac{BFmean(k) - \mu_{BFmean}}{\sigma_{BFmean}} \quad (9)$$

$BFmean(k)$ is the mean B-factor deduced from all n atoms of residue aa_k according to

$$BFmean(k) = \frac{1}{n}\sum_{i=1}^{n} BFA(atom_i) \quad (10)$$

and μ_{BFmean} and σ_{BFmean} are the mean and the standard-deviation determined individually for each 3D structure.

2.2.5 Computing the relative solvent-accessible surface area Using the software library *BALL* (Hildebrandt *et al.*, 2010), the solvent-accessible surface area ($SASA$) was deduced from the protein 3D structure for each residue aa_k to compute the relative $SASA$ ($rSASA$).

$$rSASA(aa_k) = \frac{SASA(aa_k)}{SASA_{\max}(aa_k)} \quad (11)$$

Here, $SASA_{max}(aa_k)$ is the maximally possible $SASA$ (Miller *et al.*, 1987) of the amino acid.

2.2.6 Assessing pockets As has been shown, fpocket (Le Guilloux *et al.*, 2009) is one of the best methods for the identification of pockets in proteins (Volkamer *et al.*, 2010). fpocket scores cavities of the protein surface based on a Voronoi tessellation and alpha spheres. To compensate for the protein-specific number of pockets, we determined a normalized score $nPocket$.

$$nPocket(aa_k) = \frac{\max(PocketScore)}{PocketScore(aa_k)} \quad (12)$$

$\max(PocketScore)$ is the largest score deduced for any pocket of the considered protein, and $PocketScore(aa_k)$ is the score of the pocket in which aa_k is allocated. We assigned a score of -1 to all residues that did not belong to pockets or whose $rSASA$ value was <4%.

2.2.7 Evaluation of the classification performance To assess the performance of a classification, the rates Sensitivity (Recall), Specificity and Precision

$$Sensitivity = \frac{TP}{TP+FN},\ Specificity = \frac{TN}{TN+FP},\ Precision = \frac{TP}{TP+FP} \tag{13}$$

were determined as well as areas under the precision-recall curve (PR-AUC). As a further performance measure, the Matthews correlation coefficient (MCC) has been introduced (Matthews, 1975).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{14}$$

MCC-values are considered a fair measure to assess performance on unbalanced sets of positives and negatives (Ezkurdia *et al*., 2009), as observed here. In all formulae, TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. For example, when classifying catalytic sites, positives are the selected *CAT_sites*, and negatives are all other residue-positions of the considered MSAs.

2.2.8 Classifying by means of SVMs CLIPS-4D was configured and trained as described for CLIPS-1D (Janda *et al*., 2012). We used the *libsvm* library (Chang and Lin, 2011) with a Gaussian radial basis function kernel and determined optimal parameters γ_{RBF} and C during training by means of a grid search (Schölkopf and Smola, 2002). Training and assessment was organized as an 8-fold cross validation. For each training step, the number of positive and negative cases was balanced. To eliminate sampling bias during the grid search, all parameters were deduced as means from training trials with the same positives and 50 different randomly selected sets of negative cases. In contrast to training, all positive and all negative cases were classified to compute performance measures (e.g. MCC-values). The output of the MC-SVM consists of four class-probabilities p_{CLASS} (see Wu *et al*., 2004) for each residue-position. Each residue-positions k was assigned to the class, whose p_{CLASS}-value was largest. *P*-values were determined as follows: For each class and each residue, the respective cumulative distribution was deduced from the p_{CLASS}-values of residue-positions l belonging to *NOANN_sites*. That is, the *P*-value for a Glu-residue with p_{STRUC}-value $s(k)$ is the fraction of all Glu-residues from *NOANN_sites* reaching or surpassing $s(k)$.

3 RESULTS AND DISCUSSION

3.1 Selecting an optimal combination of 1D and 3D features

For training and testing SVMs on a combination of 1D and 3D features, we used the set of MSAs prepared for CLIPS-1D (Janda *et al*., 2012) and supplemented the respective pdb-files (Dutta *et al*., 2009). In brief, all MSAs were taken from the HSSP database (Sander and Schneider, 1991), and each residue-position was assigned to one of four sets (classes) representing functional categories. The set *CAT_sites* consists of 840 catalytic residue-positions, which are listed in the manually curated part of the Catalytic Site Atlas (CSA, Version 2.2.12) (Porter *et al*., 2004) and come from 264 enzymes. For 216 of these enzymes, we found 4466 ligand-binding sites in the pdbsum database (Laskowski *et al*., 2005b), which constitute the dataset *LIG_sites*. Owing to the lack of a representative set of proteins, which are annotated with structurally important residue-positions, we regarded conserved residues in the core of proteins important for structure. Thus, we have handpicked 136 proteins without enzymatic function and identified 3703 residue-positions, which were both buried and more conserved than the mean; see Janda *et al*. (2012). This set was named *STRUC_sites*; the remaining 19 223 residue-positions were named *NOANN_sites* and represented residue-positions without crucial function. During training and testing, residue-positions from one set $CLASS \in \{ CAT\_sites, LIG\_sites, STRUC\_sites \}$ and from *NOANN_sites* served as positive or negative cases to train six two-class SVMs; for details, see Janda *et al*. (2012).

For CLIPS-1D, we have chosen seven sequence-based features for classification: $cons_{JSD}(k)$ [Formula (2)] assesses the conservation of individual residue-positions and $cons_{neib}(k)$ [Formula (3)] assesses the conservation of their neighborhood. $abund(k, CLASS)$ [Formula (4)] scores the abundance of residues at functionally or structurally important sites and $abund_{neib}(k, CAT\_sites)$ and $abund_{neib}(k, LIG\_sites)$ [Formula (5)] score the composition in the neighborhood of functionally important sites.

To this end, six more features, which require a 3D structure for their computation, were selected as candidates for a combination with the aforementioned sequence-based features. Two are 3D versions of sequence-based conservation scores: $3D - cons_{neib}(k)$ [Formula (6)] scores the conservation of residues in the 3D neighborhood of residue-position k and $3D - abund_{neib}(aa_s^k, CLASS)$ assesses the class-specific abundance of residues in the 3D neighborhood of amino acid aa_s at position k [Formula (7)]. $pMI(k)$ [Formula (8)] scores dependencies of residue distributions in the vicinity of residue k and has been reported as improving the prediction of catalytic sites (Buslje *et al*., 2010). The biochemical role of a residue may affect its flexibility, which can be estimated with the mean B-factor $BF(k)$ (Petrova and Wu, 2006) computed according to Formula (9). The relative solvent accessible surface area $rSASA(k)$ [Formula (11)] allows for the differentiation of surface and core residues. Catalytic and ligand-binding sites tend to lie in surface pockets (Volkamer *et al*., 2010); thus, we used the normalized score $nPocket(k)$ according to Formula (12) as a further feature.

Before finding an optimal combination of features, we were interested to corroborate the contribution of evolutionary information and of 3D data to classification quality. Thus, we combined those features that do not require an MSA for classification and named the resulting MC-SVM CLIPS-3D. We chose $abund(k, CLASS)$, which scores the abundance of residues in functionally or structurally important sites (Janda *et al*., 2012), and the 3D features $3D - abund_{neib}(aa_s^k, CLASS)$, $rSASA(k)$ and $nPocket(k)$. CLIPS-3D was trained and assessed by means of an 8-fold cross validation on the aforementioned classes. The output of this and all other MC-SVMs of the CLIPS-suite is for each residue-position a set of four class probabilities $p_{CLASS}(k)$ (Wu *et al*., 2004), which were taken to assign k to the class with the highest probability.

Generally, it is difficult to characterize the performance of a classifier, if the number of positive and negative cases is highly unbalanced as is also the case here. A fair measure (Ezkurdia *et al*., 2009) is the MCC [Formula (14)], which was computed for all analyses; see Table 1. Comparing the MCC-values of CLIPS-1D and CLIPS-3D shows that evolutionary information is important to predict *CAT_sites* and *STRUC_sites*, and that 3D data contribute markedly to the prediction of *LIG_sites*.

Next, we combined each of the aforementioned 3D features and all CLIPS-1D features to train and assess six different MC-SVMs

analogously. The comparison with MCC-values of CLIPS-1D made clear that only *rSASA* and *nPocket* improved classification performance and that the classification of *LIG_sites* profited most by the latter two features (Table 1). Further performance tests showed that the assessment of a residue's neighborhood in 3D space did not outperform the respective 1D features. Thus, we combined the seven sequenced-based features of CLIPS-1D with *rSASA* and *nPocket* to form the classifier CLIPS-4D. Compared with CLIPS-1D, the MCC-values increased from 0.34 to 0.43 for *CAT_sites*, from 0.12 to 0.27 for *LIG_sites* and from 0.67 to 0.68 for *STRUC_sites*. As the respective MCC-value was optimal, if *CAT_sites* with $p_{CAT}(k) > 0.64$ were selected as positives, we implemented this cutoff for functional assignment. For the output of CLIPS-4D, we additionally determined a residue-specific *P*-value, which indicates the fraction of *NOANN_sites* reaching or surpassing the considered p_{CLASS}-value (see Section 2.2.8). Using *P*-values in the range of 0.01–0.20 as cutoffs, MCC-values as well as Sensitivity (Recall), Specificity and Precision [Formulae (13)] of CLIPS-4D were determined (Table 2 and Supplementary Fig. S1). MCC-values of *CAT_sites* and *LIG_sites* reached a plateau for $P \geq 0.05$. Thus, we recommend the *P*-value of 0.05 for the selection of functional sites, as sensitivity is acceptable and specificity is then as high as 0.99, 0.95 and 0.96 for *CAT_sites*, *LIG_sites* and *STRUC_sites*; compare Table 2.

In summary, the final configuration of CLIPS-4D is an MC-SVM, which classifies based on nine features. These are the seven sequence-based features $cons_{neib}(k)$, $abund(k, CLASS)$ and $abund_{neib}(aa_s^k, CLASS)$, plus the two 3D features $rSASA(k)$ and $nPocket(k)$. CLIPS-4D is available as a web service at http://www-bioinf.uni-regensburg.de; this version was trained on the full datasets. An assessment of typical output is provided as Supplementary Data.

Table 1. MCC-values of classifiers for crucial residue-positions

Classifier	*CAT_sites*	*LIG_sites*	*STRUC_sites*
CLIPS-3D	0.31	0.22	0.43
CLIPS-1D	0.34	0.12	0.67
$3D - cons_{JSD}$	0.29	0.10	0.69
$3D - abund_{neib}$	0.32	0.11	0.66
pMI	0.34	0.09	0.64
BF	0.32	0.11	0.66
rSASA	0.34	0.13	0.63
nPocket	0.37	0.27	0.68
CLIPS-4D	0.43	0.27	0.68
ConSurf	0.30		0.46

Note: For all variants of the CLIPS classifier, MCC-values for the classification of *CAT_sites*, *LIG_sites* and *STRUC_sites* are listed. CLIPS-3D is based on seven propensities or structure-based features, which do not require an MSA for computation, and CLIPS-1D uses seven sequence-based features. Each of the lines labeled $3D - cons_{JSD}$, $3D - abund_{neib}$, *pMI*, *BF*, *rSASA* and *nPocket* gives the performance of an MC-SVM exploiting the seven CLIPS-1D features plus the listed 3D feature. CLIPS-4D is a classifier using the seven CLIPS-1D features plus *rSASA* and *nPocket*. The classifier ConSurf does not distinguish catalytic and ligand-binding sites. Therefore, we merged the sets *CAT_sites* and *LIG_sites* before classification.

3.2 Classification performance of CLIPS-4D varies in a class- and residue-specific manner

Owing to their biochemical properties, residues are not evenly distributed at functionally or structurally important positions. For example, only 11 residues are generally observed as being directly involved in catalysis (Bartlett *et al.*, 2002), and few residues are overrepresented at catalytic sites. The charged residues Lys, Glu, Arg, Asp and His as well as Cys are the only residues with an $abund(k, CAT\_sites)$-value > 0.5, whereas Pro and the hydrophobic residues Val, Ile, Leu and Ala are scored < -2.0, i.e. are drastically underrepresented. To characterize classification performance in detail, we determined in a class-specific manner MCC-values for each individual residue. In Figure 1, these MCC-values were plotted versus abundance scores. For *CAT_sites*, all of the overrepresented residues were classified with an MCC-value > 0.33. In contrast, the underrepresented residues Ala, Gly and Phe had an MCC-value of zero; no MCC-value could be computed for the underrepresented residues Pro, Val, Ile, Leu and Met due to missing values. The $abund(k, LIG\_sites)$-scores were less extreme, indicating that more types of residues are involved in ligand-binding than in catalysis. MCC-values were lowest (<0.13) for the underrepresented residues Glu and Lys but also for the overrepresented residues Asp, Arg and His. These three residues were also observed as *CAT_sites*, which might explain why some ligand-binding sites were misclassified as catalytic ones. Among *STRUC_sites*, the MCC-values were generally higher (mean 0.63) and for the hydrophobic residues Ala, Val, Ile, Leu, Met, Phe, Tyr and Trp the mean was 0.73.

Table 2. Classification performance of CLIPS-4D for different *P*-value thresholds

P-value threshold	Sensitivity (Recall)			Specificity			Precision			MCC		
	CAT	LIG	STRUC	CAT	LIG	STRUC	CAT	LIG	STRUC	CAT	LIG	STRUC
0.010	0.26	0.10	0.34	1.00	0.99	0.99	0.45	0.44	0.88	0.33	0.19	0.50
0.025	0.34	0.20	0.51	0.99	0.98	0.98	0.36	0.39	0.81	0.34	0.25	0.59
0.050	0.47	0.32	0.64	0.99	0.95	0.96	0.34	0.31	0.75	0.39	0.28	0.64
0.100	0.50	0.43	0.74	0.99	0.92	0.95	0.32	0.27	0.73	0.39	0.28	0.68
0.150	0.50	0.46	0.77	0.99	0.91	0.94	0.32	0.26	0.72	0.39	0.29	0.69
0.200	0.50	0.46	0.79	0.99	0.91	0.94	0.32	0.25	0.71	0.39	0.28	0.69

Note: All values were determined according to Formulae (13) and (14). The specific results for the classes *CAT_sites*, *LIG_sites* and *STRUC_sites* are listed in the columns labeled 'CAT', 'LIG', and 'STRUC', respectively.

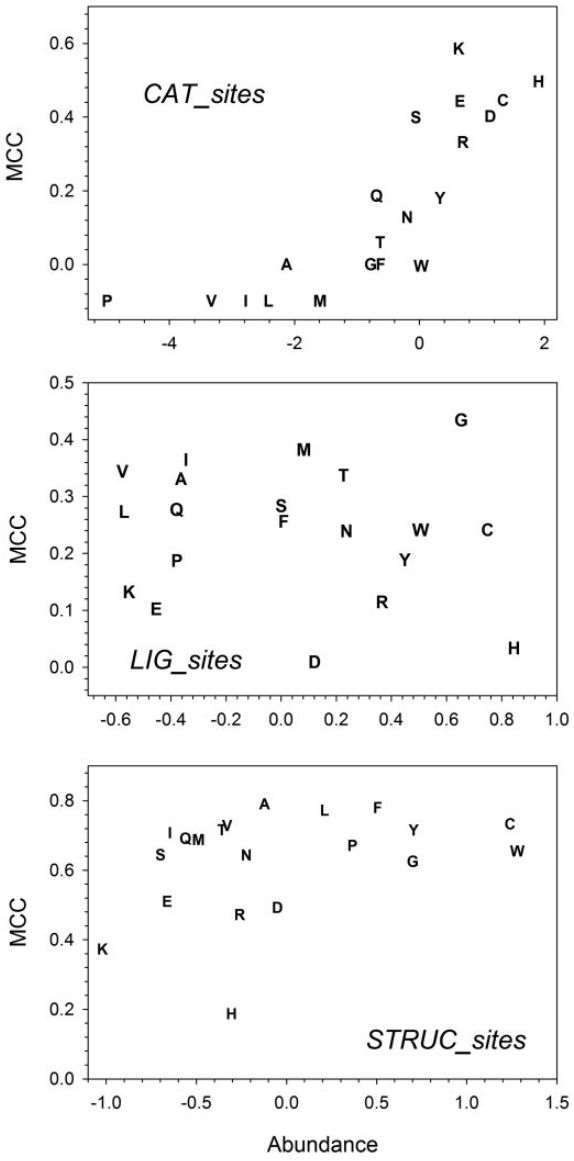

Fig. 1. Class-specific performance of CLIPS-4D for individual residues. In a class-specific manner, MCC-values were determined for each residue and plotted versus the respective abundance values *abund*(*k, CLASS*). The rareness of residues P, V, I, L and M among *CAT_sites* precluded to compute MCC-values. Thus, they were assigned an MCC-value of −0.1

Classification performance was lowest for the two less abundant residues His and Lys. In summary, these findings indicate that classification performance varies to a great extent depending on residues and their function. This is why we introduced *P*-values, which allow the user to select predictions of similar quality for all classes and sites. A detailed analysis of class-specific misclassifications was added to the Supplementary Data.

3.3 The more specific classification of CLIPS-4D reaches the same performance as less specific alternatives

Methods for the identification of functionally important residues can be divided into homology transfer methods and other approaches that do not use knowledge about binding sites in a homologous protein for a prediction. As CLIPS-4D belongs to the second group, we comprehensively compared its performance with other methods following the same approach.

DISCERN (Sankararaman *et al*., 2010) and POOL (Somarowthu *et al*., 2011) use 1D and 3D features to predict functionally important residue-positions. On a subset of the CSA, a recall of 0.50 at a precision of 0.19 was determined for DISCERN (Sankararaman *et al*., 2010); CLIPS-4D reached for this recall a precision of 0.32. For a subset of 100 enzymes from the CSA and a false-positive rate of 0.05, POOL had a recall of 0.87 at a precision of 0.15 (Somarowthu *et al*., 2011). For $P = 0.05$, the recall of CLIPS-4D was 0.47 at a precision of 0.34; compare Table 2. ConCavity, which focusses on ligand-binding residues, reaches a PR-AUC value of 0.65 for entries of the LigASite database and one of 0.32 for catalytic sites of the CSA (Capra *et al*., 2009). CLIPS-4D reaches PR-AUC values of 0.23 for *LIG_sites* and of 0.30 for *CAT_sites*, respectively (Supplementary Fig. S1). Most plausibly, the lower performance of CLIPS-4D on *LIG_sites* is due to the pdbsum-specific choice of ligands (which include metal ions) and ligand-binding residues. However, the MCC-value of 0.56 reached by CLIPS-4D for the difficult case of CASP9 target T0604 (Schmidt *et al*., 2011) suggests the prediction of a substantial number of biologically relevant ligand-bindings sites; see later in the text.

An alternative to CLIPS-4D is ConSurf, which predicts two distinct categories, namely, functionally or structurally important residue-positions. ConSurf is available as a web service and deduces a measure for evolutionary conservation from an MSA (Ashkenazy *et al*., 2010). The overall performance of ConSurf was better when we uploaded our preprocessed MSAs instead of letting ConSurf generate MSAs on its own (data not shown). Additionally, performance of ConSurf was best if we classified residues having assigned the maximal conservation score of 9 as positive and all other residues as negative cases. We presumed a structural role if the residue was buried (rSASA < 5%) and a functional role if it was exposed to the solvent (rSASA ≥ 5%). As ConSurf does not distinguish between catalytic and ligand-binding sites, we merged *CAT_sites* and *LIG_sites* before the assessment. The resulting MCC-value of 0.30 was closer to the MCC-value reached by CLIPS-4D for *LIG_sites*, which corresponds to their overrepresentation in the merged datasets. For *STRUC_sites*, the MCC-value was 0.46; see Table 1.

In summary, these comparisons confirmed state-of-the-art performance for CLIPS-4D, which offers a broader classification spectrum than alternatives.

3.4 CLIPS-4D can supplement homology transfer methods in the prediction of ligand-binding sites

In the ligand-binding site prediction category of CASP, it is the task to predict residues directly involved in ligand binding in the experimental control structure. The results of CASP9 (Schmidt *et al*., 2011) and CASP10 experiments (http://www.predictioncenter.org/casp10/) impressively demonstrate that most ligand-binding residues can be predicted with high performance by homology transfer methods. However, if the ligand is large and flexible, it is difficult to predict the full binding site, as indicated

Table 3. Classification performance on ligand-binding sites of *firestar*, CLIPS-4D and a combination of predictions determined for CASP targets

	T0526 3NRE	T0584 3NF2	T0604 3NLC	T0615 3NQW	T0632 3NWZ	T0721 4FK1
firestar						
MCC	0.49	0.69	0.45	0.52	0.49	0.73
Sens	0.44	1.00	0.36	0.36	0.38	0.74
Spec	0.99	0.96	0.99	0.99	0.98	0.97
CLIPS-4D						
MCC	0.61	0.19	0.54	0.34	0.24	0.45
Sens	1.00	0.46	0.73	0.55	0.50	0.48
Spec	0.95	0.87	0.94	0.91	0.82	0.95
Union						
MCC	0.58	0.44	0.54	0.50	0.40	0.68
Sens	1.00	1.0	0.79	0.82	0.75	0.84
Spec	0.94	0.86	0.93	0.90	0.81	0.94

Note: All MCC-, sensitivity (label 'Sens'), and specificity (label 'Spec') values were determined according to Formulae (13) and (14). The rows with label 'Union' give performance values resulting from merging positive predictions from *firestar* and CLIPS-4D. The first line gives the number of the target, and the second line gives the pdb id.

by a lower performance (MCC-value of ~0.5 for best-performing methods) on CASP9 target T0604 (Schmidt *et al*., 2011). Additionally, substrates tend not to be crystallized in proteins, and their binding residues are more family specific. Thus, we hypothesized that CLIPS-4D might supplement homology transfer by predicting additional residues that bind substrates or non-metal ligands.

First, we confirmed that CLIPS-4D predictions help to identify substrate or product binding sites of the enzymes IGPS (1A53), LgtC (1G9R), HIT (1KPF) and TIM (1M7P); pdb ids are given in brackets. In each of these and the following cases, the sets *CAT_sites* and *LIG_sites* were merged before assessing in a CASP-related manner the performance for residues binding non-metal ligands. A detailed analysis was added as Supplementary Data.

Second, we compared the outcome of CLIPS-4D and *firestar* (Lopez *et al*., 2011), a top-performing method, for those cases of CASP9 where the MCC-value of the best-performing participant of the contest was lowest (Schmidt *et al*., 2011). We selected five targets with a non-metal ligand, namely, T0526, T0584, T0604, T0615 and T0632. Owing to the lack of sufficiently large MSAs or unavailable 3D structures, we could only analyze one non-metal target of CASP10, namely T0721. Results were summarized in Table 3 and listed as Supplementary Data. In two of the six cases (T0526, T0604), the MCC-value of CLIPS-4D was superior to *firestar*. A union of the predictions generated by *firestar* and CLIPS-4D gave in comparison with *firestar* for two more cases (T0615, T0721) a higher sensitivity at the cost of a moderate loss in specificity.

Among these six CASP targets, the performance of CLIPS-4D was worst for T0584 (pdb id 3NF2). T0584 is a polyprenyl transferase generating the product from the building blocks isopentenyl diphosphate and dimethylallyl diphosphate (DMAPP) by consecutive steps of elongation, cyclopropagation, rearrangement and cyclization reactions (Wallrapp *et al*., 2013). During synthesis, the product grows into an elongation cavity, and mutagenesis studies made clear that residues protruding into the elongation cavity determine the length of the product (Liang *et al*., 2002). Two pairs of aspartates of Asp-rich regions are involved in binding DMAPP and catalysis *via* chelation of the cofactor Mg^{2+}. Five residues from these Asp-rich regions shown to be important for catalysis (Liang *et al*., 2002) were predicted by CLIPS-4D as *CAT_sites*. Four of these residues are not directly involved in ligand binding in the experimental control structure 1RQI and are thus false-positive predictions as well as 41 *LIG_sites*. Of these, 18 line the elongation cavity modeled previously (Tarshis *et al*., 1996). Three more *LIG_sites* most likely contact the ligand after an active site rearrangement; see Supplementary Data for details. Therefore, experimental evidence makes plausible some of these predictions that do not belong to the extended binding site.

Thus, although not representative due to the small number of analyses, these findings suggest to supplement the result of homology transfer with CLIPS-4D predictions in cases of active site rearrangements, flexible substrates or unknown poses of a ligand.

4 CONCLUSIONS

The combination of evolutionary and 3D data allows CLIPS-4D to predict critical residue-positions with state-of-the-art quality. As shown here, not more than nine features are sufficient to reach state-of-the-art classification performance, if features are orthogonal to each other. The sequence- and structure-based features contribute differently to the identification of functionally and structurally important residue-positions: For the identification of catalytic and structurally important sites, sequence-based features like conservation are most relevant, for ligand-binding sites 3D features indicating a position in a surface pocket contribute markedly to classification quality. Assessing the content of the CSA made clear that those residues, which were frequently found at catalytic sites could be identified with high quality. In contrast, the identification of residues, which are rare at catalytic sites, and those of ligand-binding sites is still a

difficult problem. CLIPS-4D identifies biologically relevant residue-positions and can supplement methods of homology transfer.

ACKNOWLEDGEMENT

The authors thank Patrick Löffler for assistance in implementing the web server.

Funding: The work was supported by the Deutsche Forschungsgemeinschaft (grant ME-2259/1-1).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashkenazy,H. *et al.* (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bartlett,G.J. *et al.* (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

Berezin,C. *et al.* (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322–1324.

Buslje,C.M. *et al.* (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol.*, **6**, e1000978.

Capra,J.A. *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Casari,G. *et al.* (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Sys. Technol.*, **2**, 1–27.

Dutta,S. *et al.* (2009) Data deposition and annotation at the worldwide protein data bank. *Mol. Biotechnol.*, **42**, 1–13.

Ezkurdia,I. *et al.* (2009) Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinform.*, **10**, 233–246.

Fischer,J.D. *et al.* (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.

Goyal,K. *et al.* (2007) PAR-3D: a server to predict protein active site residues. *Nucleic Acids Res.*, **35**, W503–W505.

Gutman,R. *et al.* (2005) QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Res.*, **33**, W255–W261.

Hildebrandt,A. *et al.* (2010) BALL-biochemical algorithms library 1.3. *BMC Bioinformatics*, **11**, 531.

Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res.*, **29**, 202–204.

Janda,J.O. *et al.* (2012) CLIPS-1D: Analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinformatics*, **13**, 55.

Kalinina,O.V. *et al.* (2009) Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinformatics*, **10**, 174.

Laskowski,R.A. *et al.* (2005a) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.

Laskowski,R.A. *et al.* (2005b) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.

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

Liang,P.H. *et al.* (2002) Structure, mechanism and function of prenyltransferases. *Eur. J. Biochem.*, **269**, 3339–3354.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lopez,G. *et al.* (2011) Firestar-advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Miller,S. *et al.* (1987) Interior and surface of monomeric proteins. *J. Mol. Biol.*, **196**, 641–656.

Overington,J. *et al.* (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.*, **241**, 132–145.

Panchenko,A.R. *et al.* (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

Petrova,N.V. and Wu,C.H. (2006) Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sankararaman,S. *et al.* (2009) INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Res.*, **37**, W390–W395.

Sankararaman,S. *et al.* (2010) Active site prediction using evolutionary and structural information. *Bioinformatics*, **26**, 617–624.

Schmidt,T. *et al.* (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79** (**Suppl. 10**), 126–136.

Schölkopf,B. and Smola,A.J. (2002) *Learning with kernels*. The MIT Press, London.

Shannon,C. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Somarowthu,S. *et al.* (2011) High-performance prediction of functional residues in proteins with machine learning and computed input features. *Biopolymers*, **95**, 390–400.

Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3341–3344.

Tang,K. *et al.* (2009) Prediction of functionally important sites from protein sequences using sparse kernel least squares classifiers. *Biochem. Biophys. Res. Commun.*, **384**, 155–159.

Tarshis,L.C. *et al.* (1996) Regulation of product chain length by isoprenyl diphosphate synthases. *Proc. Natl Acad. Sci. USA*, **93**, 15018–15023.

Teppa,E. *et al.* (2012) Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. *BMC Bioinformatics*, **13**, 235.

Volkamer,A. *et al.* (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.*, **50**, 2041–2052.

Wallrapp,F.H. *et al.* (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proc. Natl Acad. Sci. USA*, **110**, E1196–E1202.

Wu,T.F. *et al.* (2004) Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, **5**, 975–1005.

Yahalom,R. *et al.* (2011) Structure-based identification of catalytic residues. *Proteins*, **79**, 1952–1963.

Yao,H. *et al.* (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.