Data mining for important amino acid residues in multiple sequence alignments and protein structures

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.) der Fakultät für Biologie und vorklinische Medizin der Universität Regensburg

vorgelegt von

Jan-Oliver Janda

aus Linz, Österreich

im Jahr 2014

Das Promotionsgesuch wurde eingereicht am: 18.02.2014

Die Arbeit wurde angeleitet von: apl. Prof. Dr. Rainer Merkl

Unterschrift:

Contents

List of Figures

List of Tables

Abbreviations

Å	*Å*ngström (10^{-10} m)
A	alanine
C	cysteine
CASP	Critical Assessment of protein Structure Prediction
D	aspartic acid
DHFR	dihydrofolate reductase
DMAPP	dimethylallyl diphosphate
DNA	deoxyribonucleic acid
E	glutamic acid
F	phenylalanine
G	glycine
H	histidine
HK	hexokinase
I	isoleucine
IGP	indole-3-glycerol phosphate
K	lysine
L	leucine
M	methionine
MCC	Matthew's correlation coefficient
MSA	multiple sequence alignment

N	asparagine
P	proline
PDB	Protein Data Bank
Q	glutamine
QMI	quantum mutual information
R	arginine
rSASA	relative solvent accessible surface area
S	serine
SASA	solvent accessible surface area
SVM	support vector machine
T	threonine
TrpA	α-subunit of tryptophan synthase
TrpB	β-subunit of tryptophan synthase
TrpC	indole-3-glycerol phosphate synthase
V	valine
W	tryptophan
Y	tyrosine

Abstract

Enzymes are highly efficient bio-catalysts interesting for industries and medicine. Therefore, a goal of utmost importance in biochemical research is to understand how an enzyme catalyzes a chemical reaction. Here, the computational identification of functionally or structurally important residue positions can be of tremendous help. The datasets that are most informative for the algorithms are the 3D structure of a protein and a multiple sequence alignment (MSA) composed of homologous sequences. For example, an MSA allows for the quantification of residue conservation. Residue conservation at a given position indicates that only one type of amino acid fulfills all constraints imposed by protein structure or function. Furthermore, a detailed analysis of less strictly conserved residue positions may identify pairs, whose orchestration is mutually dependent and induces correlated mutations. Both of these conservation signals are indicative of functionally or structurally important positions.

In the first part of this thesis, methods of machine learning were used to identify and classify these residue positions. It was the aim to predict in a mutually exclusively manner a role in catalysis, ligand-binding or protein stability for each residue position of a protein. Unfortunately, for many proteins the 3D structure is unknown. For other proteins, the number of known homologs is not sufficient to compile a meaningful MSA. Therefore, three variants of a classifier were designed and implemented, named CLIPS-1D, CLIPS-3D, and

CLIPS-4D. These multi-class support vector machines allow for a classification based on an MSA (CLIPS-1D), a 3D structure (CLIPS-3D), and a combination of both (CLIPS-4D). CLIPS-1D exploits seven sequence-based features, whereas CLIPS-3D utilizes seven structure-based features. CLIPS-4D combines the seven sequence-based features of CLIPS-1D with those two structure-based features that increased its classification performance. A comparison with existing methods and a detailed analysis on a well-studied enzyme confirmed state-of-the-art prediction quality for CLIPS-1D and CLIPS-4D.

In the second part of this thesis an algorithm for the identification of correlated mutations was improved. A common method for the identification of correlated mutations is to deduce the mutual information (MI) of a pair of residue positions from an MSA. The classical MI is based on Shannon's information theory that utilizes probabilities only. Consequently, these approaches do not consider the similarity of residue pairs, which is a severe limitation. In order to improve these algorithms, H2rs was developed for this thesis. Thus, the MI-values originate from the von Neumann entropy (vNE), which takes into account amino acid similarities modeled by means of a substitution matrix. To further improve the specificity of H2rs, the significance of MI_{vNE}-values was assessed with a bootstrapping approach. The analysis of a large *in silico* testbed and the detailed assessment of five well-studied enzymes demonstrated state-of-the-art performance.

Zusammenfassung

Enzyme sind hocheffiziente Biokatalysatoren, die sowohl für industrielle als auch für medizinische Anwendungen höchst interessant sind. Deshalb ist es eines der wichtigsten Ziele biochemischer Forschung zu verstehen wie Enzyme chemische Reaktion katalysieren. Dafür ist eine computergestützte Identifikation von funktionell oder strukturell wichtigen Aminosäuren von außerordentlicher Hilfe. Die Datensätze mit dem größten Informationsgehalt für solche Algorithmen sind Protein 3D-Strukturen und multiple Sequenzalignments (MSAs), die aus homologen Sequenzen bestehen. MSAs erlauben es beispielweise die Konservierheit einzelner Aminosäuren zu quantifizieren. Die strikte Konserviertheit einer Aminosäure an einer bestimmten Position zeigt, dass nur ein Typ von Aminosäure alle Anforderungen der Struktur und Funktion erfüllt. Darüber hinaus kann die Analyse weniger strikt konservierter Aminosäuren solche Paare identifizieren, die voneinander abhängig sind und deshalb korrelierte Mutationen auslösen. Diese beiden Konserviertheitssignale deuten auf funktionell oder strukturell wichtige Aminosäuren hin.

In dieser Arbeit wurden Methoden des maschinellen Lernens dazu verwendet solche Aminosäuren zu identifizieren und zu klassifizieren. Ziel war es für jede Aminosäure eines Proteins eine Rolle in der Katalyse, der Ligandenbindung oder der Proteinstabilität vorherzusagen. Leider ist die 3D-Struktur vieler Proteine noch nicht bekannt. Für andere Proteine ist es nicht möglich

ein MSA von ausreichender Größe und Qualität zu erzeugen. Deshalb wurden drei Varianten eines Klassifikator entwickelt: CLIPS-1D, CLIPS-3D und CLIPS-4D. Diese Mehrklassen Support Vektor Maschinen ermöglichen eine Klassifikation anhand eines MSAs (CLIPS-1D), einer 3D Struktur (CLIPS-3D) oder beidem (CLIPS-4D). CLIPS-1D nutzt sieben sequenz-basierte Merkmale. CLIPS-3D hingegen nutzt sieben struktur-basierte Merkmale. CLIPS-4D wiederum kombiniert die sieben sequenz-basierten Merkmale von CLIPS-1D mit den zwei struktur-basierten Merkmalen von CLIPS-3D, die die Klassifikation verbesserten. Ein Vergleich mit etablierten Methoden und eine detaillierte Analyse eines gut untersuchten Enzyms bestätigten für CLIPS-1D und CLIPS-4D eine Vorhersagequalität auf dem Stand der Technik.

Eine weit verbreitete Methode um korrelierte Mutationen zu identifizieren, ist die Bestimmung der Transinformation (MI) eines Aminosäurepaares anhand eines MSAs. Die klassische MI basiert auf Shannon's Informationstheorie, die nur Wahrscheinlichkeiten zur Berechnung heranzieht. Folglich können diese Methoden Ähnlichkeiten von Aminosäuren nicht berücksichtigen, was eine große Einschränkung darstellt. Deshalb wurde in dieser Arbeit der Algorithmus H2rs entwickelt. Hier basieren die MI-Werte auf der von Neumann Entropie (vNE), die Aminosäureähnlichkeiten in Form einer paarweisen Ähnlichkeitsmatrix berücksichtigt. Um die Spezifität von H2rs weiter zu verbessern, wurde die Signifikanz der MI_{vNE}-Werte durch einen Bootstrapping-Ansatz bestimmt. Die Auswertung eines großen Datensatzes und eine detaillierte Analyse von fünf gut untersuchten Enzymen hat für H2rs eine Vorhersagequalität auf dem Stand der Technik bestätigt.

1 Introduction

Proteins are versatile macromolecules that are of utmost importance for almost all cellular processes. They are involved in signaling, transport, and have structural as well as defense functions. Furthermore, they act as enzymes that catalyze a multitude of chemical reactions that would otherwise not take place in a reasonable timespan [160]. For a detailed understanding of proteins, their composition and properties are described in the following section.

1.1 Proteins and enzymes

Proteins usually consist of one or more chains of amino acids. An amino acid is composed of an amine (-NH_2) and a carboxylic acid (-COOH) functional group that are connected by a C-atom. Furthermore, this C-atom binds a hydrogen atom and another group, the amino acid side chain which is responsible for the differences in physicochemical properties of amino acids. Figure 1.1 gives an overview of the most important amino acid properties like polarity, size, and charge. One amino acid molecule can react with two others and become chained by peptide bonds, thus forming peptides. Proteins are peptides consisting of more than 30 amino acids. Once linked in a chain, an individual amino acid is called a residue. Figure 1.2 shows the conformation of a residue participating in two peptide bonds. The linked residues without side chains are also known

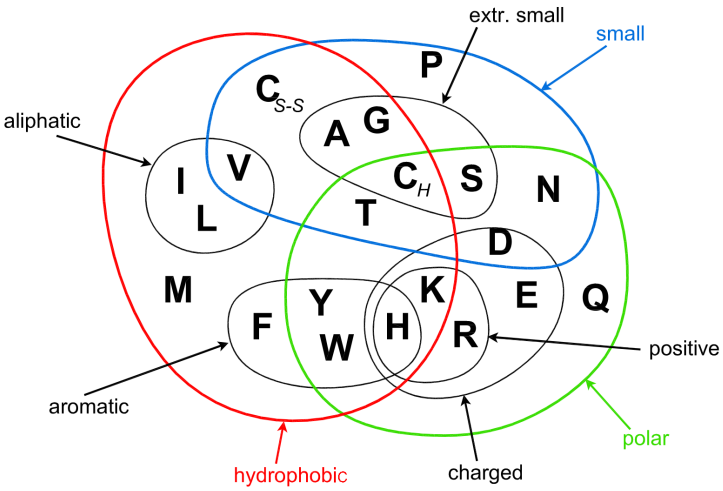


Figure 1.1: Venn-diagram of the properties of the 20 amino acids
Amino acids are grouped by their phyicochemical properties. Major properties are polar, hydrophobic, and small. This figure was taken from [96].

as the main chain or the backbone. The sequence in which residues appear in a main chain is called primary structure. Determined by the primary structure and the local environment, amino acid chains will fold to energetically favorable secondary structures like the extremely regular local sub-structures α-helix and β-sheet which are linked by flexible regions, called loops. These folding patterns are mostly stabilized by backbone interactions like hydrogen bonds in which hydrogens are bound to highly electronegative atoms such as nitrogen and oxygen. A single protein usually contains several of these secondary structure elements. Driven by non-specific hydrophobic effects, the secondary structure elements are folded into a compact three-dimensional arrangement, the tertiary structure. Furthermore, the term quaternary structure is used to describe proteins composed of multiple subunits. Such complexes of two or more subunits are called multimers. Specifically, a complex with only one subunit is called a monomer, one with two

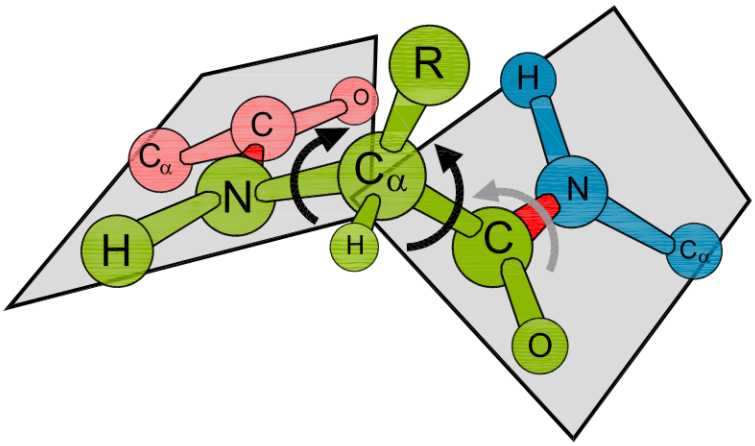


Figure 1.2: Conformation of a peptide bond

The two peptide bonds of an amino acid residue (green) are shown (red). All six atoms participating in a peptide bond lie on a plane. The amino acid side chain is denoted as R. This figure was taken from [96].

subunits a dimer. Proteins from different species which have a common ancestor are called homologs if they share the same function. Based on this evolutionary relationship these proteins are similar in sequence and structure and thus can be grouped into protein families. The Pfam protein families database (release 26.0) [117] contains a collection of over 13000 manually curated protein families. Of those, over 3500 are domains of unknown function (DUF), protein subunits that have no characterized function.

Enzymes are proteins that act as bio-catalysts. The decarboxylation of orotidine 5‘-phosphate, for example, proceeds with a half-life of 78 million years in neutral solution. In the presence of the enzyme orotidine 5‘-phosphate decarboxylase from *Saccharomyces cerevisiae*, however, the reaction is accelerated by a factor of 10^{17} to a half-life of merely 18 milliseconds [118]. The effectiveness of enzymes is due to the reduction of the activation energy E_a by stabilizing the energetically

unfavourable transition state of a substrate. This is shown in Figure 1.3 for the enzyme indole-3-glycerol phosphate synthase (TrpC) from *Sulfolobus solfataricus* which catalyzes the ring closure of an N-alkylated anthranilate (CdRP) to a 3-alkyl indole derivative (IGP). The activation energy $E_{a\ (catalyst)}$ of the reaction in the presence of TrpC is lower than the activation energy E_a in absence of a catalyst because the amino acids K53 and E51 stabilize the transition state. This illustration, however, is a simplification as there are several intermediate products which induce more than one transition state in this particular reaction. As enzymes are such outstanding bio-catalysts, one of the most important goals in biochemical research is to determine the function of an enzyme (and how exactly such an enzyme catalyzes its chemical reaction). Because most experimental methods for the characterization of molecular functions are expensive, time-consuming and hard to conduct, computational methods have become more and more popular. Only a few residues are directly involved in catalysis. In Figure 1.3 these are the two residues E51 and K53. These are close to the active site of an enzyme which is where substrate-binding and the reaction take place. Other residues do not directly participate in catalysis but are still important for binding the substrate or cofactors and yet others play a fundamental role in stabilizing the whole enzyme. Therefore, algorithms that are able to predict such crucial sites are of tremendous help in unraveling the function of an enzyme.
The aim of this work was to develop algorithms for the identification of residues important for catalysis, ligand-binding or structure. Amongst others, machine learning algorithms were utilized with features from protein structures and sequences.

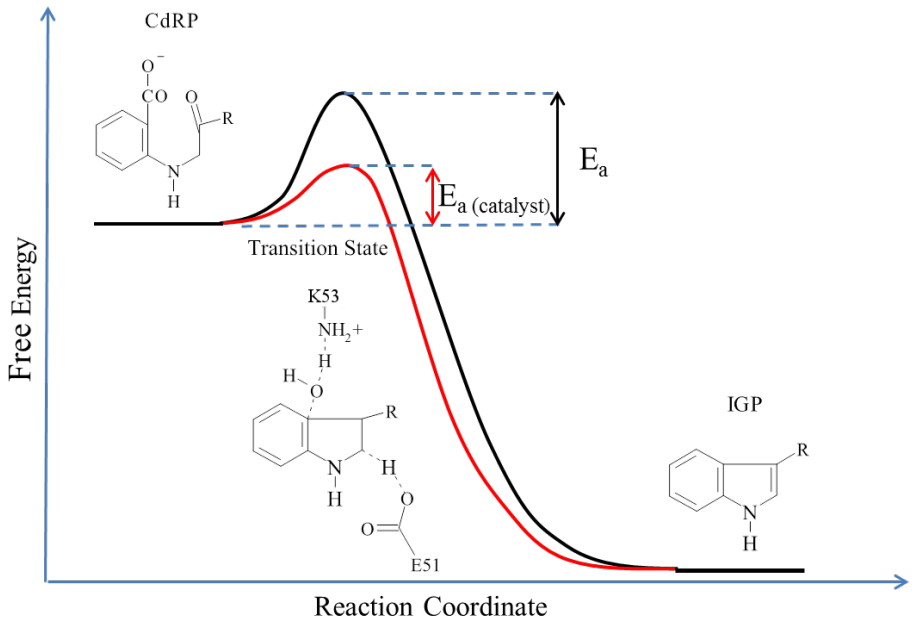


Figure 1.3: Simplified reaction coordinate diagram

The reaction from an N-alkylated anthranilate (CdRP) to indole-3-glycerol phosphate (IGP) is shown. The activation energy $E_{a\ (catalyst)}$ of the reaction in the presence of the enzyme indole-3-glycerol phosphate synthase (TrpC) from *Sulfolobus solfataricus* is lower than the activation energy E_a in absence of a catalyst because the amino acids K53 and E51 stabilize the transition state.

1.2 Machine learning

In general, machine learning is a discipline of artificial intelligence that studies systems that can learn from data without being specifically programmed. Machine learning algorithms can be used to combine overlapping or contradictory information to an optimal prediction. If there is no prior knowledge about the available data, unsupervised algorithms like principal component analysis, independent component analysis or clustering algorithms can be applied to discover structure in the data [33]. If there are labelled examples available (i.e. samples

with known class affiliation), supervised algorithms like neural networks, or support vector machines (SVMs) can be utilized. As these algorithms are able to learn from crucial information to correctly classify unknown data, they perform in general better than unsupervised procedures. Because SVMs perform exceedingly well in identifying important protein residues [113] [166] [148], they were also used in this work. Separating two sets of data can be a very difficult task. Therefore, a basic SVM projects the data into a high dimensional space and constructs a hyperplane to separate the (lower-dimensional) data, thus, efficiently achieving a non-linear classification into two classes. This procedure is also called the kernel trick. The best separation is achieved by the hyperplane that has the largest distance to the nearest training data points, also called the support vectors, of any class (Figure 1.4). SVMs can be tuned to allow for multi-class predictions

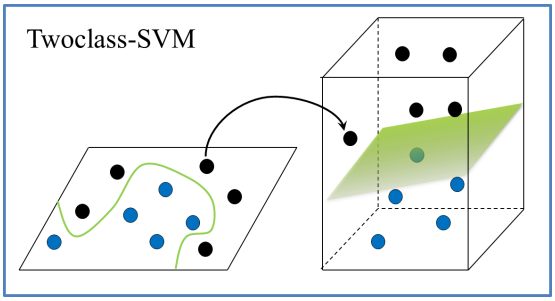


Figure 1.4: Principle of a support vector machine
Data points of class one (black dots) are separated from data points of class two (blue dots) by a hyperplane (green) in a higher dimensional space.

and soft margins for a better consideration of outliers by allowing misclassified examples. Furthermore, SVMs are able to calculate *a posteriori* probabilities for predictions which is often preferable over a binary classification. To avoid over-learning, which occurs if a statistical model is too tightly fit to a finite set of data points, the labelled data have to be partitioned for testing and training. This

ensures the generalizability of a classifier, i.e. the accurate classification of new data after having experienced the training data. Usually a k-fold cross-validation is used for this purpose. Therefore, the original data set is randomly divided into k subsets of equal size. Of these, $k-1$ subsets are used as training data and the remaining single subset is used for testing, so that each subset is used once for testing and $k-1$ times for training. As the test data are labeled samples, they can be categorized in four groups after classification. True positives and false positives are correct and incorrect positive predictions, respectively. True negatives and false negatives are defined accordingly. Based on this four categories, a multitude of measures to determine the performance of a classifier can be calculated. Sensitivity, Specificity, Precision, and Accuracy are the most widely used ones. Furthermore, Matthew's correlation coefficient (MCC) is especially useful when comparing classes with a high imbalance in the number of positive and negative samples [93].

Prior to a classification, meaningful information (features) have to be extracted and selected from the data. This is a crucial step in creating a classifier because the information content and orthogonality of the used features determines the performance of the whole classifier to a large degree. In the case of identifying important residues, features for training an SVM can be extracted from sequence data and protein structures.

1.3 Protein structures

Protein structures are represented *in silico* as rigid 3D models of actual proteins. They are either obtained by X-ray crystallography or by nuclear magnetic resonance spectroscopy. The former uses the diffraction of X-ray beams into many specific directions when hitting atoms of a previously crystallized protein [95]. By measuring the angles and intensities of the diffracted beams, the density of electrons within the crystal can be determined. Consequently, atom positions can be modeled into the electron density map. The latter uses powerful magnets to send radio frequency signals through a protein sample and measures the absorption [124]. The nuclei of individual atoms will, depending on their environment, absorb different frequencies of radio signals. Furthermore, the absorption signals can be perturbed by neighboring nuclei, which can be exploited to determine their distance. These distances in turn can be used to determine the structure of the protein. All structures are stored as text files with atom types and coordinates in the Protein Data Bank [13]. As an example, TrpC from *Sulfolobus solfataricus* is shown in Figure 1.5. The enzyme belongs to the highly populated $(\beta\alpha)_8$-topology where canonically eight β-sheets are surrounded by eight α-helices. A simplified representation of this enzyme's reaction coordinate diagram is shown in Figure 1.3.

An especially useful feature for classifying important residues which can be deduced from protein structures is the relative solvent accessible surface area (rSASA). It is used to identify catalytic residues and ligand-binding sites as these are often positioned in surface pockets to be able to interact with the substrate. Structurally important sites on the other hand are mostly buried in the core of the protein. Usually the accessible surface of a protein is determined by a small spherical probe molecule rolling over the protein *in silico* [122]. During this process, the probe touches all possible contact points of the protein [46]. The center of

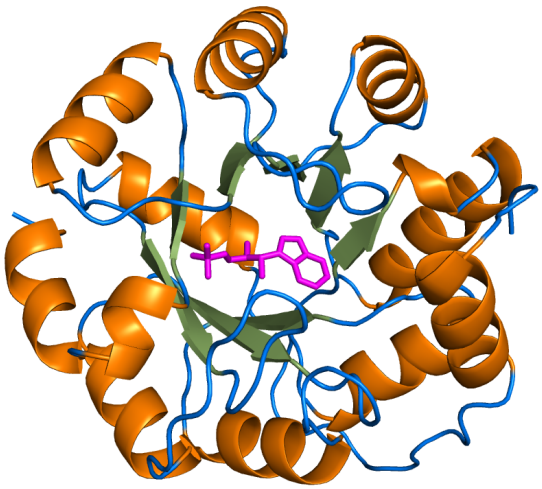


Figure 1.5: Representation of a protein structure
The protein structure of TrpC from *Sulfolobus solfataricus* is shown in cartoon representation. The bound ligand indole-3-glycerol phosphate is depicted in stick representation (magenta). α-helices are colored orange, β-sheets are green and loops blue. The PDB entry with ID 1A53 was used and visualized with PyMOL [132].

the probe is tracked continuously and, thus, outlines the solvent accessible surface area (SASA) as depicted in Figure 1.6. The resulting surface is measured analytically or numerically in $\mathring{A}^2$. The radius of the probe is usually chosen as 1.4 $\mathring{A}$. This corresponds to the size of a water molecule [82]. Therefore, the center of the probe touches about the same area of a protein surface that is accessible for a water molecule [119]. If the probe radius is chosen as zero, the Van-der-Waals-surface will be determined, which is a protein's imaginary hard shell that can not be penetrated by other molecules. The rSASA of a residue is then determined by dividing its SASA by the maximally obtainable SASA for that kind of residue, which is the SASA of a G-X-G tripeptide. For classification purposes, a wide

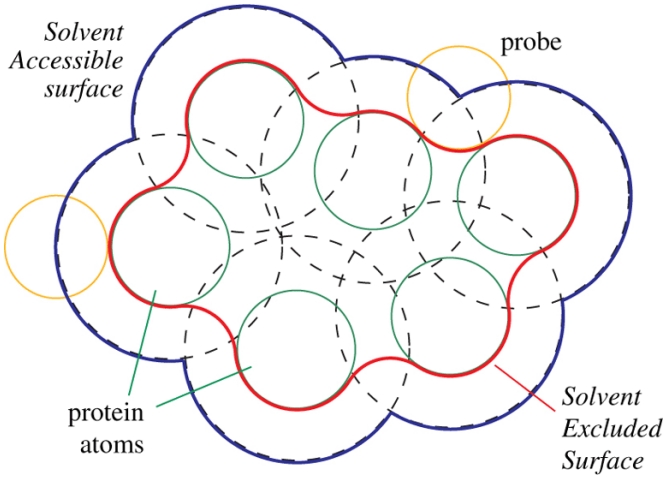


Figure 1.6: Schematics of the solvent accessible surface area
The solvent accessible surface area (blue) and the solvent excluded surface (red) are defined by a probe (orange) rolling over the molecule's atoms (green). This figure was taken from [74].

range of structure-based features like flexibility [113], distortion angles [162], and other topological constraints [151] are useful in finding important residues.

1.4 Multiple sequence alignments

Besides structural data, sequence data can be utilized for extracting features to train and assess an SVM. The most information about a protein can be deduced from multiple sequence alignments (MSAs). An MSA is an arrangement of three or more sequences of homologs, so that corresponding residues are aligned in one column. During evolution sequences might be changed by point mutations or insertions and deletions which alter amino acid occupation or protein length. Therefore, gaps have to be inserted between residues to mathematically mimic these mutations. An MSA holds much more information than a single sequence

because there is a multitude of sequence information for each residue position reflecting the evolution this protein has undergone. Figure 1.7 depicts a part of an MSA of TrpC homologs. The term residue position denotes a column of

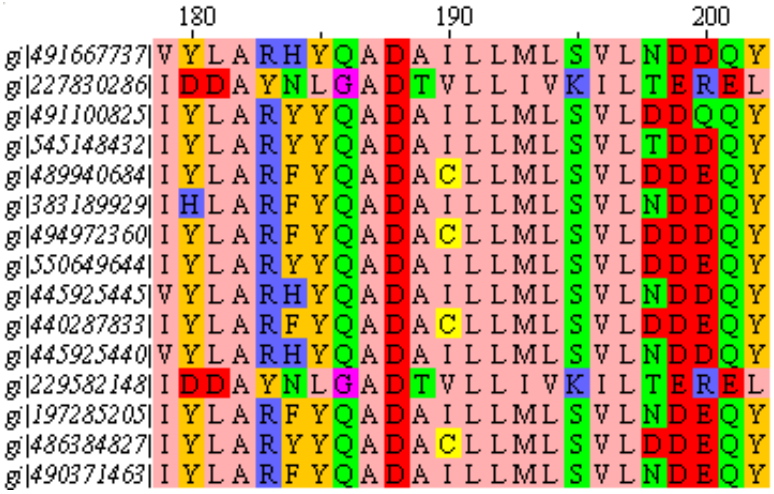


Figure 1.7: Example of a multiple sequence alignment

Positions 179 - 202 of an MSA of indole-3-glycerol phosphate synthase (TrpC) from *Sulfolobus solfataricus* are shown. Each line represents a homologous protein sequence named by a gene identifier number (e.g. gi|491667737). Each letter depicts an amino acid residue (see Abbreviations). The alignment was generated with MAFFT [70] and visualized with Jalview [156] in the *Zappo* color scheme which groups residues by their physiochemical properties.

an MSA and likewise the residue's position in a corresponding protein structure. At some residue positions only a certain residue fulfils all requirements for a properly functioning protein. Consequently, some or all mutations at such a residue position are impossible (i.e. lethal for the organism carrying the mutation) which leads to characteristically occupied columns in MSAs. An invariant column is called strictly conserved as e.g. column 188 in Figure 1.7. As important residues are subjected to the most severe requirements, the predominant feature used to

identify important residue positions *in silico* is the conservation of individual MSA columns. Strictly conserved residues are often vital for protein function [20] [110] [155] whereas a predominant but not strictly conserved amino acid may play a role in protein stability or ligand-binding [83] [3]. A multitude of scores for measuring conservation has been proposed. Reviews of conservation measures are given in [150] and [65]. One of the simplest methods used is the relative frequency $f(k)$ of an MSA column k, where the most frequent residue determines the conservation of a column. Therefore, strictly conserved columns score equal to one. For instance, column 198 in figure 1.7 has a conservation of $f(198) = 6/15$. More sophisticated and widely used are methods based on Shannon's entropy $H(k)$ [135]:

$$H(k) = -\sum_{i=1}^{20} f(a_i^k) \ln f(a_i^k) \tag{1.1}$$

Here, a strictly conserved column has an entropy of zero. The maximum value differs regarding to the used logarithm but is always reached when all 20 amino acids occur with equal frequency. Many variations of Shannon's entropy like the relative entropy or the Jensen-Shannon divergence exist.

Due to the different physicochemical properties of residues, only some can participate in catalysis [11]. Therefore, the abundance of amino acids is an especially prominent feature for catalytic residues as these can be separated in two groups. The first group (A, F, G, I, L, M, P, l, and W) consists of the nonpolar, uncharged and, thus, catalytically inert amino acids (see 1.1), which cannot directly participate in catalysis. But even these residues can in some cases be of importance for catalysis due to steric effects and substrate specificity. The residues of the second group are polar or charged (C, D, E, H, K, N, Q, R, S, T, and Y). They can, for example, donate or accept protons and are thus catalytically active. There are similar but weaker tendencies for binding sites and structurally important residues. Hydrophobic residues tend to be stabilizing, whereas hydrophilic

residues are rather unimportant for protein structure. Polar, hydrophilic residues often play an important role in ligand-binding as they tend to be exposed to the solvent. Additional features for finding important residues are the conservation of proximate residues [20] [11] and the abundance of particular amino acid residues observed at important sites [42] [11].

Due to mutual dependencies even not highly conserved residue positions can be of importance the function and stability of a protein. Such residue positions can be identified with the help of an MSA but correlations are often due to structural reasons. For instance, the occurrence of a specific amino acid at a given position may crucially depend on the local environment, which can impose restrictions with respect to the size (or chemical properties) of the neighboring residues. Thus, an amino acid replacement at one position is tolerated only together with a complementary residue substitution at a correlated site. As a consequence, the frequencies of particular residues at adjacent positions in the structure of a protein can be interrelated (Figure 1.8). Correlated mutations can appear either intermolecular [145] [159] and can, therefore, play an important role in protein-protein interfaces or intramolecular [134] [145]. Intramolecularly correlated residue positions can be important for protein structure as covarying residues are often in close proximity to each other [91]. Generally, it is difficult to predict a structural or functional role for covarying residues [31]. The identification of correlated mutations is further complicated by phylogenetic noise if an MSA is dominated by a large number of closely related homologs and small sample size bias, i.e. correlations caused randomly by the composition of sparse sequence data. Algorithms for the identification of correlated mutations can be divided into two groups: global and local approaches. Global methods were developed recently [16] [67] [102] [90]. These methods treat pairs of residues as mutually dependent entities and, thus, minimize the effects of chained covariation and noise which

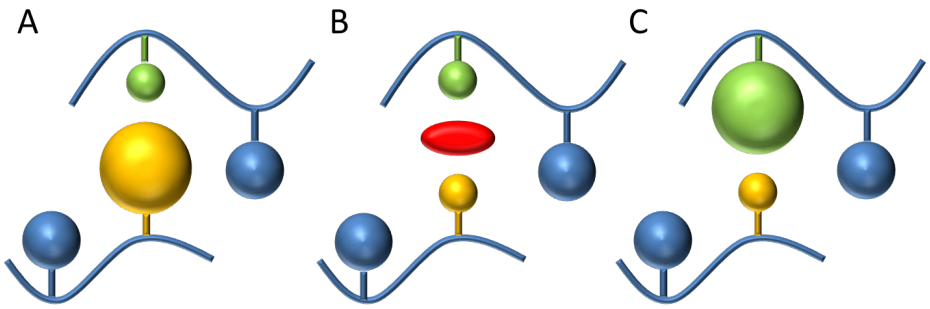


Figure 1.8: Schematics of a correlated mutation

A) The backbone (blue lines) and proximate residues (blue spheres) of a small residue (green sphere) and a large residue (orange sphere) are shown. **B)** The large residue (orange) is mutated to a small residue, which causes a destabilizing cavity (red ellipse). **C)** A compensating mutation occurs, the green residue is mutated to a larger one to stabilize the local environment defined by the neighboring residues.

is very advantageous in contact prediction. On the other hand, the older local methods [97] [88] [92] [6] [147] [44] [51] [136] assume the independence of residue pairs. Consequently, local methods tend to find structurally close and likewise distant correlations. A prerequisite for the application of global and likewise local methods is the existence of MSAs with a sufficiently large number of diverse sequences. As local methods only consider pairs of residues, the required number of sequences is considerably lower. Local methods have been successfully applied to contact prediction and the identification of important residues. For example, Lockless and Ranganathan [88] have developed a specific method named SCA since 1999 and recently proposed the existence of correlated groups of residues, called sectors, each responsible for a physicochemical property of a protein [54]. Another method is H2r, which has successfully identified residue positions important for function or structure [31]. A review can be found in [28].

There are many different measures that are used to quantify the correlation of the occupancy of residue positions. One of the most used measures in global and local methods alike is the mutual information $U(k, l)$ of two residue position k and l which is based on Shannon's entropy (Formula 1.1):

$$U(k, l) = H(k) + H(l) - H(k, l) \tag{1.2}$$

Consequently, most of these approaches only take the raw amino acid frequencies into account but not the similarities between different residues (Figure 1.1) or the required magnitude of biochemical changes it takes to mutate an amino acid to another.

1.5 Aim of this work

The aim of this work was the development of algorithms for the identification of functionally or structurally important amino acid residues by utilizing MSAs and protein structures and thus helping in elucidating the function of uncharacterized proteins. This aim was divided into two subtasks.

The first subtask was to develop a classifier for the identification of important residue positions and to assign a role in catalysis, ligand-binding or protein structure for each residue. Such a fine grained separation is to date unique. Due to its good performance a SVM was chosen as classifier. For training and assessment, datasets were collected and preprocessed. A data set for structurally important sites was manually compiled from non-enzymatic proteins as there was no other data available. The greatest challenge was to extract and select appropriate features to identify and simultaneously distinguish the three different groups of highly conserved residues (i.e. catalytic, ligand-binding, and structurally important residues). Finally, the classifier was trained and assessed.

The second subtask was the improvement of H2r, an algorithm that predicts

correlated residue positions. These are not highly conserved but still important residue positions. To increase the specificity of H2r, a method to incorporate the similarity of amino acids was developed. Therefore, the von Neumann entropy was utilized as it can serve as a framework for such similarities. However, no experimentally validated data set for correlation analysis exists and well studied proteins are still an exception. Because of that it is neither possible to construct reliable models for similarities of couplings directly nor by means of machine learning methods. Even unsupervised machine learning methods cannot be applied because the signal is too weak. To overcome this drawback a model for not correlated residues was deduced from a very large data set. As correlated mutations are rare, a coupled pair of residues can be detected by the deviation from this model. For evaluation purposes a test bed for statistical analysis was created.

2 Summary and discussion

In this chapter the results of the developed algorithms are summarized and discussed. The first part covers the identification of highly conserved, structurally and functionally important residues. Therefore, three classifiers were developed that are based on (1) sequence data, (2) structural data, and (3) both sequence and structural data. The second part covers the improvement of H2r, an algorithm for finding correlated mutations in MSAs.

2.1 Classification of highly conserved residue positions

The program CLIPS was developed in different versions in regard to the type of features used for training and testing the underlying multi-class SVM. CLIPS-1D is solely based on features extracted from sequence data. CLIPS-3D is based exclusively on features from protein structures and is thus applicable in cases where not enough sequences are available to generate a high-quality MSA. Finally, CLIPS-4D combines the best structure- and sequence-based features in one classifier. All versions were trained and assessed on three non-redundant and non-overlapping datasets. The set CAT_sites consists of 840 catalytic sites from 264 enzymes. Residues are defined as catalytic if annotated as such in the manually curated part of the Catalytic Site Atlas [114]. Of those enzymes, 216 also

have annotated binding sites in the PDBsum database [79], which constitute the 4466 ligand-binding sites of the LIG_sites dataset. As there was no representative set of structurally important sites available, conserved residues in the core of non-enzymatic proteins were utilized as these most likely play a crucial role in determining the structure of those proteins [133]. Thus, for the STRUC_sites dataset, 3703 residues that are more conserved than the average from 136 non-enzymatic proteins were merged. The complement of STRUC_sites constitutes the set of unimportant residues NOANN_sites. All structural data were acquired from the PDB database and the corresponding MSAs were taken from the HSSP database [126].

2.1.1 CLIPS-1D: A solely sequence-based classifier

Sequence-based features were extracted and selected from STRUC_sites, LIG_-sites, and CAT_sites by means of three two-class SVMs. Each SVM distinguished for one of these data sets if a residue was of the respective importance or not. The resulting performance values suggested to use the entropy-based normalized Jensen-Shannon divergence as a measure for conservation (Formula (4), Publication A) and an abundance value for scoring the occurrence of residues at crucial sites (Formula (6), Publication A). Furthermore, the conservation and abundance of a residue's sequence neighborhood was assessed by two weighted scores (Formula (5) and (7), Publication A). All abundance scores compare a residue position with the composition of all residue positions annotated as important. All features improved performance, but the conservation of a considered residue position contributed most to the classification of CAT_sites, LIG_sites, and STRUC_sites. Furthermore, the abundance of a residue position had a large influence on the classification of CAT_sites. Based on the above features the multi-class SVM CLIPS-1D was trained on all four data sets. The multi-class

SVM returned for each sample four class probabilities p_{class}. Classification was achieved based on the largest of the four p_{class}-values determined by the SVM. As tests showed that the abundance of STRUC_sites did not contribute to the performance of CLIPS-1D, the final feature set consists of the following seven features: (1) the Jensen-Shannon divergence of a residue position and (2) its neighborhood, the abundance of a residue position in regard to (3) CAT_sites, (4) LIG_sites, and (5) STRUC_sites, the abundance of a site's neighborhood in regard to (6) CAT_sites and (7) LIG_sites. Finally, residue specific p-values for funtionally and structurally important residue positions were deduced from the p_{class}-values of the residue in NOANN_sites to assess the reliability of the predictions.

2.1.1.1 Statistical Evaluation

Classifiers that identify two classes of important residues have already been developed in the past. One of these is FRpred [42] which was selected for a comparison with CLIPS-1D because it had outperformed other sequence-based methods. It assigns scores of 0-9 for each class to each residue; the higher the score, the more probable a functional role is. A classification of CAT_sites and LIG_sites with FRpred resulted in MCC-values (Formula (10), Publication A) of 0.250 and 0.197 when considering predictions scored 9 as positive cases (Table 2.1). For predictions scored $\geq$ 8, MCC-values of 0.231 and 0.219 were achieved. In contrast to CLIPS-1D's fine-grained separation of important residues, FRpred distinguishes only catalytic and ligand-binding sites. Thus, the percentage of false positives from the other two groups is given as a measure for STRUC_sites. FRpred predicted 22% (score 9) and 41% (score 8) of the STRUC_sites as catalytic sites or ligand-binding sites. The MCC-values for the binary SVMs were 0.324, 0.213, and 0.782, respectively. For CLIPS-1D, the performance in classifying CAT_-

	CAT_sites	LIG_sites	STRUC_sites
2C-SVM	0.324	0.213	0.782
CLIPS-1D	0.337	0.117	0.666
FRpred, score ≥ 8	0.231	0.219	41%
FRpred, score $= 9$	0.250	0.197	22%

Table 2.1: Performance of CLIPS-1D, the binary SVMs, and FRpred
MCC-values are given for the three binary SVMs, the multi-class SVM CLIPS-1D and FRpred on the three data sets CAT_sites, LIG_sites, and STRUC_sites. For FRpred, residues that scored at least 8 or 9 were evaluated. As FRpred does not predict structurally important sites, the percentage of false positives from the other two groups is given. Taken from Publication A, Table 1.

sites increased with respect to the binary SVM to 0.337. For STRUC_sites, the MCC-value decreased from 0.78 to 0.67 as LIG_sites and STRUC_sites share similar values of conservation and have an overlapping abundance. Because the binary SVMs were trained on only one data set each, they did not suffer from these similarities. The classification of LIG_sites was worst. The MCC-value droped from 0.21 to 0.12 due to that similarity.

Generally, performance differed to a large degree for different residue types and different classes. Class-specific MCC-values for each residue type are listed in Table 2.2. For CAT_sites, the catalytic active residues R, D, C, H, K, and S were predicted with high MCC-values. This was partially due to the larger data basis for these residues and their unique characteristics as being polar or charged. Most other MCC-values were close to zero or not determinable (P, V) due to empty sets. Catalytic residues C and T were regularly misclassified as structurally important. Both residues are not exceedingly overrepresented at catalytic sites and share similar abundance values with STRUC_sites. The performance values for

Residue	CAT_sites	LIG_sites	STRUC_sites
A	-0.002	0.164	0.774
C	0.404	0.162	0.676
D	0.302	0.016	0.315
E	0.345	0.052	0.348
F	0.058	0.041	0.771
G	0.024	0.262	0.591
H	0.424	-0.063	0.086
I	-0.001	0.135	0.701
K	0.452	0.031	0.337
L	-0.001	0.056	0.815
M	-0.002	0.127	0.666
N	0.071	0.139	0.561
P	-	0.139	0.683
Q	0.098	0.111	0.678
R	0.287	0.04	0.319
S	0.307	0.156	0.595
T	0.055	0.174	0.682
V	-	0.119	0.761
W	-0.008	0.007	0.689
Y	0.097	0.046	0.741

Table 2.2: Residue and class-specific performance values of CLIPS-1D MCC-values for each residue type and each class are given. No MCC-values could be determined for catalytic residues P and V due to missing cases. Taken from Publication A, Table 3.

LIG_sites were overall lower as conservation and abundance values did not stand out. For STRUC_sites, the mean MCC-value for the hydrophobic residues A, I, L, M, F, P, W, and V were 0.733, whereas the mean value of the hydrophilic ones was 0.494. Hydrophobic residues tend to be buried in the core of proteins and are thus often crucial for stability. But not all STRUC_sites are hydrophobic as the hydrophilic residues C, G, and T are overrepresented in STRUC_sites. Furthermore, residues important for secondary structure like L, F, and P tend to be overrepresented as well. On the other hand hydrophobic residues like A, I, M, and V are underrepresented in STRUC_sites.

2.1.1.2 A Case study that illustrates classification performance

To investigate the performance of CLIPS-1D more closely, the well-studied enzyme indole-3-glycerol phosphate synthase (TrpC) from *Sulfolobus solfataricus*

was analyzed. For TrpC, not only catalytic and binding sites are annotated in the PDB-sum database [79], but also many studies have been conducted about its folding kinetics [50] and its structure [57] [130]. The analysis was based on the HSSP-MSA related to PDB-ID 1A53. Table 2.3 lists all predictions for true functional sites. The catalytic residues E51, K53, K110, E159, and S211 were cor-

Residue	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOA}	**p-value**	**Classification**	
						CS	**LBS**
E51	0.806	0.075	0.114	0.005	0.02	CAT	
K53	0.835	0.065	0.088	0.012	0.004	CAT	
N89	0.006	0.231	0.001	0.762			NOA
K110	0.866	0.078	0.046	0.011	0.002	CAT	
F112	0.001	0.202	0.007	0.79			NOA
L131	0.001	0.071	0.92	0.008	0.006		STRUC
E159	0.899	0.048	0.05	0.003	0.005	CAT	
N180	0.098	0.77	0.116	0.016	0.016	LIG	
E210	0.866	0.059	0.068	0.007	0.008		CAT
S211	0.738	0.168	0.087	0.007	0.005	CAT	
I212	0.001	0.655	0.104	0.24	0.065		NOA
L231	0.003	0.224	0.762	0.011	0.025		STRUC
I232	0.006	0.835	0.059	0.099	0.017		LIG
S233	0.449	0.363	0.098	0.09			NOA
S234	0.133	0.289	0.006	0.572			NOA

Table 2.3: Classification of catalytic and binding sites for TrpC
Predictions for true catalytic and binding sites are given. Besides residue type and position, the four class probabilities p_{class} and p-values are shown. The last two columns give the predicted classes (CAT, LIG, STRUC, and NOA) for true catalytic and binding sites, respectively.

rectly identified. Only N180 was falsely predicted as LIG_site. Sites in contact with the ligand were classified as NOANN_sites (N89, F112, I212, S233, S234), CAT_sites (E210), and STRUC_sites (L131, L231). Only I232 was correctly

classified as LIG_site. Additionally classified as LIG_sites were K55, I179, and S181, which are all neighbors of catalytic sites. Furthermore, 20 residues were predicted as structurally important. Figure 2.1 shows that all of these are buried in the core of the protein. Nine of these 20 residue positions (and the three

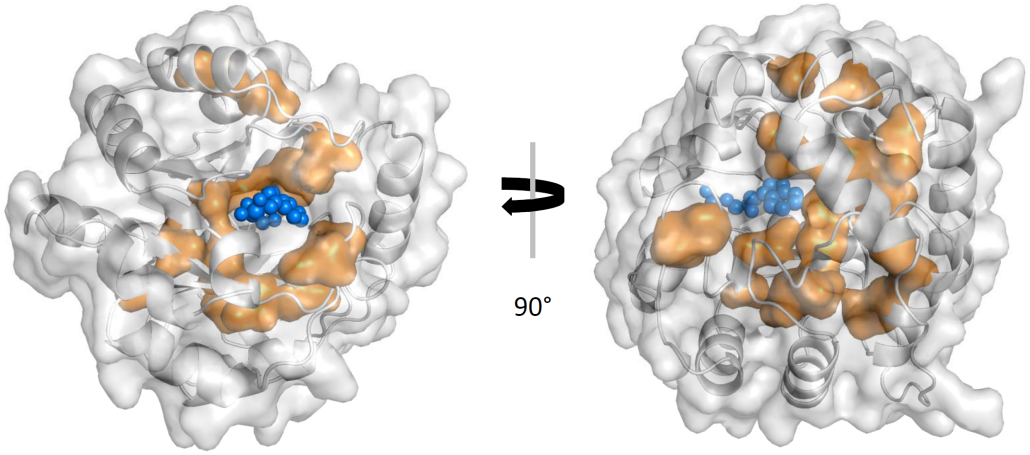


Figure 2.1: Location of predicted STRUC_sites in TrpC
The surface of TrpC (grey) and its substrate IGP (blue) are shown. All predicted STRUC_sites are buried (orange) in the core of the enzyme. The entry with PDB-ID 1A53 was used and visualized with PyMOL [132]. Taken from Publication A, Figure 2.

false-positive LIG_sites) are known to be structurally important in homologous proteins [8]. Eight of the 20 belong to fragments, which are strongly protected against deuterium exchange [49]. This indicates a crucial role of these residues in the folding mechanism. Furthermore, a molecular dynamics study [94] and a comparison of enzyme variants [130] suggest that two more STRUC_sites interact with the substrate. In summary, only four of the 20 STRUC_sites have no known structural function.

2.1.2 CLIPS-3D: A solely structure-based classifier

To investigate the performance of structure-based features, CLIPS-3D was developed. Furthermore, there are some cases where not enough sequences exist to generate a high-quality MSA (e.g. for proteins exclusively occurring in mammals). In that case CLIPS-1D cannot be applied as it is solely based on features extracted from sequence data. However, if a protein structure is available, CLIPS-3D can be used to analyze these cases. CLIPS-3D uses eight structure-based features: The amino acid composition in regard to (1) CAT_sites, (2) LIG_sites, and (3) STRUC_sites and the composition of the 3D neighborhoods (4-6). Furthermore, the (7) rSASA (Formula (12), Publication B) and (8) the location of residues in pockets (Formula (13), Publication B) is utilized. Pocket detection is based on the cavity detection algorithm fpocket [81].

In comparison to CLIPS-1D, the classification of functionally important residues was improved (Table 2.4). In fact the MCC-value for CAT_sites decreased slightly from 0,337 to 0,307 but the MCC-value for LIG_sites increased drastically from 0.117 to 0.221. This is due to the fact that functional residues have to be close to the surface to be able to interact with the ligands. As CLIPS-1D only uses sequence data, it is ignorant about that property and is thus outperformed by CLIPS-3D. The MCC-value for STRUC_sites, however, dropped by 38% from 0.666 to 0.426, because CLIPS-3D lacks crucial information about conservation.

2.1.3 CLIPS-4D: A sequence- and structure-based classifier

It is to be expected that the exploitation of information from sequence and structure combines the advantages of both approaches. Therefore, CLIPS-4D which uses the same sequence-based information as CLIPS-1D and, additionally, structural information was developed. The existing pool of CLIPS-1D's seven sequence-based features was extended by two of CLIPS-3D's structure-based features consisting of the rSASA and the pocket score as those turned out to be the most informative. Other structure-based features did not improve the performance of CLIPS-4D.

2.1.3.1 Statistical analysis

For a large scale evaluation CLIPS-1D, CLIPS-3D, CLIPS-4D, and ConSurf [4] were considered. ConSurf is a well-established tool for calculating the evolutionary conservation of residue positions in proteins using an empirical Bayesian inference or (optionally) a maximum likelihood method. Amongst other values, it outputs a conservation score in the range of 0-9. In this analysis, a residue with a high conservation score was assigned a structural role if it was buried or a functional role if it was exposed to the solvent. As ConSurf cannot distinguish between catalytic and ligand-binding residues, the union of CAT_sites and LIG_sites was used for this part of the evaluation. Residues with score 9 were considered as positives as Consurf performed best with this threshold. The resulting MCC-values can be found in Table 2.4 which also lists the MCC-values of CLIPS-1D, CLIPS-3D, CLIPS-4D. ConSurf performed marginally better than CLIPS-1D and CLIPS-3D for functionally relevant sites due to the access to structural data and sequence information. The MCC-value of 0.46 for structurally important sites, however, was inferior to CLIPS-1D. In comparison with these algorithms, CLIPS-4D performed best for both catalytic and ligand-binding

	CAT_sites	LIG_sites	STRUC_sites
CLIPS-1D	0.34	0.12	0.67
CLIPS-3D	0.31	0.22	0.43
CLIPS-4D	0.43	0.27	0.68
Consurf	0.30		0.46

Table 2.4: Performance of CLIPS-1D, CLIPS-3D, CLIPS-4D, and Consurf

MCC-values are given for CLIPS-1D, CLIPS-3D, CLIPS-4D, and Consurf on the three data sets CAT_sites, LIG_sites, and STRUC_sites. As ConSurf does not distinguish between catalytic and ligand-binding sites, the sets CAT_sites and LIG_sites were merged before classification.

sites with MCCs of 0.43 and 0.27, respectively. Additionally, it performed slightly better than CLIPS-1D for structurally important sites. This improvement was small because the MCC value of CLIPS-1D was already 0.67. Furthermore, the additional structural data available to CLIPS-4D is not specifically useful for the recognition of structurally relevant sites, as these only contain binary information about a residue's position in a surface pocket. It contains no information that allows to distinguish between surface and core residues in general.

2.1.3.2 Case studies that illustrate classification performance

For a more detailed analysis, targets were chosen from the ligand-binding site prediction category of the Critical Assessment of protein Structure Prediction (CASP) experiments. Corresponding MSAs were fetched from the HSSP database. The best performing methods of CASP9 [129] and CASP10 [22] were based on homology transfer which maps the annotations of a similar, already annotated enzyme to the enzyme under study. Therefore, methods like firestar [89] reached

MCC-values of up to 0.7 in CASP10. However, if the ligand is large and flexible, it is particularly difficult for these methods to predict the full binding site as a mapping from a known enzyme might be incomplete or even false. Therefore, the six most difficult CASP9 and CASP10 targets binding biological ligands were chosen for a detailed comparison of CLIPS-4D and firestar. As shown in Table 2.5 the MCC-value of CLIPS-4D was higher than firestar's for targets T0526 and T0604. Additionally, a union of the predictions generated by firestar and CLIPS-4D resulted in a higher sensitivity at the cost of a moderate loss in specificity for targets T0615 and T0721 (Publication B, Table 3 and Supplement). On targets T0584 and T0632 firestar performed better than CLIPS-4D.

Among these targets, the performance of CLIPS-4D was worst for T0584, a

	T0526 3NRE	**T0584 3NF2**	**T0604 3NLC**	**T0615 3NQW**	**T0632 3NWZ**	**T0721 4FK1**
firestar	0.49	0.69	0.45	0.52	0.49	0.73
CLIPS-4D	0.61	0.19	0.54	0.34	0.24	0.45
Union	0.58	0.44	0.54	0.50	0.40	0.68

Table 2.5: Classification performance of firestar, CLIPS-4D and an ensemble classifier on ligand-binding sites of six CASP targets

The row with label 'Union' gives MCC-values resulting from merging positive predictions from *firestar* and CLIPS-4D. The first two lines give the target number and the PDB-ID, respectively.

polyprenyl transferase. It generates the product from isopentenyl diphosphate and dimethylallyl diphosphate by consecutive steps of elongation, cyclopropagation, rearrangement and cyclization reactions [153]. Consequently, the product grows into an elongation cavity until reaching residues protruding into the cavity [84] and thus determining the length of the product. Five residues shown to

be important for catalysis [84] were predicted exclusively by CLIPS-4D. Four of these residues are not directly involved in ligand-binding in the CASP control structure with PDB-ID 1RQI and are thus handled as false positive predictions. Furthermore, 18 CLIPS-4D predictions line the elongation cavity modeled previously [143] and three more contact the ligand after an active site rearrangement. Therefore, experimental evidence confirmed that at least some of these predictions were not false positives.
In summary, these findings suggest that CLIPS-4D can supplement homology transfer methods in difficult cases like active site rearrangements, flexible substrates or unknown poses of a ligand.

2.1.3.3 Comparison with CLIPS-1D

To underline the benefits of incorporating structural data into CLIPS, class- and residue-specific MCC-values are listed for CLIPS-1D and CLIPS-4D in Table 2.6. No values could be calculated for CAT_sites I, L, M, V, and P because of missing cases. On the one hand, the catalytically inert amino acids did not benefit from structural data as MCC-values were still close to zero. This might be due to a lack of data as on average only nine samples per amino acid were available. On the other hand the catalytically active residues were predicted by CLIPS-4D with MCC-values of up to 0.585, which is a large improvement. CLIPS-1D performed slightly (0.01) better on catalytic phenylalanines and glycines but was outperformed on all other residues. Thus, the prediction of these amino acids benefits from the incorporation of structural data. In the case of ligand-binding sites the improvement was even larger. CLIPS-4D predicted 19 residues with a mean improvement of 0.15, only the MCC-value for aspartic acid droped slightly. The improvement was due to the fact that binding sites tend to lie in surface cavities which CLIPS-1D is ignorant about. The recognition of structurally relevant

Residue	CAT_sites	LIG_sites	STRUC_sites
A	-0.001 (0.001)	0.331 (0.167)	0.790 (0.016)
C	0.445 (0.041)	0.241 (0.079)	0.732 (0.056)
D	0.402 (0.100)	0.009 (-0.007)	0.491 (0.176)
E	0.443 (0.098)	0.102 (0.052)	0.509 (0.161)
F	-0.001 (-0.059)	0.256 (0.216)	0.779 (0.008)
G	-0.001 (-0.025)	0.434 (0.172)	0.624 (0.033)
H	0.496 (0.072)	0.033 (0.096)	0.186 (0.100)
I	- (-)	0.364 (0.229)	0.707 (0.006)
K	0.585 (0.133)	0.132 (0.101)	0.372 (0.035)
L	- (-)	0.273 (0.217)	0.772 (-0.043)
M	- (-)	0.382 (0.255)	0.687 (0.021)
N	0.128 (0.057)	0.239 (0.100)	0.642 (0.081)
P	- (-)	0.187 (0.048)	0.670 (-0.013)
Q	0.186 (0.088)	0.277 (0.166)	0.691 (0.013)
R	0.331 (0.044)	0.114 (0.074)	0.471 (0.152)
S	0.399 (0.092)	0.283 (0.127)	0.643 (0.048)
T	0.059 (0.004)	0.337 (0.164)	0.715 (0.033)
V	- (-)	0.344 (0.225)	0.727 (-0.034)
W	-0.005 (0.018)	0.241 (0.234)	0.654 (-0.035)
Y	0.179 (0.082)	0.188 (0.146)	0.714 (-0.027)

Table 2.6: Residue and class-specific performance values of CLIPS-4D
MCC-values for each residue and each class are given. The change in regard to CLIPS-1D is given in brackets.

sites is generally sound in both methods. Nevertheless, CLIPS-4D performed slightly better in recognizing these buried and highly conserved amino acids. For 15 residues, the MCC-value increased by 0.06 on average, the decrease for the other five residues was 0.03. Taken together, the classification of ligand-binding sites benefited most from the structural data assessed by CLIPS-4D. The improvement for structurally and catalytically important sites was less pronounced but still significant.

2.2 Identification of correlated mutations

Unlike highly conserved residues, correlated residue positions cannot be identified by machine learning methods as the data basis is too sparse. Consequently, no model for correlated mutations exists as the substitution frequencies for pairs of amino acids are unknown. Under the assumption that correlated mutations are rare, the consideration of a very large amount of residue pairs describes the normal, uncorrelated case, in other words, a null model. Thus, a coupled pair of residues can be detected by the deviation from this null model. In analogy to matrices like P2PMAT [39] in inter residue contact prediction, the null model is a similarity matrix A for residue pair substitutions. The similarity values were deduced from a large, non-redundant set of protein structures and corresponding MSAs from the HSSP database. The matrix entries were calculated in analogy to the BLOSUM series [56] but the concept was adapted to pairs of residues. To integrate this null model in the previously developed algorithm H2r [97], the von Neumann entropy (Formula (5), Publication C) from quantum physics was adopted as it is a generalization of Shannon's entropy and can regard amino acid similarities in an appropriate manner. Thus, the mutual information ($U_{vNE}(k,l)$) can be deduced for two residue positions k and l (Formula (9), Publication C). Only significant values of $U_{vNE}(k,l)$ were further regarded by calculating a bootstrapping-based statistical measure for the strength of pairwise correlations. Bootstrapping is a resampling method for estimating the distribution function of a random variable by deducing an empirical distribution from just one sample [37]. Here, this sample is the given arrangement of residues in a pair of columns k and l. By shuffling the content column-wise, the entropy (conservation) of the two individual columns remains constant; however, the putative correlation between the two residue positions degrades. Thus, by comparing the $U_{vNE}(k,l)$-value deduced from the unaltered combination of residue pairs with a

distribution of $U_{vNE}(k^*, l^*)$-values resulting from many shuffling rounds, the correlation strength for this specific combination of residue pairs observed in columns k and l can be rated. Consequently, if the $U_{vNE}(k^*, l^*)$-values are similarly large or surpass the $U_{vNE}(k, l)$-value, the correlation is statistically not significant. On the other hand, if all $U_{vNE}(k^*, l^*)$-values are significantly lower, then this specific $U_{vNE}(k, l)$-value signals a pronounced dependency in the orchestration of the two residue positions, which indicates correlated mutations. Finally, the conn(k)-value for all residue positions k is determined by counting the number of significantly correlated pairs k is part of. To alleviate the comparison of different proteins, conn(k) was further transformed into z-scores conz(k). This way the importance of a single residue position k can be assessed and measured in a robust manner.

2.2.1 Statistical analysis

Based on the observation that many correlated mutations are in close proximity to functional sites [115] [75], a test bed for a large scale evaluation was created. A dataset of 200 non-redundant enzymes (PDB structures and corresponding HSSP MSAs) with known catalytic and binding sites served as a basis for the assessment. Residues in a proximity of 1 $\AA$ to a functional site were regarded as positives, all other residues (including the functional sites) as negatives. This is a crude approach as not all residues in the proximity of a functional site are covarying (and vice versa) but there is no generally accepted alternative for a large scale evaluation. Although H2rs does not use any information other than an MSA it yielded a specificity between 0.97 and 0.98, a precision between 0.18 and 0.19, and a balanced accuracy between 0.51 and 0.52 for a conz(k)-threshold of 2 and p-values between 10^{-2} and 10^{-4} as can be seen in Table 2.7. Smaller p-values and higher conz(k)-thresholds, however, yielded a precision of up to 0.3,

	p-value	z-score	Specificity	Precision	Accuracy
	10^{-2}	4.0	1.00	0.30	0.50
	10^{-2}	2.0	0.97	0.18	0.51
	10^{-3}	2.0	0.97	0.18	0.51
H2rs	10^{-4}	2.0	0.98	0.19	0.52
	10^{-5}	2.0	0.98	0.18	0.51
	10^{-10}	2.0	0.98	0.17	0.51
	10^{-11}	2.0	0.98	0.17	0.51
H2r			0.95	0.17	0.5

Table 2.7: Performance of H2rs

Specificity, precision and balanced accuracy for H2rs and H2r are shown. Residues in a proximity of less than 1 $\mathring{A}$ of at least one functional site in a dataset of 200 enzymes were regarded as positives. Negatives were the functional sites itself and residues with a distance of more than 1 $\mathring{A}$ to any functional site. For H2rs, the performance for different p-values and z-scores is shown.

a specificity of over 0.99, and a balanced accuracy of 0.5 at the cost of fewer predictions made by H2rs. Nevertheless, the introduction of p-values and z-scores allows for a flexible analysis, which can be adapted to the user's needs. Although the previously developed algorithm, H2r, predicted important residue positions with high specificity [31], H2rs performed better. Furthermore, H2rs attained slightly higher values in precision and balanced accuracy. This is due to a better consideration of the physicochemical properties of amino acids and the calculation of a significance threshold for each individual residue position pair which in combination leads to less false positive predictions.

2.2.2 Case studies that illustrate classification performance

To investigate the performance of H2rs more closely, a detailed analysis on three well studied enzymes of the tryptophan biosynthesis TrpA, TrpB, and TrpC and, additionally, on the dihydrofolate reductase (DHFR) and a hexokinase (HK) was done. The results were compared with the global contact prediction method PSICOV [67] (the best scoring $L/5$ predictions, where L is the sequence length) and H2r. MSAs for each enzyme were created by using DELTA-BLAST [15] with a *max target threshold* of 2000 and an *expect threshold* of 10^{-10} and aligned with Mafft [70] in linsi mode. To concentrate on the most promising candidates, p-value thresholds of 10^{-11} and a conz(k)-thresholds of 2.0 were chosen for all evaluations. Using less conservative thresholds would lead to more predicted residues but, presumably, to more false positives as well.

TrpA and TrpB constitue the tetrameric tryptophan synthase complex, which catalyzes the final reaction from indole-3-glycerole phosphate and serine to tryptophan. TrpA cleaves indoleglycerol-3-phosphate to glyceraldehyde-3-phosphate and indole, which is transported to TrpB through a hydrophobic tunnel. There it is condensed with serine to yield tryptophan. TrpA from *Salmonella typhimurium* consists of 268 residues of which only two surpassed the chosen thresholds. L100 had a conz(100)-value of 2.2 and L127 had a conz(127)-value of 2.0. Both residues are in close proximity to the substrate. Among others, both residue positions were part of the $L/5$ predictions of PSICOV (Table 2.8). H2r predicted only L100. For TrpB from *Salmonella typhimurium*, 13 of the 397 residues were predicted by H2rs as being important. T88, Q90, and V91 are in close proximity to the substrate-binding residue K87 [99]. C170 and F280 are at the end of the hydrophobic tunnel [125]. T190 and S308 are metal-binding sites [79]. Furthermore, experiments in [121] have shown that S308 and G268 are important for the coordination of ion binding. S297 and P257 are in close proximity to the bound sodium ion. M282

Protein	Residue	PSICOV	H2rs	H2r	Residue's role
stTrpA	L100	1	2.2	3.23	Near binding site
	L127	2	2		Near binding site
stTrpB	C62	0	2.2	7.31	ND
	T88	1	2.4		Near binding site
	Q90	0	2.4	7.46	Near binding site
	V91	0	2.1		Near binding site
	C170	4	4.5		End of substrate tunnel
	T190	6	2.2		Metal-binding site
	P257	0	2.2	6.69	Near metal ion
	G268	0	2.3		Coordination of ion binding
	F280	0	2.4	2.81	End of substrate tunnel
	M282	4	2.6		Near binding site
	S297	3	4.2		Near metal ion
	S308	0	2.4	8.54	Metal-binding site
	Q312	0	2.9		ND
ssTrpC	I48	3	2.4		ND
	I133	3	2.6	9.77	Catalytically important
	V134	2	2.3		Near active site
	I136	1	2.1		ND
	L142	1	2.7		Catalytically important
	A209	3	2.1		Near binding site
	S234	4	2.1	9.54	Phosphate-binding site
ecDHFR	A9	2	2.2		Near active site
	W30	0	2.3		Binding site
	K32	0	2.3		Binding site
	M92	0	3.4		Near active site
	G121	0	2.7	2.77	Near active site
	H149	0	2.1	4.38	Coupled motion
smHK	T69	1	2.8		Domain interface
	A215	2	2.6		End of domain 1
	C217	0	2.7	13.9	End of domain 1
	A218	0	2.3		End of domain 1
	C224	0	2.2		Start of domain 2
	V230	3	2.1		Near binding site
	V256	2	2.1		Domain interface
	K290	0	2.2		Near binding site
	T409	1	2.4		Near C224
	V412	0	2.0		Near binding site

Table 2.8: Overlapping predictions of H2rs with H2r and PSICOV on five case studies

All residues predicted by H2rs with conz(k)-values ≥ 2 on the five case studies are shown. For each residue, corresponding PSICOV and H2r predictions are noted. For PSICOV the occurrence of a residue in the top $L/5$ contact predictions is shown, for H2r mean conn(k)-values from 25 bootstrapping runs are noted. The last column lists the role of the residues that could be found in literature. "ND" indicates that no role for this residue could be found.

is in contact with F280 and S308. There was no information available for C62 and Q312. Of those 13 predictions, PSICOV identified five as well (Table 2.8). Consequently, H2rs identified six residue positions of importance that were not

recognized by PSICOV. This might be due to the fact that important residues are not necessarily in contact with each other. H2r predicted C62, Q90, P257, S308, and F280.

TrpC from *Sulfolobus solfataricus* catalyzes the ring closure of an N-alkylated anthranilate to a 3-alkyl indole derivative, which is the fourth step in the tryptophan biosynthesis. H2rs predicted seven of the 248 residues as important. The highest conz(k)-values were assigned to L142 and I133. After individually mutating those residues to alanine, the activity of TrpC dropped 30-fold in each case [30]. I133 and V134 are in close proximity to the substrate-binding site L132. A209 lies next to the substrate-binding site E210 and the catalytic residue S211 [79]. S234 is known to be a phosphate-binding site. No information was found for I48 and I136. All those residue positions were detected by PSICOV as well, even the potential false positives I48 and I136. H2r, however, only detected two of those residue positions, the binding site S234 and the catalytically important residue I133.

DHFR from *Escherichia coli* catalyzes the reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate utilizing the nucleotide cofactor 5,10-methylenetetrahydrofolate reductase. It can be found in most organisms and plays a critical role for cell proliferation and cell growth. H2rs predicted seven of 159 residue positions as important. A9 and M92 are in close proximity to the binding site A7 and the catalytic site I94, respectively [79]. W30 and K32 are in contact with the substrate. H149 is known to play a significant role in the network of coupled motions that induce configurations allowing for the hydride transfer [157]. Finally, mutations of G121, which lies in proximity of NADP, are known to reduce the hydride transfer rate [146]. Among those only A9 was predicted by PSICOV. H2r identified G121 and H149. The low overlap shows a significant improvement over H2r which only predicted two residue positions for DHFR at all, because it is known that there

are many correlated motions which most likely involve correlated mutations in DHFR. PSICOV, however, predicted 54 residue positions which is about one third of all residues. Nevertheless, the overlap to H2rs' predictions was very small.
HK from *Schistosoma mansoni* (PDB-ID 1BDG) is the first enzyme in the glycolytic pathway and catalyzes the transfer of a phosphoryl group to alpha-6-glucose. The 3D crystal structure contains SO_4 anions in the catalytic cleft [77]. It consists of two domains, a hexokinase type-1 (residues 18 – 218) and a hexokinase type-2 domain (residues 221 – 457). H2rs identified 10 residues. A215, C217, and A218 are located at the very end of domain 1, whereas C224 occurs at the very beginning of domain 2. Furthermore, these four residues are flanking a β-turn [79]. K290 and V230 are neighbors of the binding sites Q291 and I229, respectively. V412 is a neighbor of the SO_4-binding sites G414 and S415. T409 is close to C224 (see above). The role of only two residues (T69 and V256) is unknown. However, both residues are located close to the domain interface ($\leq$ 5.2 $\AA$). H2r predicted C217 and additionally D376, whose function is not known. Five of H2rs' predicted residue positions were also predicted by PSICOV. The results obtained for HK emphasize that not all correlated mutations are caused by functional constraints: 4 of 10 residues with high conz(k)-values are located at the domain interface and two of them (C217, C224) belong to a disulfide bond that stiffens the orientation of the two domains in some of the homologous proteins. This positions were occupied in only 43% of the sequences by cysteines. The orchestration of these two residue positions fits well to the idea of mutual dependencies and pairwise correlations.
Generally, this detailed analysis of five enzymes as well as the assessment of the *in silico* testbed signals the improved specificity gained by introducing the von Neumann entropy and by integrating a more sensitive statistical approach that adapts to the composition of each pair of MSA columns. However, this improve-

ment suffers by a much longer execution time as the calculation of eigenvalues is computationally very demanding. The predictions of H2rs and PSICOV overlap only marginally, which can be explained by the scope of the methods. Global methods aim at identifying contacting residue pairs which are not all necessarily located near functional sites and therefore eliminate transitive correlations. Local methods do not eliminate transitive correlations and are thus able to predict correlation far apart in the structure. Therefore, global methods are not able to substitute local methods.

3 Bibliography

[1] D Altschuh, AM Lesk, AC Bloomer, and A Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193:693–707, 1987.

[2] SF Altschul, TL Madden, AA Schäffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[3] N Amin, AD Liu, S Ramer, W Aehle, D Meijer, M Metin, S Wong, P Gualfetti, and V Schellenberger. Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Engineering Design and Selection*, 17:787–793, 2004.

[4] H Ashkenazy, E Erez, E Martz, T Pupko, and N Ben-Tal. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, 38(suppl 2):W529–W533, 2010.

[5] H Ashkenazy, R Unger, and Y Kliger. Optimal data collection for correlated mutation analysis. *Proteins: Structure, Function, and Bioinformatics*, 74:545–555, 2009.

[6] WR Atchley, KR Wollenberg, WM Fitch, W Terhalle, and AW Dress. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular Biology and Evolution*, 17:164–178, 2000.

[7] D Baccanari, A Phillips, S Smith, D Sinski, and J Burchall. Purification and properties of Escherichia coli dihydrofolate reductase. *Biochemistry*, 14:5267–5273, 1975.

[8] B Bagautdinov and K Yutani. Structure of indole-3-glycerol phosphate synthase from Thermus thermophilus HB8: implications for thermal stability. *Acta Crystallographica Section D: Biological Crystallography*, 67:1054–1064, 2011.

[9] A Bairoch and R Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45–48, 2000.

[10] E Balog, D Perahia, JC Smith, and F Merzel. Vibrational softening of a protein on ligand binding. *The Journal of Physical Chemistry B*, 115:6811–6817, 2011.

[11] GJ Bartlett, CT Porter, N Borkakoti, and JM Thornton. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, 324:105–121, 2002.

[12] C Berezin, F Glaser, J Rosenberg, I Paz, T Pupko, P Fariselli, R Casadio, and N Ben-Tal. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, 20:1322–1324, 2004.

[13] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[14] FC Bernstein, TF Koetzle, GJ Williams, EF Meyer, MD Brice, JR Rodgers, O Kennard, T Shimanouchi, and M Tasumi. The Protein Data Bank. A computer-based archival file for macromolecular structures. *European Journal of Biochemistry*, 80:319–324, 1977.

[15] GM Boratyn, AA Schäffer, R Agarwala, SF Altschul, DJ Lipman, and TL Madden. Domain enhanced lookup time accelerated BLAST. *Biology Direct*, 7:12, 2012.

[16] L Burger and E van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology*, 6:e1000633, 2010.

[17] CM Buslje, E Teppa, T Di Doménico, JM Delfino, and M Nielsen. Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Computational Biology*, 6:e1000978, 2010.

[18] G Caetano-Anollés, HS Kim, and JE Mittenthal. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences*, 104:9358–9363, 2007.

[19] JA Capra, RA Laskowski, JM Thornton, M Singh, and TA Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology*, 5:e1000585, 2009.

[20] JA Capra and M Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23:1875–1882, 2007.

[21] G Casari, C Sander, and A Valencia. A method to predict functional residues in proteins. *Nature Structural Biology*, 2:171, 1995.

[22] TG Cassarino, L Bordoli, and T Schwede. Assessment of ligand binding site predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82:154–163, 2013.

[23] A Ceroni, A Passerini, A Vullo, and P Frasconi. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Research*, 34:W177–W181, 2006.

[24] CC Chang and CJ Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology.*, 2:27, 2011.

[25] PY Chou and GD Fasman. Empirical predictions of protein conformation. *Annual Review of Biochemistry*, 47(1):251–276, 1978.

[26] NJ Davidson and X Wang. *2010 Ninth International Conference on Machine Learning and Applications.*, chapter Non-Alignment Features Based Enzyme/Non-Enzyme Classification Using an Ensemble Method., pages 546–551. Institute of Electrical and Electronics Engineers, 2010.

[27] J Davis and M Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

[28] D de Juan, F Pazos, and A Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14:249–261, 2013.

[29] JP Dekker, A Fodor, RW Aldrich, and G Yellen. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, 20:1565–1572, 2004.

[30] S Dietrich. *Mutationsanalyse und kinetische Untersuchungen zum Reaktionsmechanismus der Indolglycerinphosphat-Synthase aus Sulfolobus solfataricus.* PhD thesis, Universität Regensburg, 2011.

[31] S Dietrich, N Borst, S Schlee, D Schneider, JO Janda, R Sterner, and R Merkl. Experimental assessment of the importance of amino acid positions identified by an entropy-based correlation analysis of multiple-sequence alignments. *Biochemistry*, 51:5633–5641, 2012.

[32] Y Dou, X Geng, H Gao, J Yang, X Zheng, and J Wang. Sequence conservation in the prediction of catalytic sites. *The Protein Journal*, 30:229–239, 2011.

[33] RO Duda, PE Hart, and DG Stork. *Pattern Classification.* Wiley-Interscience, 2000.

[34] OJ Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56:52–64, 1961.

[35] SD Dunn, LM Wahl, and GB Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24:333–340, 2008.

[36] S Dutta, K Burkhardt, J Young, GJ Swaminathan, T Matsuura, K Henrick, H Nakamura, and HM Berman. Data deposition and annotation at the worldwide protein data bank. *Molecular Biotechnology*, 42:1–13, 2009.

[37] B Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[38] S Erdin, RM Ward, E Venner, and O Lichtarge. Evolutionary trace annotation of protein function in the structural proteome. *Journal of Molecular Biology*, 396:1451–1473, 2010.

[39] E Eyal, M Frenkel-Morgenstern, V Sobolev, and S Pietrokovski. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 67:142–153, 2007.

[40] I Ezkurdia, L Bartoli, P Fariselli, R Casadio, A Valencia, and ML Tress. Progress and challenges in predicting protein–protein interaction sites. *Briefings in Bioinformatics*, 10:233–246, 2009.

[41] RD Finn, J Mistry, B Schuster-Böckler, S Griffiths-Jones, V Hollich, T Lassmann, S Moxon, M Marshall, A Khanna, Ri Durbin, RS Eddy, EL Sonnhammer, and A Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34:D247–D251, 2006.

[42] JD Fischer, CE Mayer, and J Söding. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24:613–620, 2008.

[43] I Friedberg, M Jambon, and A Godzik. New avenues in protein function prediction. *Protein Science*, 15:1527–1529, 2006.

[44] H Gao, Y Dou, J Yang, and J Wang. New methods to measure residues coevolution in proteins. *BMC Bioinformatics*, 12:206, 2011.

[45] JA Gerlt, KN Allen, SC Almo, RN Armstrong, PC Babbitt, JE Cronan, D Dunaway-Mariano, HJ Imker, MP Jacobson, W Minor, CD Poulter, FM Raushel, A Sali, BK Shoichet, and JV Sweedler. The enzyme function initiative. *Biochemistry*, 50:9950–9962, 2011.

[46] M Gerstein and FM Richards. A Manuscript for inclusion in: The International Tables for Crystallography.

[47] U Göbel, C Sander, R Schneider, and A Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18:309–317, 1994.

[48] K Goyal, D Mohanty, and SC Mande. PAR-3D: a server to predict protein active site residues. *Nucleic Acids Research*, 35:W503–W505, 2007.

[49] Z Gu, MK Rao, WR Forsyth, JM Finke, and CR Matthews. Structural analysis of kinetic folding intermediates for a TIM barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and gō model simulation. *Journal of Molecular Biology*, 374:528–546, 2007.

[50] Z Gu, JA Zitzewitz, and CR Matthews. Mapping the Structure of Folding Cores in TIM Barrel Proteins by Hydrogen Exchange Mass Spectrometry: The Roles of Motif and Sequence for the indole-3-glycerol Phosphate Synthase from Sulfolobus solfataricus. *Journal of Molecular Biology*, 368:582–594, 2007.

[51] M Gültas, M Haubrock, N Tueysuez, and S Waack. Coupled mutation finder: a new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *BMC Bioinformatics*, 13:225, 2012.

[52] EJ Gumbel. *Statistics of Extremes*. Columbia University Press, 1958.

[53] R Gutman, C Berezin, R Wollman, Y Rosenberg, and N Ben-Tal. QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Research*, 33:W255–W261, 2005.

[54] N Halabi, O Rivoire, S Leibler, and R Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138:774–786, 2009.

[55] I Halperin, H Wolfson, and R Nussinov. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins: Structure, Function, and Bioinformatics*, 63:832–845, 2006.

[56] S Henikoff and JG Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89:10915–10919, 1992.

[57] M Hennig, B Darimont, R Sterner, K Kirschner, and JN Jansonius. 2.0 Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile Sulfolobus solfataricus: possible determinants of protein stability. *Structure*, 3:1295–1306, 1995.

[58] A Hildebrandt, AK Dehof, A Rurainski, A Bertsch, M Schumann, NC Toussaint, A Moll, D Stöckel, S Nickels, SC Mueller, HP Lenhof, and O Kohlbacher. BALL-biochemical algorithms library 1.3. *BMC Bioinformatics*, 11:531, 2010.

[59] DJ Hosfield, Y Zhang, DR Dougan, A Broun, LW Tari, RV Swanson, and J Finn. Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis. *Journal of Biological Chemistry*, 279:8526–8529, 2004.

[60] FL Hsieh, TH Chang, TP Ko, and AH Wang. Enhanced specificity of mint geranyl pyrophosphate synthase by modifying the R-loop interactions. *Journal of Molecular Biology*, 404:859–873, 2010.

[61] FL Hsieh, TH Chang, TP Ko, and AH Wang. Structure and mechanism of an Arabidopsis medium/long-chain-length prenyl pyrophosphate synthase. *Plant Physiology*, 155:1079–1090, 2011.

[62] JY Huang and DL Brutlag. The EMOTIF database. *Nucleic Acids Research*, 29:202–204, 2001.

[63] JO Janda, M Busch, F Kück, M Porfenenko, and R Merkl. 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, 2012.

[64] JO Janda, A Meier, and R Merkl. CLIPS-4D: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3D data. *Bioinformatics*, 29:3029–3035, 2013.

[65] F Johansson and H Toh. A comparative study of conservation and variation scores. *BMC Bioinformatics*, 11:388, 2010.

[66] F Johansson and H Toh. Relative von Neumann entropy for evaluating amino acid conservation. *Journal of Bioinformatics and Computational Biology*, 8:809–823, 2010.

[67] DT Jones, DW Buchan, D Cozzetto, and M Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28:184–190, 2012.

[68] OV Kalinina, MS Gelfand, and RB Russell. Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinformatics*, 10:174, 2009.

[69] I Kass and A Horovitz. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins: Structure, Function, and Bioinformatics*, 48:611–617, 2002.

[70] K Katoh, K Kuma, H Toh, and T Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33:511–518, 2005.

[71] K Katoh and DM Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780, 2013.

[72] G Kiss, N Celebi-Ölcüm, R Moretti, D Baker, and KN Houk. Computational enzyme design. *Angewandte Chemie International Edition*, 52:5700–5725, 2013.

[73] T Knöchel, A Pappenberger, JN Jansonius, and K Kirschner. The Crystal Structure of indoleglycerol-phosphate Synthase from Thermotoga maritima. Kinetic stabilization by salt bridges. *Journal of Biological Chemistry*, 277:8626–8634, 2002.

[74] M Krone, K Bidmon, and T Ertl. Interactive visualization of molecular surface dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 15:1391–1398, 2009.

[75] RK Kuipers, HJ Joosten, E Verwiel, S Paans, J Akerboom, J van der Oost, NG Leferink, WJ van Berkel, G Vriend, and PJ Schaap. Correlated

mutation analyses on super-family alignments reveal functionally important residues. *Proteins: Structure, Function, and Bioinformatics*, 76:608–616, 2009.

[76] V Kulik, E Hartmann, M Weyand, M Frey, A Gierl, D Niks, MF Dunn, and I Schlichting. On the structural basis of the catalytic mechanism and the regulation of the alpha subunit of tryptophan synthase from Salmonella typhimurium and BX1 from maize, two evolutionarily related enzymes. *Journal of Molecular Biology*, 352:608–620, 2005.

[77] PR Kuser, S Krauchenco, OA Antunes, and I Polikarpov. The high resolution crystal structure of yeast hexokinase PII with the correct primary sequence provides new insights into its mechanism of action. *Journal of Biological Chemistry*, 275:20814–20821, 2000.

[78] SM Larson, AA Di Nardo, and AR Davidson. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *Journal of Molecular Biology*, 303:433–446, 2000.

[79] RA Laskowski, VV Chistyakov, and JM Thornton. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Research*, 33:D266–D268, 2005.

[80] RA Laskowski, JD Watson, and JM Thornton. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research*, 33:W89–W93, 2005.

[81] V Le Guilloux, P Schmidtke, and P Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168, 2009.

[82] B Lee and FM Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, 55:379–400, 1971.

[83] M Lehmann, C Loch, A Middendorf, D Studer, SF Lassen, L Pasamontes, AP van Loon, and M Wyss. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Engineering*, 15:403–411, 2002.

[84] PH Liang, TP Ko, and AH Wang. Structure, mechanism and function of prenyltransferases. *European Journal of Biochemistry*, 269:3339–3354, 2002.

[85] O Lichtarge, HR Bourne, and FE Cohen. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257:342–358, 1996.

[86] O Lichtarge, H. Yao, DM Kristensen, S Madabushi, and I Mihalek. Accurate and scalable identification of functional sites by evolutionary tracing. *Journal of Structural and Functional Genomics*, 4:159–166, 2003.

[87] CD Lima, MG Klein, and WA Hendrickson. Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science*, 278:286–290, 1997.

[88] SW Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286:295–299, 1999.

[89] G Lopez, P Maietta, JM Rodriguez, A Valencia, and ML Tress. firestar-advances in the prediction of functionally important residues. *Nucleic Acids Research*, 39:W235–W241, 2011.

[90] DS Marks, LJ Colwell, R Sheridan, TA Hopf, A Pagnani, R Zecchina, and C Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6:e28766, 2011.

[91] DS Marks, TA Hopf, and C Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30:1072–1080, 2012.

[92] LC Martin, GB Gloor, SD Dunn, and LM Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21:4116–4124, 2005.

[93] BW Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta - Protein Structure*, 405:442–451, 1975.

[94] D Mazumder-Shivakumar and TC Bruice. Molecular dynamics studies of ground state and intermediate of the hyperthermophilic indole-3-glycerol phosphate synthase. *Proceedings of the National Academy of Sciences of the United States of America*, 101:14379–14384, 2004.

[95] A McPherson. *Introduction to Macromolecular Crystallography*. Wiley-Liss, 2002.

[96] R Merkl and S Waack. *Bioinformatik Interaktiv: Grundlagen, Algorithmen und Anwendungen*. Wiley-VCH, 2009.

[97] R Merkl and M Zwick. H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. *BMC Bioinformatics*, 9:151, 2008.

[98] A Messiah. *Quantum Mechanics*. Dover Publications, Incorporated, 1999.

[99] EW Miles, H Kawasaki, SA Ahmed, H Morita, H Morita, and S Nagata. The β subunit of tryptophan synthase. Clarification of the roles of histidine 86, lysine 87, arginine 148, cysteine 170, and cysteine 230. *Journal of Biological Chemistry*, 264:6280–6287, 1989.

[100] S Miller, J Janin, AM Lesk, and C Chothia. Interior and surface of monomeric proteins. *Journal of Molecular Biology*, 196:641–656, 1987.

[101] S Mitternacht and IN Berezovsky. A geometry-based generic predictor for catalytic and allosteric sites. *Protein Engineering Design and Selection*, 24:405–409, 2011.

[102] F Morcos, A Pagnani, B Lunt, A Bertolino, DS Marks, C Sander, R Zecchina, JN Onuchic, T Hwa, and M Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108:E1293–E1301, 2011.

[103] NJ Mulder, R Apweiler, TK Attwood, A Bairoch, A Bateman, D Binns, P Bork, V Buillard, L Cerutti, R Copley, E Courcelle, U Das, L Daugherty, M Dibley, R Finn, W Fleischmann, J Gough, D Haft, N Hulo, S Hunter, D Kahn, A Kanapin, A Kejariwal, A Labarga, PS Langendijk-Genevaux, D Lonsdale, R Lopez, I Letunic, M Madera, J Maslen, C McAnulla, J McDowall, J Mistry, A Mitchell, AN Nikolskaya, S Orchard, C Orengo, R Petryszak, JD Selengut, CJ Sigrist, PD Thomas, F Valentin, D Wilson, CH Wu, and C Yeats. New developments in the InterPro database. *Nucleic Acids Research*, 35:D224–D228, 2007.

[104] E Neher. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America*, 91:98–102, 1994.

[105] J Overington, MS Johnson, A Sali, and TL Blundell. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 241:132–145, 1990.

[106] CN Pace, H Fu, KL Fryar, J Landua, SR Trevino, BA Shirley, MM Hendricks, S Iimura, K Gajiwala, JM Scholtz, and GR Grimsley. Contribution of hydrophobic interactions to protein stability. *Journal of Molecular Biology*, 408:514–528, 2011.

[107] AR Panchenko, F Kondrashov, and S Bryant. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Science*, 13(4):884–892, 2004.

[108] S Parthasarathy, H Balaram, P Balaram, and MR Murthy. Structures of Plasmodium falciparum triosephosphate isomerase complexed to substrate analogues: observation of the catalytic loop in the open conformation in the ligand-bound state. *Acta Crystallographica Section D: Biological Crystallography*, 58(12):1992–2000, 2002.

[109] F Pazos, M Helmer-Citterich, G Ausiello, and A Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271:511–523, 1997.

[110] J Pei and NV Grishin. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17:700–712, 2001.

[111] K Persson, HD Ly, M Dieckelmann, WW Wakarchuk, SG Withers, and NC Strynadka. Crystal structure of the retaining galactosyltransferase LgtC from Neisseria meningitidis in complex with donor and acceptor sugar analogs. *Nature Structural & Molecular Biology*, 8:166–175, 2001.

[112] D Petrey, M Fischer, and B Honig. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proceedings of the National Academy of Sciences of the United States of America*, 106(41):17377–17382, 2009.

[113] NV Petrova and CH Wu. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, 7:312, 2006.

[114] CT Porter, GJ Bartlett, and JM Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32:D129–D133, 2004.

[115] EA Proctor, P Kota, SJ Demarest, JA Caravella, and NV Dokholyan. Highly covarying residues have a functional role in antibody constant domains. *Proteins: Structure, Function, and Bioinformatics*, 81:884–895, 2013.

[116] OB Ptitsyn and KL Ting. Non-functional conserved residues in globins and their possible role as a folding nucleus. *Journal of Molecular Biology*, 291:671–682, 1999.

[117] M Punta, PC Coggill, RY Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, K Forslund, G Ceric, J Clements, A Heger, L Holm, EL Sonnhammer, SR Eddy, A Bateman, and RD Finn. The Pfam protein families database. *Nucleic Acids Research*, 40:D290–D301, 2012.

[118] A Radzicka and R Wolfenden. A proficient enzyme. *Science*, 267:90–93, 1995.

[119] S Raychaudhuri. *Computational Text Analysis: For Functional Genomics and Bioinformatics.* Oxford University Press, USA, 2006.

[120] A Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

[121] S Rhee, KD Parris, SA Ahmed, EW Miles, and DR Davies. Exchange of K^+ or Cs^+ for Na^+ induces local and long-range changes in the three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ complex. *Biochemistry*, 35:4211–4221, 1996.

[122] FM Richards. Areas, volumes, packing and protein structure. *Annual Review of Biophysics and Bioengineering*, 6:151–176, 1977.

[123] TH Rod, JL Radkiewicz, and CL Brooks, 3rd. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the United States of America*, 100:6980–6985, 2003.

[124] GS Rule and TK Hitchens. *Fundamentals of Protein NMR Spectroscopy (Focus on Structural Biology).* Springer, 2005.

[125] SB Ruvinov, XJ Yang, KD Parris, U Banik, SA Ahmed, EW Miles, and DL Sackett. Ligand-mediated changes in the tryptophan synthase indole tunnel probed by nile red fluorescence with wild type, mutant, and chemically modified enzymes. *Journal of Biological Chemistry*, 270:6357–6369, 1995.

[126] C Sander and R Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9:56–68, 1991.

[127] S Sankararaman, B Kolaczkowski, and K Sjölander. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Research*, 37:W390–W395, 2009.

[128] S Sankararaman, F Sha, JF Kirsch, M. I.I Jordan, and K Sjölander. Active site prediction using evolutionary and structural information. *Bioinformatics*, 26:617–624, 2010.

[129] T Schmidt, J Haas, TG Cassarino, and T Schwede. Assessment of ligand-binding residue predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*, 79:126–136, 2011.

[130] B Schneider, T Knöchel, B Darimont, M Hennig, S Dietrich, K Babinger, K Kirschner, and R Sterner. Role of the N-terminal extension of the $\beta\alpha_8$-barrel enzyme indole-3-glycerol phosphate synthase for its fold, stability, and catalytic activity. *Biochemistry*, 44:16405–16412, 2005.

[131] B Schölkopf and AJ Smola. *Learning with kernels.* The MIT Press, 2002.

[132] Schrödinger Inc. PyMOL. Schrödinger.

[133] O Schueler-Furman and D Baker. Conserved residue clustering and protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 52:225–235, 2003.

[134] G Shackelford and K Karplus. Contact prediction using mutual information and neural nets. *Proteins: Structure, Function, and Bioinformatics*, 69 Suppl 8:159–164, 2007.

[135] CE Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5:3–55, 2001.

[136] FL Simonetti, E Teppa, A Chernomoretz, M Nielsen, and CM Buslje. MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Research*, 41:W8–W14, 2013.

[137] MS Singer, G Vriend, and RP Bywater. Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Engineering*, 15:721–725, 2002.

[138] N Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19:279–281, 1948.

[139] J Söding. Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 21:951–960, 2005.

[140] S Somarowthu, H Yang, DG Hildebrand, and MJ Ondrechen. High-performance prediction of functional residues in proteins with machine learning and computed input features. *Biopolymers*, 95:390–400, 2011.

[141] A Stark and RB Russell. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Research*, 31:3341–3344, 2003.

[142] K Tang, G Pugalenthi, PN Suganthan, CJ Lanczycki, and S Chakrabarti. Prediction of functionally important sites from protein sequences using sparse kernel least squares classifiers. *Biochemical and Biophysical Research Communications*, 384:155–159, 2009.

[143] LC Tarshis, Philip J Proteau, BA Kellogg, JC Sacchettini, and CD Poulter. Regulation of product chain length by isoprenyl diphosphate synthases.

Proceedings of the National Academy of Sciences of the United States of America, 93:15018–15023, 1996.

[144] E Teppa, A Wilkins, M Nielsen, and CM Buslje. Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. *BMC Bioinformatics*, 13:235, 2012.

[145] J Thomas, N Ramakrishnan, and C Bailey-Kellogg. Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. *Proteins: Structure, Function, and Bioinformatics*, 76:911–929, 2009.

[146] IF Thorpe and CL Brooks, 3rd. The coupling of structural fluctuations to hydride transfer in dihydrofolate reductase. *Proteins: Structure, Function, and Bioinformatics*, 57:444–457, 2004.

[147] ER Tillier and TW Lui. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, 19:750–755, 2003.

[148] W Tong, RJ Williams, Y Wei, LF Murga, J Ko, and MJ Ondrechen. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Protein Science*, 17:333–341, 2008.

[149] UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41:D43–D47, 2013.

[150] WS Valdar. Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, 48:227–241, 2002.

[151] A Volkamer, A Griewel, T Grombacher, and M Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, 50:2041–2052, 2010.

[152] J von Neumann. *Mathematical foundations of quantum mechanics.* Princeton University Press, 1996.

[153] FH Wallrapp, JJ Pan, G Ramamoorthy, DE Almonacid, BS Hillerich, R Seidel, Y Patskovsky, PC Babbitt, SC Almo, MP Jacobson, and CD Poulter. Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proceedings of the National Academy of Sciences*, 110:E1196–E1202, 2013.

[154] G Wang and RL Dunbrack. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33:W94–W98, 2005.

[155] K Wang and R Samudrala. Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, 7:385, 2006.

[156] AM Waterhouse, JB Procter, DM Martin, M Clamp, and GJ Barton. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25:1189–1191, 2009.

[157] JB Watney and S Hammes-Schiffer. Comparison of coupled motions in Escherichia coli and Bacillus subtilis dihydrofolate reductase. *The Journal of Physical Chemistry B*, 110:10130–10138, 2006.

[158] E Weber-Ban, O Hur, C Bagwell, U Banik, LH Yang, EW Miles, and MF Dunn. Investigation of allosteric linkages in the regulation of tryptophan synthase: the roles of salt bridges and monovalent cations probed

by site-directed mutation, optical spectroscopy, and kinetics. *Biochemistry*, 40:3497–3511, 2001.

[159] M Weigt, RA White, H Szurmant, JA Hoch, and T Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the Uni*, 106:67–72, 2009.

[160] R Wolfenden and MJ Snider. The depth of chemical time and the power of enzymes as catalysts. *Accounts of Chemical Research*, 34:938–945, 2001.

[161] TF Wu, CJ Lin, and RC Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.

[162] R Yahalom, D Reshef, A Wiener, S Frankel, N Kalisman, B Lerner, and C Keasar. Structure-based identification of catalytic residues. *Proteins: Structure, Function, and Bioinformatics*, 79:1952–1963, 2011.

[163] H Yao, DM Kristensen, I Mihalek, ME Sowa, C Shaw, M Kimmel, L Kavraki, and O Lichtarge. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of Molecular Biology*, 326:255–261, 2003.

[164] H Zellner, M Staudigel, T Trenner, M Bittkowski, V Wolowski, C Icking, and R Merkl. Prescont: Predicting protein-protein interfaces utilizing four residue properties. *Proteins: Structure, Function, and Bioinformatics*, 80:154–168, 2012.

[165] SW Zhang, YL Zhang, Q Pan, YM Cheng, and KC Chou. Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. *Amino Acids*, 35:495–501, 2008.

[166] T Zhang, H Zhang, K Chen, S Shen, J Ruan, and L Kurgan. Accurate sequence-based prediction of catalytic residues. *Bioinformatics*, 24:2329–2338, 2008.

[167] Y Zhang. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40, 2008.

4 List of publications and personal contribution

A JO Janda, M Busch, F Kück, M Porfenenko, R Merkl. *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, 2012.
doi: 10.1186/1471-2105-13-55.

Markus Busch, Fabian Kück, and Mikhail Porfenenko prepared datasets and were involved in programming and assessment. The algorithms were designed and implemented by myself. The SVMs were trained and assessed by myself. The publication was written by Rainer Merkl.

B JO Janda, A Meier, R Merkl. *CLIPS-4D: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3D data.*
Bioinformatics, 29(23):3029-35, 2013.
doi: 10.1093/bioinformatics/btt519. Epub 2013 Sep 18.

Andreas Meier and myself implemented the algorithms. Andreas Meier trained and assessed the program. The manuscript was written by Rainer Merkl and myself.

C JO Janda, A Popal, J Bauer, M Busch, M Klocke, W Spitzer, J Keller, R Merkl. *H2rs: Deducing evolutionary and functionally important residue-positions by means of an entropy and similarity based analysis of multiple sequence alignments*
BMC Bioinformatics, 15:118, 2014.
doi:10.1186/1471-2105-15-118

The algorithm H2rs was implemented and assessed by myself. Ajmal Popal deduced the matrix A. Jochen Bauer implemented and assessed the algorithm for the computation of the p-values. Markus Busch was involved in the assessment. Michael Klocke, Wolfgang Spitzer, and Jörg Keller designed and assessed the method to compute the quantum mutual information. Rainer Merkl conceived of the project. The manuscript was written by Rainer Merkl and myself.

5 Publications

5.1 Publication A

CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure.

JO Janda, M Busch, F Kück, M Porfenenko, R Merkl

BMC Bioinformatics, 13:55, 2012

doi: 10.1186/1471-2105-13-55

Janda et al. BMC Bioinformatics 2012, 13:55
http://www.biomedcentral.com/1471-2105/13/55

BMC Bioinformatics

RESEARCH ARTICLE Open Access

CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure

Jan-Oliver Janda[1], Markus Busch[1], Fabian Kück[2], Mikhail Porfenenko[1] and Rainer Merkl[1*]

Abstract

Background: One aim of the *in silico* characterization of proteins is to identify all residue-positions, which are crucial for function or structure. Several sequence-based algorithms exist, which predict functionally important sites. However, with respect to sequence information, many functionally and structurally important sites are hard to distinguish and consequently a large number of incorrectly predicted functional sites have to be expected. This is why we were interested to design a new classifier that differentiates between functionally and structurally important sites and to assess its performance on representative datasets.

Results: We have implemented CLIPS-1D, which predicts a role in catalysis, ligand-binding, or protein structure for residue-positions in a mutually exclusive manner. By analyzing a multiple sequence alignment, the algorithm scores conservation as well as abundance of residues at individual sites and their local neighborhood and categorizes by means of a multiclass support vector machine. A cross-validation confirmed that residue-positions involved in catalysis were identified with state-of-the-art quality; the mean MCC-value was 0.34. For structurally important sites, prediction quality was considerably higher (mean MCC = 0.67). For ligand-binding sites, prediction quality was lower (mean MCC = 0.12), because binding sites and structurally important residue-positions share conservation and abundance values, which makes their separation difficult. We show that classification success varies for residues in a class-specific manner. This is why our algorithm computes residue-specific *p*-values, which allow for the statistical assessment of each individual prediction. CLIPS-1D is available as a Web service at http://www-bioinf.uni-regensburg.de/.

Conclusions: CLIPS-1D is a classifier, whose prediction quality has been determined separately for catalytic sites, ligand-binding sites, and structurally important sites. It generates hypotheses about residue-positions important for a set of homologous proteins and focuses on conservation and abundance signals. Thus, the algorithm can be applied in cases where function cannot be transferred from well-characterized proteins by means of sequence comparison.

Background

It is of general interest to identify important sites of a protein, for example when elucidating the reaction mechanism of an enzyme. To support this task, classifiers have been developed, which utilize different kinds of information about the protein under study. Some algorithms are based on sequences [1-11], other ones make use of 3D-data [12,13], and a third class combines both approaches [14-18].

A strong argument in favor of sequence-based methods is their broad applicability and their potential to characterize proteins with a novel fold. Additionally, some signals seem to be more pronounced in sequence- than in 3D-space [19]. Commonly, these methods depend on a multiple sequence alignment (MSA) composed of a sufficiently large number of homologs. Based on the assumption that critical residues are not altered during evolution, the canonical feature to identify important residue-positions in an MSA is the conservation of individual columns. The

* Correspondence: Rainer.Merkl@biologie.uni-regensburg.de
[1]Institute of Biophysics and Physical Biochemistry, University of Regensburg, 93040 Regensburg, Germany
Full list of author information is available at the end of the article

degree of conservation can help to predict a role: In many cases, strictly conserved residues are essential for protein function [7,20,21]. In contrast, a prevalent but not exclusively found amino acid is often important for protein stability [22,23], which similarly holds for ligand-binding sites. Thus, for a precise discrimination, several properties have to be interpreted. Features that improve prediction of functionally important sites are the conservation of proximate residues [7,24] and the abundance of amino acid residues observed at catalytic sites [8,24]. In addition, implicit features deduced from protein sequences have been utilized, like the predicted secondary structure and the predicted solvent accessible surface of residues [5,8].

Most of the existing algorithms focus on the identification of sites relevant for protein function. In order to broaden the classification spectrum, we implemented the sequence-based algorithm CLIPS-1D, which predicts functionally important sites in addition to residue-positions crucial for protein structure in a mutually exclusive manner. It is based on a multiclass support vector machine, which assesses not more than seven properties deduced from residue-positions and their local neighborhood in sequence space. Our approach compares favorably with state-of-the-art classifiers and predicts catalytic residue-positions with a mean MCC-value of 0.34. The mean MCC-value is for structurally important sites 0.67 and for ligand-binding sites it is 0.12. Our findings show that separating ligand-binding sites and structurally important sites is difficult due to their similar properties and that classification quality depends on the residue type.

Results and discussion

Analysis of local conservation and abundance signals allows for a state-of-the-art classification

High-quality datasets consisting of catalytic sites, ligand-binding sites, and sites important for protein structure are required to train and assess support vector machines (SVMs), which predict the respective roles of residue-positions. Based on the content of EBI-databases, we prepared the redundancy-free and non-overlapping sets *CAT_sites* and *LIG_sites*, which consist of 840 catalytic sites and 4466 ligand-binding sites deduced from a set of 264 enzymes named *ENZ* (see Methods). Whereas the full set of functionally important sites is known for many enzymes, residues that crucially determine structure have not been identified for a representative set of proteins. Thus, to compile such sites, we had to follow an indirect approach [25] by assuming that residues in the core of proteins lacking enzymatic function are conserved due to their relevance for structure. This notion is supported by the fact that conserved hydrophobic core-residues can contribute substantially to protein stability [26]. By re-annotating a comprehensive set of non-enzymes from reference [27], we culled the dataset *NON_ENZ*, which consists of 136 proteins. *NON_ENZ* contains 3703 buried residue-positions, which are more conserved than the mean (see Methods); we designated these sites *STRUC_sites*. For all proteins under study, MSAs were taken from the HSSP database [28] and filtered prior to analysis.

Next, we identified features, which allow for a state-of-the-art classification of *CAT_sites*, *LIG_sites*, and *STRUC_sites*. Thus, we trained three two-class (2C-) SVMs to predict for each residue-position *k*, whether it is important for catalysis (SVM_{CAT}), ligand-binding (SVM_{LIG}), or protein structure (SVM_{STRUC}) and compared performance values. In the end, the features used to characterize each *k* were in the case of SVM_{CAT} a normalized Jensen-Shannon divergence $cons_{JSD}$ (*k*) (formula (4)) and an abundance-value *abund*(*k, CAT_sites*) scoring the occurrence of residues at *CAT_sites* according to formula (6). The proximity of *k* was assessed by means of a weighted score $cons_{neib}(k)$ (formula (5)) and a novel abundance-value $abund_{neib}(aa_s^k, CAT\_sites)$, deduced from conditional frequencies in the ± 3 neighborhood [8] of *CAT_sites* (formula (7)). Thus, $abund_{neib}(aa_s^k, CAT\_sites)$ compares the local environment of site *k* with the one observed for residues aa_s^k at positions annotated as catalytic sites. In order to quantify the contribution of individual features to classification quality, performance was determined for SVMs exploiting either all four features or a combination of three features, respectively. Analogously, scores for *LIG_sites* were computed, and SVM_{LIG} was trained and assessed.

It is difficult to unambiguously determine a classifier's performance, if the numbers of positive and negative cases differ to a great extent, as is here the case. This is why we computed a battery of performance values, which are given in Additional file 1: Table S1. Their comparison confirms for our problem that the performance measures support each other, thus we focus on MCC-values [29], which are also listed in Table 1. The MCC-values for SVM_{CAT} and SVM_{LIG} were 0.324 and 0.213, respectively. MCC-comparison makes clear that for *CAT_sites* and *LIG_sites* all four features add to classification quality. For *CAT_sites*, $cons_{JSD}$ (*k*) and *abund*(*k, CAT_sites*) contributed most, for *LIG_sites*, the conservation score $cons_{JSD}(k)$ was most relevant; compare Additional file 1: Table S1 and Additional file 1: Figure S1, which shows ROC and PROC curves.

Can SVM_{CAT} and SVM_{LIG} compete with state-of-the-art classifiers? For the assessment, we selected FRpred, which has outperformed other approaches and which additionally exploits the predicted secondary structure and solvent accessibility [8]. It has reached 40% precision at 20% sensitivity for the identification of catalytic

Table 1 Classification performance of SVMs and FRpred on functionally and structurally important residue-positions

	CAT_sites	*LIG_sites*	*STRUC_sites*
2C-SVM	0.324	0.213	0.782
CLIPS-1D	0.337	0.117	0.666
FRpred, score ≥ 8	0.231	0.219	41%
FRpred, score = 9	0.250	0.197	22%

The line "2C-SVM" gives MCC-values resulting from a classification of catalytic sites (*CAT_sites*) with SVM_{CAT}, of ligand-binding sites (*LIG_sites*) with SVM_{LIG}, and of structurally important sites (*STRUC_sites*) with SVM_{STRUC}. The line "CLIPS-1D" shows the performance of the MC-SVM. For FRpred, performance resulting from the analysis of HSSP-MSAs is given. For *CAT_sites* and *LIG_sites*, MCC-values are listed resulting from FRcons-cat or FRcons-lig scores of at least 8 or 9, respectively. For *STRUC_sites*, the same percentage of false positives resulted from FRcons-cat and FRcons-lig predictions.

residues and is accessible as a Web service [8]. FRpred lists two subtypes of predictions, FRcons-cat for catalytic sites and FRcons-lig for ligand-binding sites. All results are scored with values of 0-9; the higher the score, the more probable is a functional role of the residue. A classification of *CAT_sites* and *LIG_sites* with FRpred resulted in MCC-values of 0.250 (FRcons-cat) and 0.197 (FRcons-lig), when considering predictions scored 9 as positive cases. For predictions scored at least 8, the MCC-values were 0.231 and 0.219, respectively. Interestingly, performance was better, when we uploaded our preprocessed HSSP-MSAs than when FRpred compiled MSAs on itself (compare Additional file 1: Table S1), which indicates the high quality of these specifically filtered MSAs. In summary, the comparison of performance values for FRpred, SVM_{CAT}, and SVM_{LIG} confirmed that the four features selected by us account for a state-of-the-art classification.

Using corresponding features and the set *STRUC_sites*, we analogously trained SVM_{STRUC} for the prediction of residue-positions important for structure, which gave an MCC-value of 0.761. Classification quality was determined to the greatest extent by $cons_{JSD}$ (k). When classifying without this feature, MCC was lowered to 0.346. Utilizing the feature $abund_{neib}$(*k, STRUC_sites*) deteriorated performance; a higher MCC-value (0.782) was gained by an SVM trained on the remaining three features. Even *abund* (*k, STRUC_sites*) had only a marginal effect, although the respective scores differ considerably from those of *abund* (*k, CAT_sites*) and *abund*(*k, LIG_sites*); compare Table 2 and Additional file 1: Figure S2. Thus, in proteins without enzymatic function, the assessment of conservation contributed most to separate the conserved buried residues from all other ones, which constitute the negative cases. FRpred predicted with score 9 22% and with score 8 41% of the *STRUC_sites* as catalytic sites or ligand-binding sites; see Table 1.

Table 2 *abund*(*k, CLASS*)-values for amino acid residues

Residue	*CAT_sites*	*LIG_sites*	*STRUC_sites*
A	-2.0424	-0.3537	-0.1210
C	1.3255	0.7376	1.2398
D	1.1178	0.0426	-0.0498
E	0.6536	-0.3856	-0.6615
F	-0.7708	-0.0081	0.5057
G	-0.7533	0.4195	0.7020
H	1.8883	0.8279	-0.3044
I	-2.8164	-0.3026	-0.6449
K	0.6051	-0.3615	-1.0215
L	-2.4503	-0.5416	0.2116
M	-1.4026	0.1374	-0.4882
N	-0.1972	0.3566	-0.2254
P	-5.0000	-0.4542	0.3643
Q	-0.7243	-0.1841	-0.5615
R	0.6834	0.3879	-0.2593
S	0.0027	-0.0125	-0.7006
T	-0.5435	0.2314	-0.3363
V	-2.9568	-0.4130	-0.3294
W	0.1927	0.5548	1.2811
Y	0.3265	0.4572	0.7058

The score-values were deduced from residues belonging to the respective classes. See formula (6) for a definition of the scores.

CLIPS-1D: Towards a more diversified prediction of residue function

In order to elaborate the subtle differences distinguishing functionally and structurally important residue-positions, all combinations of the above training sets have to be exploited. This is why we prepared a multi-class support vector machine (MC-SVM) for CLIPS-1D, which was trained on the four classes *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites*, *i.e.*, all residue-positions from *NON_ENZ* not selected as *STRUC_sites*. Due to the above findings on 2C-SVMs, we chose the following seven features: $cons_{JSD}$ (k), $cons_{neib}$(k), *abund*(*k, CAT_sites*), *abund*(*k, LIG_sites*), *abund*(*k, STRUC_sites*), $abund_{neib}$(*k, CAT_sites*), and $abund_{neib}$(*k, LIG_sites*). The MC-SVM outputs a list of four class-specific probability values p_{class}. Based on the largest p_{class}-values, residue-positions were assigned one of the four classes; the resulting distributions are shown in Figure 1. 65% of the *CAT_sites* and 76% of the *STRUC_sites* were correctly assigned. 64% of the *LIG_sites* and 19% of *NOANN_sites* were misclassified, and each class contributed a noticeable fraction of false positives. 13% of the *STRUC_sites* were classified as *CAT_sites* and 10% as *LIG_sites*. Although the algorithm frequently failed to assign the correct class, separating positions with and without a crucial role was more successful: 96% of the *CAT_sites*,

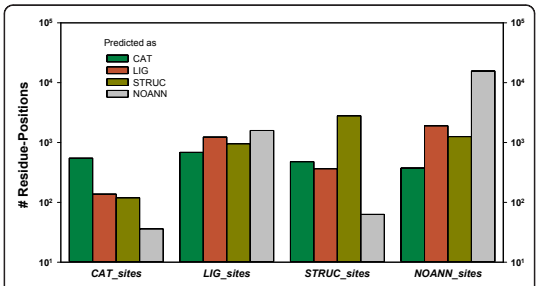


Figure 1 Classification performance of CLIPS-1D in predicting functionally and structurally important residue-positions. Based on the maximal class-probability p_{class} all members of the classes *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* were categorized. *NOANN_sites* are all residue-positions not selected as *STRUC_sites* in the *NON_ENZ* dataset, *i.e.* positions without assigned function. Note that the absolute numbers of residue-positions are plotted with a logarithmic scale.

65% of the *LIG_sites*, and 98% of the *STRUC_sites* were classified as structurally or functionally important and 81% of the *NOANN_sites* were classified as having no crucial function. It turned out that the respective MCC-value was optimal, if *CAT_sites* with $p_{CAT}(k) > 0.61$ were selected as positives. In summary, the corresponding MCC-values were 0.337, 0.117, and 0.666 for *CAT_sites*, *LIG_sites*, and *STRUC_sites*; see Table 1. In comparison with 2C-SVMs, the performance on *CAT_sites* improved moderately. However, the performance on *LIG_sites* and *STRUC_sites* dropped, which indicates that the separation of *LIG_sites* and *STRUC_sites* is difficult.

The comparison of *abund*()-values (compare Table 2) makes clear that residues are unevenly distributed among the classes, which must influence the residue-specific classification quality. Thus, we determined class-specific MCC-values for each residue, which are listed in Table 3. As expected, performance differs drastically for individual residues and between classes. Among *CAT_sites*, Arg, Asp, Cys, His, Lys, and Ser were predicted with high quality. Most of the other MCC-values were near zero and no MCC-value could be computed for Pro and Val due to empty sets. The performance-values for *LIG_sites* were generally lower. Among *STRUC_sites*, the mean MCC-value for the hydrophobic residues Ala, Ile, Leu, Met, Phe, Pro, Trp, and Val was 0.733; the mean of all hydrophilic ones was 0.494. In summary, these findings proposed to determine classification quality in more detail by computing class- and residue-specific *p*-values (see Methods). Thus, the user can assess the statistical significance of each individual prediction. Table 4 lists the resulting performance for *p*-value cut-offs of 0.01, 0.025, and 0.05. As can be seen, specificity is high in all cases; sensitivity and precision are lower and class-dependent.

Table 3 Residue-specific MCC-values

Residue	*CAT_sites*	*LIG_sites*	*STRUC_sites*
A	-0.002	0.164	0.774
C	0.404	0.162	0.676
D	0.302	0.016	0.315
E	0.345	0.052	0.348
F	0.058	0.041	0.771
G	0.024	0.262	0.591
H	0.424	-0.063	0.086
I	-0.001	0.135	0.701
K	0.452	0.031	0.337
L	-0.001	0.056	0.815
M	-0.002	0.127	0.666
N	0.071	0.139	0.561
P	-	0.139	0.683
Q	0.098	0.111	0.678
R	0.287	0.040	0.319
S	0.307	0.156	0.595
T	0.055	0.174	0.682
V	-	0.119	0.761
W	-0.008	0.007	0.689
Y	0.097	0.046	0.741

The MCC-values were determined in a class- and residue-specific manner. Due to missing cases, MCC-values could not be determined for Pro and Val residues at *CAT_sites*.

An alternative to CLIPS-1D is the algorithm ConSeq, which predicts functionally or structurally important residue-positions but does not distinguish catalytic and ligand-binding sites. Based on the analysis of five proteins, a success rate of 0.56 has been reported [5]. In order to estimate the performance of the latest ConSeq version [30], we have uploaded one sequence for each of the first five *ENZ* and *NO_ENZ* entries (see Additional file 1: Tables S3 and S4 for PDB-IDs) and used the Web server with default parameters. As ConSeq does not differentiate between catalytic sites and ligand-binding sites, the union of *CAT_sites* and *LIG_sites* was considered as positives in this case. For the combination of these residue-positions, sensitivity was 0.41, specificity 0.84, and precision 0.16; for *STRUC_sites* the values were 0.30, 0.86, and 0.31, respectively. A comparison of the performance values indicates that CLIPS-1D can compete with ConSeq.

Utilizing CLIPS-1D as a web service

A version of CLIPS-1D trained on the full datasets is available as a Web service at http://www-bioinf.uni-regensburg.de/. Its usage requires to upload an MSA in multiple Fasta-format; the result will be sent to the user *via* email.

Table 4 Performance of CLIPS-1D for different *p*-values

Cut-off	Sensitivity			Specificity			Precision		
	CAT	LIG	STRUC	CAT	LIG	STRUC	CAT	LIG	STRUC
0.010	0.170	0.030	0.225	0.996	0.991	0.991	0.316	0.176	0.827
0.025	0.276	0.077	0.445	0.992	0.977	0.977	0.270	0.178	0.789
0.050	0.401	0.137	0.582	0.987	0.954	0.961	0.246	0.165	0.742

The three performance measures were determined (see Methods) by selecting as positive cases all residue-positions with a *p*-value not greater than the given cut-off. Labels: "CAT" *CAT_sites*, "LIG" *LIG_sites*, "STRUC" *STRUC_sites*.

To illustrate the application of CLIPS-1D, we present an analysis of the enzyme indole-3-glycerol phosphate synthase (IGPS), which is found in many mesophilic and thermophilic species. IGPS belongs to the large and versatile family of $(\beta\alpha)_8$-barrel proteins, which is one of the oldest folds [31]. Additionally, folding kinetics [32] and 3D-structure of IGPS [33,34] have been studied in detail.

We analyzed the HSSP-MSA related to PDB-ID 1A53, *i.e.* the IGPS from *Sulfolobus solfataricus*. Table 5 lists all CLIPS-1D predictions with a *p*-value ≤ 0.025. According to the respective PDB-sum page [35], E51, K53, K110, E159, N180, and S211 are the catalytic residues. Besides N180, which was predicted as *LIG_site*, the other 5 sites were correctly identified as *CAT_sites*. The sites which have contact to the ligand were classified as follows: *CAT_sites* E210, *LIG_sites* I232, *STRUC_sites* F112, L131, L231, *NOANN_sites* G212, G233, S234. Classified as *LIG_sites* were also K55, I179, and S181, which are all neighbors of catalytic sites. 20 residues were predicted as *STRUC_sites*; Figure 2 shows that all belong to the core of the protein. Their function will be discussed below.

Strengths and weaknesses of CLIPS-1D

Adding the class *STRUC_sites* allowed us to compare properties of functionally and structurally important residue-positions and to assess their impact on classification quality.

For *CAT_sites*, the abundance scores indicate a strong bias of Arg, Asp, Glu, His, and Lys towards catalytic residue-positions, which is in agreement with previous findings [24]. *CAT_sites*, which were classified as structurally important, were most frequently Cys and Tyr residues. Both residues are not exceedingly overrepresented at catalytic sites and *abund*(*k, CAT_sites*)- and *abund*(*k, STRUC_sites*)-values are similarly high; compare Table 2. For extracellular proteins, structurally important Cys residues are frequently involved in disulphide bonds. Thus, algorithms like DISULFIND [40] can help to clarify CLIPS-1D's Cys classification.

Least specific was the classification of *LIG_sites*, which also suffered the most drastic loss of performance. The MCC-value dropped from 0.21 (gained with SVM_{LIG}) to 0.12, and most misclassifications gave *STRUC_sites*, which is due to the similarity of these sites with respect to the features used for classification: For both classes, $cons_{JSD}(k)$ is most relevant for classification success, and among all combinations of abundance-values the pairs *abund*(*k, LIG_sites*) and *abund*(*k, STRUC_sites*) differ least; compare Table 2. The similarity of these residue-positions is further confirmed by the large number of *STRUC_sites* classified as functionally important by FRpred, which additionally suggests that the assessment of the predicted secondary structure and the predicted solvent accessibility contributes little to discriminate functionally and structurally important sites. It follows that *LIG_sites* and *STRUC_sites* span a fuzzy continuum, which cannot be divided by means of the considered sequence-based features. On the other hand, each MCC-value characterizes a binary classification and underestimates the performance of CLIPS-1D. For example, when assessing the performance of *LIG_sites via* an MCC-value, residue-positions classified as *STRUC_sites* were counted as false-negatives. A more detailed analysis of Figure 1 and the findings on sIGPS illustrate that *LIG_sites* were often classified as *CAT_sites* or *STRUC_sites* and not as sites without any function (*NOANN_sites*), which is a drastic difference not considered by an MCC-value.

For *STRUC_sites*, the MCC-value decreased from 0.78 to 0.67 for the above reasons; however, the MCC-value is still considerably high. Can one make plausible, why these buried residue-positions are preferentially occupied by a specific set of residues? At mean, hydrophobic interactions contribute 60% and hydrogen bonds 40% to protein stability; for the stability of larger proteins, hydrophobic interactions are even more important [41]. The fraction of misclassified hydrophobic *STRUC_sites* was low; compare MCC-values of Table 3. Thus, CLIPS-1D identifies with high reliability conserved residues of the protein's core, which are most likely important for protein stability. On the other hand, the analysis of *abund*(*k, STRUC_sites*)-values (compare Table 2) shows that not all *STRUC_sites* are conserved hydrophobic residues: The hydrophobic residues Ala, Ile, Met, and Val are underrepresented, whereas the hydrophilic residues Cys, Gly, and Tyr are overrepresented. Additionally, the comparison of abundance scores indicates a preference of Leu, Phe, and Pro for structurally relevant sites. These preferences reflect the specific function of these residues for secondary structure

Table 5 CLIPS-1D predictions for residue-positions in sIGPS (PDB-ID 1A53)

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p*-value	Classification		
							CS	LBS	STRUC
I	49	0.001	0.154	0.824	0.022	0.003			SC
E	51	0.806	0.075	0.114	0.005	0.020	CAT		
K	53	0.835	0.065	0.088	0.012	0.004	CAT		
K	55	0.051	0.544	0.197	0.208	0.011		SC	
S	56	0.017	0.170	0.801	0.012	0.004			SC
L	60	0.002	0.128	0.829	0.041	0.019			IA
A	77	0.006	0.172	0.810	0.011	0.018			FC
I	82	0.002	0.259	0.667	0.073	0.011			SR
T	84	0.002	0.111	0.881	0.007	0.003			N
L	108	0.006	0.106	0.863	0.024	0.012			SR
K	110	0.866	0.078	0.046	0.011	0.002	CAT		
F	112	0.146	0.053	0.788	0.014	0.020		STRUC	FC
Q	118	0.007	0.114	0.872	0.008	0.002			FC
A	122	0.001	0.066	0.882	0.051	0.010			FC
A	127	0.024	0.193	0.776	0.008	0.022			N
L	131	0.001	0.071	0.920	0.008	0.006		STRUC	SR
L	132	0.004	0.164	0.794	0.038	0.023			SR,FC
I	133	0.005	0.169	0.790	0.036	0.005			FC
L	137	0.007	0.151	0.813	0.029	0.020			SC,FC
L	157	0.001	0.105	0.886	0.008	0.010			SC,FC
E	159	0.899	0.048	0.050	0.003	0.005	CAT		
D	165	0.189	0.071	0.699	0.040	0.007			N
I	179	0.001	0.819	0.068	0.112	0.021		SCE	
N	180	0.098	0.770	0.116	0.016	0.016	LIG		
S	181	0.011	0.774	0.134	0.081	0.019		SCE	
L	184	0.009	0.157	0.818	0.016	0.020			IA
L	197	0.003	0.130	0.818	0.049	0.020			N
E	210	0.866	0.059	0.068	0.007	0.008		CAT	
S	211	0.738	0.168	0.087	0.007	0.005	CAT		
L	231	0.003	0.224	0.762	0.011	0.025		STRUC	SC
I	232	0.006	0.835	0.059	0.099	0.017		LIG	

The first two columns give the residue and its position in sIGPS. The following four columns list the probabilities for the residue's membership with *CAT_sites*, *LIG_sites*, *STRUC_sites*, or *NOANN_sites*. The column labeled "*p*-value" lists the *p*-value for the class with max(p_{CLASS}). The columns "CS" and "LBS" indicate the classification of known catalytic and ligand-binding sites. The last column lists the annotation deduced for residues predicted as *STRUC_sites*. Meaning of labels: "CAT", "LIG", "STRUC", residues predicted as *CAT_sites*, *LIG_sites*, or *STRUC_sites*, respectively. "SC" element of a stabilization center pair in sIGPS, "SCE" *ditto* in eIGPS, "SR" stabilization residue in sIGPS; see [36]. "FC" element of the folding core; see [37]. "IA" interaction with substrate; see [38]. "N" no function assigned.

[42]. Additionally, the score-values demonstrate that CLIPS-1D does not exclusively select ILV-residues, which are considered important for protein folding [32]. *STRUC_sites*, misclassified as catalytic ones, were often Arg, Asp, and Glu, which shows that the *abund*(*k, CAT_sites*)-values have a strong effect on classification. *NOANN_sites* predicted as *CAT_sites* were frequently Arg, Asp, and His; Gly, Ser, and Thr were often predicted as *LIG_sites*. Most likely, at least some of these residue-positions belong to binding sites on the protein-surface *e.g.* protein-protein interfaces. Identifying these residues is possible [43], but beyond the scope of this study.

STRUC_sites are crucial elements of the sIGPS structure

A detailed comparison of the two thermostable variants sIGPS from *S. solfataricus* [33], tIGPS from *Thermotoga maritima*, and the thermolabile eIGPS from *Escherichia coli* has made clear that these thermostable proteins have 7 strong salt bridges more than eIGPS, and that only 3 of 17 salt bridges in tIGPS and sIGPS are

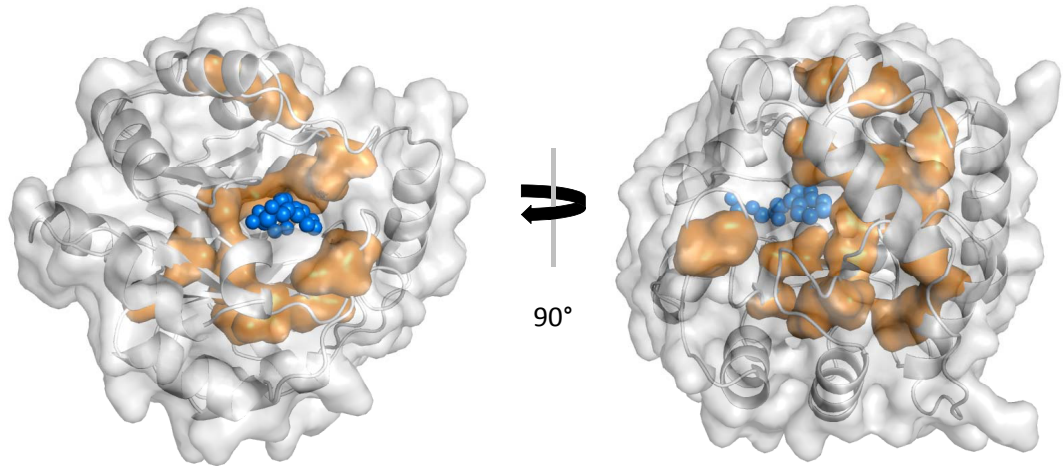


Figure 2 Localization of *STRUC_sites* in sIGPS. Based on PDB-ID 1A53, the surface of the whole protein (grey) and of residues predicted as *STRUC_sites* (orange) is shown. The substrate indole-3-glycerole phosphate is plotted in dark blue. The picture was generated by means of PyMOL [39].

topologically conserved [44]. It follows that CLIPS-1D can only identify the specific subset of structurally important residue-positions which are relevant for most of the homologous proteins constituting the MSA under study. For sIGPS, tIGPS, and eIGPS stabilization centers (*SC*) and stabilization residues (*SR*) have been determined [36]. Residues of *SCs* form tight networks of cooperative interactions which are energetically stabilized; *SRs* are embedded into a conserved hydrophobic 3D-neigborhood. 20 residue-positions of sIGPS were classified as *STRUC_sites* by CLIPS-1D. 9 of these 20 residue-positions as well as the 3 false-positive *LIG_sites* are a *SC* or *SR* residue in one of the three homologous enzymes; compare Table 5. For sIGPS, the structure of folding cores, *i.e.* local substructures, which form early during protein folding has been determined by means of HD exchange experiments [37]. 8 of the *STRUC_sites* belong to fragments, which are strongest protected against deuterium exchange (> 84%, see Table 3 in reference [37]), which indicates their significant role in the partially folded protein. A molecular dynamics study [38] and a comparison of enzyme variants [34] have made clear that two more *STRUC_sites* belong to loops interacting with the substrate. When combining the above findings, only 4 of the 20 *STRUC_sites* have no accentuated function, which confirms the relevance of these sites for the enzyme's structure.

Main application of CLIPS-1D: Predicting important sites of uncharacterized proteins

For the test cases of the CASP 7 contest, the *firestar* [17] and the I-TASSER [45] server have reached MCC-values of 0.7 when predicting functionally important residues; the performance of other servers has been substantially lower [17]. Both servers utilize the transfer of information from evolutionary related and well-characterized proteins. If applicable, this approach allows for a superior prediction quality. However, it fails completely if the function of homologous proteins is unknown. For such cases, methods are required that identify functionally and structurally important sites by analyzing conservation signals and propensity values. In contrast to ConSeq [5] and FrPred [8], CLIPS-1D predicts a specific role in catalysis, ligand-binding, or structure for each residue-position. The only prerequisite for its application is the existence of a sufficiently large number of homologous sequences, which can easily be combined to an MSA and which should be filtered according to our experience.

The number of genes which lack annotated homologs is huge: In mid 2011, the Pfam database [46] contained nearly 4000 domains of unknown function. Additionally, a comparison of databases for protein-coding genes and their products unravels a tremendous deficit of knowledge by indicating that function is unknown for more than 40% of all protein-coding genes [47]. These genes may code for unknown folds and novel enzymatic capabilities. However, if computational biology fails to identify function, an enormous battery of experiments have to be accomplished, due to the number of distinct enzymatic activities and other protein functions observed in Nature; see *e.g.* [48]. Therefore, all plausible hypotheses generated by CLIPS-1D and similar methods are of value and help to reduce the number of experimental analyses.

One might expect that exploiting the 3D-structure of a protein contributes a lot to functional assignment. This is not necessarily the case: Structure-based algorithms have failed to outperform MSA-based approaches in predicting catalytic sites and have maximally reached the same MCC-value; see [18] and references therein. However, if 3D-data and an MSA are at hand, features deduced from structure and from homologous sequences can be utilized in a concerted manner. In addition to the above features, signals caused by correlated mutations [3,49] can then be utilized to further characterize catalytic sites, which are surrounded by residues spanning a network of mutual information [50]. This is why we work on exploiting a combination of these features and the near future will show, whether this approach further improves classification quality. There is an urgent need for such methods: In mid 2011, no function has been attributed to more than 4% of the protein structures deposited in the Protein Data Bank [51].

Conclusions

By analyzing an MSA by means of CLIPS-1D, residue-positions involved in catalysis can be identified with acceptable quality. In contrast, ligand-binding sites and residue-positions important for protein structure are hard to distinguish due to their similar patterns of conservation and residue propensities. Our MC-SVM can be applied to cases where the function of all homologs is unknown. The algorithm supports the user's decisions by computing a *p*-value for each prediction.

Methods

CAT_sites and *LIG_sites*, datasets of catalytic and ligand-binding residue-positions

To compile a test set of functionally important sites, we processed the content of the Catalytic Site Atlas (CSA) [52]. We exclusively utilized the manually curated entries of CSA and did not consider sites that have been annotated by means of PSI-BLAST alignments. In order to eliminate redundancy of proteins, we used the PISCES server [53] with a sequence-similarity cut-off of 25%. For each protein, an MSA was taken from the HSSP database [28] and selected for further analyses, if it contained at least 125 sequences. The resulting dataset consists of 264 enzymes and related MSAs, which we named *ENZ*. These proteins contain 840 catalytic residues, which we denominated *CAT_sites*. For these proteins we also deduced ligand-binding sites by exploiting PDBsum pages [35]. The resulting dataset consists of 216 proteins and contains 4466 binding sites, which we named *LIG_sites*. The datasets *CAT_sites* and *LIG_sites* do not overlap; their content is listed in Additional file 1: Tables S2 and S3.

In order to eliminate too similar and too distant sequences which might introduce a bias, the number of identical residues $ident(s_i, s_j)$ was determined for each pair of sequences s_i, s_j belonging to the same MSA. Sequences were removed until the fraction of identical residues was in the range $0.25 \le ident(s_i, s_j) \le 0.90$. Additionally, sequences deviating from the first one in length by more than 30% were deleted.

STRUC_sites, a set of conserved residue-positions in proteins lacking enzymatic function

A set of 480 non-enzyme proteins has been compiled in reference [27]. Based on PDBsum and CSA, we re-annotated all entries and prepared a redundancy-free set of MSAs as explained above. The resulting dataset *NON_ENZ* consists of 136 proteins and related MSAs from HSSP with at least 50 sequences. In order to exclude residues from interfaces and other binding sites, we did not consider residue-positions lying at the protein surface by eliminating all sites with a relative solvent accessible surface area of at least 5% (see [43] and references therein). Among the remaining sites were 3703 with a conservation value $cons_{ident}(k) > 1.0$ (see formula (2)). For lack of a more biochemically motivated classification scheme, these conserved sites were regarded as important for structure. We named this set *STRUC_sites*, its content is listed in Additional file 1: Table S4. We designated the complement *NO_ANN* sites; these are the remaining 19,223 residue-positions of the *NON_ENZ* dataset.

Conservation of an individual site

An instructive measure to assess conservation of a single residue-position k is $max\_frequ(k)$, the largest amino acid frequency $f_k(aa_i)$ observed in column k of an MSA:

$$max\_frequ(k) = \max_{i=1..20}(f_k(aa_i)) \quad (1)$$

To normalize for MSA-specific variations of conservation, we computed $cons_{ident}(k)$, which is a z-score deduced from $max\_frequ(k)$ according to

$$cons_{ident}(k) = \frac{max\_frequ(k) - \mu_{ident}}{\sigma_{ident}} \quad (2)$$

Mean μ_{ident} and standard deviation σ_{ident} values were determined individually for each MSA under study. An alternative conservation measure is the Jensen-Shannon divergence [8] of site k:

$$JSD(k) = H\left(\frac{f_K^{obs} - f^{backgr}}{2}\right) - \frac{1}{2}H(f_K^{obs}) - \frac{1}{2}H(f^{backgr}) \quad (3)$$

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 SwissSProt database [54] were taken as background frequencies f^{backgr}. $H(.)$ is Shannon's entropy [55]. For classification, we used the z-score $cons_{JSD}$ (k):

$$cons_{JSD}(k) = \frac{JSD(k) - \mu_{JSD}}{\sigma_{JSD}} \quad (4)$$

Mean μ_{JSD} and standard deviation σ_{JSD} values were determined individually for each MSA. For the prediction of functionally important residues, $JSD(k)$ has performed better than other conservation measures [7].

Conservation of a sequence neighborhood

To characterize the conservation of a sequence neighborhood, $cons_{neib}(k)$ was computed in analogy to [8]:

$$cons_{neib}(k) = \frac{1}{|Neib|} \sum_{l \in Neib} w_l \, cons_{JSD}(k + l) \quad (5)$$

$Neib$ = {-3,-2,-1,+1,+2,+3} determined the set of neighboring positions. The weights were: $w_{-1} = w_{+1} = 3$, $w_{-2} = w_{+2} = 2$, $w_{-3} = w_{+3} = 1$. Note that conservation of position k was not considered to compute $cons_{neib}(k)$.

Propensities of catalytic sites, ligand-binding sites, and positions important for structure

Inspired by [24], three scores $abund(k, CLASS)$ were computed as:

$$abund(k, CLASS) = \sum_{i=1}^{20} f_k(aa_i) \log \frac{f^{CLASS}(aa_i)}{f^{backgr}(aa_i)} \quad (6)$$

$f^{backgr}(aa_i)$ were the above background frequencies. $f^{CLASS}(aa_i)$ were the frequencies of residues from one set $CLASS \in$ {CAT_sites, LIG_sites, $STRUC\_sites$}.

Scoring propensities of a neighborhood

To assess the class-specific neighborhood of a site k, we introduced:

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

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 $CLASS \in$ {CAT_sites,LIG_sites,$STRUC\_sites$}. $Neib$ is the ± 3 neighborhood.

Evaluating classification performance

To assess the performance of a classification, the rates TPR (*Sensitivity*), FPR, *Specificity*, and *Precision*

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}, Specificity = \frac{TN}{TN + FP}, Precision = \frac{TP}{TP + FP} \quad (8)$$

as well as ROC and PROC curves were determined [56]. For a ROC curve, depending on a cut-off for one parameter (here it is p_{class} (k)), the TPR values are plotted *versus* the FPR values. For a PROC curve, *Precision* is plotted *versus* TPR. As a further performance measure, the Matthews correlation coefficient (MCC) has been introduced [29]:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (9)$$

MCC-values are considered a fair measure to assess performance on unbalanced sets of positives and negatives, as observed here [57]. 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 with SVM_{CAT}, positives are the selected CAT_sites and negatives are all other residue-positions of the considered MSAs.

Classifying by means of support vector machines

We utilized the *libsvm* library [58] with a Gaussian radial basis function kernel and determined during training optimal parameters γ_{RBF} and C by means of a grid search [59]. Prior to presenting features to the SVM, they were normalized according to

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (10)$$

Here, $V_e(k)$ is for residue k the value of feature e, and $\min(V_e)$ and $\max(V_e)$ are the smallest and the largest value determined for this feature.

Our 2C-SVMs predict for each residue-position k, whether it is a catalytic site (SVM_{CAT}), a ligand-binding site (SVM_{LIG}), or a site important for structure (SVM_{STRUC}). Taking SVM_{CAT} as an example, an *a posteriori* probability p_{class} (k), here it is p_{CAT} (k), for the label "k is a catalytic site" was deduced from the distance of the feature set for k and the hyperplane separating catalytic and non-catalytic residue-positions [60].

We utilized p_{class} (k) to assess performance and to assign classes. Training and assessment was organized as an 8-fold cross validation. For each training step, the number of positive and negative cases was balanced, *i.e.* for SVM_{CAT}, residue-positions from CAT_sites and the same number of non-catalytic sites was selected. In order to eliminate sampling bias during the grid search, each parameter was deduced as means from training trials with the same positives and 50 different, randomly selected sets of negative cases. To compute the

performance measures (*e.g.* MCC-values), all positive and all negative cases belonging to the selected subset of MSAs were classified.

Analogously, an MC-SVM was applied to the four classes *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites*. The output of the MC-SVM consists of four class-probabilities p_{class} (see [60]) for each residue-position. These were deduced from the *a posteriori* probabilities of the six 2C-SVMs, which were trained on one specific combination of two classes, each. 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 all residue-positions k not belonging to the considered class. *I. e.*, 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 $s(k)$.

Additional material

Additional file 1: A plot comparing *abund*(*k, CLASS*)-values, Figures and Tables giving performance-values of 2C-SVMs, and Tables listing the composition of datasets. (PDF 327 kb).

Acknowledgements
The work was supported by DFG grant ME-2259/1-1.

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

Authors' contributions
JOJ designed and implemented algorithms, and trained and assessed the SVMs. MB, FK, and MP prepared datasets and were involved in programming and assessment. RM conceived of and coordinated the study, and wrote the manuscript. All authors read and approved the manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 22 December 2011 Accepted: 5 April 2012
Published: 5 April 2012

References

1. Overington J, Johnson MS, Sali A, Blundell TL: **Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction.** *Proc Biol Sci* 1990, **241(1301)**:132-145.
2. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2(2)**:171-178.
3. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257(2)**:342-358.
4. Huang JY, Brutlag DL: **The EMOTIF database.** *Nucleic Acids Res* 2001, **29(1)**:202-204.
5. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N: **ConSeq: the identification of functionally and structurally important residues in protein sequences.** *Bioinformatics* 2004, **20(8)**:1322-1324.
6. Gutman R, Berezin C, Wollman R, Rosenberg Y, Ben-Tal N: **QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns.** *Nucleic Acids Res* 2005, **33**:W255-261, Web Server issue.
7. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics* 2007, **23(15)**:1875-1882.
8. Fischer JD, Mayer CE, Söding J: **Prediction of protein functional residues from sequence by probability density estimation.** *Bioinformatics* 2008, **24(5)**:613-620.
9. Sankararaman S, Kolaczkowski B, Sjölander K: **INTREPID: a web server for prediction of functionally important residues by evolutionary analysis.** *Nucleic Acids Res* 2009, **37**:W390-395, Web Server issue.
10. Tang K, Pugalenthi G, Suganthan PN, Lanczycki CJ, Chakrabarti S: **Prediction of functionally important sites from protein sequences using sparse kernel least squares classifiers.** *Biochem Biophys Res Commun* 2009, **384(2)**:155-159.
11. Erdin S, Ward RM, Venner E, Lichtarge O: **Evolutionary trace annotation of protein function in the structural proteome.** *J Mol Biol* 2010, **396(5)**:1451-1473.
12. Petrey D, Fischer M, Honig B: **Structural relationships among proteins with different global topologies and their implications for function annotation strategies.** *Proc Natl Acad Sci USA* 2009, **106(41)**:17377-17382.
13. Mitternacht S, Berezovsky IN: **A geometry-based generic predictor for catalytic and allosteric sites.** *Protein Eng* 2011, **24(4)**:405-409.
14. Panchenko AR, Kondrashov F, Bryant S: **Prediction of functional sites by analysis of sequence and structure conservation.** *Prot Sci* 2004, **13(4)**:884-892.
15. Laskowski RA, Watson JD, Thornton JM: **ProFunc: a server for predicting protein function from 3D structure.** *Nucleic Acids Res* 2005, **33**:W89-93, Web Server issue.
16. Kalinina OV, Gelfand MS, Russell RB: **Combining specificity determining and conserved residues improves functional site prediction.** *BMC Bioinformatics* 2009, **10**:174.
17. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML: **Firestar-advances in the prediction of functionally important residues.** *Nucleic Acids Res* 2011, , **39 Web Server:** W235-241.
18. Yahalom R, Reshef D, Wiener A, Frankel S, Kalisman N, Lerner B, Keasar C: **Structure-based identification of catalytic residues.** *Proteins* 2011, **79(6)**:1952-1963.
19. Dou Y, Geng X, Gao H, Yang J, Zheng X, Wang J: **Sequence conservation in the prediction of catalytic sites.** *Prot J* 2011, **30(4)**:229-239.
20. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics* 2001, **17(8)**:700-712.
21. Wang K, Samudrala R: **Incorporating background frequency improves entropy-based residue conservation measures.** *BMC Bioinformatics* 2006, **7**:385.
22. Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M: **The consensus concept for thermostability engineering of proteins: further proof of concept.** *Protein Eng* 2002, **15(5)**:403-411.
23. Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, Wong S, Gualfetti P, Schellenberger V: **Construction of stabilized proteins by combinatorial consensus mutagenesis.** *Protein Eng Des Sel* 2004, **17(11)**:787-793.
24. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites.** *J Mol Biol* 2002, **324(1)**:105-121.
25. Ptitsyn OB, Ting KL: **Non-functional conserved residues in globins and their possible role as a folding nucleus.** *J Mol Biol* 1999, **291(3)**:671-682.
26. Schueler-Furman O, Baker D: **Conserved residue clustering and protein structure prediction.** *Proteins* 2003, **52(2)**:225-235.
27. Davidson NJ, Wang X: **Non-alignment features based enzyme/non-enzyme classification using an ensemble method.** *Proc Int Conf Mach Learn Appl* 2010, 546-551.
28. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9(1)**:56-68.
29. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405(2)**:442-451.
30. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N: **ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.** *Nucleic Acids Res* 2010, , **38 Web Server:** W529-533.

31. Caetano-Anollés G, Kim HS, Mittenthal JE: **The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture.** *Proc Natl Acad Sci USA* 2007, **104(22)**:9358-9363.
32. Gu Z, Rao MK, Forsyth WR, Finke JM, Matthews CR: **Structural analysis of kinetic folding intermediates for a TIM barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and Gō model simulation.** *J Mol Biol* 2007, **374(2)**:528-546.
33. Hennig M, Darimont B, Sterner R, Kirschner K, Jansonius JN: **2.0 Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile *Sulfolobus solfataricus*: possible determinants of protein stability.** *Structure* 1995, **3(12)**:1295-1306.
34. Schneider B, Knöchel T, Darimont B, Hennig M, Dietrich S, Babinger K, Kirschner K, Sterner R: **Role of the N-terminal extension of the $(\beta\alpha)_8$-barrel enzyme indole-3-glycerol phosphate synthase for its fold, stability, and catalytic activity.** *Biochemistry* 2005, **44(50)**:16405-16412.
35. Laskowski RA, Chistyakov VV, Thornton JM: **PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.** *Nucleic Acids Res* 2005, , **33 Database:** D266-268.
36. Bagautdinov B, Yutani K: **Structure of indole-3-glycerol phosphate synthase from *Thermus thermophilus* HB8: implications for thermal stability.** *Acta Crystallogr D: Biol Crystallogr* 2011, **67(Pt 12)**:1054-1064.
37. Gu Z, Zitzewitz JA, Matthews CR: **Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry: the roles of motif and sequence for the indole-3-glycerol phosphate synthase from *Sulfolobus solfataricus*.** *J Mol Biol* 2007, **368(2)**:582-594.
38. Mazumder-Shivakumar D, Bruice TC: **Molecular dynamics studies of ground state and intermediate of the hyperthermophilic indole-3-glycerol phosphate synthase.** *Proc Natl Acad Sci USA* 2004, **101(40)**:14379-14384.
39. **Schrödinger: PyMOL.** Schrödinger Inc.
40. Ceroni A, Passerini A, Vullo A, Frasconi P: **DISULFIND: a disulfide bonding state and cysteine connectivity prediction server.** *Nucleic Acids Res* 2006, , **34 Web Server:** W177-181.
41. Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley BA, Hendricks MM, Iimura S, Gajiwala K, Scholtz JM, *et al*: **Contribution of hydrophobic interactions to protein stability.** *J Mol Biol* 2011, **408(3)**:514-528.
42. Chou PY, Fasman GD: **Empirical predictions of protein conformation.** *Annu Rev Biochem* 1978, **47**:251-276.
43. Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, Icking C, Merkl R: **Prescont: Predicting protein-protein interfaces utilizing four residue properties.** *Proteins* 2012, **80(1)**:154-168.
44. Knöchel T, Pappenberger A, Jansonius JN, Kirschner K: **The crystal structure of indoleglycerol-phosphate synthase from *Thermotoga maritima*. Kinetic stabilization by salt bridges.** *J Biol Chem* 2002, **277(10)**:8626-8634.
45. Zhang Y: **I-TASSER server for protein 3D structure prediction.** *BMC Bioinformatics* 2008, **9**:40.
46. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al*: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**:D247-D251, Database issue.
47. Friedberg I, Jambon M, Godzik A: **New avenues in protein function prediction.** *Prot Sci* 2006, **15(6)**:1527-1529.
48. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W, *et al*: **The enzyme function initiative.** *Biochemistry* 2011, **50(46)**:9950-9962.
49. Merkl R, Zwick M: **H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments.** *BMC Bioinformatics* 2007, **9**:151.
50. Marino Buslje C, Teppa E, Di Domenico T, Delfino JM, Nielsen M: **Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification.** *PLoS Comp Biol* 2010, **6(11)**:e1000978.
51. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1)**:235-242.
52. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, , **32 Database:** D129-133.
53. Wang G, Dunbrack RL Jr: **PISCES: recent improvements to a PDB sequence culling server.** *Nucleic Acids Res* 2005, , **33 Web Server:** W94-98.
54. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28(1)**:45-48.
55. Shannon C: **A mathematical theory of communication.** *Bell Sys Tech J* 1948, **27**:379-423.
56. Davis J, Goadrich M: **The relationship between precision-recall and ROC curves.** *ICML* NewYork: Pittsburgh; 2006, 233-240.
57. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML: **Progress and challenges in predicting protein-protein interaction sites.** *Brief Bioinform* 2009, **10(3)**:233-246.
58. Chang CC, Lin CJ: **LIBSVM: a library for support vector machines.** *ACM Trans Int Sys Tech* 2011, **2(27)**:1-27.
59. Schölkopf B, Smola AJ: *Learning with kernels* London: The MIT Press; 2002.
60. Wu TF, Lin CJ, Weng RC: **Probability estimates for multi-class classification by pairwise coupling.** *J Mach Learn Res* 2004, **5**:975-1005.

doi:10.1186/1471-2105-13-55
Cite this article as: Janda *et al.*: CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinformatics* 2012 **13**:55.

5.2 Publication B

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

JO Janda, A Meier, R Merkl

Bioinformatics, 29(23):3029-35, 2013

doi: 10.1093/bioinformatics/btt519. Epub 2013 Sep 18.

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

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

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).

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

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.

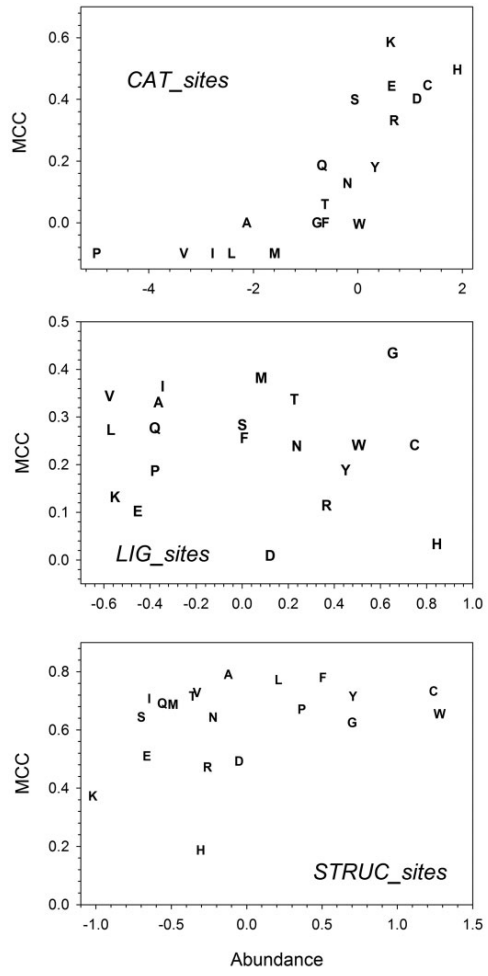


Fig. 1. Class-specific performance of CLIPS-4D for individual residues. In a class-specific manner, MCC-values were determined for each residue and plotted versus the respective abundance values *abund(k, CLASS)*. The rareness of residues P, V, I, L and M among *CAT_sites* precluded to compute MCC-values. 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 $\geq 5\%$). 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.

Supplementary Data for

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§], Rainer Merkl[1]*

[1]Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany

[2]Faculty of Mathematics and Computer Science, University of Hagen, 58084 Hagen, Germany

[§]The first two authors contributed equally to this work

* To whom correspondence should be addressed:

Phone: +49-941-943 3086;

Fax: +49-941-943 2813;

Email: Rainer.Merkl@ur.de

A class-specific assessment of classification performance

The class-specific MCC-values signal a noticeable fraction of misclassifications. To study the kind of misclassifications in detail, all residues of our four datasets were classified with CLIPS-4D and based on the largest p_{class}-values, residue-positions were assigned one of the four classes; the resulting distributions are shown in Supplementary Figure 2. 68% of the *CAT_sites* and 79% of the *STRUC_sites* were assigned correctly. 57% of the *LIG_sites* and 21% of the *NOANN_sites* were misclassified, and each class contributed a noticeable fraction of false positives. 9% of the *STRUC_sites* were classified as *CAT_sites* and 10% as *LIG_sites.* However, although the algorithm failed to assign the correct class to a certain extent, separating positions with and without a crucial role was more successful: 97% of the *CAT_sites,* 73% of the *LIG_sites*, and 98% of the *STRUC_sites* were classified as structurally or functionally important and 79% of the *NOANN_sites* were classified as having no crucial function.

Ten case studies: classifying the residues from proteins with bound substrate or product analogs and from six CASP targets

To illustrate the performance of CLIPS-4D, we first analyzed data related to the enzyme indole-3-glycerol phosphate synthase (IGPS) from the hyperthermophilic archaeon *Sulfolobus solfataricus*. We chose this enzyme as many functionally and structurally important residues have been identified previously. CLIPS-4D was provided with the MSA from the HSSP database and the pdb-file 1A53.

Supplementary Table 1 lists the predictions of CLIPS-4D for all functional sites and all sites with a *p*-value ≤ 0.05, which is our recommended setting. CLIPS-4D predicted 24 *STRUC_sites*; eight of which are known to be part of the folding core (Pace *et al.*, 2011) and another six have been identified as stabilization residues or stabilization centers (Bagautdinov and Yutani, 2011). L131 is a *LIG_site* and L60 is known to interact with the substrate (Mazumder-Shivakumar and Bruice, 2004). No structural role has so far been assigned to the remaining eight buried residues T84 (98%), E85 (67%), G91 (100%), L96 (82%), A127 (97%), I160 (78%), D165 (89%), and L197 (84%). The conservation (given in brackets) of most residues strikingly exceeds the mean

value of all IGPS residues, which is 50%. Of the six catalytic residues listed in CSA (E51, K53, K110, E159, N180, S211), CLIPS-4D correctly classified five as *CAT_sites* and N180 was predicted as a *LIG_site*. According to pdbsum, nine residues are in contact with the ligand. The three residues G212, L231, and G233 were classified correctly as *LIG_sites*. E210 was predicted as *CAT_site* and L131 as *STRUC_site*. For K55 the p_{LIG_sites}-value was highest, but the p-value was 0.095. For F89, F112, and S234 p_{CAT_sites}-values were highest, but below the cut-off of 0.64. Of the false positive predicted eight *LIG_sites*, three are in contact with the substrate, see (Mazumder-Shivakumar and Bruice, 2004) and two are important for structure.

Additionally, we analyzed this dataset and three more ones in a manner similar to the CASP approach. To begin with, we identified by means of PyMol (Schrödinger) those residues with a distance of at most 4 Å from one heavy atom belonging to a product, or substrate analog. These residues were taken as positive cases and all other residues were negative cases. From the union of *CAT_sites* and *LIG_sites*, the numbers of TP, TN, FP, and FN residues were deduced, performance values were determined as described and listed in Supplementary Table 2. The last column of Supplementary Table 1 lists the resulting classification and Supplementary Figure 3 shows the localization of these residues in IGPS. The MCC-value was 0.56 and all positive predictions were in the vicinity of the product IGP, compare Supplementary Figure 3.

For the following three enzymes, performance values were added to Supplementary Table 2 and for each case, a Supplementary Table and a Supplementary Figure shows the results in more detail. Again, CLIPS-4D was provided with a pdb-file and an MSA from the HSSP database.

LgtC from *Neisseria meningitidis* is a glycosyltransferase which was crystallized with a donor sugar analog (UPF) and an acceptor sugar analog (ACY). The MCC-value of the CLIPS-4D prediction was 0.51. Note that the crystal structure is lacking the C-terminal 25 residues (Persson *et al.*, 2001). Thus the environment of the residues 260-282 might not be the natural one and this might explain some FP predictions of CLIPS-4D. Results are listed in Supplementary Table 3 and illustrated in Supplementary Figure 4.

The histidine triad protein (HIT) family consist of relatively small proteins and is among the most ubiquitous in nature. The pdb dataset 1KPF contains the structure of this nucleotidyl hydrolase from *Homo sapiens* plus a substrate analog, namely AMP (Lima *et al.*, 1997). The MCC-value of the CLIPS-4D prediction was 0.42. Results are listed in Supplementary Table 4 and illustrated in Supplementary Figure 5.

The *Plasmodium falciparum* triosephosphate isomerase (TIM) catalyzes the isomerisation between dihyroxyacetone phosphate and d-glyceraldehyde-3-phosphate. The crystal structure (pdb id 1M7P) contains the enzyme and the substrate analog G3H. The MCC-value for CLIPS-4D predictions was 0.45. However, according to (Parthasarathy et al., 2002) several residues classified as FP are involved in catalysis or ligand binding; see Supplementary Table 5: Residues 166 - 176 belong to a catalytic loop, which undergoes a structural rearrangement during catalysis and plays a crucial role in preventing phosphate elimination, E165 is a catalytic base of the catalytic triad, S73 provides an anchoring hydrogen bond to the ligand and residues G209 - V212 are anchoring the phosphate group of the ligand. These findings make plausible that at least some of the CLIPS-4D predictions which are far apart from the substrate are important for the enzyme's function; compare also Supplementary Figure 6. On the other hand, comparing these and the following results makes also clear that each output will contain a certain number of false positive predictions. It is the similarity of these residue-positions to catalytic or ligand binding sites that brings CLIPS-4D to distinguish them from the rest.

In a second round of performance tests, we concentrated on CASP targets and a comparison with the outcome of homology transfer methods. Supplementary Tables 6 – 11 list the output for the CASP targets T0526, T0584, T0604, T0615, T0632, and T0721, which possess a nonmetal ligand and are the five most difficult cases of CASP9 (Schmidt*, et al.*, 2011) or are from CASP10. In all cases, CLIPS-4D was provided with an MSA from the HSSP database and the corresponding pdb file. The tables list all predicted *CAT_sites* and *LIG_sites* with a *p*-value ≤ 0.05. The union of these predictions was considered as ligand-binding and classified according to the definition of the extended binding sites of the CASP contest in order to deduce

performance values for nonmetal ligands. For comparison, the predictions of *firestar* (Lopez*, et al.*, 2011) were listed as well. The MCC-values of CLIPS-4D predictions varied in the range of 0.19 - 0.61.

Note that F156, R232, T270, and E506 of T0604 have been under predicted binding sites in CASP9 (Schmidt*, et al.*, 2011) but were correctly predicted as ligand-binding by CLIPS-4D. The performance on T0584 is discussed in detail in the next paragraph.

A detailed analysis of CLIPS-4D predictions for CASP target T0584

According to the definition of the extended ligand binding site (Schmidt*, et al.*, 2011) 9 *CAT_sites* and 31 *LIG_sites* are false positive predictions for target T0584 (pdb id 3NF2), which results in an MCC-value of not more than 0.19. This target is a polyprenyl transferase and synthesizes a C15 isoprenoid from the fundamental building blocks IPP and DMAPP by consecutive steps of elongation, cyclopropagation, rearrangement and cyclization reactions (Wallrapp *et al*., 2013). Aspartates of two Asp-rich regions are involved in binding DMAPP and catalysis *via* chelation of Mg^{2+}, a cofactor required for enzyme activity. Five Asp-residues from these regions confirmed to be important for catalysis by mutagenesis studies (Liang *et al.*, 2002) were predicted by CLIPS-4D as *CAT_sites*; compare Supplementary Table 7. Supplementary Figure 7 shows the orientation of these functionally important Asp-residues, namely D128, D129, D132, D258, and D259 relative to the position of three Mg^{2+} atoms and the ligands, whose positions were transferred from the experimental control structure 1RQI *via* 3D superposition using PyMol (Schrödinger). According to the definition of the extended ligand binding site, only D128 is a positive prediction.

During synthesis, the substrate grows into an elongation cavity, which was modeled previously (Tarshis *et al.*, 1996). 18 of the *LIG_sites* line this elongation cavity; compare Supplementary Table 7 and Supplementary Figure 8. The homologous enzyme farnesyl pyrophosphate synthetase from *Escherichia coli* undergoes significant substrate induced active site rearrangements in the C terminus, the α4-α5 loop, and the α9-α10 loop, which contain residues contacting the substrate (Hosfield *et al.*, 2004). A similar mechanism has been deduced for the mint geranyl pyrophosphate synthase (Hsieh *et al.*, 2010) and a prenyltransferase from *Arabidopsis thaliana* (Hsieh *et al.*, 2011). Three *LIG_sites* belong to the α4-α5 loop in

3NF2. Six *LIG_sites* belong to a loop not found in other polyprenyl transferases (see Supplementary Figure 9). It would be interesting to see whether this loop alters conformation upon substrate binding and shields the reaction center in a similar manner as the α4-α5 loop. All mentioned residues are not directly involved in ligand binding in the experimental control structure 1RQI and are thus false positive predictions according to CASP. However, the function of corresponding residues in homologous enzymes suggests that at least some of the predicted *CAT_sites* or *LIG_sites* are involved in substrate-binding or catalysis.

References for Supplementary Data

Bagautdinov, B. and Yutani, K. (2011) Structure of indole-3-glycerol phosphate synthase from *Thermus thermophilus* HB8: implications for thermal stability, *Acta Crystallogr D Biol Crystallogr*, **67**, 1054-1064.

Cassarino, T.G., Bordoli, L. and Schwede, T. (under review, draft version) Assessment of ligand binding site predictions in CASP10.

Gu, Z., Zitzewitz, J.A. and Matthews, C.R. (2007) Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry: the roles of motif and sequence for the indole-3-glycerol phosphate synthase from *Sulfolobus solfataricus*, *Journal of Molecular Biology*, **368**, 582-594.

Hosfield, D.J. *et al.* (2004) Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis, *Journal of Biological Chemistry*, **279**, 8526-8529.

Hsieh, F.L. *et al.* (2010) Enhanced specificity of mint geranyl pyrophosphate synthase by modifying the R-loop interactions, *Journal of Molecular Biology*, **404**, 859-873.

Hsieh, F.L. *et al.* (2011) Structure and mechanism of an Arabidopsis medium/long-chain-length prenyl pyrophosphate synthase, *Plant Physiol*, **155**, 1079-1090.

Liang, P.H., Ko, T.P. and Wang, A.H. (2002) Structure, mechanism and function of prenyltransferases, *Eur J Biochem*, **269**, 3339-3354.

Lima, C.D., Klein, M.G. and Hendrickson, W.A. (1997) Structure-based analysis of catalysis and substrate definition in the HIT protein family, *Science*, **278**, 286-290.

Lopez, G. *et al.* (2011) Firestar-advances in the prediction of functionally important residues, *Nucleic Acids Research*, **39**, W235-241.

Mazumder-Shivakumar, D. and Bruice, T.C. (2004) Molecular dynamics studies of ground state and intermediate of the hyperthermophilic indole-3-glycerol phosphate synthase, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14379-14384.

Pace, C.N. *et al.* (2011) Contribution of hydrophobic interactions to protein stability, *Journal of Molecular Biology*, **408**, 514-528.

Parthasarathy, S. *et al.* (2002) Structures of *Plasmodium falciparum* triosephosphate isomerase complexed to substrate analogues: observation of the catalytic loop in the open conformation in the ligand-bound state, *Acta Crystallogr D Biol Crystallogr*, **58**, 1992-2000.

Persson, K. *et al.* (2001) Crystal structure of the retaining galactosyltransferase LgtC from *Neisseria meningitidis* in complex with donor and acceptor sugar analogs, *Nature Structural Biology*, **8**, 166-175.

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

Schrödinger PyMOL. Schrödinger Inc.

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

Wallrapp, F.H. *et al.* (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily, *Proc Natl Acad Sci U S A*, **110**, E1196-1202.

Waterhouse, A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench, *Bioinformatics*, **25**, 1189-1191.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	Classification			Contact
								CS	LBS	STRUC	
I	49	0.001	0.053	0.906	0.040	0.000	STRUC			SC	
A	50	0.011	0.095	0.487	0.406	0.045	STRUC			SR	
E	51	0.660	0.059	0.262	0.019	0.032	CAT	CAT			TP
TPK	53	0.888	0.084	0.016	0.013	0.001	CAT	CAT			TP
K	55	0.049	0.410	0.267	0.274	0.095	NOANN		FN		
S	56	0.084	0.723	0.144	0.048	0.026	LIG				FP
S	58	0.107	0.803	0.069	0.020	0.011	LIG		(IA)		FP
L	60	0.004	0.073	0.879	0.044	0.011	STRUC			IA	
Y	76	0.057	0.151	0.566	0.227	0.041	STRUC			FC	
A	77	0.009	0.064	0.907	0.021	0.008	STRUC			FC	
S	81	0.394	0.098	0.493	0.015	0.022	STRUC			SR	
I	82	0.000	0.062	0.835	0.103	0.003	STRUC			SR	
L	83	0.012	0.767	0.185	0.036	0.014	LIG		(IA)		FP
T	84	0.004	0.072	0.907	0.016	0.004	STRUC				
E	85	0.358	0.093	0.409	0.140	0.025	STRUC				
F	89	0.530	0.306	0.149	0.015		NOANN		FN		FN
G	91	0.002	0.039	0.948	0.011	0.030	STRUC				
L	96	0.002	0.073	0.790	0.135	0.019	STRUC				
P	106	0.000	0.044	0.902	0.053	0.022	STRUC			SC	
L	108	0.005	0.021	0.962	0.012	0.002	STRUC			SR	
K	110	0.893	0.085	0.012	0.010	0.001	CAT	CAT			TP
F	112	0.523	0.252	0.198	0.026		NOANN		FN		FN
Q	118	0.009	0.046	0.930	0.016	0.001	STRUC			FC	
A	122	0.001	0.038	0.882	0.079	0.010	STRUC			FC	
A	127	0.027	0.047	0.914	0.012	0.007	STRUC				
L	131	0.000	0.013	0.983	0.004	0.000	STRUC		STRUC		FN
L	132	0.004	0.039	0.925	0.032	0.005	STRUC			SR,FC	
I	133	0.033	0.715	0.185	0.067	0.028	LIG			FC,FP	FP
L	137	0.005	0.046	0.920	0.029	0.005	STRUC			FC,SC	
L	157	0.001	0.024	0.968	0.008	0.002	STRUC			FC,SC	
E	159	0.918	0.061	0.014	0.007	0.001	CAT	CAT			TP
I	160	0.000	0.100	0.456	0.444	0.045	STRUC				
D	165	0.091	0.035	0.839	0.036	0.002	STRUC				
A	169	0.008	0.067	0.523	0.403	0.041	STRUC			FC	
N	180	0.172	0.735	0.080	0.013	0.026	LIG	LIG			TP
L	184	0.065	0.596	0.296	0.043	0.043	LIG		(IA)		FP
L	197	0.001	0.065	0.857	0.076	0.012	STRUC				
E	210	0.764	0.061	0.148	0.026	0.018	CAT		CAT		TP
S	211	0.748	0.203	0.041	0.009	0.004	CAT	CAT			TP
G	212	0.118	0.660	0.177	0.045	0.049	LIG		LIG		TP
I	213	0.001	0.825	0.084	0.091	0.012	LIG				
S	214	0.037	0.730	0.005	0.228	0.023	LIG				FP
L	231	0.012	0.706	0.253	0.028	0.025	LIG		LIG		TP
I	232	0.018	0.234	0.307	0.441		NOANN				FN
G	233	0.090	0.777	0.107	0.026	0.023	LIG		LIG		TP
S	234	0.615	0.319	0.026	0.040		NOANN		FN		FN
R	238	0.129	0.662	0.025	0.184	0.028	LIG			SC	FP

Supplementary Table 1: CLIPS-4D predictions for residue-positions in *Sulfolobus solfataricus* IGPS (pdb id 1A53).
The table lists results for known catalytic and ligand binding sites and all predictions with a *p*-value ≤ 0.05. The first two columns give the residue and its position in sIGPS. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. Column "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and column "Prediction" the resulting prediction. "FN" are false negative predictions. The columns "CS" and "LBS" give the classification of known catalytic and ligand-binding sites which were deduced from CSA and pdbsum. Residues that interact with the substrate are indicated by "(IA)"; see (Mazumder-Shivakumar and Bruice, 2004). In contrast, column "STRUC" indicates the annotation deduced for residues predicted as *STRUC_sites*. Labels "CAT", "LIG" symbolize residues predicted as *CAT_sites* or *LIG_sites*, respectively. Meaning of labels in column "*STRUC*" is as follows: "IA" interaction with substrate; see (Mazumder-Shivakumar and Bruice, 2004) "SC" element of a stabilization center pair in sIGPS, "SR" stabilization residue in sIGPS; see (Bagautdinov and Yutani, 2011). "FC" element of the folding core, see (Gu *et al.*, 2007). The column "Contact" gives a classification of all predictions with respect to their distance to the product IGP, which is an approach similar to that of the CASP contest. The resulting MCC-value is 0.56

	IGPS 1A53	LgtC 1G9R	HIT 1KPF	TIM 1M7P
MCC	0.56	0.51	0.42	0.45
Sens	0.67	0.50	0.53	0.88
Spec	0.96	0.96	0.91	0.92
Prec	0.53	0.61	0.47	0.26

Supplementary Table 2: CLIPS-4D performance for enzymes with bound substrate analogs or products

All MCC-, sensitivity (label "Sens"), specificity (label "Spec"), and precision ("Prec") values were determined according to Formulae (13) and (14) given in the manuscript. The first two lines list the name and the pdb id of the enzyme.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	Contact
D	2	0.199	0.576	0.049	0.176	LIG	0.029	FP
A	6	0.040	0.671	0.035	0.254	LIG	0.031	TP
A	7	0.015	0.177	0.009	0.799			FN
D	8	0.801	0.141	0.036	0.021	CAT	0.020	TP
N	10	0.037	0.698	0.024	0.241	LIG	0.037	TP
Y	11	0.397	0.407	0.113	0.083	LIG	0.138	FN
H	78	0.151	0.532	0.026	0.292	LIG	0.058	FN
I	79	0.003	0.540	0.042	0.416	LIG	0.064	FN
S	80	0.037	0.769	0.029	0.165	LIG	0.018	TP
T	82	0.085	0.685	0.015	0.215	LIG	0.052	FN
T	83	0.014	0.757	0.123	0.107	LIG	0.030	TP
R	86	0.641	0.252	0.070	0.037	CAT	0.032	TP
D	103	0.871	0.080	0.032	0.017	CAT	0.007	TP
I	104	0.000	0.478	0.025	0.497			FN
D	105	0.861	0.080	0.039	0.020	CAT	0.008	TP
D	130	0.491	0.295	0.108	0.106			FN
F	132	0.014	0.518	0.013	0.455	LIG	0.087	FN
V	133	0.000	0.187	0.001	0.812			FN
N	153	0.321	0.574	0.054	0.051	LIG	0.074	FN
A	154	0.028	0.237	0.245	0.489			FN
G	155	0.013	0.119	0.817	0.051	STRUC	0.094	FN
Y	186	0.209	0.532	0.045	0.214	LIG	0.080	FN
Q	187	0.005	0.143	0.005	0.847			FN
D	188	0.766	0.137	0.063	0.034	CAT	0.032	TP
Q	189	0.098	0.727	0.120	0.055	LIG	0.019	TP
D	227	0.824	0.099	0.058	0.018	CAT	0.015	FP
Y	230	0.777	0.073	0.125	0.026	CAT	0.002	FP
R	231	0.716	0.132	0.117	0.035	CAT	0.017	FP
H	244	0.852	0.128	0.007	0.013	CAT	0.038	TP
C	246	0.635	0.302	0.027	0.036	CAT	0.012	TP
G	247	0.845	0.101	0.035	0.019	CAT	0.000	TP
K	250	0.946	0.030	0.007	0.016	CAT	0.000	TP
E	262	0.704	0.122	0.125	0.048	CAT	0.027	FP
E	266	0.738	0.081	0.133	0.048	CAT	0.022	FP
T	272	0.115	0.701	0.153	0.031	LIG	0.046	FP
E	276	0.646	0.142	0.156	0.055	CAT	0.035	FP
K	281	0.205	0.553	0.137	0.105	LIG	0.042	FP

Supplementary Table 3: CLIPS-4D predictions for residue-positions in galactosyltransferase LgtC from *Neisseria meningitidis* (pdb id 1G9R).

The table lists all residue positions predicted by CLIPS-4D as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding in analogy to the definition introduced for CASP. Column "Contact" indicates the classification similar to that of the CASP definition. Residues, with a distance of at most 4 Å to the substrate analog AMP are positive all other ones are negative cases. The MCC-value is 0.51.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	Contact
I	18	0.001	0.770	0.072	0.157	LIG	0.018	TP
F	19	0.004	0.662	0.304	0.030	LIG	0.045	TP
I	22	0.001	0.757	0.207	0.035	LIG	0.022	FP
I	27	0.002	0.670	0.188	0.140	LIG	0.040	FP
A	29	0.001	0.738	0.030	0.230	LIG	0.023	FP
F	41	0.044	0.730	0.192	0.034	LIG	0.025	TP
H	42	0.082	0.483	0.009	0.426	LIG	0.080	FN
D	43	0.723	0.204	0.056	0.016	CAT	0.040	TP
I	44	0.001	0.726	0.109	0.165	LIG	0.027	TP
S	45	0.013	0.406	0.009	0.572			FN
L	53	0.021	0.055	0.897	0.027	STRUC	0.009	FN
Q	62	0.010	0.619	0.008	0.363	LIG	0.044	FP
R	95	0.707	0.151	0.113	0.029	CAT	0.018	FP
V	98	0.000	0.681	0.005	0.314	LIG	0.029	FP
N	99	0.497	0.373	0.110	0.020			FN
E	100	0.059	0.709	0.010	0.222	LIG	0.006	FP
G	105	0.049	0.699	0.111	0.141	LIG	0.039	TP
Q	106	0.128	0.769	0.079	0.024	LIG	0.011	TP
S	107	0.047	0.740	0.015	0.197	LIG	0.021	TP
V	108	0.002	0.521	0.385	0.092	LIG	0.061	FN
V	111	0.000	0.871	0.029	0.100	LIG	0.006	FP
H	112	0.671	0.276	0.020	0.033	CAT	0.092	FN
H	114	0.670	0.275	0.021	0.034	CAT	0.092	FN
L	116	0.121	0.617	0.040	0.222	LIG	0.039	FP

Supplementary Table 4: CLIPS-4D predictions for residue-positions in the histidine triad (HIT) protein from *Homo sapiens* (pdb id 1KPF).

The table lists all residue positions predicted by CLIPS-4D as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding in analogy to the definition introduced for CASP. Column "Contact" indicates the classification similar to that of the CASP definition. Residues, with a distance of at most 4 Å to the sugar analog UPF and the acceptor analog ACY are positive all other ones are negative cases. The MCC-value is 0.42.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	Contact	Function
N	10	0.711	0.134	0.141	0.014	CAT	0.003	TP	
K	12	0.936	0.041	0.011	0.012	CAT	0.000	TP	
S	45	0.038	0.648	0.008	0.306	LIG	0.047	FP	
Q	64	0.195	0.679	0.072	0.054	LIG	0.031	FP	
N	65	0.048	0.711	0.159	0.082	LIG	0.032	FP	
S	73	0.046	0.774	0.154	0.026	LIG	0.016	FP	SUB_INTERACTION
G	76	0.112	0.720	0.129	0.038	LIG	0.031	FP	
E	77	0.771	0.173	0.027	0.029	CAT	0.017	FP	
V	78	0.003	0.779	0.022	0.196	LIG	0.016	FP	
H	95	0.838	0.128	0.012	0.021	CAT	0.042	TP	
F	96	0.021	0.865	0.087	0.027	LIG	0.005	TP	
E	97	0.718	0.054	0.190	0.038	CAT	0.025	FP	
E	104	0.798	0.123	0.057	0.021	CAT	0.014	FP	
E	165	0.877	0.082	0.028	0.012	CAT	0.004	FP	CAT_BASE
A	169	0.001	0.577	0.387	0.035	LIG	0.047	FP	CAT_LOOP
I	170	0.019	0.867	0.089	0.026	LIG	0.007	FP	CAT_LOOP
G	171	0.250	0.659	0.074	0.016	LIG	0.049	FP	CAT_LOOP
T	175	0.063	0.722	0.025	0.191	LIG	0.040	FP	CAT_LOOP
Q	180	0.054	0.640	0.052	0.253	LIG	0.039	FP	
L	183	0.001	0.623	0.031	0.345	LIG	0.037	FP	
E	187	0.027	0.487	0.003	0.483	LIG	0.046	FP	
G	209	0.022	0.853	0.106	0.019	LIG	0.010	FP	ANCHOR
S	211	0.260	0.691	0.036	0.012	LIG	0.034	TP	
V	212	0.020	0.730	0.077	0.173	LIG	0.023	FP	ANCHOR
L	230	0.002	0.785	0.188	0.025	LIG	0.010	TP	
G	232	0.134	0.768	0.081	0.018	LIG	0.024	TP	
N	233	0.377	0.509	0.092	0.022	LIG	0.099	FN	
A	234	0.125	0.756	0.096	0.024	LIG	0.020	FP	

Supplementary Table 5: CLIPS-4D predictions for residue-positions of the *Plasmodium falciparum* triosephosphate isomerase (pdb id 1M7P).

The table lists all residue positions predicted by CLIPS-4D as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding in analogy to the definition introduced for CASP. Column "Contact" indicates the classification similar to that of the CASP definition. Residues, with a distance of at most 4 Å to the sugar analog UPF and the acceptor analog G3H are positive all other ones are negative cases. The MCC-value is 0.43. The last column "Function" lists the function of FP predictions according to (Parthasarathy, *et al.*, 2002). S73 provides an anchoring hydrogen bond to the ligand, E165 is a catalytic base, residues 166 - 176 belong to the catalytic loop and residues G209 - V212 are anchoring the phosphate group of the ligand.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	CASP	*firestar*
P	35	0.001	0.631	0.047	0.321	0.027	LIG	FP	TN
K	37	0.031	0.679	0.006	0.284	0.012	LIG	FP	TN
T	43	0.040	0.696	0.016	0.248	0.049	LIG	TP	FN
D	44	0.051	0.614	0.004	0.330	0.022	LIG	FP	TN
N	55	0.365	0.521	0.101	0.014	0.091	LIG	TN	FP
R	56	0.792	0.176	0.027	0.005	0.008	CAT	TP	TP
R	67	0.156	0.603	0.039	0.203	0.044	LIG	FP	TN
W	77	0.031	0.637	0.017	0.316	0.035	LIG	TP	FN
L	82	0.088	0.303	0.099	0.510			TN	FP
H	83	0.877	0.105	0.008	0.009	0.031	CAT	TP	TP
W	87	0.052	0.764	0.125	0.059	0.010	LIG	FP	TN
W	147	0.007	0.901	0.015	0.077	0.000	LIG	FP	TN
H	148	0.896	0.088	0.007	0.009	0.020	CAT	TP	TP
Y	150	0.058	0.768	0.084	0.090	0.013	LIG	FP	FP
W	173	0.006	0.607	0.012	0.376	0.045	LIG	TP	FN
W	198	0.018	0.792	0.016	0.175	0.008	LIG	FP	TN
N	200	0.099	0.711	0.111	0.078	0.032	LIG	TP	FN
W	206	0.043	0.731	0.199	0.027	0.017	LIG	FP	TN
N	207	0.036	0.691	0.011	0.262	0.037	LIG	FP	TN
E	224	0.023	0.519	0.003	0.455	0.037	LIG	FP	TN
F	235	0.007	0.783	0.039	0.171	0.016	LIG	FP	TN
F	241	0.011	0.752	0.005	0.232	0.021	LIG	TP	FN
L	252	0.002	0.742	0.077	0.178	0.018	LIG	FP	TN
E	253	0.932	0.058	0.007	0.003	0.000	CAT	TP	TP
E	282	0.029	0.532	0.006	0.432	0.033	LIG	FP	TN

Supplementary Table 6: CLIPS-4D and *firestar* predictions for residue-positions of CASP target T0526 (pdb id 3NRE).

The table lists all residue positions predicted by CLIPS-4D or by *firestar* (Lopez*, et al.*, 2011) as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding according to CASP specification. Column "CASP" indicates the rating according to the CASP definition of the extended ligand binding site (Schmidt*, et al.*, 2011). The column "*firestar*" lists the corresponding classification for predictions generated by *firestar* during the CASP contest.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	CASP	*firestar*	*Localization*
R	47	0.192	0.598	0.055	0.156	0.049	LIG	FP	TN	
L	54	0.010	0.644	0.189	0.158	0.034	LIG	FP	TN	
Y	71	0.155	0.661	0.105	0.080	0.037	LIG	FP	TN	
F	73	0.004	0.814	0.029	0.153	0.011	LIG	FP	TN	
A	83	0.004	0.783	0.009	0.204	0.016	LIG	FP	TN	ADD_LOOP
D	84	0.025	0.714	0.005	0.256	0.009	LIG	FP	TN	ADD_LOOP
G	85	0.033	0.857	0.014	0.096	0.010	LIG	FP	TN	ADD_LOOP
D	86	0.046	0.661	0.007	0.286	0.015	LIG	FP	TN	ADD_LOOP
G	87	0.010	0.867	0.032	0.091	0.009	LIG	FP	TN	ADD_LOOP
G	88	0.087	0.751	0.133	0.029	0.026	LIG	FP	FP	ADD_LOOP
K	89	0.685	0.274	0.028	0.014	0.014	CAT	TP	TP	
A	90	0.001	0.182	0.011	0.806			TN	FP	
V	91	0.004	0.596	0.034	0.366	0.048	LIG	FP	FP	CAVITY
R	92	0.614	0.284	0.084	0.018			FN	TP	
E	118	0.863	0.112	0.016	0.009	0.004	CAT	FP	FP	CAVITY
H	121	0.634	0.287	0.035	0.043	0.102	CAT	FN	TP	
S	124	0.164	0.544	0.142	0.150	0.076	LIG	FN	TP	
L	125	0.039	0.659	0.266	0.036	0.033	LIG	TP	TP	
L	126	0.016	0.785	0.034	0.165	0.010	LIG	FP	TN	
H	127	0.940	0.044	0.008	0.008	0.002	CAT	FP	FP	
D	128	0.923	0.050	0.021	0.007	0.001	CAT	TP	TP	
D	129	0.831	0.139	0.024	0.006	0.013	CAT	FP	FP	CAVITY
M	131	0.039	0.625	0.096	0.240	0.050	LIG	TN	FP	
D	132	0.873	0.084	0.036	0.008	0.007	CAT	FP	FP	CAVITY
D	134	0.666	0.199	0.086	0.049	0.051	CAT	TN	FP	
R	137	0.778	0.169	0.043	0.009	0.009	CAT	TP	TP	
R	138	0.604	0.302	0.077	0.017			FN	TP	
R	140	0.111	0.737	0.064	0.088	0.010	LIG	FP	TN	α4-α5 LOOP
D	141	0.133	0.590	0.076	0.201	0.024	LIG	FP	TN	α4-α5 LOOP
T	142	0.004	0.810	0.115	0.072	0.019	LIG	FP	TN	α4-α5 LOOP
W	144	0.030	0.741	0.094	0.134	0.015	LIG	FP	TN	
D	157	0.692	0.079	0.220	0.009	0.045	CAT	FP	TN	
L	189	0.004	0.638	0.200	0.158	0.036	LIG	TP	TP	
Q	193	0.157	0.687	0.128	0.028	0.028	LIG	FP	FP	CAVITY
D	196	0.776	0.141	0.065	0.018	0.029	CAT	FP	TN	CAVITY
E	200	0.359	0.516	0.057	0.069	0.037	LIG	FP	TN	CAVITY
M	213	0.001	0.672	0.285	0.042	0.041	LIG	FP	TN	CAVITY
E	214	0.074	0.583	0.028	0.316	0.021	LIG	FP	TN	CAVITY
K	217	0.911	0.071	0.012	0.006	0.000	CAT	TP	TP	
T	218	0.352	0.522	0.112	0.014	0.100	LIG	FN	TP	
L	221	0.054	0.673	0.246	0.028	0.029	LIG	FP	TN	CAVITY
L	222	0.003	0.805	0.019	0.174	0.008	LIG	FP	TN	CAVITY
F	254	0.150	0.620	0.207	0.024	0.058	LIG	TN	FP	
Q	255	0.407	0.443	0.120	0.029	0.097	LIG	FN	TP	
D	258	0.801	0.125	0.061	0.013	0.020	CAT	FP	FP	CAVITY
D	259	0.696	0.245	0.049	0.010	0.045	CAT	FP	TN	CAVITY
L	261	0.003	0.731	0.180	0.087	0.020	LIG	FP	TN	CAVITY
G	262	0.021	0.745	0.156	0.078	0.028	LIG	FP	TN	
G	265	0.069	0.716	0.132	0.082	0.033	LIG	FP	TN	
A	269	0.002	0.597	0.006	0.395	0.043	LIG	FP	TN	
T	270	0.036	0.711	0.130	0.123	0.043	LIG	FP	TN	
K	272	0.923	0.062	0.009	0.006	0.000	CAT	FP	FP	CAVITY
Q	273	0.085	0.684	0.045	0.185	0.028	LIG	FP	TN	CAVITY
D	277	0.814	0.131	0.046	0.009	0.018	CAT	FP	TN	CAVITY
R	281	0.118	0.629	0.068	0.184	0.039	LIG	FP	TN	
K	282	0.614	0.322	0.040	0.024			FN	TP	
K	283	0.162	0.597	0.026	0.215	0.028	LIG	FP	TN	CAVITY
V	373	0.001	0.730	0.039	0.230	0.023	LIG	FP	TN	

Supplementary Table 7: CLIPS-4D and *firestar* predictions for residue-positions of CASP target T0584 (pdb id 3NF2).

For the meaning of the first ten columns, see legend to Supplementary Table 2. The column "Localization" indicates residues that line the elongation cavity (label CAVITY), belong to the α4-α5 loop (label α4-α5 LOOP) or to an additional loop (label ADD_LOOP) not found in other polyprenyl transferases.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	CASP	*firestar*
R	3	0.643	0.228	0.086	0.043	0.032	CAT	FP	TN
E	6	0.729	0.046	0.171	0.055	0.023	CAT	FP	TN
E	14	0.813	0.072	0.081	0.033	0.010	CAT	FP	TN
E	15	0.755	0.080	0.120	0.044	0.019	CAT	FP	TN
D	20	0.786	0.071	0.110	0.033	0.027	CAT	FP	TN
K	24	0.754	0.069	0.121	0.057	0.008	CAT	FP	TN
S	35	0.056	0.731	0.164	0.049	0.023	LIG	FP	TN
N	37	0.091	0.661	0.205	0.043	0.046	LIG	FP	TN
R	40	0.688	0.249	0.042	0.020	0.023	CAT	FP	TN
G	42	0.021	0.735	0.187	0.056	0.029	LIG	FP	TN
R	46	0.733	0.182	0.061	0.025	0.016	CAT	FP	TN
I	103	0.041	0.815	0.026	0.119	0.012	LIG	TP	TP
G	104	0.047	0.871	0.056	0.026	0.008	LIG	TP	TP
F	105	0.002	0.285	0.002	0.710			TN	FP
G	106	0.200	0.728	0.048	0.024	0.031	LIG	TP	TP
P	107	0.000	0.641	0.284	0.076	0.025	LIG	TP	TP
C	108	0.323	0.497	0.044	0.136	0.053	LIG	FN	TP
L	110	0.002	0.686	0.047	0.264	0.028	LIG	FP	TN
V	126	0.001	0.116	0.029	0.854			TN	FP
E	127	0.844	0.118	0.019	0.019	0.007	CAT	TP	TP
R	128	0.481	0.402	0.075	0.042			FN	TP
R	135	0.082	0.075	0.810	0.033	0.002	STRUC	TN	FP
T	136	0.001	0.123	0.003	0.873			TN	FP
F	156	0.047	0.781	0.132	0.040	0.016	LIG	TP	FN
G	157	0.027	0.290	0.634	0.049	0.157	STRUC	FN	FN
G	159	0.023	0.877	0.080	0.020	0.008	LIG	FP	TN
G	160	0.035	0.893	0.051	0.020	0.005	LIG	TP	TP
A	161	0.017	0.919	0.048	0.016	0.004	LIG	TP	TP
G	162	0.073	0.856	0.050	0.021	0.010	LIG	TP	FN
F	164	0.047	0.674	0.107	0.172	0.042	LIG	TP	FN
S	165	0.634	0.287	0.054	0.025	0.010	CAT	TP	FN
G	167	0.007	0.159	0.769	0.065	0.105	STRUC	FN	FN
K	168	0.787	0.171	0.018	0.025	0.006	CAT	TP	FN
Y	170	0.080	0.707	0.024	0.188	0.025	LIG	TP	FN
E	189	0.024	0.534	0.002	0.439	0.033	LIG	FP	TN
A	190	0.023	0.711	0.010	0.257	0.026	LIG	FP	TN
H	203	0.800	0.134	0.021	0.045	0.048	CAT	FP	TN
G	205	0.042	0.124	0.787	0.047	0.102	STRUC	TN	FP
K	215	0.009	0.592	0.008	0.391	0.028	LIG	FP	TN
T	231	0.013	0.646	0.014	0.327	0.063	LIG	FN	FN
R	232	0.089	0.711	0.010	0.190	0.014	LIG	TP	FN
V	233	0.002	0.546	0.045	0.407	0.058	LIG	FN	FN
A	262	0.062	0.796	0.103	0.040	0.015	LIG	TP	TP
V	263	0.000	0.819	0.008	0.173	0.011	LIG	TP	TP
G	264	0.111	0.778	0.083	0.029	0.023	LIG	TP	TP
H	265	0.715	0.208	0.020	0.057	0.079	CAT	TN	FP
A	267	0.001	0.136	0.500	0.364	0.043	STRUC	FN	FN
T	270	0.009	0.713	0.050	0.228	0.043	LIG	TP	FN
G	289	0.005	0.084	0.861	0.049	0.078	STRUC	TN	FP
H	294	0.865	0.063	0.025	0.047	0.035	CAT	FP	TN
A	315	0.041	0.621	0.039	0.299	0.038	LIG	FP	TN
Y	331	0.087	0.686	0.143	0.084	0.030	LIG	FP	TN
F	333	0.004	0.861	0.115	0.021	0.005	LIG	FP	TN
C	334	0.902	0.056	0.016	0.026	0.001	CAT	FP	TN
C	336	0.827	0.131	0.016	0.025	0.004	CAT	FP	TN
P	337	0.000	0.804	0.133	0.063	0.004	LIG	FP	TN
G	338	0.690	0.180	0.111	0.018	0.000	CAT	FP	TN
V	342	0.003	0.422	0.152	0.424			FN	FN
N	354	0.075	0.778	0.119	0.028	0.014	LIG	TP	FN
G	355	0.079	0.770	0.115	0.036	0.023	LIG	TP	FN
R	361	0.648	0.211	0.099	0.041	0.032	CAT	FP	TN
R	480	0.646	0.114	0.207	0.033	0.032	CAT	FP	TN
G	505	0.025	0.223	0.690	0.062	0.137	STRUC	FN	FN
E	506	0.799	0.165	0.018	0.018	0.014	CAT	TP	FN
G	512	0.048	0.853	0.075	0.025	0.010	LIG	FP	TN
G	513	0.060	0.838	0.077	0.025	0.012	LIG	TP	FN
I	514	0.001	0.886	0.089	0.024	0.005	LIG	TP	FN
L	515	0.010	0.643	0.007	0.340	0.034	LIG	FP	TN
A	517	0.006	0.844	0.105	0.046	0.010	LIG	TP	FN

Supplementary Table 8: CLIPS-4D and *firestar* predictions for residue-positions of CASP target T0604 (pdb-file 3NLC).
The table lists all residue positions predicted by CLIPS-4D or by *firestar* (Lopez *et al.*, 2011) as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding according to CASP specification. Column "CASP" indicates the rating according to the CASP definition of the extended ligand binding site (Schmidt *et al.*, 2011). The column "*firestar*" lists the corresponding classification for predictions generated by *firestar* during the CASP contest.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	CASP	*firestar*
Q	23	0.022	0.783	0.132	0.062	0.009	LIG	FP	TN
R	25	0.566	0.325	0.083	0.026			TN	FP
K	26	0.054	0.496	0.010	0.440	0.058	LIG	FN	TP
D	27	0.201	0.591	0.060	0.148	0.024	LIG	FP	TN
Y	33	0.156	0.676	0.138	0.030	0.033	LIG	TP	FN
E	66	0.777	0.186	0.025	0.012	0.017	CAT	FP	TN
D	67	0.861	0.100	0.031	0.008	0.008	CAT	FP	TN
R	88	0.165	0.598	0.024	0.213	0.049	LIG	FP	TN
E	89	0.256	0.603	0.101	0.039	0.019	LIG	FP	TN
V	90	0.001	0.865	0.058	0.076	0.007	LIG	FP	TN
D	93	0.090	0.562	0.081	0.268	0.030	LIG	FP	TN
K	98	0.020	0.709	0.008	0.262	0.007	LIG	TP	FN
R	101	0.071	0.662	0.012	0.255	0.028	LIG	FP	TN
K	102	0.040	0.583	0.006	0.372	0.032	LIG	TP	FN
Q	105	0.014	0.746	0.081	0.159	0.016	LIG	FP	TN
N	108	0.016	0.659	0.073	0.252	0.049	LIG	FP	TN
K	111	0.202	0.636	0.026	0.137	0.017	LIG	FP	TN
K	120	0.810	0.149	0.020	0.022	0.005	CAT	FP	TN
D	126	0.815	0.075	0.023	0.087	0.018	CAT	FP	TN
N	127	0.306	0.594	0.086	0.013	0.067	LIG	FN	TP
L	128	0.002	0.334	0.310	0.354			FN	FN
D	130	0.281	0.452	0.168	0.100	0.063	LIG	FN	TP
L	131	0.012	0.618	0.236	0.135	0.039	LIG	TP	TP
W	139	0.016	0.593	0.030	0.362	0.046	LIG	TP	FN
R	143	0.014	0.267	0.034	0.685			FN	FN
Y	147	0.034	0.674	0.025	0.267	0.033	LIG	TP	FN
W	150	0.022	0.599	0.071	0.308	0.046	LIG	FP	TN

Supplementary Table 9: CLIPS-4D and *firestar* predictions for residue-positions of CASP target T0615 (pdb id 3NQW).

The table lists all residue positions predicted by CLIPS-4D or by *firestar* (Lopez*, et al.*, 2011) as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding according to CASP specification. Column "CASP" indicates the rating according to the CASP definition of the extended ligand binding site (Schmidt*, et al.*, 2011). The column "*firestar*" lists the corresponding classification for predictions generated by *firestar* during the CASP contest.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	CASP	*firestar*
D	34	0.033	0.517	0.004	0.446	0.043	LIG	FP	TN
K	39	0.027	0.577	0.005	0.391	0.036	LIG	FP	TN
R	43	0.034	0.600	0.004	0.362	0.049	LIG	FP	TN
L	48	0.008	0.697	0.065	0.230	0.027	LIG	FP	TN
A	49	0.004	0.593	0.052	0.352	0.043	LIG	FP	TN
Q	53	0.013	0.673	0.056	0.258	0.031	LIG	FP	TN
E	55	0.065	0.652	0.009	0.273	0.012	LIG	FP	TN
N	76	0.121	0.303	0.559	0.016	0.028	STRUC	FN	TP
V	81	0.003	0.773	0.065	0.159	0.016	LIG	FP	FP
H	82	0.918	0.072	0.005	0.005	0.013	CAT	FP	TN
G	83	0.010	0.904	0.078	0.008	0.003	LIG	TP	TP
I	85	0.001	0.710	0.013	0.276	0.028	LIG	FP	TN
L	90	0.003	0.605	0.079	0.313	0.041	LIG	FP	TN
D	91	0.929	0.055	0.014	0.002	0.001	CAT	FP	TN
T	92	0.082	0.717	0.036	0.165	0.043	LIG	FP	TN
G	95	0.021	0.780	0.175	0.024	0.021	LIG	FP	TN
Q	96	0.020	0.760	0.006	0.214	0.014	LIG	FP	TN
N	99	0.059	0.668	0.013	0.260	0.046	LIG	FP	TN
S	107	0.038	0.752	0.016	0.195	0.020	LIG	FP	TN
A	108	0.002	0.843	0.028	0.127	0.010	LIG	FP	TN
V	109	0.001	0.756	0.085	0.158	0.020	LIG	TP	FN
T	110	0.196	0.572	0.215	0.016	0.083	LIG	FN	FN
Y	117	0.139	0.302	0.511	0.049	0.046	STRUC	FN	TP
V	118	0.000	0.645	0.029	0.325	0.037	LIG	TP	TP
K	119	0.120	0.458	0.026	0.395	0.073	LIG	FN	TP
P	120	0.006	0.326	0.319	0.349			FN	TP
H	134	0.441	0.311	0.059	0.189			FN	FN
G	136	0.090	0.443	0.434	0.033	0.107	LIG	FN	FN
K	137	0.106	0.637	0.017	0.240	0.017	LIG	TP	FN
Q	138	0.006	0.612	0.012	0.371	0.044	LIG	TP	FN
R	139	0.026	0.629	0.015	0.330	0.039	LIG	TP	FN
E	143	0.652	0.181	0.140	0.026	0.034	CAT	FP	TN
E	152	0.017	0.218	0.007	0.758			TN	FP
G	157	0.053	0.784	0.096	0.067	0.021	LIG	FP	TN
T	158	0.053	0.744	0.115	0.088	0.032	LIG	FP	TN
G	159	0.060	0.669	0.133	0.137	0.047	LIG	FP	TN
V	163	0.002	0.645	0.104	0.250	0.037	LIG	FP	TN
L	164	0.003	0.600	0.010	0.387	0.043	LIG	TP	FN
R	165	0.073	0.595	0.017	0.315	0.049	LIG	FP	TN
S	166	0.024	0.687	0.042	0.247	0.036	LIG	TP	FN
R	167	0.271	0.524	0.059	0.145	0.074	LIG	FN	FN

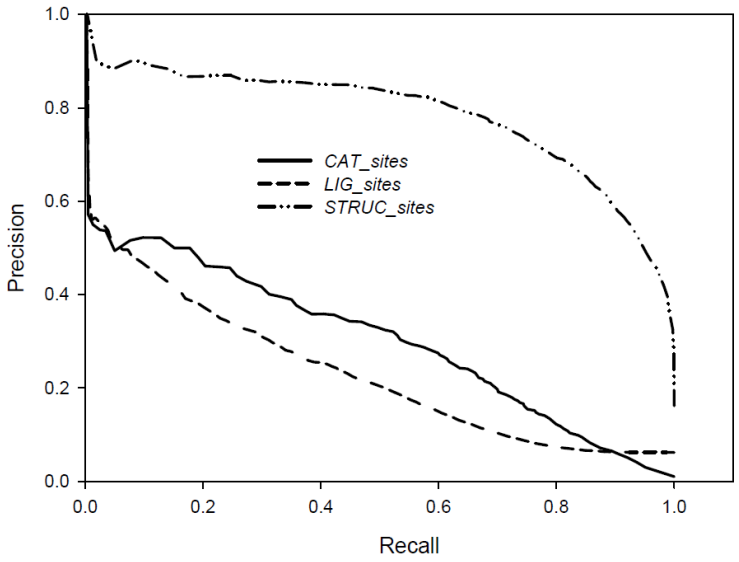
Supplementary Table 10: CLIPS-4D and *firestar* predictions for residue-positions of CASP target T0632 (pdb id 3NWZ).

The table lists all residue positions predicted by CLIPS-4D or by *firestar* (Lopez*, et al.*, 2011) as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding according to CASP specification. Column "CASP" indicates the rating according to the CASP definition of the extended ligand binding site (Schmidt*, et al.*, 2011). The column "*firestar*" lists the corresponding classification for predictions generated by *firestar* during the CASP contest.

Residue	Position	p_{CAT}	p_{LIG}	p_{STRUC}	p_{NOANN}	*p-value*	Prediction	CASP	*firestar*
I	9	0.378	0.550	0.028	0.044	0.062	LIG	TN	FP
G	10	0.084	0.865	0.037	0.014	0.009	LIG	TP	TP
A	11	0.041	0.749	0.170	0.040	0.022	LIG	FP	FP
G	12	0.400	0.559	0.028	0.013	0.076	LIG	FN	TP
P	13	0.000	0.865	0.057	0.077	0.002	LIG	TP	TP
A	14	0.018	0.893	0.075	0.014	0.005	LIG	TP	TP
V	21	0.000	0.783	0.022	0.195	0.014	LIG	FP	TN
R	24	0.739	0.131	0.121	0.009	0.016	CAT	FP	TN
F	32	0.004	0.199	0.016	0.780			TN	FP
D	33	0.919	0.048	0.028	0.005	0.002	CAT	TP	TP
N	34	0.020	0.650	0.010	0.319	0.052	LIG	FN	TP
N	35	0.009	0.275	0.061	0.655			FN	FN
T	36	0.003	0.635	0.003	0.359	0.066	LIG	FN	FN
N	37	0.109	0.458	0.177	0.255	0.124	LIG	FN	TP
R	38	0.276	0.549	0.075	0.100	0.065	LIG	FN	TP
N	39	0.059	0.719	0.203	0.019	0.032	LIG	TP	TP
V	41	0.002	0.611	0.016	0.370	0.043	LIG	FP	TN
T	42	0.013	0.663	0.179	0.145	0.058	LIG	FN	TP
S	45	0.106	0.682	0.016	0.196	0.036	LIG	TP	TP
H	46	0.477	0.406	0.038	0.080			FN	TP
G	47	0.022	0.760	0.162	0.055	0.024	LIG	FP	TN
K	60	0.103	0.541	0.027	0.328	0.045	LIG	TP	FN
K	78	0.026	0.658	0.006	0.311	0.015	LIG	TP	FN
T	79	0.018	0.520	0.005	0.457	0.100	LIG	FN	FN
V	80	0.003	0.519	0.119	0.359	0.064	LIG	FN	TP
A	109	0.262	0.608	0.111	0.019	0.041	LIG	TP	TP
T	110	0.004	0.835	0.083	0.078	0.014	LIG	TP	TP
G	111	0.142	0.664	0.165	0.029	0.047	LIG	TP	TP
Q	113	0.016	0.721	0.006	0.257	0.019	LIG	FP	TN
E	114	0.170	0.669	0.044	0.117	0.010	LIG	TP	TP
Y	126	0.324	0.503	0.095	0.079	0.093	LIG	FN	TP
G	127	0.057	0.665	0.241	0.037	0.047	LIG	FP	FP
F	131	0.012	0.800	0.052	0.136	0.012	LIG	FP	TN
S	132	0.051	0.727	0.016	0.207	0.026	LIG	FP	FP
C	133	0.296	0.632	0.031	0.041	0.019	LIG	FP	FP
Y	135	0.043	0.835	0.074	0.048	0.004	LIG	FP	TN
C	136	0.067	0.848	0.045	0.040	0.001	LIG	TP	TP
D	137	0.589	0.312	0.035	0.064			TN	FP
R	233	0.139	0.596	0.032	0.234	0.049	LIG	FP	TN
N	235	0.012	0.645	0.009	0.334	0.052	LIG	FN	FN
F	237	0.104	0.545	0.110	0.241	0.076	LIG	FN	FN
G	268	0.024	0.140	0.819	0.017	0.094	STRUC	FN	TP
E	269	0.530	0.350	0.066	0.054			FN	TP
Q	273	0.005	0.656	0.002	0.337	0.034	LIG	FP	TN
S	277	0.113	0.479	0.016	0.392	0.101	LIG	FN	TP
L	278	0.008	0.586	0.024	0.381	0.044	LIG	TP	FN
A	281	0.036	0.805	0.047	0.111	0.014	LIG	TP	TP

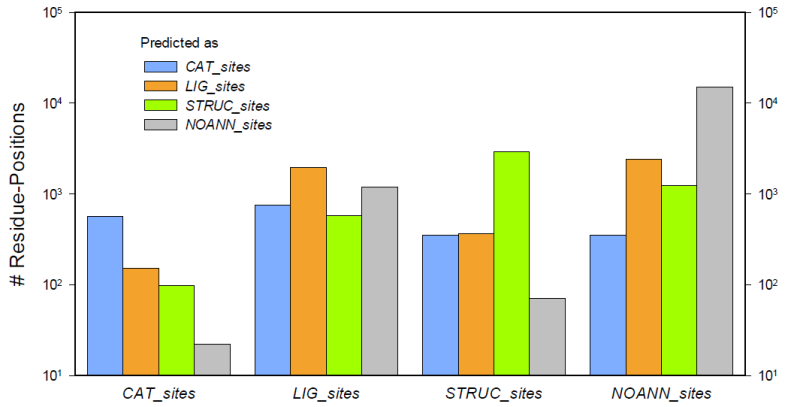
Supplementary Table 11: CLIPS-4D and *firestar* predictions for residue-positions of CASP target T0721 (pdb id 4FK1).

The table lists all residue positions predicted by CLIPS-4D or by *firestar* (Lopez, *et al.*, 2011) as being involved in ligand binding. The first two columns give the residue and its position in the pdb file. The following four columns list the probabilities for the assignment to *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* calculated by CLIPS-4D. "*p*-value" lists the *p*-value for the class with the highest probability p_{CLASS} and all residue positions with a *p*-value ≤ 0.05 for *CAT_sites* or *LIG_sites* were predicted as ligand binding according to CASP specification. Column "CASP" indicates the rating according to the CASP definition of the extended ligand binding site (Cassarino *et al.*, under review, draft version). The column "*firestar*" lists the corresponding classification for predictions generated by *firestar* during the CASP contest.



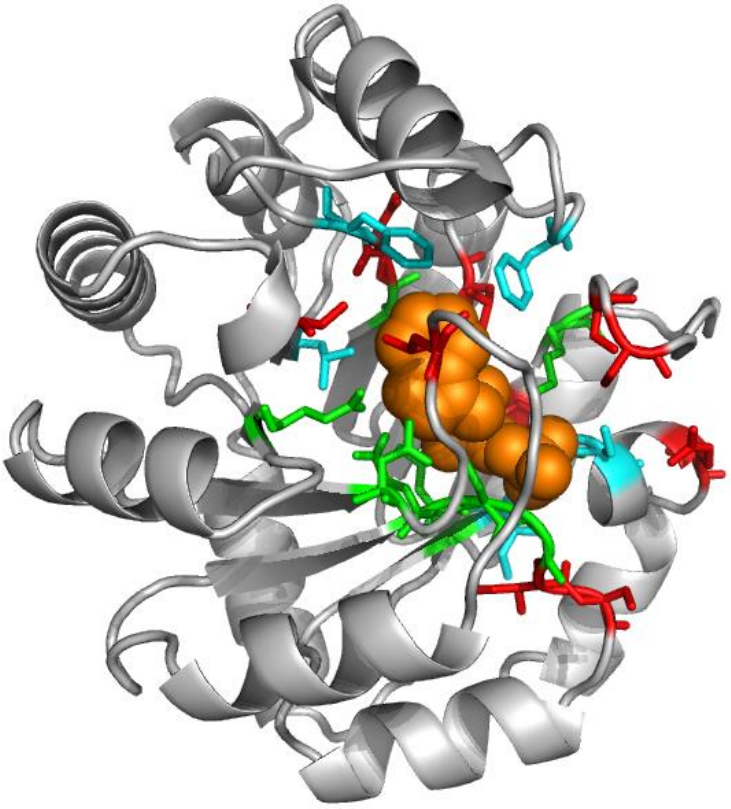
Supplementary Figure 1: Class-specific performance of CLIPS-4D.

Precision-recall curves were plotted for *CAT_sites*, *LIG_sites*, and *STRUC_sites*, respectively. The corresponding PR-AUC values are 0.30, 0.23, and 0.78, respectively.



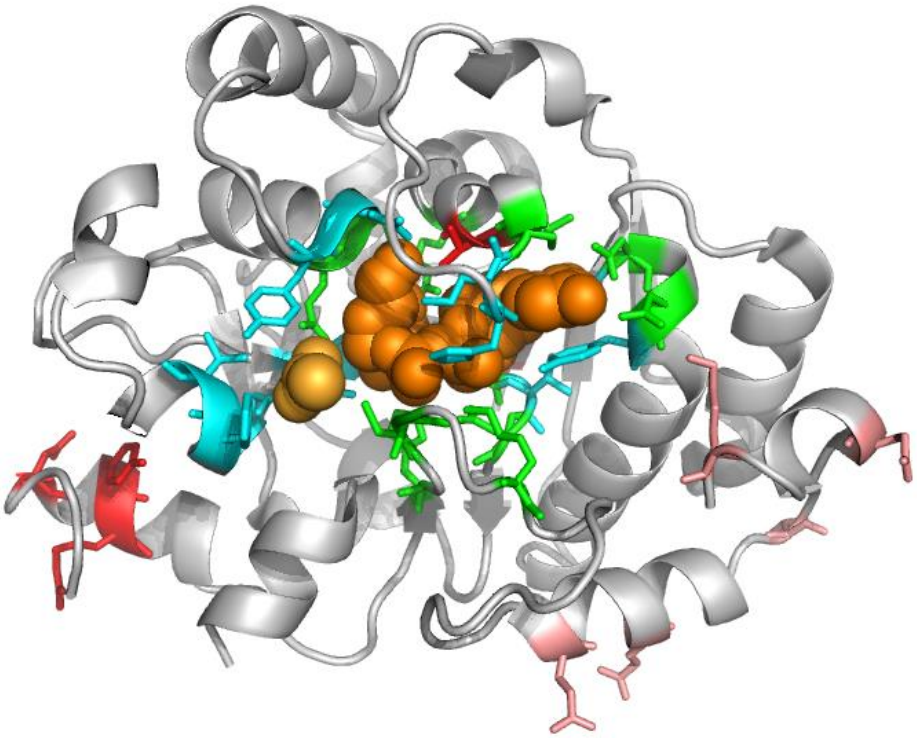
Supplementary Figure 2: Classification performance of CLIPS-4D in predicting functionally and structurally important residue-positions.

Based on the largest class-probability p_{class} all members of the classes *CAT_sites*, *LIG_sites*, *STRUC_sites*, and *NOANN_sites* were categorized. Note that the absolute numbers of residue-positions are plotted with a logarithmic scale.



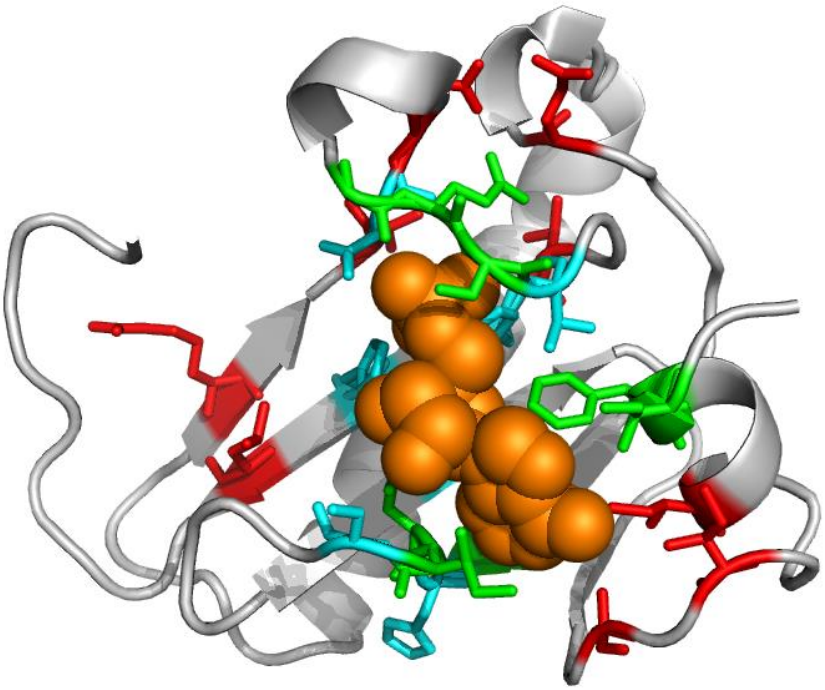
Supplementary Figure 3: CLIPS-4D prediction of ligand-binding sites in *Sulfolobus solfataricus* IGPS (pdb id 1A53).

The protein is displayed in gray (cartoon mode); the sets *CAT_sites* and *LIG_sites* were merged prior to the assessment. Residues with a distance of at most 4 Å to the product IGP (orange spheres) are positive, all other ones are negative cases. TP predictions are green, FP red, and FN are cyan.



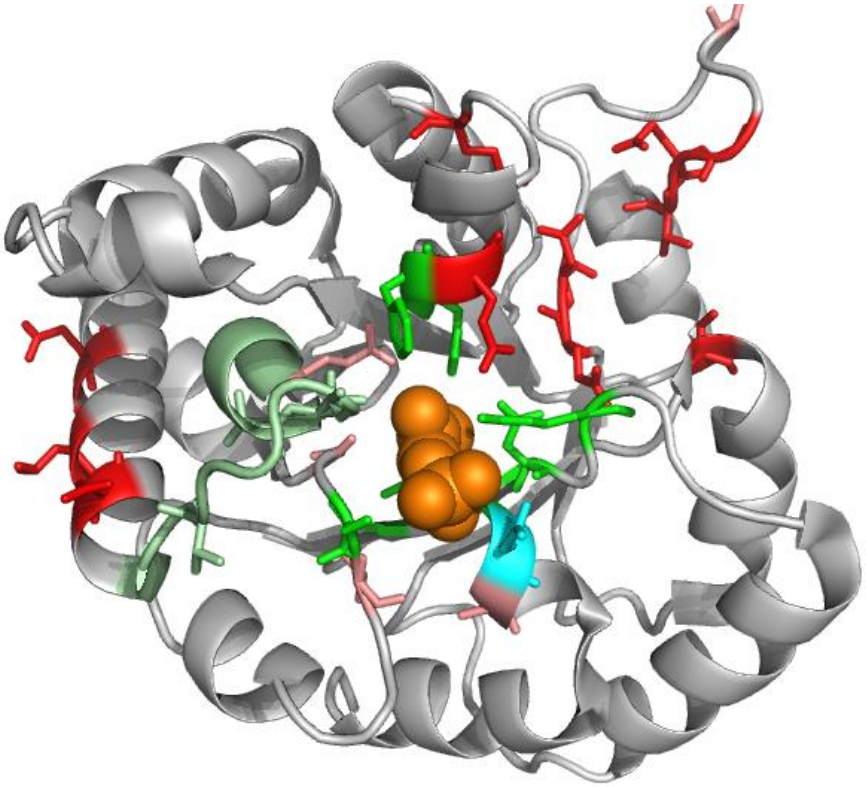
Supplementary Figure 4: CLIPS-4D prediction of substrate-binding sites in LgtC (pdb id 1G9R).

The protein is displayed in gray (cartoon mode); the sets *CAT_sites* and *LIG_sites* were merged prior to the assessment. Residues with a distance of at most 4 Å to the substrate analogs ACY (light orange spheres) and UPF (orange spheres) are positive, all other ones are negative cases. TP predictions are green, FP red, and FN are cyan. 25 C-terminal residues are missing in the crystal structure, thus the environment of residues 260 - 282 might not be the natural one; FP predictions from this region are shown in pink.



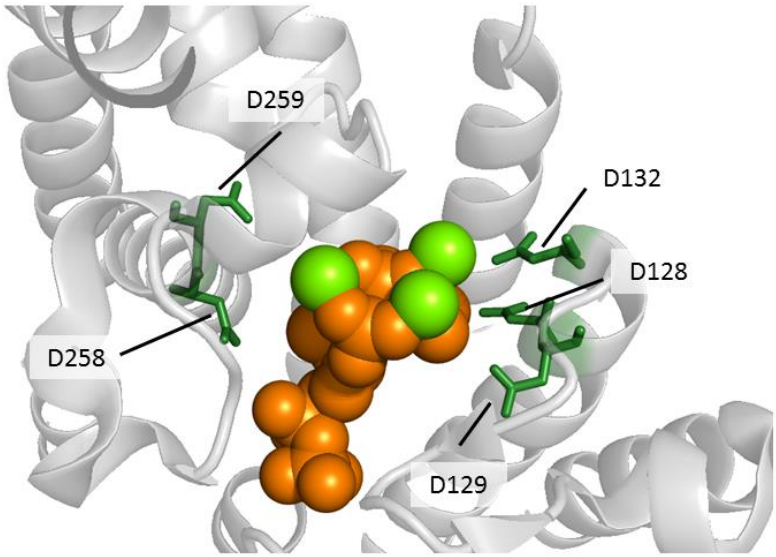
Supplementary Figure 5: CLIPS-4D prediction of substrate-binding sites in HIT protein family (pdb id 1KPF).

The protein is displayed in gray (cartoon mode); the sets *CAT_sites* and *LIG_sites* were merged prior to the assessment. Residues with a distance of at most 4 Å to the substrate analog AMP (orange spheres) are positive, all other ones are negative cases. TP predictions are green, FP red, and FN are cyan.



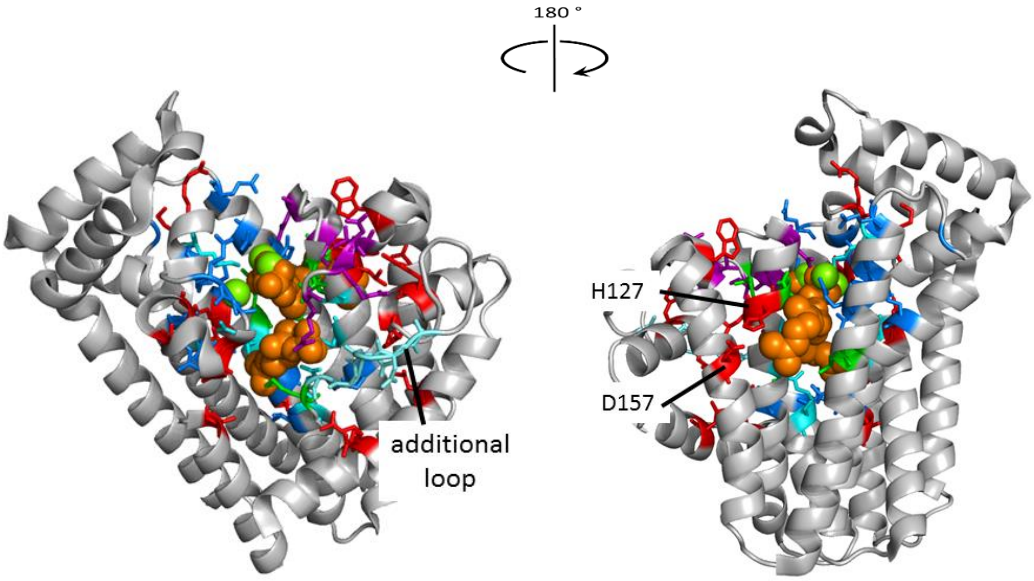
Supplementary Figure 6: CLIPS-4D prediction of substrate-binding sites in *Plasmodium falciparum* triosephosphate isomerase (pdb id 1M7P).

The protein is displayed in gray (cartoon mode); the sets *CAT_sites* and *LIG_sites* were merged prior to the assessment. Residues with a distance of at most 4 Å to the substrate analog G3H (orange spheres) are positive, all other ones are negative cases. TP predictions are green, FP red, and FN are cyan. Several FP residues are involved in catalysis or ligand binding according to (Parthasarathy*, et al.*, 2002) and are shown in a different color: Residues 166 - 176 (pale green) belong to the catalytic loop, E165 is a catalytic base, S73 provides an anchoring hydrogen bond to the ligand and residues G209-V212 are anchoring the phosphate group of the ligand. These residues are shown in pink.



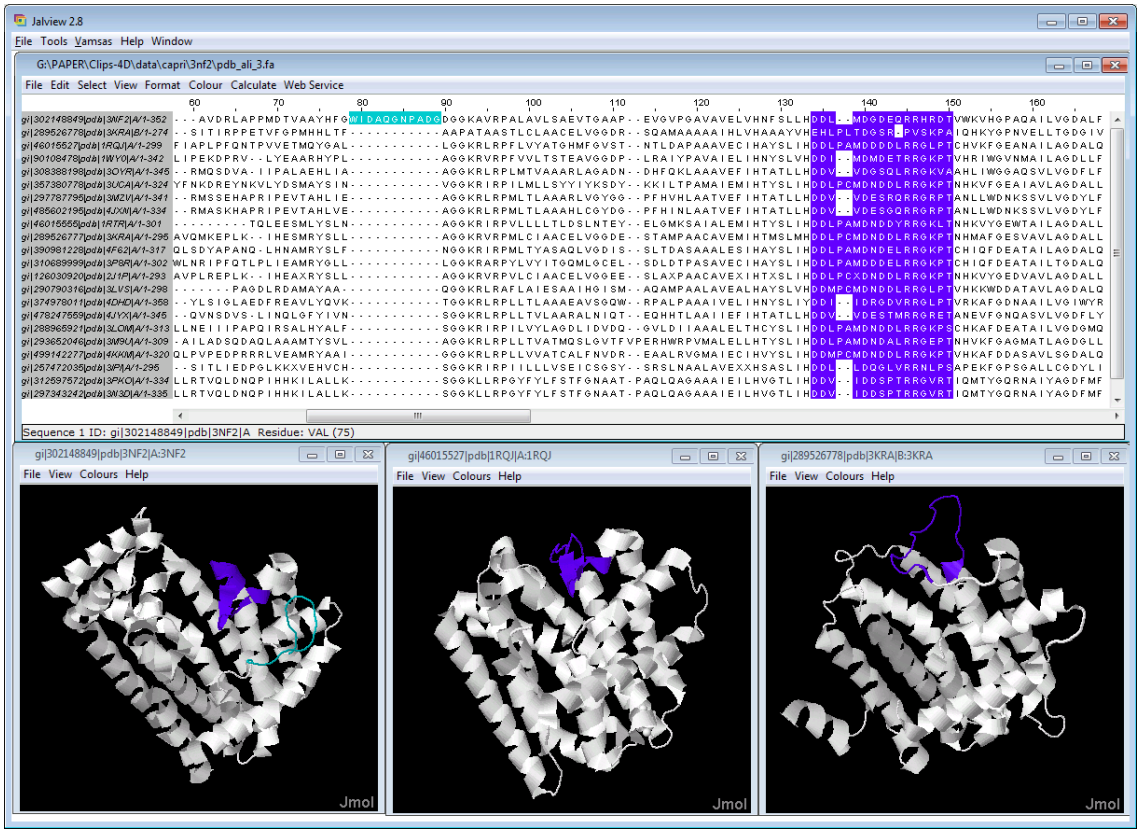
Supplementary Figure 7: CLIPS-4D prediction of *CAT_sites* for CASP target T0584 (pdb id 3NF2).

The protein is displayed in gray (cartoon mode). The five Aps-residues D128, D129, D132, D258, and D259 involved in binding DMAPP and catalysis *via* chelation of Mg^{2+} (Wallrapp*, et al.*, 2013) are shown as dark green sticks. Only D128 is a TP according to the definition of the extended ligand binding site. 3 Mg^{2+} atoms (light green) and the ligands DST and IPR (orange) are shown as spheres, their position was transferred *via* 3D superposition from the experimental control structure 1RQI by means of PyMol (Schrödinger).



Supplementary Figure 8: CLIPS-4D prediction of all ligand-binding sites for CASP target T0584 (pdb id 3NF2).

The protein is displayed in gray (cartoon mode). TP are shown in green, FP residues which line the elongation cavity as deduced from the avian geranyltransferase (Tarshis*, et al.*, 1996) are dark blue, FN residues are cyan. Left image: Residues belonging to the α4-α5 loop are purple. Six more *LIG_sites* (colored light blue) belong to a loop not found in other polyprenyl transferases, compare Supplementary Figure 5. All other FP predictions are red. 3 Mg^{2+} atoms are show as green spheres and farnesyl diphosphate, which is a C15 compound and mimics the product, is shown in orange; their orientation was transferred from 1RQI or 3AQ0 in analogy to (Wallrapp*, et al.*, 2013) by means of PyMol (Schrödinger). Right image: H127 and D157 are close to farnesyl diphosphate, but their function is unknown.



Supplementary Figure 9: Sequence and structure comparison of polyprenyl transferases.

The MSA contains sequences of polyprenyl transferases whose structure is known. The first column of the MSA (grey) lists the GI number and the pdb id of the proteins. Residues of the α4-α5 loops are shown in purple. The MSA indicates that polyprenyl synthase from *Streptomyces coelicolor* (target T0584, pdb id 3NF2, first line of the MSA and left image) contains an insert (labeled light blue) not found in other polyprenyl transferases. 1RQJ shows the active conformation of farnesyl pyrophosphate synthase from *E. coli* and 3KRA the geranyl pyrophosphate synthase from mint. The figure was generated by means of Jalview (Waterhouse *et al.*, 2009).

5.3 Publication C

H2rs: Deducing evolutionary and functionally important residue-positions by means of an entropy and similarity based analysis of multiple sequence alignments

JO Janda, A Popal, J Bauer, M Busch, M Klocke, W Spitzer, J Keller, R Merkl

BMC Bioinformatics, 15:118, 2014.

doi:10.1186/1471-2105-15-118

Janda et al. BMC Bioinformatics 2014, 15:118
http://www.biomedcentral.com/1471-2105/15/118

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

H2rs: Deducing evolutionary and functionally important residue positions by means of an entropy and similarity based analysis of multiple sequence alignments

Jan-Oliver Janda[1], Ajmal Popal[2], Jochen Bauer[2], Markus Busch[1], Michael Klocke[2], Wolfgang Spitzer[2], Jörg Keller[2] and Rainer Merkl[1*]

Abstract

Background: The identification of functionally important residue positions is an important task of computational biology. Methods of correlation analysis allow for the identification of pairs of residue positions, whose occupancy is mutually dependent due to constraints imposed by protein structure or function. A common measure assessing these dependencies is the mutual information, which is based on Shannon's information theory that utilizes probabilities only. Consequently, such approaches do not consider the similarity of residue pairs, which may degrade the algorithm's performance. One typical algorithm is H2r, which characterizes each individual residue position *k* by the *conn*(*k*)-value, which is the number of significantly correlated pairs it belongs to.

Results: To improve specificity of H2r, we developed a revised algorithm, named H2rs, which is based on the von Neumann entropy (*vNE*). To compute the corresponding mutual information, a matrix **A** is required, which assesses the similarity of residue pairs. We determined **A** by deducing substitution frequencies from contacting residue pairs observed in the homologs of 35 809 proteins, whose structure is known. In analogy to H2r, the enhanced algorithm computes a normalized *conn*(*k*)-value. Within the framework of H2rs, only statistically significant *vNE* values were considered. To decide on significance, the algorithm calculates a *p*-value by performing a randomization test for each individual pair of residue positions. The analysis of a large *in silico* testbed demonstrated that specificity and precision were higher for H2rs than for H2r and two other methods of correlation analysis. The gain in prediction quality is further confirmed by a detailed assessment of five well-studied enzymes. The outcome of H2rs and of a method that predicts contacting residue positions (PSICOV) overlapped only marginally. H2rs can be downloaded from www-bioinf.uni-regensburg.de.

Conclusions: Considering substitution frequencies for residue pairs by means of the von Neumann entropy and a *p*-value improved the success rate in identifying important residue positions. The integration of proven statistical concepts and normalization allows for an easier comparison of results obtained with different proteins. Comparing the outcome of the local method H2rs and of the global method PSICOV indicates that such methods supplement each other and have different scopes of application.

* Correspondence: rainer.merkl@ur.de
[1]Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany
Full list of author information is available at the end of the article

Background

An important objective of molecular biochemistry is a detailed analysis of protein characteristics like functionality, stability, and dynamics. This is a laborious and time consuming task due to the many aspects of protein function and the large spectrum of experimental methods required for their determination. Ideally, one would characterize experimentally the contribution of each individual amino acid residue, which is however not feasible for larger proteins. This is why the biochemical assessment of proteins has to concentrate on a relatively small number of residues. In enzymes, these are the residues directly involved in catalysis and substrate binding; resulting annotations can be found in dedicated databases like PDBsum [1]. However, there are no equivalent databases available when one has to identify residues which are important for stability or other characteristics.

Due to the enormous success of genome sequencing projects, the sequences of more than 17 000 protein families (InterPro Version 45, [2]) are known at date and thus, methods of computational biology are of utmost importance to support their characterization. A large number of *in silico* approaches are at hand to identify important residues. Often, a family-specific multiple sequence alignment (MSA) is the main data source to elucidate the role of the residues; for latest reviews see refs. [3,4]. Most effective is the assessment of residue variation deduced from the corresponding MSA columns. The success of these analyses can be explained with the biochemical properties of the residues: For example, in most cases only one residue-type fulfills all critical requirements at catalytic sites, which prohibits a mutation. Accordingly, a strict residue conservation is a strong indicator signaling functionally important residues [5-8]. In contrast, a prevalent but not exclusively found amino acid is often important for protein stability [9,10], which similarly holds for ligand-binding sites [8]. Interestingly, these less conserved residue positions may bear a pattern indicative of dependencies in the occupancy of two or more positions. The importance of these correlation signals and their consequences have long been realized [11]. Quite different approaches have been introduced to identify correlated residue pairs; see e.g. refs. [12-24]. Unfortunately, these correlation signals, which are due to constraints imposed by the local environment of a residue, can be disturbed by neutral mutations. If an MSA contains sequences from many closely related species, neutral mutations in a predecessor may give rise to a strong correlation signal. Thus, the elimination of highly similar sequences improves the quality of correlation analysis [25,26]. Additionally, other approaches have been proposed to eliminate signals induced by a common evolutionary path of the proteins [27-29].

All these methods for the analysis of correlation patterns are aimed at the identification of pairs of residues, which are functionally or structurally important. More specific methods enable us to predict residue contacts. For the latter application, transitive dependencies, which by definition interlink several pairs of residues, have to be eliminated as well [30]. Different approaches have proven applicable and these algorithms have been named global methods [4]. Among them are PSICOV [31], DCA [32], and EVfold [33]. The common idea of global methods is to treat pairs of residues as mutually dependent entities and to minimize the effects of transitive covariation and phylogenetic noise.

In contrast, most algorithms like those described in refs. [12-24,34] do not correct for transitive dependencies. These approaches have been named local methods [4] as they assume that pairs of residue positions are statistically independent of other pairs. Due to chaining effects, the identified residue positions constituting a pair, can be near to each other or far apart in the protein's structure.

Most of the local methods rely in one way or another on assessing the mutual information, which is commonly based on Shannon's entropy [35]. Thus, these local methods deduce a measure for mutual dependencies solely from the amino acid frequencies observed at the positions under study. Consequently, the biochemical properties of the residues are ignored, which may degrade the performance of the algorithm.

One of these local methods is the algorithm H2r [34], which identifies in a first step mutual dependencies between pairs of residue positions and scores in a second step each residue position k by the $conn(k)$-value, which is the number of significant pairwise correlations it is involved in. Mutagenesis studies with two enzymes demonstrated that positions with high $conn(k)$-values have an increased probability of being important for enzyme function or stability [36].

As we were interested to further improve performance of H2r in terms of specificity, we implemented H2rs, which additionally takes into account substitution frequencies for residue pairs. Moreover, H2rs determines a specific p-value for each analysis of a residue pair, which facilitates the selection of significant correlation signals. To further standardize the analyses, H2rs normalizes the resulting $conn(k)$-values to z-scores, which we named $conz(k)$-values. Using a testbed consisting of 200 enzymes, we demonstrated in a comparison with the predecessor algorithm H2r and two alternative algorithms that a larger fraction of residues endowed by H2rs with high $conz(k)$-values are located near ligand binding sites. Additionally, we studied in detail the predictions of H2r, H2rs, and the global method PSICOV for five well characterized enzymes. It turned out that the outcome of local and global methods overlapped only marginally and that residues with high $conz(k)$-values are functionally or structurally significant.

Results

Utilizing the von Neumann entropy to improve the identification of correlated mutations

A classification or regression problem can be solved optimally by means of sophisticated classifiers like support vector machines, given that positive and negative examples are at hand during training. However, there is no clear definition of a correlated mutation. This is why we cannot model the positive cases and can only characterize as precisely as possible the standard situation. Thus, to create a null model, we can deduce mean substitution frequencies for residue pairs from a large number of samples by analyzing known proteins. These substitution frequencies reflect the expected case and will allow us to identify more precisely deviations, which indicate mutual dependencies. Based on this argument, we anticipated an improvement in the identification of correlated mutations, if we additionally take into account the similarity of residue pairs together with their frequencies. Note that frequencies are the only source of information in the standard approach.

The algorithm H2r is based on Shannon's information theory [35] and computes for each pair of residue positions k, l the term $U(k, l)$ according to

$$U(k,l) = 2\frac{H(k) + H(l) - H(k,l)}{H(k) + H(l)} \quad (1)$$

Here, $H(k)$ is the entropy of an individual column k

$$H(k) = -\sum_{i=1}^{20} p\left(a_i^k\right) \ln p\left(a_i^k\right) \quad (2)$$

and $p\left(a_i^k\right)$ is the probability of amino acid a_i at position k. The entropy $H(k, l)$ of two variables (columns) k and l is

$$H(k,l) = -\sum_{i,j} p\left(a_i^k, a_j^l\right) \ln p\left(a_i^k, a_j^l\right) \quad (3)$$

and $p\left(a_i^k, a_j^l\right)$ is the probability of the amino acid pair (a_i, a_j) at positions k and l. In this context, frequency values deduced from the columns of an MSA served as estimates for probabilities.

Due to normalization, $U(k, l)$ is a more reliable indicator of co-evolution than a raw mutual information value [14]. As we were interested to improve specificity, we searched for an information theoretical concept allowing the integration of substitution frequencies determined for residue pairs.

The von Neumann entropy (*vNE*) is a generalization of the classical Shannon entropy and has been introduced in quantum statistical mechanics [37]. In computational biology, the *vNE* has been used successfully to characterize the conservation of individual residue positions [38,39]. Extending this concept to residue pairs, we aimed at a novel $U_{vNE}(k, l)$ term to replace $U(k, l)$.

The core concept of the *vNE* is the utilization of a so-called density matrix $\boldsymbol{\rho}_{k,l}$, that is, a positive definite matrix whose trace (the sum of the diagonal elements) equals to 1. $\boldsymbol{\rho}_{k,l}$ can be computed for each pair k, l according to:

$$\boldsymbol{\rho}_{k,l} = \mathbf{P}_{k,l}\,\mathbf{A}\,\mathbf{P}_{k,l} \quad (4)$$

Here, $\mathbf{P}_{k,l} = \mathrm{diag}\left(\sqrt{p_1}, \ldots, \sqrt{p_{400}}\right)$ and $p_1 \ldots p_{400}$ are the pairwise amino acid probabilities $p\left(a_i^k, a_j^l\right)$ specified in Formula (3). These probabilities satisfy the normalization condition $\sum_{i=1}^{400} p_i = 1$. $\mathbf{A}$ is a 400 × 400 matrix that assesses the similarity of residue pairs and it is this matrix that allows us to model substitutions more precisely. If $\mathbf{A}$ is equal to the identity matrix, then the *vNE* is equal to the Shannon entropy, that is, $vNE(k, l) = H(k, l)$; see below. Based on $\boldsymbol{\rho}_{k,l}$, the von Neumann entropy $vNE(k, l)$ can be calculated as

$$vNE(k,l) = vNE\left(\boldsymbol{\rho}_{k,l}\right) = -\sum_{i=1}^{400} \lambda_i \log \lambda_i \quad (5)$$

by means of the eigenvalues λ_i of $\boldsymbol{\rho}_{k,l}$. Normalization analogous to Formula (1), which reduces phylogenetic crosstalk, requires corresponding values $vNE(k)$ and $vNE(l)$. For their determination, we applied partial traces [40] on $\boldsymbol{\rho}_{k,l}$ to deduce two density matrices $\boldsymbol{\rho}_k^{k,l}$ and $\boldsymbol{\rho}_l^{k,l}$, which are specific for a pair of columns k, l. The elements of $\boldsymbol{\rho}_k^{k,l}$ and $\boldsymbol{\rho}_l^{k,l}$ were named $s_{i,j}$ and $t_{i,j}$, respectively, and were computed according to

$$s_{i,j} = \sum_{u=1}^{20} r_{20(i-1)+u,\ 20(j-1)+u} \quad (6)$$

and

$$t_{i,j} = \sum_{u=1}^{20} r_{20(u-1)+i,\ 20(u-1)+j} \quad (7)$$

where $r_{i,j}$ denotes the appropriate entry in the density matrix $\boldsymbol{\rho}_{k,l}$. Thus, this approach allows us to deduce all entropy terms from the density matrix $\boldsymbol{\rho}_{k,l}$, which eliminates normalization problems. We calculate the $vNE\left(\boldsymbol{\rho}_m^{k,l}\right)$ for the residue positions $m \in \{k, l\}$ analogously to equation (5) based on the eigenvalues λ_i of the 20 × 20 matrix $\boldsymbol{\rho}_m^{k,l}$:

$$vNE\left(\boldsymbol{\rho}_m^{k,l}\right) = -\sum_{i=1}^{20} \lambda_i \log \lambda_i \quad (8)$$

Finally, we define the normalized $U_{vNE}(k, l)$-value:

$$U_{vNE}(k,l) = \frac{vNE\left(\boldsymbol{\rho}_k^{k,l}\right) + vNE\left(\boldsymbol{\rho}_l^{k,l}\right) - vNE\left(\boldsymbol{\rho}_{k,l}\right)}{vNE\left(\boldsymbol{\rho}_k^{k,l}\right) + vNE\left(\boldsymbol{\rho}_l^{k,l}\right)} \quad (9)$$

Computing these values is straightforward, if a matrix **A** is at hand.

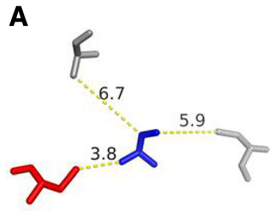
Computing a matrix A to assess the similarity of residue pairs

In the case of correlated mutations, the matrix **A** is a prerequisite to assess the similarity of residue pairs that occur in homologous proteins at corresponding positions. To determine the 400 × 400 values of **A**, we followed the concept introduced for the BLOSUM approach to score the similarity of amino acid residues based on substitution frequencies [41]. Here, we extended this concept to pairs of residues, as similarly used in P2PMAT [42]. A pre-compiled and redundancy free set of 35 809 protein 3D structures [43] offered by the PISCES server [44] was used as a representative sample. For each protein, the corresponding MSA was taken from the HSSP database [45] to deduce pairwise substitution frequencies. Based on the 3D structure, those residue pairs *k, l* were identified which contacted each other in the protein. The distances between the centers of any two heavy atoms belonging to one residue each were determined and alternatively the cut-offs 3.5 Å and 5.0 Å were chosen to select contacting pairs. These values correspond to the interval of distances used during CASP9 to identify contacts between residues and ligands [46]. For these cut-offs, we deduced 7 752 286 and 27 283 508 contacts from 15 062 205 sequences, respectively. Then, substitution frequencies were determined by analyzing the corresponding columns of the MSAs; see Figure 1 and Methods. The values of the two corresponding matrices $\mathbf{A_{3.5}}$ and $\mathbf{A_{5.0}}$ were normalized to affirm symmetry. Their comparison indicated highly similar values indicating that this distance is no critical parameter, which is in agreement with findings of CASP9 [46]. As we wanted to consider the larger number of contacts for the determination of the similarity values, we chose $\mathbf{A} = \mathbf{A_{5.0}}$ for all further computations. This matrix is available as Additional file 1.

A *p*-value for the strength of correlation signals deduced from a randomization test

Our next goal was to introduce a universally applicable statistical measure for the strength of the pairwise correlations, and we opted for a randomization test. Here, the null hypothesis is that there is no dependency in the pairwise frequencies. Thus, we can assess the strength of each pairwise correlation by shuffling the content of the two columns *k, l* under study [47]. As we shuffle the content column-wise, the entropy (conservation) of the two individual columns remains constant; however, we simultaneously degrade the putative correlation between the two residue positions. Then, we can compare the $U_{vNE}(k, l)$ value deduced from the unaltered combination of residue pairs with a distribution of $U_{vNE}(k^*, l^*)$ values resulting from many shuffling rounds. Thus, we can rate the correlation strength for this specific combination of residue pairs observed in columns *k* and *l*. Consequently, if the $U_{vNE}(k^*, l^*)$ values are similarly large or surpass the $U_{vNE}(k, l)$ value, the correlation is statistically not significant. On the other hand, if all $U_{vNE}(k^*, l^*)$ values are significantly lower, then this specific $U_{vNE}(k, l)$ value signals a pronounced dependency in the occupancy of the two residue positions, which indicates correlated mutations.

To compute this *p*-value efficiently, the number of randomized samples has to be minimized. Moreover, we need a statistical model which has to be valid, if the number of residue types is relatively small which may cause a skewed distribution. The more conserved the residue positions are, the fewer pairwise frequencies occur and the more the distribution of pairwise frequencies deviates from a normal distribution; compare Figure 2. As we wanted to assess the extremeness of the $U_{vNE}(k, l)$ values, we selected



B

	Pos k	Pos l
Seq 1	A	A
Seq 2	A	C
Seq 3	C	A
Seq 4	C	A
Seq 5	C	C

C

	AA	AC	CA	CC
AA	0	1	2	1
AC	1	0	2	1
CA	2	2	1	2
CC	1	1	2	0

Figure 1 Computation of a pairwise similarity matrix A. (A) For each residue (*k*, blue) of our dataset, all neighbors with a distance of at most 5 Å measured between the centers of heavy atoms were determined. Here, it is one residue *l* marked red. **(B)** Residue positions *k, l* were linked with the corresponding columns of the MSA and transition frequencies were deduced from a comparison of the residue pairs. **(C)** In this illustrative example, we observe one transition from AA to AC, two transitions from AA to CA and one transition from AA to CC. Transition frequencies were used to construct the 400 × 400 matrix **A** of substitution frequencies for residue pairs.

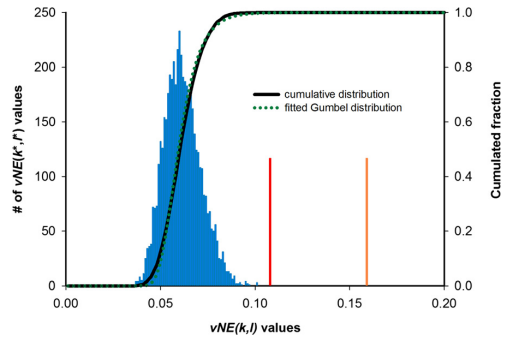


Figure 2 Distribution of U_{vNE}() values for one pair of residue positions. The histogram (blue) shows the distribution of the U_{vNE} (k^*, l^*) values of the first two residue positions of ssTrpC resulting from shuffling the content of columns k and l of the MSA. A normality test on this distribution failed (P = 0.991), which indicates that the distribution is not Gaussian. The corresponding cumulative distribution is shown in black. The cumulative *Gumbel* distribution with parameters μ and β deduced from 25 randomization tests is shown in green. The red line depicts the actual U_{vNE} value of this pair of residue positions. The orange line shows the U_{vNE} value this pair would need to surpass a p-value of 0.01.

a Gumbel distribution [48] for modeling. This distribution is specified by only two parameters μ and β that can be determined in a straightforward manner; see Methods and Formulae 12–14. To confirm that the Gumbel distribution is a proper model, we determined histograms consisting of 1000 $U_{vNE}(k^*, l^*)$ values each for all of 2 646 726 pairs of residue positions in our dataset. Prior to the computation of the next $U_{vNE}(k^*, l^*)$ value, columns were shuffled $100M$ times, where M is the number of sequences in the respective MSA. A Kolmogorov Smirnov test [49] with $\alpha = 0.01$ confirmed that the distributions of these $U_{vNE}(k^*, l^*)$ values and the deduced Gumbel distribution did not differ significantly for 99.14% of all cases. Using the same dataset, we additionally made clear that the two parameters μ and β can be estimated with adequate precision after 25 instances of randomization. Thus, to compute a specific p-value for each residue pair, it is sufficient to compute 25 $U_{vNE}(k^*, l^*)$ values and to determine one value of the fitted cumulative Gumbel distribution.

For a protein of length L, we apply this test $N = L(L + 1)/2$ times, which suggests to introduce the Bonferroni correction [50] in order to reduce the number of false positive results caused by the frequent application of the test. Thus, a corrected cut-off c_o for the corresponding p-value p is

$$c\_o(k,l) = \mu - \beta \log\left(\log\left(\frac{1}{1-p/N}\right)\right). \qquad (10)$$

$c\_o(k, l)$ allows for a statistically meaningful and content specific selection of correlated residue positions. μ and β are defined by Formulae (13) and (14); see Methods.

For the identification of correlated mutations, a p-value p has to be selected beforehand. Then, all pairs of residue positions with $U_{vNE}(k, l) \geq c\_o(k, l)$ are utilized to compute $conn(k)$-values by counting the number of significantly correlated pairs k (or analogously l) is part of. To further alleviate the comparison of different test sets, $conn(k)$-values were transformed to z-scores $conz(k)$; see Formula (15).

An *in silico* testbed for the assessment of correlation methods

The ultimate validation of a correlated mutation is a biochemical experiment, which is frequently based on the replacement of residues by the standard amino acid alanine. However, the detailed experimental analysis of a large number of mutations introduced in one protein like dihydrofolate reductase [51,52] is still the exception. This lack of reliable results impedes establishing a *bona fide* testbed for correlation methods and enforces the use of *in silico* surrogates. It is known that many correlated mutations are in close proximity to functional sites [19,47,53-55]. Thus, a testbed has been created that consists of 44 enzymes whose structure and active site residues are well characterized [54]. To assess the quality of correlation analysis, residue positions around functional sites have been counted as positives and all others as negatives [54]. To broaden the statistical basis, we compiled a non-redundant dataset of 200 enzymes, whose functional sites, i.e. catalytic and binding sites, are known and which are represented by a PDB structure and a corresponding MSA in the HSSP database; see Materials. To determine performance values, 64 575 residues were classified and the distances between van der Waals spheres were determined. We regarded all 6192 residues with a maximal distance of 1 Å to a functional site as positive cases and all other 58 383 residues as negative cases. The classification and the resulting performance depends on the chosen p-value and the cut-off for $conz(k)$. This is why we tested several combinations and summarized results in Table 1. For a p-value between 10^{-2} and 10^{-4} and a $conz(k)$-threshold of 2.0, the specificity was between 0.97 and 0.98 and precision was between 0.18 and 0.19. For the p-value 10^{-2} and the $conz(k)$-threshold of 4.0, specificity was 1.0 and precision 0.30. For p-values $\leq 10^{-5}$ and $conz(k) = 2.0$ the performance reached a plateau. The comparison with the predecessor algorithm H2r made clear that the novel algorithm performed better: Specificity and precision were up to 3% higher. Additionally, we analyzed the same dataset with the algorithms CMAT [56] and SCA [16], which predict pairs of correlated residue positions.

Table 1 Performance of four local methods deduced from an *in silico* testbed

	Cut-off	z-score	Specificity	Precision
	10^{-2}	4.0	1.00	0.30
	10^{-2}	2.0	0.97	0.18
	10^{-3}	2.0	0.97	0.18
H2rs	10^{-4}	2.0	0.98	0.19
	10^{-5}	2.0	0.98	0.18
	10^{-10}	2.0	0.98	0.17
	10^{-11}	2.0	0.98	0.17
H2r			0.95	0.17
CMAT			0.77	0.13
SCA	0.7		0.53	0.12
	1.5		0.84	0.15
	3.0		0.99	0.15

For all programs, specificity and precision were deduced from the analysis of 200 enzymes with known catalytic and binding sites. Residues with a maximal distance of 1 Å to a functional site were regarded as positives. All other residues were regarded as negatives. H2r and CMAT were used with default settings. For H2rs, the cut-off was applied to the *p*-value. For SCA, three cut-off values were chosen.

Standalone versions as of February 2014 were downloaded and applying the same criteria as above, performance was determined. CMAT was used with default parameters. For SCA, we selected three cut-off values 0.7, 1.5, and 3.0. Performance values were added to Table 1. CMAT reached a specificity of 0.77 and a precision of 0.13. For SCA, the specificity increased from 0.53 to 0.99, and the precision from 0.12 to 0.15, for the cut-offs 0.7 and 3.0. These results indicate that residue positions predicted by H2rs are more likely close to functional sites. Moreover, the number of false positives is lower, as indicated by the higher precision values determined for H2rs. These numbers are a rough estimate of the algorithm's performance due to the limitations of the *in silico* testbed. However, all other alternative methods of performance evaluation [57] are not applicable here: These are the analysis of simulated MSAs, the determination of the residues' spatial distance or an assessment of free energy differences derived from double mutants.

An assessment of predicted coevolving residues in well-characterized enzymes

To evaluate performance of our algorithm in more detail, we analyzed the H2rs predictions for five well studied enzymes: three enzymes from tryptophan biosynthesis, named TrpA, TrpB, TrpC, dihydrofolate reductase (DHFR), and hexokinase (HK). TrpA and TrpB constitute the heteromeric tryptophan synthase complex, which catalyzes the final reaction of indole-3-glycerole phosphate and serine to tryptophan. TrpA cleaves indole-3-glycerol phosphate to glyceraldehyde-3-phosphate and indole, which is transported through a hydrophobic tunnel to the active center of TrpB. There, tryptophan is synthesized from serine and indole [58]. For the localization of predicted residue positions, we utilized the 3D dataset with PDB ID 1KFC, which is the TrpA/TrpB complex from *Salmonella typhimurium* (stTrpA, stTrpB). The enzyme indole-3-glycerol phosphate synthase (TrpC) catalyzes the ring closure of an N-alkylated anthranilate to a 3-alkyl indole derivative, which is the fourth step in the tryptophan biosynthesis. It adopts the widespread $(\beta\alpha)_8$-barrel fold and has been studied in detail [59]. Here, we utilized the structure of TrpC from *Sulfolobus solfataricus* (ssTrpC, PDB ID 1A53). DHFR catalyzes the reduction of dihydrofolate to tetrahydrofolate via hydride transfer from NADPH. It has been found in most organisms and plays a critical role for cell proliferation and cell growth [60]. We utilized the structure determined for DHFR from *Escherichia coli* (ecDHFR, PDB ID 7DFR). The hexokinase from *Schistosoma mansoni* (smHK, PDB ID 1BDG) is the first enzyme in the glycolytic pathway and catalyzes the transfer of a phosphoryl group to alpha-6-glucose (GLC). The 3D crystal structure contains SO_4 anions in the catalytic cleft [61]. smHK is the only enzyme of a larger set that has been analyzed previously by correlation analysis and for which the MSA (*smHK_CMA*) was available online. To generate *smHK_CMA*, the authors have used a sophisticated protocol to merge several structure based MSAs [19].

Although local and global methods of correlation analysis have different objectives, we were interested to determine the overlap of their predictions. This is why we also compared the outcome of H2rs and PSICOV [31], which is a global method predicting residue contacts. For PSICOV we analyzed the top $L/5$ predictions, which is the recommended default for a protein sequence of length L. An MSA was created for each enzyme by using DELTA-BLAST [62] with the options `max target threshold 2000` and `expect threshold` 10^{-10}. The resulting sequences were realigned by means of MAFFT [63] in `linsi` mode. We were interested in an assessment of the most specific H2rs predictions. This is why we chose the low cut-off 10^{-11} for the *p*-value and a *conz*(k)-threshold of 2.0. To allow for a comparison, we also listed the *conz*(k)-values for all residues predicted by H2r in Table 2. Residues were regarded as functionally important, if they were close to a functional site specified in PDBsum [1]. Thus, all direct neighbors in the sequence were chosen and all residues with a 3D distance of maximally 5 Å (determined between heavy atoms).

stTrpA consists of 268 residues, and H2rs predicted two important residues, namely L100 and L127. Both residues are in close proximity to the substrate; see Figure 3. H2r predicted L100, S125, A129, I153 and L162. S125 stabilizes the inactive conformation of the

Table 2 Annotation of residue positions predicted in five enzymes as being important by H2rs and H2r

Protein	Residue	H2rs	H2r	PSICOV	Residue's role
stTrpA	L100	2.2	3.2	1	Near binding site
	S125	1.1	6.8	1	Stabilizes the active site
	L127	2.0		2	Near binding site
	A129	1.9	5.7	5	Near active site
	I153	0.9	4.6	1	Near active site
	L162	0.7	6.1	0	TrpA/TrpB interface
stTrpB	P7	1.3	6.8	0	ND
	C62	2.2	7.3	0	ND
	G83	1.8	7.2	2	Near binding site
	T88	2.4		1	Near binding site
	Q90	2.4	7.5	0	Near binding site
	V91	2.1		0	Near binding site
	L121	1.8	6.3	1	ND
	C170	4.5		4	End of substrate tunnel
	T190	2.2		6	Metal binding site
	P257	2.2	6.7	0	Near metal ion
	G268	2.3		0	Coordination of ion binding
	F280	2.4	2.8	0	End of substrate tunnel
	M282	2.6		4	Near binding site
	S297	4.2		3	Near metal ion
	F306	−0.8	5.0	0	Metal binding site
	S308	2.4	8.5	0	Metal binding site
	Q312	2.9		0	ND
ssTrpC	I48	2.4		3	ND
	A50	1.4	6.1	1	Near active site
	Y76	1.1	4.0	1	ND
	M109	1.9	4.3	2	Near active site
	I133	2.6	9.8	3	Catalytically important
	V134	2.3		2	Near active site
	I136	2.1		1	ND
	L142	2.7		1	Catalytically important
	N161	1.4	6.9	2	Near active site
	L187	1.8	4.6	1	Mutation L187A is neutral
	A209	2.1		3	Near binding site
	S234	2.1	9.5	4	Phosphate binding site
ecDHFR	A9	2.2		2	Near active site
	W30	2.3		0	Binding site
	K32	2.3		0	Binding site
	M92	3.4		0	Near active site
	G121	2.7	2.8	0	Near active site
	D144	1.9	5.1	0	ND
	H149	2.1	4.4	0	Coupled motion
smHK	T69	2.8		1	Domain interface
	A215	2.6		2	End of domain 1
	C217	2.7	13.9	0	End of domain 1
	A218	2.3		0	End of domain 1
	C224	2.2		0	Begin of domain 2
	V230	2.1		3	Near binding site
	V256	2.1		2	Domain interface
	K290	2.2		0	Near binding site
	D367	1.5	9.8	2	ND
	T409	2.4		1	Near C224
	V412	2.0		0	Near binding site

For the enzymes stTrpA, stTrpB, ssTrpC, ecDHFR, and smHK, H2r and H2r were used to identify important residue positions. For these residues, annotation was deduced from literature. The first column lists the name of the enzyme. The second column gives the residue and its position. The third column gives the *conz*(*k*)-value deduced by H2rs from all U_{vNE}()-values based on a *p*-value of 10^{-11}. The column H2r lists mean *conn*(*k*)-values resulting from 25 randomization tests. The column PSICOV lists the number of contacting pairs the residue belonged to. The last column lists the role of the residues, for details see Results. "ND" indicates that we did not find clues to the function of this residue.

active center [64]. A129 and I153 are near the active site and L162 belongs to the TrpA/TrpB interface [1]. L100 and L127 also belong to the 80 *L*/5 predictions of PSICOV; see Table 2.

For stTrpB, H2rs predicted 13 of the 397 residues as being important; see Figure 3. T88, Q90, and V91 are in close proximity to the substrate binding residue K87 [65]. C170 and F280 are located at the end of the hydrophobic tunnel [66] and T190 and S308 are metal binding sites [1]. G268 is important for the coordination of ion binding [67], and S297 and P257 are in close proximity

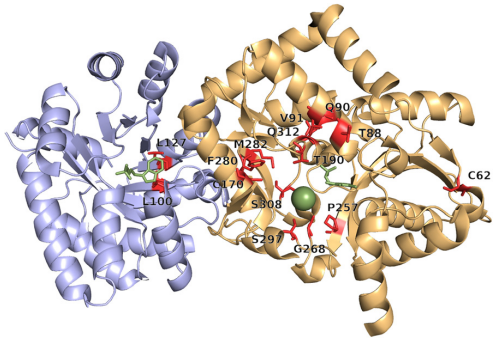


Figure 3 Residues of the stTrpA/stTrpB complex possessing highest *conz*(*k*)-values. For stTrpA (light blue) and stTrpB (gold), residues with *conz*(*k*)-values ≥ 2.0 and *p*-values ≤ 10^{-11} are plotted in red as sticks. H2rs predicted for stTrpA 2, and for stTrpB 13 important residue positions. Ligands indole-3-glycerol phosphate and pyridoxal phosphate are plotted as green sticks. The sodium ion is shown as a green ball.

to the bound sodium ion. M282 is in contact with F280 and S308; see above. The role of the two residues C62 and Q312 is unknown to us. In contrast, H2r predicted five of these residues, namely C62, Q90, P257, F280, S308, and additionally P7, G83, L121, and F306. F306 is a metal binding site, G83 is near the binding site for the substrate and the function of P7 and L121 is unknown to us. Of the 13 H2rs predictions, 5 belong to the 80 $L/5$ contacting residues predicted by PSICOV; see Table 2.

For ssTrpC, H2rs predicted 7 important positions; see Figure 4. V134 is near the active site. I133 and L142 are catalytically important: After replacing each of these two residues by alanine, the activity of TrpC dropped 30-fold [68]. A209 is located next to the substrate binding site E210 and the catalytic residue S211 [1]; S234 is known to be a phosphate binding site [1]. The role of the two residues I48 and I136 is unknown to us. H2r detected the phosphate binding site S234, the catalytically important residue I133, plus the residues A50, Y76, M109, N161, and L187. A50, M109, and N161 are near the active site. The role of L187 is unknown however, the L187A mutation has no drastic effect on function and stability [36]. The function of Y76 is unknown to us. All of the residue positions predicted by H2rs belonged to the 50 $L/5$ contacting residue pairs predicted by PSICOV; see Table 2.

For ecDHFR, H2rs predicted six important residue positions; see Figure 5. W30 and K32 are contacting the substrate, whereas A9 and M92 are in close proximity to the binding site A7 and the catalytic site I94, respectively [1]. H149 plays a significant role in the network of coupled motions required for a hydride transfer [69] and a mutation of G121, which lies in proximity of NADPH, reduced the hydride transfer rate [70]. The predecessor algorithm, H2r, identified G121, H149, plus D144, whose function is unknown to us. Of the above sites, only A9 was an element of the 32 $L/5$ predictions of PSICOV; see Table 2.

smHK consists of a HK type-1 (residues 18 – 218) and a HK type-2 domain (residues 221 – 457); see entry Q26609 of Uniprot [71]. H2rs identified 10 suspicious residues (Figure 6), which we number according to the

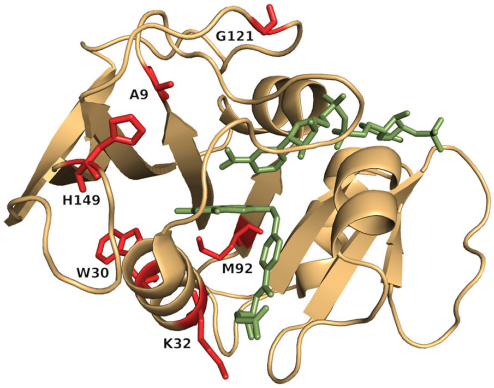


Figure 5 ecDHFR residues with highest *conz*(*k*)-values. For ecDHFR, H2rs predicted 6 residues with *conz*(*k*)-values ≥ 2.0 and *p*-values ≤ 10^{-11}, which are shown as red sticks. The ligands folic acid and NADP are shown as green sticks.

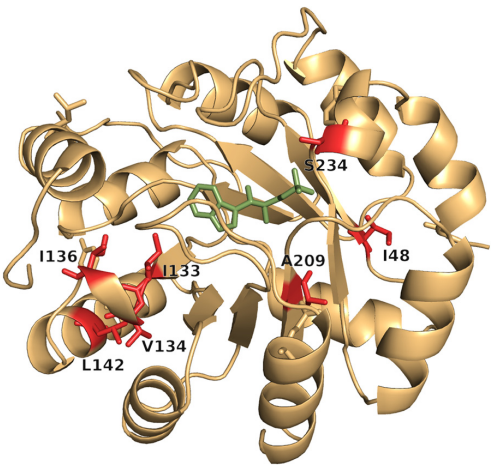


Figure 4 Residues of ssTrpC with highest *conz*(*k*)-values. For ssTrpC, H2rs identified 7 residues with *conz*(*k*)-values ≥ 2.0 and *p*-values ≤ 10^{-11}, which are shown as red sticks. The ligand indole-3-glycerol phosphate is shown as green sticks.

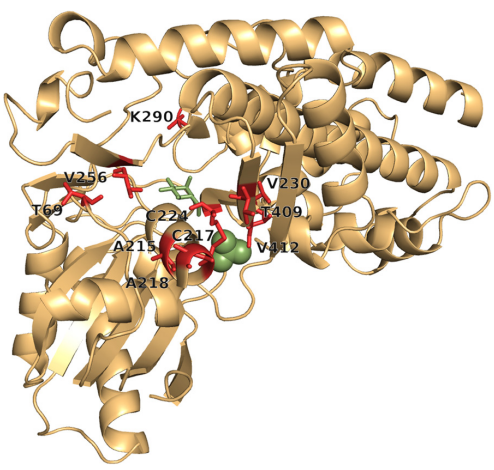


Figure 6 smHK residues with highest *conz*(*k*)-values. For smHK, H2rs predicted 10 residues with *conz*(*k*)-values ≥ 2.0 and *p*-values ≤ 10^{-11}, which are shown as red sticks. The ligand GLC is shown as green sticks and the SO_4 ion in the catalytic cleft as green balls.

PDBsum [1] entry 1BDG. A215, C217, and A218 are located at the very end of domain 1, whereas C224 occurs at the very beginning of domain 2 and these four residues are flanking a ß-turn [1]. K290 is a neighbor of Q291 that binds GLC, V230 is a neighbor of I229 (binds GLC) and of T232 (binds SO_4) [1]. V412 is a neighbor of G414 and S415 that both bind SO_4 [1]. T409 is close to C224 (see above). Only for two residues, namely T69 and V256, their role is unknown to us; however both residues are located at the domain interface at a distance of not more than 5.2 Å. H2r found C217 and additionally D376, whose function is unknown to us. 5 of the H2rs predictions were in the 91 $L/5$ predictions of PSICOV. When utilizing the MSA *smHK_CMA*, H2rs predicted only three residues with a positive *conz*(k)-value, which is given in brackets: K295 (3.0), T172 (0.71), and C217 (0.71). T172 binds GLC, and K295 is located next to the GLC binding E294 [1]. For C217, see above. Interestingly, in the 668 sequences remaining in the MSA after filtering, residue positions 217 and 224 were occupied in not more than 43% by cysteines, which form a disulfide bridge that stiffens the orientation of the two domains [1]. Alternatively, the following residue pairs were observed with more than 2% frequency: ST (12.7%), GV (7.8%), SM (6.1%), RT (5.1%), HP (2.7%), AV (2.4%) and RA (2.1%). These distinct pairwise combinations support nicely the idea of mutual dependencies and pairwise correlations.

Although the number of cases is small, these well characterized proteins allow for a more realistic assessment of the prediction performance. Altogether, H2rs predicted 38 important residues and H2r 26, respectively. False positives were 4 (11%) in the case of H2rs and 6 (23%) in the case of H2r. Thus, the resulting precision is 0.89 for H2rs and 0.77 for H2r. These results emphasize the relatively high specificity reached by computing *conn*(k)-values and additionally suggest a considerable improvement for the novel algorithm.

Discussion

H2rs is a major improvement over H2r

For all well-characterized enzymes studied in Results, H2rs predicted a larger number and a higher fraction of residue positions for which we could rationalize an important role in function or stability. Here, we concentrated on the analysis of residues with a *conz*(k)-value ≥ 2.0. Generally, this detailed analysis of five enzymes signals more precisely than the assessment of our *in silico* testbed the improved specificity of H2rs. It was achieved *i*) by replacing Shannon's entropy by the von Neumann entropy and *ii*) by integrating a more sensitive statistical approach that adapts to the composition of each pair of MSA columns. Based on this dataset, we can expect a 10% increase in specificity to nearly 90%. However, this improvement has to be paid with a much longer execution time: Computing the von Neumann entropy requires the determination of eigenvalues, which is time-consuming and the determination of p-values further increases the execution time by a factor of 25. One way of accelerating the calculation of entropy values might be an application of the Rényi entropy [72], which is a generalization of the von Neumann entropy.

For $0 < \alpha \neq 1$, the α-Rényi entropy is given by $\alpha\text{–}RE(k,l) = \frac{1}{1-\alpha} \log \sum_{i=1}^{400} \lambda_i^{\alpha}$ and for $\alpha \to 1$, we recover the Neumann entropy *vNE*(k, l). Interestingly, for $\alpha = 2$, the calculation of the α-Rényi entropy does not require the eigenvalues of the matrix $\boldsymbol{\rho}_{k,l}$ but only the diagonal entries of the square of $\boldsymbol{\rho}_{k,l}$, which drastically speeds up the computation. However, it has not been tested yet whether the Rényi entropy allows the adequate modeling of biological phenomena like residue substitutions.

Global and local methods of correlation analysis complement each other

One goal in the design of H2r, which is a local method, was the identification of individual residue positions important for protein function or stability. This is why we introduced the *conn*(k)-value. For two enzymes it has been shown that positions with high *conn*(k)-values have an increased probability of being important for enzyme function or stability [36]. The results presented here further confirm the high specificity to be gained with local methods, which is in agreement with data from the literature; see e.g. refs. [19,73]. The results obtained for smHK emphasize that not all correlated mutations are due to functional constraints: 4 of 10 residues with high *conz*(k)-values were located at the domain interface and two of them (C217, C224) belong to a disulfide bond that interlinks the domains in some of the homologous proteins. The other residue combinations observed at these two positions illustrate nicely that they were to a great extent occupied by unique residue pairs. Moreover, these findings emphasize a limitation of the *in silico* testbed. Structurally important residues often lay far apart from the catalytic center [74]. As shown above, some bear a strong correlation signal and are identified by H2rs. However, these hits are regarded as false positives and deteriorate the performance values deduced from the testbed.

Whereas local methods consider transitive correlations as well, global methods aim at eliminating these dependencies. The outcome of H2rs and the $L/5$ predictions of the global method PSICOV overlapped only for 22 of 53 residue positions; see Table 2. This result can be explained by the scope of the methods: According to the desired function, global methods identify contacting

residue pairs which are not necessarily enriched near functional sites.

Using the MSA *smHK_CMA*, H2rs predicted only three residues known to be functionally important, albeit two with low *conz*(*k*)-values. Using the same dataset, the algorithm Comulator, which aims at identifying perturbations [16], detected a network of six residue positions that surround the active site. Their occupancy almost perfectly separated the two main groups of glucokinases [19]. In summary, these findings highlight the pros and cons of the different approaches and suggest that they supplement each other quite well.

MSAs have to be prepared carefully

A critical parameter of correlation analysis is the preparation of the input, i.e. the MSA. For the prediction of intra-protein residue contacts, a strong correlation between the number of homologs and the prediction strength has been shown, which further increased, if orthologs and paralogs were included in the MSA [25]. For the sake of standardization, we used in all cases studied here the same methods of MSA preparation without human intervention. Additionally we chose identical and very rigorous cut-offs for the identification of important residue positions. This rigid protocol might be the reason for the considerably differing number of predictions: Using the cut-off $conz(k) \geq 2.0$ and a *p*-values of 10^{-11}, H2rs predicted for stTrpA only 2, but for stTrpB 13 important residue positions. These differences suggest for the user an individual adjustment of the parameters for each protein family in order to optimize the benefit of correlation analysis.

Conclusions

The various global and local methods of correlation analysis have their own field of application and supplement each other. We made plausible that residues in the vicinity of functional sites, which are a large portion of H2rs predictions, do not necessarily belong to residue pairs with the strongest global correlation signal. The predictions of global methods can be assessed by the 3D distance of the involved residue pairs. In contrast, the evaluation of local methods is more ambiguous. Due to the lack of a precise definition of a correlated mutation, it is difficult to specify positive cases. This circumstance has drastic consequences and imposes restrictions to the design and the evaluation of algorithms. With this in mind, we developed an algorithm that considers pairwise substitution frequencies and assesses the strength of the correlation signal statistically. We made plausible that *in silico* testbeds only allow for a rough performance evaluation. Favorable is the detailed analysis of well characterized model systems, which is only feasible for a small number of cases.

Methods

Similarity of amino acid pairs and density matrices

Our approach requires for the assessment of two amino acid pairs $i = (aa_n, aa_s)$ and $j = (aa_b, aa_u)$ a similarity matrix **A** of size 400 × 400 such that each entry $a_{i,j}$ gives a normalized measure for the similarity of the two pairs. To create **A**, we utilized a precompiled and redundancy free list of 35 809 PDB entries [43] offered by the PISCES server [44]. For each protein structure, we analyzed the corresponding MSA from the HSSP database [45]. These MSAs were further processed to eliminate unrelated sequences and closely related ones, which is known to improve the quality of the predictions [25]. This is why we compared for each MSA all pairs of sequences s_n, s_s and eliminated sequences s_s until all sequences contained in pairwise comparison at least 20% and not more than 90% identical residues.

Next, we determined for each protein all pairs of residue positions *k*, *l* which are close in 3D space. Distances were determined by using the BALL software library [75] and the cut-off was a maximal distance of 5.0 Å between the centers of any two heavy atoms belonging to one of the corresponding residues. Alternatively a cut-off of 3.5 Å was used. Contacting residues were mapped to the corresponding MSA columns and pairwise amino acid transitions were counted for all sequence pairs to determine substitution frequencies $f(i, j)$. We adapted a concept, which was introduced for the determination of the BLOSUM matrices [41]; see Figure 1. Each matrix element $a_{i,j}$ was normalized [38]:

$$a_{i,j} = \frac{f(i,j)}{\sqrt{f(i,i)f(j,j)}} \tag{11}$$

The result is a positive semi-definite similarity matrix **A** with $a_{i,i} = 1$ and $0 \leq a_{i,j} \leq 1$ $(i \neq j)$ elsewhere. **A** can then be used to calculate density matrices $\boldsymbol{\rho}_{k,l}$ for residue positions *k* and *l*, see Formula (4). The matrix $\boldsymbol{\rho}_{k,l}$ fulfills all requirements of being a density matrix: First, $\boldsymbol{\rho}_{k,l}$ is positive semi-definite since **A** is positive definite. Second, by the cyclicity of the trace, the trace of $\boldsymbol{\rho}_{k,l}$ equals the sum of all probabilities, which is 1 due to our normalization.

A *p*-value for the significance of pairwise correlations

In order to determine the statistical significance of correlations, we utilized a randomization test and shuffled the columns of the MSA. Consequently, the entropy at each individual position was unchanged, but the correlation between pairs of positions was randomized. Subsequently, we re-calculated a distribution *X* of U_{vNE} values *x* and repeated this process 25 times, which was sufficient to estimate the mean $\bar{x}$ and the standard deviation σ of *X* needed to approximate a Gumbel

distribution [48]. The cumulative Gumbel distribution F has the form

$$F(x,\mu,\beta) = e^{-e^{-(x-\mu)/\beta}} \quad (12)$$

and requires two parameters

$$\beta = \frac{(\sigma\sqrt{6})}{\pi} \quad (13)$$

$$\mu = \bar{x} + \gamma\beta \quad (14)$$

β and μ result from $\bar{x}$ and σ of X and γ is the Euler–Mascheroni constant (≈0.5772). Using $F(.)$, we determined a Bonferroni corrected p-value; see Formula (10).

Characterization of individual residues

In analogy to H2r, H2rs calculates a $conn(k)$-value by counting the occurrence of each residue k in the set of all significantly correlated pairs of residues. Furthermore, the $conn(k)$-values are transformed into z-scores $conz(k)$ by

$$conz(k) = \frac{conn(k) - \overline{conn(k)}}{\sigma_{conn(k)}} \quad (15)$$

where $\overline{conn(k)}$ and $\sigma_{conn(k)}$ are the mean and standard deviation of the distribution of all $conn(k)$-values > 0 determined for the protein under study.

In silico testbed and assessment of performance

To statistically evaluate algorithms, we utilized parts of the datasets *CAT_sites* and *LIG_sites* consisting of known catalytic and ligand binding sites, which we have introduced recently [76]. In short, the dataset consists of 200 non redundant PDB entries with corresponding HSSP MSAs [45], each containing at least 125 sequences. Functional sites were identified by means of annotations from the literature entries of the catalytic site atlas [77] and binding site annotations from the PDBsum database [1]. All residues within a maximal distance of 1 Å to a functional site were taken as positives, all other residues as negatives. Subsequently, we determined specificity, and precision:

$$Specificity = \frac{TN}{TN + FP} \quad (16)$$

$$Precision = \frac{TP}{TP + FP} \quad (17)$$

In both 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.

Additional file

Additional file 1: Similarity Matrix A. Format Excel. The file contains raw substitution frequencies and normalized values.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JOJ: Implemented and validated the algorithm H2rs and wrote a first draft of the manuscript. AP deduced the matrix **A**. JB implemented and assessed the algorithm for the computation of the p-value. MB was involved in implementing the testbed and determined the performance of CMAT and SCA. MK, WS, and JK designed and assessed the method to compute the $U_{vNE}(k, l)$-values. RM conceived of and managed the project and wrote the final version of the manuscript. All authors read and approved the final version.

Acknowledgements
This work was supported by the Deutsche Forschungsgemeinschaft within the priority program SPP 1395 (ME 2259/1-1).

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

Received: 13 January 2014 Accepted: 17 April 2014
Published: 27 April 2014

References

1. Laskowski RA, Chistyakov VV, Thornton JM: **PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.** *Nucleic Acids Res* 2005, **33**(Database issue):D266–D268.
2. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J: **New developments in the InterPro database.** *Nucleic Acids Res* 2007, **35**(Database issue):D224–228.
3. de Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution.** *Nat Rev Genet* 2013, **14**(4):249–261.
4. Marks DS, Hopf TA, Sander C: **Protein structure prediction from sequence variation.** *Nat Biotechnol* 2012, **30**(11):1072–1080.
5. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics* 2001, **17**(8):700–712.
6. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics* 2007, **23**(15):1875–1882.
7. Wang K, Samudrala R: **Incorporating background frequency improves entropy-based residue conservation measures.** *BMC Bioinformatics* 2006, **7**:385.
8. Janda JO, Busch M, Kuck F, Porfenenko M, Merkl R: **CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure.** *BMC Bioinformatics* 2012, **13**:55.
9. Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M: **The consensus concept for thermostability engineering of proteins: further proof of concept.** *Prot Eng* 2002, **15**(5):403–411.
10. Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, Wong S, Gualfetti P, Schellenberger V: **Construction of stabilized proteins by combinatorial consensus mutagenesis.** *Protein Eng Des Sel* 2004, **17**(11):787–793.
11. Altschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.** *J Mol Biol* 1987, **193**(4):693–707.
12. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.** *Mol Biol Evol* 2000, **17**(1):164–178.
13. Neher E: **How frequent are correlated changes in families of protein sequences?** *Proc Natl Acad Sci U S A* 1994, **91**(1):98–102.
14. Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using information theory to search for co-evolving residues in proteins.** *Bioinformatics* 2005, **21**(22):4116–4124.
15. Larson SM, Di Nardo AA, Davidson AR: **Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions.** *J Mol Biol* 2000, **303**(3):433–446.

16. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286**(5438):295–299.
17. Dekker JP, Fodor A, Aldrich RW, Yellen G: **A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.** *Bioinformatics* 2004, **20**(10):1565–1572.
18. Kass I, Horovitz A: **Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations.** *Proteins* 2002, **48**(4):611–617.
19. Kuipers RK, Joosten HJ, Verwiel E, Paans S, Akerboom J, van der Oost J, Leferink NG, van Berkel WJ, Vriend G, Schaap PJ: **Correlated mutation analyses on super-family alignments reveal functionally important residues.** *Proteins* 2009, **76**(3):608–616.
20. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins* 1994, **18**(4):309–317.
21. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein-protein interaction.** *J Mol Biol* 1997, **271**(4):511–523.
22. Halperin I, Wolfson H, Nussinov R: **Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.** *Proteins* 2006, **63**(4):832–845.
23. Singer MS, Vriend G, Bywater RP: **Prediction of protein residue contacts with a PDB-derived likelihood matrix.** *Protein Eng* 2002, **15**(9):721–725.
24. Lichtarge O, Yao H, Kristensen DM, Madabushi S, Mihalek I: **Accurate and scalable identification of functional sites by evolutionary tracing.** *J Struct Funct Genomics* 2003, **4**(2–3):159–166.
25. Ashkenazy H, Unger R, Kliger Y: **Optimal data collection for correlated mutation analysis.** *Proteins* 2009, **74**(3):545–555.
26. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24**(3):333–340.
27. Tillier ER, Lui TW: **Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments.** *Bioinformatics* 2003, **19**(6):750–755.
28. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C: **MISTIC: Mutual information server to infer coevolution.** *Nucleic Acids Res* 2013, **41**(Web Server issue):W8–W14.
29. Gültas M, Haubrock M, Tüysüz N, Waack S: **Coupled mutation finder: a new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations.** *BMC Bioinformatics* 2012, **13:**225.
30. Burger L, van Nimwegen E: **Disentangling direct from indirect co-evolution of residues in protein alignments.** *PLoS Comp Biol* 2010, **6**(1):e1000633.
31. Jones DT, Buchan DW, Cozzetto D, Pontil M: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.** *Bioinformatics* 2012, **28**(2):184–190.
32. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein-protein interaction by message passing.** *Proc Natl Acad Sci U S A* 2009, **106**(1):67–72.
33. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C: **Protein 3D structure computed from evolutionary sequence variation.** *PLoS One* 2011, **6**(12):e28766.
34. Merkl R, Zwick M: **H2r: identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments.** *BMC Bioinformatics* 2008, **9:**151.
35. Shannon C: **A mathematical theory of communication.** *Bell Syst Technical J* 1948, **27:**379–423.
36. Dietrich S, Borst N, Schlee S, Schneider D, Janda JO, Sterner R, Merkl R: **Experimental assessment of the importance of amino acid positions identified by an entropy-based correlation analysis of multiple-sequence alignments.** *Biochemistry* 2012, **51**(28):5633–5641.
37. von Neumann J: *Mathematical Foundations of Quantum Mechanics.* Princton: Princeton University Press; 1996.
38. Johansson F, Toh H: **Relative von Neumann entropy for evaluating amino acid conservation.** *J Bioinform Comput Biol* 2010, **8**(5):809–823.
39. Zhang SW, Zhang YL, Pan Q, Cheng YM, Chou KC: **Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach.** *Amino Acids* 2008, **35**(2):495–501.
40. Messiah A: *Quantum mechanics.* Dover: Dover Publications; 1999.
41. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci U S A* 1992, **89**(22):10915–10919.
42. Eyal E, Frenkel-Morgenstern M, Sobolev V, Pietrokovski S: **A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction.** *Proteins* 2007, **67**(1):142–153.
43. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank. A computer-based archival file for macromolecular structures.** *Eur J Biochem* 1977, **80**(2):319–324.
44. Wang G, Dunbrack RL Jr: **PISCES: recent improvements to a PDB sequence culling server.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W94–W98.
45. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**(1):56–68.
46. Schmidt T, Haas J, Gallo Cassarino T, Schwede T: **Assessment of ligand-binding residue predictions in CASP9.** *Proteins* 2011, **79**(Suppl 10):126–136.
47. Proctor EA, Kota P, Demarest SJ, Caravella JA, Dokholyan NV: **Highly covarying residues have a functional role in antibody constant domains.** *Proteins* 2013, **81**(5):884–895.
48. Gumbel EJ: *Statistics of Extremes.* New York: Columbia University Press; 1958.
49. Smirnov N: **Table for estimating the goodness of fit of empirical distributions.** *Ann Math Stat* 1948, **19:**279–281.
50. Dunn OJ: **Multiple comparisons among means.** *J Am Stat Assoc* 1961, **56**(293):52–64.
51. Rod TH, Radkiewicz JL, Brooks CL 3rd: **Correlated motion and the effect of distal mutations in dihydrofolate reductase.** *Proc Natl Acad Sci U S A* 2003, **100**(12):6980–6985.
52. Balog E, Perahia D, Smith JC, Merzel F: **Vibrational softening of a protein on ligand binding.** *J Phys Chem B* 2011, **115**(21):6811–6817.
53. Travers SA, Fares MA: **Functional coevolutionary networks of the Hsp70-Hop-Hsp90 system revealed through computational analyses.** *Mol Biol Evol* 2007, **24**(4):1032–1044.
54. Lee BC, Park K, Kim D: **Analysis of the residue-residue coevolution network and the functionally important residues in proteins.** *Proteins* 2008, **72**(3):863–872.
55. Wang ZO, Pollock DD: **Coevolutionary patterns in cytochrome c oxidase subunit I depend on structural and functional context.** *J Mol Evol* 2007, **65**(5):485–495.
56. Jeong CS, Kim D: **Reliable and robust detection of coevolving protein residues.** *Protein Eng Des Sel* 2012, **25**(11):705–713.
57. Xu H, Li X, Zhang Z, Song J: **Identifying coevolution between amino acid residues in protein families: advances in the improvement and evaluation of correlated mutation algorithms**. In *Current Bioinformatics*, Volume 8. Bentham Science Publishers Ltd. Netherlands; 2013:148–160.
58. Weber-Ban E, Hur O, Bagwell C, Banik U, Yang LH, Miles EW, Dunn MF: **Investigation of allosteric linkages in the regulation of tryptophan synthase: the roles of salt bridges and monovalent cations probed by site-directed mutation, optical spectroscopy, and kinetics.** *Biochemistry* 2001, **40**(12):3497–3511.
59. Schneider B, Knöchel T, Darimont B, Hennig M, Dietrich S, Babinger K, Kirschner K, Sterner R: **Role of the N-terminal extension of the $(\beta\alpha)_8$-barrel enzyme indole-3-glycerol phosphate synthase for its fold, stability, and catalytic activity.** *Biochemistry* 2005, **44**(50):16405–16412.
60. Baccanari D, Phillips A, Smith S, Sinski D, Burchall J: **Purification and properties of *Escherichia coli* dihydrofolate reductase.** *Biochemistry* 1975, **14**(24):5267–5273.
61. Kuser PR, Krauchenco S, Antunes OA, Polikarpov I: **The high resolution crystal structure of yeast hexokinase PII with the correct primary sequence provides new insights into its mechanism of action.** *J Biol Chem* 2000, **275**(27):20814–20821.
62. Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL: **Domain enhanced lookup time accelerated BLAST.** *Biol Direct* 2012, **7:**12.
63. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.** *Mol Biol Evol* 2013, **30**(4):772–780.
64. Kulik V, Hartmann E, Weyand M, Frey M, Gierl A, Niks D, Dunn MF, Schlichting I: **On the structural basis of the catalytic mechanism and the regulation of the alpha subunit of tryptophan synthase from *Salmonella typhimurium* and BX1 from maize, two evolutionarily related enzymes.** *J Mol Biol* 2005, **352**(3):608–620.
65. Miles EW, Kawasaki H, Ahmed SA, Morita H, Morita H, Nagata S: **The beta subunit of tryptophan synthase. Clarification of the roles of histidine 86,**

lysine 87, arginine 148, cysteine 170, and cysteine 230. *J Biol Chem* 1989, **264**(11):6280–6287.
66. Ruvinov SB, Yang XJ, Parris KD, Banik U, Ahmed SA, Miles EW, Sackett DL: **Ligand-mediated changes in the tryptophan synthase indole tunnel probed by nile red fluorescence with wild type, mutant, and chemically modified enzymes.** *J Biol Chem* 1995, **270**(11):6357–6369.
67. Rhee S, Parris KD, Ahmed SA, Miles EW, Davies DR: **Exchange of K^+ or Cs^+ for Na^+ induces local and long-range changes in the three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ complex.** *Biochemistry* 1996, **35**(13):4211–4221.
68. Dietrich S: **Mutationsanalyse und kinetische Untersuchungen zum Reaktionsmechanismus der Indolglycerinphosphat-Synthase aus *Solfolobus solfataricus*.** *PhD thesis.* University of Regensburg, Biochemistry II; 2010.
69. Watney JB, Hammes-Schiffer S: **Comparison of coupled motions in *Escherichia coli* and *Bacillus subtilis* dihydrofolate reductase.** *J Phys Chem B* 2006, **110**(20):10130–10138.
70. Thorpe IF, Brooks CL 3rd: **The coupling of structural fluctuations to hydride transfer in dihydrofolate reductase.** *Proteins* 2004, **57**(3):444–457.
71. UniProt C: **Update on activities at the Universal Protein Resource (UniProt) in 2013.** *Nucleic Acids Res* 2013, **41**(Database issue):D43–D47.
72. Rényi A: **On measures of information and entropy.** In *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960*; 1961:547–561.
73. Teppa E, Wilkins AD, Nielsen M, Buslje CM: **Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction.** *BMC Bioinformatics* 2012, **13**(1):235.
74. Wierenga RK: **The TIM-barrel fold: a versatile framework for efficient enzymes.** *FEBS Lett* 2001, **492**(3):193–198.
75. Hildebrandt A, Dehof AK, Rurainski A, Bertsch A, Schumann M, Toussaint NC, Moll A, Stöckel D, Nickels S, Mueller SC, Hildebrandt A, Dehof AK, Rurainski A, Bertsch A, Schumann M, Toussaint NC, Moll A, Stöckel D, Nickels S, Mueller SC, Lenhof HP, Kohlbacher O: **BALL-biochemical algorithms library 1.3.** *BMC Bioinformatics* 2010, **11**:531.
76. Janda JO, Meier A, Merkl R: **CLIPS-4D: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3D data.** *Bioinformatics* 2013, **29**(23):3029–3035.
77. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32**(Database issue):D129–D133.

doi:10.1186/1471-2105-15-118
Cite this article as: Janda *et al.*: H2rs: Deducing evolutionary and functionally important residue positions by means of an entropy and similarity based analysis of multiple sequence alignments. *BMC Bioinformatics* 2014 **15**:118.

6 Acknowledgement

At this point I would like to thank everyone that helped finishing this thesis. Special thanks goes to my supervisor Prof. Dr. Rainer Merkl, who has been always there to listen and give advice.
I am grateful to Prof. Dr. Reinhard Sterner and Prof. Dr. Stephan Waack for their expertise and advice as my mentors.

I would also like to thank all the current and former colleagues at the University of Regensburg, for their support and for some much needed humor and entertainment. In particular, I would like to thank Patrick Löffler and Dietmar Birzer for interesting discussions and Linux support, as well as Hermann Zellner for advanced Linux support.

Thanks also go to the former students Ajmal Popal, Jochen Bauer, Thomas Beisiegel, Markus Busch, Mareike Lück, Clemens Zvacek, Florian Kück, Mikhail Porfenenko, Andreas Meier, Michael Klocke, Andre Seidenspinner, and Benjamin Gathmann.

Special thanks go to my family and friends for their support. Finally, I would like to thank Anja for her faith in me and providing me with unending encouragement.