Published in final edited form as:

J Mol Biol. 2011 March 11; 406(5): 713–729. doi:10.1016/j.jmb.2010.12.005.

Structural Signatures of Enzyme Binding Pockets from Order-Independent Surface Alignment: A Study of Metalloendopeptidase and NAD Binding Proteins

Joe Dundas, Larisa Adamian, and Jie Liang

Department of Bioengineering, University of Illinois at Chicago, 835 S. Wolcott, Chicago, Illinois, 60612

Joe Dundas: jdunda1@uic.edu; Larisa Adamian: larisa@uic.edu; Jie Liang: jliang@uic.edu

Abstract

Detecting similarities between local binding surfaces can facilitate identification of enzyme binding sites, prediction of enzyme functions, as well as aid in our understanding of enzyme mechanisms. A challenging task is to construct a template of local surface characteristics for a specific enzyme function or binding activity, as the size and shape of binding surfaces of a biochemical function often varies. Here we introduce the concept of signature binding pockets, which captures information about preserved and varied atomic positions at multi-resolution levels. For proteins with complex enzyme binding and activity, multiple signatures arise naturally in our model, which form a signature basis set that characterize this class of proteins. Both signatures and signature basis set can be automatically constructed by a method called SOLAR (Signature Of Local Active Regions). This method is based on a sequence order independent alignment of computed binding surface pockets. SOLAR also provides a structure based multiple sequence fragment alignment (MSFA) to facilitate interpretation of computed signatures. For studying a family of evolutionary related proteins, we show that for metzincin metalloendopeptidase, which has a broad spectrum of substrate binding, signature and basis set pockets can be used to discriminate metzincins from other enzymes, to predict the subclass of enzyme functions, and to identify the specific binding surfaces. For studying unrelated proteins which have evolved to bind to the same NAD co-factor, signatures of NAD binding pockets can be constructed and can be used to predict NAD binding proteins and to locate NAD binding pockets. By measuring preservation ratio and location variation, our method can identify residues and atoms important for binding affinity and specificity. In both cases, we show that signatures and signature basis set reveal significant biological insight.

Keywords

functional pockets; signature pockets; signature basis set; metalloendopeptidase; NAD binding proteins

© 2011 Elsevier Ltd. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

A widely used method for inferring protein function is to transfer functional information based on homology analysis of shared characteristics between proteins. If a protein shares a high level of sequence identity to a well characterized family of proteins, frequently the biological functions of the family can frequently be accurately transferred onto that protein^{1–3}. However, limitations to sequence-based homology transfer for function prediction arise when sequence identity between a pair of proteins is less than 60%⁴. An alternative to sequence analysis is to infer protein functions based on structural similarity, as protein structure and protein function are strongly correlated⁵. It is now well known that protein structures are much more conserved than protein sequences, and proteins with little sequence identity often fold into similar three-dimensional structures^{6,7}.

Although comparison of global structural fold may offer insight into remote and complex evolutionary relationships⁸, the overall similarity in three dimensional structure is a poor predictor of protein function^{9–13}, as the structure-function relationship of proteins may be continuous and not restricted by overall fold¹⁴.

In order to obtain an accurate assessment of protein function, a number of emerging studies have shown that detecting similarity of local surface where substrate binding occurs can be very effective^{14–27}. One method for predicting protein function based on local surface similarities is to structurally compare a surface pocket from the query protein against a database of template surface pockets from proteins with known functions. If a certain level of similarity is detected, the function of the query protein can be inferred from that of the template protein^{28,29}.

To facilitate such detections, several methods have been developed to construct local structural templates or motifs, which capture structural characteristics of a class of proteins of a specific function. A well-known method is TESS²³, which uses the geometric hashing technique³⁰ to create a database of enzyme active sites³¹. Based on a graph representation of the protein structure, Huan *et.al.* developed a frequent subgraph mining algorithm to identify isomorphic subgraphs as spatial motifs common in an enzyme family ($\geq 80\%$) and absent ($\leq 5\%$) in other proteins²⁶. Goyal *et.al.* are able to create metal binding structural templates consisting of three to four residues based on similarities by several geometrical measures. CONSURF²⁴ is another well-known method that uses conservation scores derived from a phylogenetic tree of homologous sequences, which are then mapped onto the Protein Data Bank (PDB) structure. Shatky *et al* formulated the problem of binding pattern detection as that of the multiple common point set problem, and developed a branch-and-bound algorithm³². In a recent study using the PEVO_{SOAR} method^{28,33}, representative local surface pockets from protein structures of similar function are used to predict functions of uncharacterized protein structures^{33,34}. When further combined with estimated amino acid residue substitution patterns that solely reflect the selection pressure experienced by binding surfaces, enzyme functions can be predicted across 100 different enzyme families²⁹.

Defining a representative template of local surfaces for a specific functional class of proteins is a challenging task. Local structural motifs are often constructed using only a few spatially conserved residues that are likely to be functionally important, as in the well known example of the catalytic triad^{23,35}. These structural templates have been found very useful³⁶. However, when querying a small template against a large number of protein structures, false positives often result, as the small size of the template may not contain sufficient discriminating information³⁷. This problem is exacerbated in a database search, in which a large number of protein structures need to be queried against. Since the small template consists of only a few residues, too many unrelated protein surfaces may have strong

similarity by random chance. Another problem with small spatial motifs is that important structural information such as the overall shape of the binding pocket and the full physicochemical nature of the microenvironment of the binding surface is not reflected.

On the other hand, if too many residues are included in a local structural template for a functional class of proteins, an overall loss of sensitivity may result, namely, many proteins of similar function may go undetected. Using spherical harmonic expansions, Kahraman *et al* compared the shape of bound ligand molecule and the shape of the binding pocket, and found that binding pockets often are more variable in their shapes than the bound ligand³⁸. These authors also pointed out that the overall shape of the binding pocket itself is not sufficiently informative, as the binding surfaces may experience significant changes when flexible ligands are encountered. Therefore, insisting on matching a template of a full local binding surfaces can be problematic, as not all of the residues that make up one binding surface are always present in another binding surface. If binding regions experience conformational change, binding surfaces with similar function but with some residues in different spatial configurations will not be detected using a fixed template. Tseng *et al* found that two conformationally different templates are necessary to predict the function and to identify binding surfaces for 97 known structures of α -amylase²⁹.

In this study, we describe a computational method that automatically generate structural templates of local surfaces, called *signature pockets*, for an enzyme function or for a binding activity from known structures, in an effort to reconcile the requirements of including broad structural information without losing discriminating ability. A signature pocket is derived from optimal alignment of precomputed surface pockets in a sequence-order-independent fashion, in which atoms and residues are aligned based on their spatial correspondence when maximal similarity is obtained, regardless how they are ordered in the underlying primary sequences. Our method, called SOLAR (for Signatures Of Local Active Regions), does not require the atoms of the signature pocket to be present in all member structures. Instead, signature pockets can be created at varying degrees of partial structural similarity, and can be organized hierarchically at different level of binding surface similarity. Our method can automatically construct a minimal set of signature pockets, which further form a *basis set* of signatures for a specific enzyme function that may have complex binding activities. This basis set can represent many possible shapes and chemical textures of functional pockets of an enzyme class seen in known structures. It can be used to accurately predict enzymes function.

We study two problems using SOLAR. To characterize functional surfaces of enzymes with broad spectrum of substrate binding and catalytic activities, we study the family of metzincin metalloendopeptidase. To characterize proteins of unrelated evolutionary origin but converged to bind to the same cofactor, we study the NAD binding enzymes. For metalloendopeptidase, we first give an overview of the structural features of the active site pocket. We then demonstrate that key structural determinants can be automatically extracted into signatures and basis sets and can be used for automated classification of metalloendopeptidase, with results largely coinciding with known biological knowledge reported in the literature^{39,40}. We further show that signature pockets and the basis set can be used to predict the function of new metalloendopeptidase enzymes. For proteins of diverse evolutionary origin but bind to the same co-factor, we study NAD binding enzymes. We find that two basis sets of signature pockets can be automatically generated, with one for extended NAD conformations, and another for compact conformations. These signatures and basis sets then can be used to locate the NAD binding surfaces and predict NAD binding proteins.

2. Results

We study general issues of how to obtain signature pockets automatically and how they can be used for detecting binding surface and for predicting enzyme function. We use metalloendopeptidase as our example.

2.1. Metzincin metalloendopeptidase

The metzincin family of metalloendopeptidase is responsible for degradation of a wide range of peptide targets. They are found in all kingdoms of living organisms and are commonly involved in direct extracellular matrix metabolism and cell proliferation. Depending on the identity of the Z residue in a conserved catalytic motif (HEXXHXXGXXHZ), metzincins are further divided into 6 subclasses: serralsins, astacins, adamalysins, snapalysins, leishmanolysins, and matrix metalloproteinase (MMPs)⁴¹. Among these, the MMP subclasses is diverse and contains the largest number of members. Based on substrate specificity and cellular localization, MMPs are commonly grouped into five subfamilies⁴⁰: collagenases (MMPs-1,8,13, and 18), gelatinases (MMPs-2 and 9), membrane anchored MMPs (MMP-14, 15, 16, 17, 23a, 23b, 24 and 25), stromelysins (MMP-3, 10, 11, 7, and 26), and other MMPs (MMP-12, 19, 20, 21, 22, 27, and 28). Nevertheless, such a grouping is subjective, as there exists significant substrate binding promiscuity and cross-reactivity among the MMPs. For example, aggrecan, collagen I-XI, and collagen XIV can all be cleaved by multiple MMPs, That is, each can be cleaved by a different but overlapping set of MMPs⁴⁰. Conversely, each MMP has a different promiscuity profile. As an example, stromelysin can catalyze reactions with 16 different matrix substrates⁴⁰.

2.1.1. Active Site Pockets and Automated Classification

The metzincin metalloendopeptidase active site pocket: The active site pocket of metzincin is a large groove located centrally on the catalytic domain (Fig 1A). A zinc binding consensus sequence (HEXXHXXGXXH/D) consists of either 3 or 2 His, and 1 Asp, which coordinate substrate with a catalytic zinc ion. These His/Asp residues and the zinc ion are all located within the active site pocket (Figure 1), which is formed by the β IV-strand from a 5 stranded β -sheet (Fig 1B, green), the central α -helix (Figure 1, blue), and an extended variable loop (Figure 1B, red) of the catalytic domain. The β strand contains a conserved pattern of side chains located in the central region of the binding pocket, whose consensus is “small-bulky-small-bulky”⁴¹. This pattern suggests a shared common structural requirement for substrate binding⁴¹. The central α -helix forms the back-wall of the active site pocket and embeds the first half of the zinc-binding metzincin motif (HEXXHXXGXXHZ). The extended loop, also called the *specificity loop*⁴², forms the bottom wall of the active site pocket.

Classification of metalloendopeptidase enzymes by surface similarity: An important question for understanding enzyme function is what characteristics the local binding surfaces have and how well they dictate the different biological functions of enzyme subfamilies. For metzincins, the question is whether the structure of the active site pocket alone provides sufficient information for accurate classification of their functions. Although enzyme function is the product of evolution and evolutionary information can greatly facilitate protein function prediction^{15,18,24}, the interaction between substrate and enzyme is a physical event. We therefore examine in this study whether shape and physio-chemical texture of the binding pocket alone provides adequate information for metzincin function classification. Detailed studies of evolutionary pattern of protein surfaces can be found in ref^{28,29}.

We have collected through manual selection the active site pockets of 156 metzincin metalloendopeptidases from the precomputed CASTp database⁴³. We then perform all-against-all pair-wise alignment of surfaces of the 156 catalytic pockets. Unlike previous studies^{28,29,33}, the surfaces of two active-site pockets are aligned in a sequence-order-independent fashion (see Materials and Methods for details), regardless of their ordering in the primary sequence. We then calculate the *distance* between aligned surface pockets, combining the measures of cRMSD, the normalized number of aligned atoms, and a sequence identity related measure, which incorporates further atomic details (see Materials and Methods, Equation 4). This distance measure is then used to cluster the surface pockets, with results organized as a hierarchical tree (see Materials and Methods for detail).

The hierarchical tree (Figure 2A) shows that metzincins can be broadly classified into eight main clusters (shaded rectangles) based on the characteristics of the active site pockets. Each separately highlighted cluster largely coincides with a biological function based on the Enzyme Commission number. The 17 members of the two bontoxilysin clusters consists of 15 structures of the 16 structures in our data set that are known to be bontoxilysin. The serratia-protease cluster has 5 members and contains 5 of the 6 serratia-protease structures in our data set. The astacin cluster has 6 members and contains all 6 of the astacin structures in our data set. The thermolysin cluster has 21 members and contains 18 of the 19 thermolysin structures in our data set. The Pep-Lys MEPase cluster has 5 members and contains all 5 of the MEPase structures in our data set. The reprolysin cluster has 11 members, all of which are structures that are considered to be reprolysin from our data set. The MMP cluster consists entirely of MMP structures and contains 71 out of the 73 MMP structures in our data set.

The MMP cluster (Figure 2A, rightmost) contains many sub-clusters, which correspond to different known subclasses of MMPs as well (Figure 2). Specifically, the 27 members of the combined MMP-3A and MMP-3B clusters are all MMP-3 enzymes. The 5 members of the MMP-12,13 cluster are all indeed members of MMP-12 or MMP-13 family. The 6 members of the MMP-11,7 cluster are all annotated as either members of MMP-11 or MMP-7. The 4 members of the MMP-9 cluster are all MMP-9s. 11 of the 12 members of the MMP-8 cluster are MMP-8s, and all of the 6 members of the MMP-1 cluster are indeed MMP-1s.

Our clustering results also reveal inconsistency due to mislabeling and other mistakes. For example, the unshaded clusters located to the far left in both Figure 2A and 2B are groupings of domains that were mislabeled as metalloendopeptidase catalytic domains. No consensus in enzyme classification emerges from these two unshaded clusters, and the member structures are subsequently examined manually. It was found that of the 11 surfaces in the unshaded region of Fig 2A, 6 of them are domains other than the catalytic domain from the MEP enzyme. The other 5 were mislabeled during manual curation of the data set. It was also found that of the 9 surfaces in the unshaded region of Fig 2B, 4 of them were not the catalytic domain of the MEP enzyme, and the remaining 5 were mislabeled during manual curation. Instead of removing these 20 proteins from the data set, they were retained to test the ability of the sequence order independent structural comparison in detecting falsely labeled surface pockets. The results from clustering validated this approach, as these mislabeled surfaces are not clustered with true enzyme surfaces.

2.1.2. Signature Pockets and Structure-Based MSA

Constructing signature pockets of metzincin metalloendopeptidase: To generate a comprehensive description of binding pockets of metzincins, we quantify the degree of preservations of spatial locations of residues and atoms across different members of the metzincin family. The set of spatially preserved residues and atoms identified are used to compute the *signature pocket*.

We explicitly generate the coordinates of atoms for the signature pocket. This is achieved by computing the geometric center of the multiple spatially aligned atoms from member pockets. These member pockets can be of different size, and the atoms are aligned regardless of the ordering of the underlying residues in the protein sequences. Signature pockets can be computed at different clades at different level of similarity along the classification tree. Depending on the amount of details demanded, an enzyme family or sub-family can be depicted by a single signature pocket, or by multiple signature pockets at lower cut-distances (see Figure 1 in Supporting Information). The signature pockets for each shaded clade in Figure 2A are shown in Figure 3.

Preservation and variation in signature pockets: Two parameters are recorded for each atom in the signature pocket. The *preservation ratio* ρ (Equation 6) describes how often that particular atom was present in the underlying set of pockets. The *location variation* ν (Equation 9) is the mean distance of the coordinates of the aligned atoms to their geometric center. At a given clade or height of the hierarchical tree, different signature pockets can be created at varying level of structural preservation by selecting locations with a preservation ratio ρ above a threshold θ , with θ chosen to be between 50% and 100 %.

An atom with a relatively low preservation ratio ρ does not necessarily mean that the atom was not present in most binding pockets from which the signature pocket was derived. Instead, it can signify a large change in the three-dimensional coordinates, such that this atom is excluded from consideration for some structures in the structural alignment.

Although aligned atoms that enter the signature pocket have smaller fluctuation in spatial positions, those that are more varied in location have larger ν values. Atoms with small ν values vary little in their locations. Regions with low location variation ν can be interpreted as areas in the signature pocket that are rigid and have little spatial variations.

Biological Implications of Metzincin Signatures: The automatically generated signature pockets can reveal useful biological insight. In Figure 3B, a signature pocket of bontoxilysin has a subsite consisting of high preservation atoms (indicated by a high ρ value). Interestingly, it also forms structurally the patch of more dynamic atoms on the surface of the signature pocket, which is indicated by a high ν value of location variation as well. This sub-site holds the substrate but still experience moderate dynamic fluctuations.

The signature pocket of serratia-protease enzymes (Figure 3C) is of relatively low preservation ratio (low ρ , although $> 50\%$) and has high variability (high ν). This is consistent with the large conformational changes of the specificity loop of the serratia-protease enzyme structures.

The MMP enzyme family binds and cleaves a broad spectrum of substrates. We have constructed an overall signature pocket for the whole MMP enzyme subfamilies (Figure 3H), which takes on the general shape of the S1' pocket. To explore the detailed relationship of the signature pockets and the bound inhibitors, and to examine how such relationship determines substrate specificity, we have also computed the signature pockets for each MMP subfamily (Figure 4A–G), which are located at a lower cut-distance in the classification tree. These MMP subfamily signature pockets are relatively similar in overall shape, varying mostly in the width and length of the pocket. They all have a core of atoms with high preservation ratio ρ at the front opening where zinc ligation occurs, as well as along the side wall of the pocket made by the central α -helix. In general, the atoms with low preservation ratio tend to occur towards the rear of the signature pocket.

Biological Implications of MMP Signatures: For the matrix metalloendopeptidases (MMPs) class of metzincins, the active site pocket is relatively flat at the left side shown in Figure 1A and C (non-primed side of the cleavage). The S1' subsite is on the surface to the right of the catalytic zinc (*i.e.*, the primed side), often forming a tunnel that varies in size among the MMPs (Figure 1C). The entrance to the S1' subsite consists of an initial residue of varying identity on the α -helix, called residue 198 following its position in the structure of PDB 1ciz (Figure 1A, blue), the side chain of a conserved Tyr223, the main chain of the wall-forming segment Pro221-X222-Tyr223 located on the specificity loop; and the zinc coordinating His201 residue located on the α -helix. The interior of the S1' subsite is formed mostly by residues located on the specificity loop, as well as residue 197 from the α -helix. The side chain of residue 197 points into the S1' subsite and often determines the depth of the S1' subsite, affecting the allowable length of the P' side of the substrate. Residue 198 located toward the beginning of the central α -helix is also known to be important for substrate specificity⁴⁴.

According to current understanding, the main determinants of substrate specificity and cleavage position include: the interaction between residue in P1' position of the substrate (proximal to the scissile bond) and the S1' subsite of the MMP⁴⁵, the nature of residue 222 in the Pro221-X222-Tyr223 segment⁴⁵, and conformational differences at the tail end of the specificity loop⁴⁴.

Arg198 is identified in the signature of MMP-1 subfamily but not in the other MMP subfamilies. This is consistent with several biological observations. Arg198 is important for substrate specificity of MMPs. The binding pocket on MMP-1 is known to be shorter than those on other MMPs, and is more accommodating towards small inhibitors. Arg198 cuts across the S1' tunnel, causing a shallower pocket (Fig. 6C, orange). MMP-1 can bind inhibitors with long P1 groups, but Arg198 must first adopt a new position⁴⁴.

There are two signature pocket for MMP-3, one for inhibitors containing a long side chain of the P' group (Fig 5A–B), the other for inhibitors containing a short or no side chain of the P' group (Fig 5C–D). The signature pocket for inhibitors with long P' sidechain is characterized by large regions of high preservation ratio ρ (Fig 4A, red regions). The long sidechain of the inhibitor reaches into the S1' subsite and brings these regions into place. A large number of atoms appear in the signature pocket, with still variability in the locations of these atoms, as reflected by the high ν values. It also narrows the back opening of the S1' subsite pocket (Fig 4A). The signature pocket for inhibitors with short or no sidechain is characterized by large regions of low preservation ratio (Fig 4B, blue regions). Without a long sidechain, fewer atoms in these regions enter the signature pocket. These signature pockets help to identify the structural basis of the catalytic specificity of MMP enzymes.

In general, we find atoms with high preservation ratio ρ are most likely to be functionally important due to their persistent occurrence. These atoms experience no drastic changes in locations, regardless of the substrate or ligand bound. In contrast, atoms with relatively low preservation ratio most likely are important for substrate binding specificity. Overall, we find that the automatically generated signature pockets of the metzincin family contain preserved substructures in the active site, many of which are known to be important for substrate specificity and biological function.

Structure-based multiple sequence-fragment alignment: A signature pockets can be conveniently illustrated by a structure-based multiple surface sequence fragment alignment (sbMSFA) of the pocket residues based on different preservation ratio of atoms. sbMSFA align pocket residues by the degree of structural preservation of their atoms. Aside from the important difference that only binding pocket residues are included, sbMSFA resolves

ambiguities of alternate alignment possibilities when using sequence information alone, and can aid in understanding of enzyme binding activity.

The initial step for constructing sbMSFA is to map the signature pocket onto each member structure, and record for each member whether a residue appears at a position in the signature pocket. If so, the identity of the residue is recorded in the column of the sbMSFA for this member structure. If multiple atoms from different residues types in the member structure are mapped to the same residue on a signature pocket, the choice of the identity of the binding residue aligned to a position is resolved by visual inspection. Figure 6A shows the results of the sbMSFA of the concatenated residues from the signature pocket of MMP. It reveals the positions of preserved residue-types (Fig. 6A blue, completely preserved; green, highly preserved) are all located in a ring structure around the position of the scissile bond of the inhibitor, which are proximal to the chelating group of the inhibitor (Fig. 6B). They are located across the β IV strand, the central α -helix, and the front-end of the specificity loop. Located within the ring of the invariant atoms is a single residue highly variable in residue type (Fig. 6A–D, red). This residue and the flanking P and Y residues in the signature correspond to the well-known Pro221-X22-Y223 wall-forming motif of the specificity loop discussed earlier, and has been found to be important in determining substrate specificity^{42,46}. The tail-end of the specificity loop is poorly aligned as expected, because this part of the loop is known to be structurally variable, and likely contributes to the substrate specificity⁴⁷.

2.1.3. Basis Set of Signatures and Handprint of Enzyme Functions

Necessity of multiple signature pockets: As discussed earlier, we find that it is necessary to have two signature pockets for the MMP-3 subgroup. Figures 5A–D show the signature pocket MMP-3A and MMP-3B mapped onto a representative structure with inhibitor bound (Figures 5A,C) and with inhibitor removed (Figures 5B,E). These two are unexpectedly different in both the preservation ratio (ρ) and the variance of locations (ν) of pocket atoms. The physical basis for two signature pockets lies in the diversity of the conformations of the inhibitors bound to this group of MMPs. The MMP-3A signature pocket (Figure 4A) accommodates inhibitors that have long P' groups, which sticks deep into the S1' subsite (Fig. 5C, rectangle). The MMP-3B signature pocket (Fig. 4B), in contrast, accommodates inhibitors that have short or no P' groups (Fig. 4F).

Basis set of signature pockets of enzyme function: As seen in the MMP-3 subgroup and in α -amylase from an earlier study²⁸, a single signature pocket may not be sufficient for accurate characterization of the binding surfaces for the biochemical function of some enzyme families. As different substrates often take on different conformations, more than one characteristic signature pocket are needed. Collectively, we can think of these signature pockets forming a basis set of binding surfaces for a specific biological function. For the MMP subgroup of enzymes, the seven signature pockets form such a structural basis set. An MMP binding surface may be thought of as taking the shape of an interpolated version of these basis set of binding surfaces, with each contributing at a different weight.

2.1.4. Binding site and function prediction by similarity to signature pockets—

To objectively quantify how well the signature pockets capture key determinants of the biological functions of metzincin enzymes, we test whether the correct binding pocket on metzincin structures can be identified, and whether the correct functional class can be predicted. To accomplish this, we align structurally each surface pocket from a test set of protein structures, which are not used in the construction of the signature pockets, to each of the metzincin signature pockets.

Discriminating binding surfaces from non-binding surfaces on metzincins: We first determine whether the signature pockets are able to distinguish metzincin binding pockets from a background of non-binding surface pockets. We have structurally aligned all top three surface pockets ranked by size from the 148 metzincin structures in the test set, with a total of 444 pockets, to each of the seven signature pockets of metzincin (Figure 3) and the seven MMP signature pockets (Figure 3).

We find that signature pockets can be used to distinguish binding surface from non-binding surfaces, with an overall sensitivity and specificity of 0.93 and 0.89, respectively. The results for each metzincin subfamily are shown in Table 1. These results indicate that the metalloendopeptidase signature pockets are able to distinguish metzincin binding pockets from other non-binding pockets on metzincin.

Discriminating metzincin vs non-metzincin enzymes: Next, we determine whether a signature pockets can be used to detect metzincin enzymes against a set of non-metzincin enzymes, to predict their functions, and to locate their binding pockets. We took the top three pockets by size from 300 non-metzincin enzyme structures in the Protein Data Bank. These surface pocket are then combined with a set of 100 metzincin substrate binding pockets, with a total of 1,000 surface pockets. A surface pocket in this test set was assigned the functional class of a signature pocket if the distance from the structural alignment was below the threshold $\theta = 0.40$. The functional class of an enzyme was left unassigned if none of the pockets had a distance below θ .

The prediction results for each metzincin subfamily except the MMPs are shown in Table 2. These results suggest that the signature pockets of metalloendopeptidase capture important information that can distinguish this class of enzyme binding pockets from surface pockets on other enzymes. They can also be used to predict metzincin function and to distinguish the metzincin binding surfaces from surfaces on others enzymes, with an overall sensitivity and specificity of 0.90 and 0.88, respectively.

Classification of MMPs by basis set signature pockets: We find that predicting MMPs using the overfall signature (Figure 3H) has poor results, with low sensitivity (0.34) and specificity (0.29). This poor performance points to the need of a different approach.

We then explore if using the basis set pockets as a whole lead to better prediction of the membership of the MMP family. For this, we use the average distance of the query protein to the 7 signature pockets of MMPs, and declare a protein to be an MMP if the measured distance is less than $\theta = 0.4$. The sensitivity and specificity for MMP predictions have improved to 0.90 and 0.89, respectively. This also leads to an overall improvement in predicting functions of different classes of metzincin, as the number of MMPs previously incorrectly labeled as other metzincins is significantly reduced. The sensitivity and specificity of our predictions of overall metzincins now are improved to 0.95 and 0.89, respectively (Table 2). The high sensitivity and specificity of the prediction indicate that collectively, the seven MMP subgroup signature pockets do form a basis set that describes well the MMP subfamily of metalloendopeptidase.

2.2. NAD Binding Pockets

The study of metzincin provides an example of the view of the binding surfaces of an enzyme family with diverse biological functions. Although members of the metzincin family can accommodate very different substrates, they are all related evolutionarily and descended from a common ancestor. An equally important question is whether the binding surfaces on proteins that do not descend from a common ancestor but have evolved convergently to bind

the same substrate or cofactor share common characteristics. To address this question, we study the NAD binding pockets on proteins with unrelated evolutionary origins.

Nicotinamide adenine dinucleotide (NAD) consists of two nucleotides, nicotinamide and adenine, joined by two phosphate groups. NAD plays essential roles in redox reactions, including those in glycolysis and the citric acid cycle⁴⁸. A large number of oxidoreductases incorporate NAD as coenzyme⁴⁸. Although many NAD binding proteins contain the Rossman fold⁴⁹, they are diverse in fold structures (Table 1 in Supporting Information) and have independent evolutionary origins.

2.2.1. NAD binding signature pockets—We have constructed signature pockets of NAD binding sites from a collection of 457 structures of NAD binding proteins. Since our method requires no evolutionary information, it is well-suited for studying convergently evolved proteins. Based on sequence-order-independent alignments of the NAD binding pockets and the resulting distance measures, these 457 NAD binding surfaces can be hierarchically clustered into nine distinct clusters. The signature pockets representing each cluster are shown in Figure 7. Among these, six are for oxidoreductases (Fig 7E–J), two for lyase (Fig 7B and D), and one for isomerase (Fig 7C). We find that the main chain fold family and the conformation of the bound NAD are the two major factors that influence the formation of the clusters and the signatures.

These signature pockets contain information on substrate specificity. For oxidoreductase, three signature pockets (Fig 7E, H, and I) are for clusters of oxidoreductases that act on the CH-OH group of donors (alcohol oxidoreductases), one signature pocket (Fig 7J) is for a cluster that act on the aldehyde group of donors, and the remaining two signature pockets (Fig 7F and G) for oxidoreductases that act on the CH-CH group of donors. For lyase, one of the two signature pockets (Fig 7D) represent lyase that cleave both C-O and P-O bonds. The other signature pocket (Fig 7B) represent lyases that cleave both C-O and C-C bonds. These two signatures come from two clusters of lyase conformations, each with a very different class of conformations of the bound NAD cofactor.

2.2.2. Multiple Signature Pockets for Compact and Extended NAD

Conformations—The conformations of bound NAD cofactors (Fig 7B–J) take either an extended conformation (Fig 7D, E, and I, marked with an X), or a compact conformation (Fig 7B, C, F, G, H, and J, marked with a C). However, each requires multiple signature pockets. We find that a basis set of four signature pockets are needed to represent the diverse conformations of binding pockets that accommodating the extended conformation of NAD, and five for the compact conformation of NAD.

The need for more than a single signature pocket is exemplified in the two structurally distinct C-C oxidoreductase signature pockets for the compact conformation of NAD (Fig 7F and G). The conformations of both clusters of protein have the same SCOP fold, the same function with identical E.C. number, and bind the same NAD conformation, yet these two signature pockets are substantially different in shape and in dimension of the mouth openings. For binding surfaces of diverse origin, a single signature NAD binding pocket is often inadequate, and we believe a basis set of multiple signature pockets are needed to fully characterize the properties of the binding surfaces of NAD.

2.2.3. Predicting NAD-Binding Proteins by Similarity to Signature Pockets—To test how well the basis set of signature pockets captures the properties of NAD binding proteins, we examine whether these signature pockets can correctly identify NAD-binding proteins from other enzymes. We constructed a test data set by taking the top 3 largest pockets from 142 randomly chosen proteins and proteins that have NAD bound in the PDB

structure and were not in the data set used in deriving the signatures. This results in a set of 576 pockets. We then structurally aligned all of these pockets against each of the nine NAD signature pockets. The pocket in test was assigned to be an NAD binding pocket if it structurally aligned to one of the nine NAD signature pockets, with the distance under a threshold θ of 0.4. Otherwise it was classified as non-NAD binding.

We first examine whether comparing an uncharacterized surface pocket to individual signature pockets is sufficient for identifying NAD-binding enzymes. We chose a representative pocket from one of the 9 clusters that were used to construct the 9 signature pockets and compare each with a subset of the test data, with the additional requirement that only NAD binding proteins belonging to an enzyme class that is represented during the construction of the signature pockets are included. We repeat this process nine times, each with a different cluster. The prediction results are rather poor, with an average sensitivity and specificity of only 0.36 and 0.23, respectively.

However, when we take the 9 signature pockets that form the basis set and comparing a query surface to each of them in turn using the distance measure described in Eqn (10) to assess binding surface similarity, NAD binding pockets can be correctly identified with a much improved sensitivity of 0.91 and specificity of 0.89 for the same data set, respectively.

3. Discussion

In this study, we have introduced the approach of using signature pockets and basis set to characterize the binding surface of a class of enzyme function and a class of cofactor binding proteins. We showed that signatures capturing structural features that are biologically important can be automatically extracted. Our approach, called SOLAR, can be used to identify members of an enzyme family and proteins binding to a specific cofactor. Our approach works at multi-resolutions, and can capture relevant information at different level of surface similarity. Our finding further suggest that residues and atoms with high preservation ratio ρ are likely to be important for binding, and those with low ρ and high v are likely to be important for specificity.

Both this work and a previous study²⁹ address the problem of inferring protein functions from structures. The current study, however, has an important different focus. While representatives of binding surfaces were used for database search in²⁹, there was no derivation of an explicit template for the binding surface of an enzyme. The study of²⁹ was based on the assumption of the availability of a surface template, and the binding surface of a representative structure is simply taken as the template. In contrast, we now construct explicitly a template characterizing the preserved and varied regions of the binding surface. The identified preserved and varied spatial regions help to understand the structural basis and mechanism of enzyme function. Furthermore, the study of²⁹ already pointed out that single template may be inadequate, and this is now fully addressed by the model of basis set developed in this study. Namely, multiple signature pockets need to be derived, which collectively form a basis set for the binding surfaces of an enzyme class.

There are significant advantages using a sequence-order-independent method for surface alignment. First, it enables alignment of surfaces at the atomic level beyond the residue level achievable using sequence order dependent method, as atoms in a residue do not have a natural sequential ordering. Second, it can detect many instances of surface similarities previously unrecognizable. Third, it improves the accuracy of similarity assessment (*e.g.*, more aligned atoms and smaller cRMSD value). There are many cases in which an order dependent method underestimates the similarity. Fig 8 shows an example of aligned binding pockets that would not be detected by a sequence-order-dependent method.

Although our method is fully automated and the resulting signatures and basis sets characterize enzyme binding surfaces well, signatures and basis sets are not unique. Depending on different choices of similarity measure, threshold, and clustering methods, the specific signature and basis set may be different.

Although surface comparison in a sequence-order-independent manner takes longer time (about 5–10 seconds/comparison), it generates more accurate alignment. We have shown how signatures and basis sets can be computed automatically, and our method is amenable to large scale calculations. Nevertheless, as an accurate and reliable set of input binding surfaces is essential, manual curation and inspection is needed at this stage to ensure a reliable input data set. Once a large database of carefully constructed signature pockets for different enzyme classes are built, large scale systematic comparison of protein structures against such a database of signature pockets can then be carried out.

The approach of signatures and basis set strongly depends on the availability of sampled structures of binding surfaces. Without sufficient structural data, they cannot be obtained. A question that has no simple answer is how many structures are required for an accurate signature pocket and basis set to be constructed, which can comprehensively describe a particular enzyme class. Our study suggests that for metzincin, accurate signatures and basis sets can be obtained from available structures. In contrast, it is likely that our data do not encompass all enzyme conformations that bind NAD, and there may exist more than nine signature pockets for NAD binding proteins. This is evidenced by the fact that when additional NAD binding proteins of different EC classes unused in constructing NAD binding signatures are included in the test set, the sensitivity and specificity of our predictions deteriorates to 0.78 and 0.72, respectively.

Another challenging issue is to ensure that diverse structures are chosen so the underlying proteins have experienced sufficient evolution time and preservation of atoms signifies biological functional constraints rather than a mere consequence of inadequate evolving time. It is possible that the observation of the same atom type (for example, O) in two different amino acids (e.g., Glu and Asn) from two different structures may be indicative of functional importance. A more systematic approach is to construct the underlying phylogenetic tree explicitly using a carefully designed evolutionary model^{28,29} and weight the preserved atoms and residues based on their evolution time as recorded in the branch lengths of the phylogenetic trees.

As the construction of signatures and basis set requires well-represented samples of enzyme structures, an intriguing possibility is to generate model structures from homology models. This may present a viable solution that is worth further exploration.

4. Materials and Methods

A three-step process is used to compute the signature pocket of a functional class of proteins. First, all pairs of functional surface pockets from different structures of enzymes of the same function are aligned structurally in a sequence-order-independent manner. Second, surface pockets are then organized into a hierarchical tree based on measured similarity. This tree is computed from clustering a matrix of pairwise distances, which are obtained from the structural alignments. Third, the hierarchical tree is used as a guide tree and surface pockets of members of a clade are combined into a single signature pocket, whose atomic coordinates are generated explicitly. This is repeated recursively at clades of higher level. Below we discuss details of each step.

4.1. Pairwise alignment of surface pockets in sequence independent order

Similar to protein structural alignment, comparing two surface pockets p_A and p_B requires first the establishment of correspondency of atoms in both structures. Once the equivalence is established, the optimal superposition of the corresponding elements needs to be found so the root mean square distance (RMSD) between these two pockets is minimized. The optimal superposition is described by a rotation matrix R and a translation vector t . The latter moves p_B to superimpose on p_A by coinciding their centers of mass. R and t can be found by solving the optimization problem:

$$\min |p_A - R(p_B - t)|^2, \quad (1)$$

which can be achieved exactly using Singular Value Decomposition (SVD)⁵⁰.

The challenging task is to establish equivalence relationship between atoms and residues from two surface pockets in a sequence independent manner. We follow reference⁵¹ and formulate this as a problem of multi-objective optimization, *i.e.*, minimizing Eqn (1) while maximizing the total number of equivalent elements. Following⁵¹, we decompose the multi-objective optimization problem into two sub-problems. The first is to hold the equivalences constant and solve Eqn (1) using Singular Value Decomposition⁵⁰. This results in a rotation matrix R and a translation vector t for optimal rigid transformation. The second is to find the best equivalences while holding the current superposition constant. This can be formulated as a maximum weight bipartite matching problem, in which graph nodes represent atoms (or residues) from the two proteins. Directed and weighted edges connect nodes from the query protein to nodes from the target protein, if the two nodes share some similarity. The task is to find a set of edges connecting nodes of the query pocket to nodes of the target pocket, with maximized total edge weight, while insisting only at most one edge is selected for each residue⁵².

We use the Hungarian algorithm^{51,53} to solve the maximum weight bipartite matching problem. In addition, we introduce biased weighting for the edges based on aligned atom types. After representing atoms from each pocket as a set of nodes and connecting each node in p_A with an edge to each node in p_B that lies within certain distance threshold, we add a fictitious source node s that connects to every node in pocket p_A (called query node) with 0-weight. We then add a fictitious destination node d that connects to every node in p_B (called target node) with 0-weight. All other edge weights are initially computed as:

$$w(i, j) = d(i, j) - S(i, j), \quad i \in p_A, \text{ and } j \in p_B. \quad (2)$$

Here $d(i, j)$ is the euclidean distance between the two superimposed atoms and $S(i, j)$ is a scoring matrix that rewards the alignment of C to C, O to O, N to N, and S to S. A smaller reward is also given for the alignment of O to N, O to S, and N to S. The specifics are discussed in next section.

Instead of the Dijkstra algorithm used in⁵¹, we apply the Bellman-Ford algorithm to compute the distance $F(i)$ of the shortest path(s) from the source node s to each remaining node i . The weight for each edge that does not contain the source node is then updated. The new weight $w'(i, j)$ for edge $e(i, j)$ starting from node i to node j is calculated as:

$$w'(i, j) = w(i, j) + [F(i) - F(j)]. \quad (3)$$

An overall score F_{all} , initialized to 0, is also updated as $F'_{all} = F_{all} - F(d)$. Next, we flip the directions of all edges in the shortest path from the source s to the destination d . The Dykstra algorithm is not appropriate as cycles may result.

We again apply the Bellman-Ford algorithm on this new graph, and update the weights and the overall score F_{all} . This is repeated until either there is no directed path from s to d as edges have been flipped, or the shortest distance $F(d)$ to the destination node d is greater than the current overall score F_{all} . The output of the Hungarian method includes a set of directed edges starting from the target nodes to query nodes. These edges provide the equivalence relationship, namely, which atoms in the target pocket should be aligned to which atoms in the query pocket. This equivalence relationship takes no sequence ordering information into account, and is therefore sequence order independent.

To obtain an alignment of two surface pockets, we begin with an arbitrary initial equivalence relationship. We then apply the optimal rotation matrix and translation vector derived using the SVD technique to align these two pockets. The euclidean distance between atoms of the query pocket to atoms of the target pocket after optimal superposition is then calculated. The Hungarian method is applied to find the new equivalences. This process is repeated iteratively. The iterative process is stopped when the improvement in $cRMSD$ is less than a predefined threshold.

4.2. Hierarchical clustering

A distance measure is calculated between each pair of aligned pockets p_A and p_B as:

$$d(p_A, p_B) = \frac{r(p_A, p_B)}{n^{1/3}} + \alpha \cdot [1 - s(p_A, p_B)], \quad (4)$$

where $r(p_A, p_B)$ is the root means squared deviation (RMSD) of p_A and p_B obtained after optimal superposition of p_A on p_B , and n is the number of aligned atoms. The score $s(p_A, p_B)$ is a measure of sequence similarity between the structurally aligned atoms and is calculated as follows. The i -th pair of aligned atoms is assigned a score of $a_i = 0, 1$ or 2 . A score of 2 is assigned if both the atom type and the residue type are the same, 1 is assigned if there is either a matching atom type or a matching residue type, and 0 is assigned if neither the atom type nor residue type matches. The maximum score $s_{\max} = 2 \cdot n$ occurs when all atom types and residue types are the same for each equivalent pairs of atoms in the structural alignment.

The sequence similarity score $s(p_A, p_B)$ is then calculated as $s(p_A, p_B) = \frac{\sum_i a_i}{s_{\max}}$.

Using $d(p_A, p_B)$ as the distance measure, we construct a matrix recording distances for each pair of surface pockets. We then apply the average linkage hierarchical clustering method⁵⁴ on the distance matrix, and obtain a hierarchical tree reflecting the distances among surface pockets on the enzyme structures.

4.3. Computing signature pockets

To construct signature pockets at different level of similarity, the hierarchical tree is used as a guide to recursively combine sibling pockets along the paths from leaf nodes to the current node on the clustering tree. In the beginning, each leaf pocket is its own signature pocket. A new signature pocket S_C is computed as the average of two child signature pockets S_A and S_B , and the original two child nodes are replaced with a new single leaf node on the hierarchical tree. We first align the two child signature pockets S_A and S_B . The resulting

equivalence relationship $S_A \sim S_B$ is then used to create a new pocket. For an equivalent pair of atoms at the k -th position, the new atomic coordinate $x_C(k)$ are computed as:

$$x_C(k) = \frac{m_A(k)x_A(k) + m_B(k)x_B(k)}{m_A(k) + m_B(k)}, \quad (5)$$

where $x_A(k)$ and $x_B(k)$ are the atomic coordinates of atoms from signature pockets S_A and S_B , respectively, $m_A(k)$ and $m_B(k)$ are the number of pockets that contributed atoms at position k and were used in the creation of signature pocket S_A and signature pocket S_B , respectively. The atomic coordinates from S_A and S_B that were not aligned are also transferred to the combined pocket.

Two statistics are recorded for each atom in the newly created pocket. The preservation ratio $\rho(k)$ indicates how often an atom was found at position k in a structural alignment. The location variation $v(k)$ indicates the structural variability of the atom at k . The preservation ratio $\rho_A(k)$ for signature S_A is defined as:

$$\rho_A(k) = \frac{m_A(k)}{|\mathcal{A}|} \quad (6)$$

where $\mathcal{A}$ is the set of underlying pockets for the signature at branching point A , $|\mathcal{A}|$ is the total number of pockets in this set. When combining two signature pockets S_A and S_B to form a new signature pocket S_C , the preservation ratio $\rho_C(k)$ for the equivalent atom at position k is computed as:

$$\rho_C(k) = \frac{|\mathcal{A}|\rho_A(k) + |\mathcal{B}|\rho_B(k)}{|\mathcal{A}| + |\mathcal{B}|} \quad (7)$$

It is the weighted sum of the preservation frequencies from S_A and S_B . In principle, the value of ρ ranges from 0 to 1.

At each instance when two signature pockets are combined, only the equivalent atoms are used to compute the coordinate of the new atom. Atoms that do not have equivalent counterparts in the other signature pocket are carried over to the new structure, and we set their preservation ratio to $1/|\mathcal{A}|$. We maintain these non-aligned atoms, as they may enter the alignment at a later stage at a different similarity cut-off distance, where an enlarged portion of the pocket is incorporated. Once the root of the clustering tree is reached, the algorithm terminates.

The preservation ratio ρ is also used in structural alignment of two surface pockets. Since there often exists a large number of atoms, it is challenging to obtain a good structural alignment. To overcome this problem, we use ρ to bias the alignment towards atoms that appear in a larger number of pockets. Specifically, we update the weights in the bipartite graph (Eqn (2)) using:

$$w(i, j) = d(i, j) - \rho_i \rho_j S(i, j), \quad (8)$$

The atom location variance v is calculated as the average euclidean distance of the constituent atoms to their center of mass. For the k -th position in the signature S_A with the underlying participating set of pockets $\mathcal{A}(k)$, we have:

$$v_A(k) = \frac{\sum_{p_i \in \mathcal{A}(k)} |x_i(k) - \bar{x}_A(k)|}{m_A(k)} \quad (9)$$

where $\bar{x}_A(k)$ is the center of mass of the atoms at position k in the signature of A , $m_A(k)$ is the number of participating pockets at position k . In implementation, we use an incremental formulation to gain computational speed (see Supporting Information).

4.4. Computing distance to a basis set

The distance,

$$d(p_A, \mathcal{A}) = \frac{\sum_{k \in \mathcal{A}} d(p_A, \mathcal{A}(k))}{|\mathcal{A}|}, \quad (10)$$

between a testing pocket p_A to a basis set of signature pockets $\mathcal{A}$ is computed as the average of the distances to the individual component signature pockets

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants GM079804, GM081682, GM086145, NSF grant DMS-0800257, and ONR grant N00014-09-1-0028.

References

1. Shah I, Hunter L. Predicting enzyme function from sequence: a systematic appraisal. ISMB. 1997; 5:276–283. [PubMed: 9322050]
2. Altschul S, Warren G, Miller W, EW M, DJ L. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]
3. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lip-man D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 1997; 25(17): 3389–3402. [PubMed: 9254694]
4. Weidong T, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity. J Mol Biol. 2003; 333:863–882. [PubMed: 14568541]
5. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol. 1999; 288:147–164. [PubMed: 10329133]
6. Holm L, Sander C. Mapping the protein universe. Science. 1996; 273:595. [PubMed: 8662544]
7. Rost B. Twilight zone of protein sequence alignments. Prot Engr. 1999; 12:85–94.
8. Dundas J, Binkowski T, DasGupta B, Liang J. Topology independent protein structural alignment. BMC Bioinformatics. 2007; 8:388. [PubMed: 17937816]
9. Meng E, Polacco B, Babbitt P. Superfamily active site templates. Proteins. 2004; 55:962–967. [PubMed: 15146493]
10. Wistow G, Mulders J, De J. The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. Nature. 1987; 326:622–624. [PubMed: 3561501]

11. Acharya K, Ren J, Stuart D, Phillips D, Fenna R. Crystal structure of human alpha-lactalbumin at 1.7 Å resolution. *J Mol Biol.* 1991; 221:571–581. [PubMed: 1920433]
12. Orengo C, Todd A, Thornton J. From protein structure to function. *Curr Opin Struct Biol.* 1999; 9:374–382. [PubMed: 10361094]
13. Jeffery C. Molecular mechanisms for multi-tasking: recent crystal structures of moonlighting proteins. *Curr Opin Struct Biol.* 2004; 14:663–668. [PubMed: 15582389]
14. Petrey D, Fischer F, Honig B. Structural relationships among protein with different global topologies and their implications for function annotation strategies. *Proc Natl Acad Sci USA.* 2009; 106(41):17377–17382. [PubMed: 19805138]
15. Lichtarge O, Bourne H, Cohen F. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol.* 1996; 236:412–420.
16. Norel R, Fischer D, Wolfson H, Nussinov R. Molecular surface recognition by computer vision-based technique. *Protein Eng.* 1994; 7:39–46. [PubMed: 8140093]
17. Fischer D, Norel R, Wolfson H, Nussinov R. Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition. *Proteins.* 1993; 16:278–292. [PubMed: 8394000]
18. Gold N, Jackson R. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J Mol Biol.* 2006; 355:1112–1124. [PubMed: 16359705]
19. Russell R. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol.* 1998; 279:1211–1227. [PubMed: 9642096]
20. Laskowski R. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graphics.* 1995; 13:323–330.
21. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* 1998; 7:1884–1897. [PubMed: 9761470]
22. Laskowski R, Luscombe N, Swindells M, Thornton J. Protein clefts in molecular recognition and function. *Protein Sci.* 1996; 5:2438–2452. [PubMed: 8976552]
23. Wallace A, Borkakoti N, Thornton J. Tess: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. application to enzyme active sites. *Protein Sci.* 1997; 6:2308–2323. [PubMed: 9385633]
24. Glaser F, Pupko T, Paz I, Bell R, Bechor-Shental D, Martz E, Ben-Tal N. Consurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics.* 2003; 19:163–164. [PubMed: 12499312]
25. Binkowski T, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol.* 2003; 332:505–526. [PubMed: 12948498]
26. Huan J, Bandyopadhyay D, Wang W, Snoeyink J, Prins J, Tropsha A. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J Comput Biol.* 2005; 12:657–671. [PubMed: 16108709]
27. Chen B, Bryant D, Cruess A, Bylund J, Fofanov V, Kimmel M, Lichtarge O, Kavraki L. Composite motifs integrating multiple protein structures increases sensitivity for function prediction. *Computational Systems Bioinformatics Conference.* 2007; CBS2007:343–355.
28. Tseng Y, Liang J. Predicting enzyme functional surfaces and locating key residues automatically from structures. *Ann Biomed Eng.* 2007; 35(6):1037–1042. [PubMed: 17294116]
29. Tseng Y, Dundas J, Liang J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J Mol Biol.* 2009; 387(2):451–464. [PubMed: 19154742]
30. Lamdan Y, Schwartz J, Wolfson H. On recognizing 3d objects from 2d images. *Proceedings of IEEE INT Conf on Robotics and Automation, Philadelphia, Pennsylvania.* 1988:1407–1413.
31. Chou K, Cai Y. A novel approach to predict active sites of enzyme molecules. *Proteins.* 2004; 55:77–82. [PubMed: 14997541]
32. Shatsky M, Shulman-Peleg A, Nussinov R, Wolfson H. The multiple common point set problem and its application to molecule binding pattern detection. *J Comput Biol.* 2006; 13(2):407–428. [PubMed: 16597249]

33. Binkowski T, Freeman P, Liang J. pvSOAR: Detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res.* 2004; 32:W555–W558. [PubMed: 15215448]
34. Binkowski T, Joachimiak A, Liang J. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci.* 2005; 14:2972–2981. [PubMed: 16322579]
35. Goyal K, Mande S. Exploiting 3D structural templates for detection of metal-binding sites in protein structures. *Proteins.* 2008; 70:1206–1218. [PubMed: 17847089]
36. Meng E, Polacco B, Babbitt P. Superfamily active site templates. *Proteins.* 2004; 55:962–976. [PubMed: 15146493]
37. Watson J, Laskowski R, Thornton J. Predicting protein function from sequence and structural data. *Curr Opin Struc Biol.* 2005; 15:275–284.
38. Kahraman A, Morris R, Laskowski R, Thornton J. Shape variation in protein binding pockets and their ligands. *J Mol Biol.* 2007; 368:283–301. [PubMed: 17337005]
39. Stocker W, Grams F, Baumann U, Reinemer P, et al. FGR. The metzincins - topological and sequential relations between the astcins, adamalysins, serralysins, and matrixins (collagenases) define a super-family of zinc-peptidases. *Protein Sci.* 1995; 4:823–840. [PubMed: 7663339]
40. Overall C. Molecular determinants of metalloproteinase substrate specificity. *Mol Biotechnol.* 2002; 22:51–86. [PubMed: 12353914]
41. Gomis-Ruth F. Structural aspects of the metzincin clan of metalloen-dopeptidases. *Mol Biotechnol.* 2003; 24:157–202. [PubMed: 12746556]
42. Lang R, Kocourek A, Braun M, Tschesche H, Huber R, Bode W, Maskos K. Substrate specificity determinants of human macrophage elastase (MMP-12) based on the 1.1 Å crystal structure. *J Mol Biol.* 2001; 312:731–742. [PubMed: 11575928]
43. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. Castp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucl Acids Res.* 2006; 34:W116–W118. [PubMed: 16844972]
44. Lovejoy B, Welch A, Carr S, Luong C, Broka C, Hendricks R, Campbell J, Walker K, Martin R, Wart HV, Browner M. Crystal structures of mmp-1 and -13 reveal the structural basis for selectivity of collagenase inhibitors. *Nat Struct Bio.* 1999; 6(3):217–221. [PubMed: 10074939]
45. Bode W, Fernandez-Catalan C, Tschesche H, Grams F, Nagase H, Maskos K. Structural properties of matrix metalloproteinases. *CMLS -Cell Mol Life S.* 1999; 55:639–652.
46. Engel C, Pirard B, Schimanski S, Kirsch R, Habermann J, Klingler O, Schlotte V, Weithmann K, Wendt K. Structural basis for the highly selective inhibition of mmp-13. *Chem Biol.* 2005; 12:181–189. [PubMed: 15734645]
47. Sternlicht M, Werb Z. How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol.* 2001; 17:463–516. [PubMed: 11687497]
48. Berg, J.; Tymoczko, J.; Stryer, L. *Biochemistry.* WH Freeman and Company; 2002.
49. Duax W, Huether R, Pletnev V, Umland T, Weeks C. Divergent evolution of a Rossmann fold and identification of its oldest surviving ancestor. *Int J of Bioinformatics Research and Applications.* 2009; 5(3):280–294.
50. Umeyama S. Least-squares estimation of transformation parameters between two point patterns. *IEEE T Pattern Anal.* 1991; 13(4):376–380.
51. Chen L, Zhou T, Tang Y. Protein structure alignment by deterministic annealing. *Bioinformatics.* 2005; 21:51–62. [PubMed: 15308541]
52. Cormen, T.; Leiserson, C.; Rivest, R.; Stein, C. *Introduction to Algorithms.* MIT Press; 2001.
53. Kuhn H. The Hungarian method for the assignment problem. *Nav Res Logist Q.* 1955; 2:83–97.
54. Baxevanis, A.; Ouellette, BF. *Bioinformatics: A practical guide to the analysis of genes and proteins.* Third edition. John Wiley & Sons; 2005.

Figure 1.
The catalytic domain of a matrix metalloproteinase (MMP). A) Surface view of the catalytic pocket of MMP; B) The catalytic pocket is formed by a β strand (green), an α -helix (blue), and an extended loop (red); and C) View of the catalytic pocket of MMP facing down the S1' pocket.

Figure 2. Classification of metalloendopeptidase structures and MMPs by sequence-order-independent surface similarity. A). Hierarchical tree of the binding pockets of 156 metalloendopeptidase enzymes based on their structural alignment in a sequence-order-independent fashion. Eight distinct clusters of binding pockets can be seen (shaded rectangles) and are labeled accordingly. From left to right: bontoxilysin #1, bontoxilysin #2, serratia-protease, astacin, thermolysin, pep-lys mepase, reprolysin, and MMP. The left most unshaded cluster consist of mostly other proteins or domains mislabeled as metzincins. B). A detailed view of the MMP clusters. These clusters are largely in agreement with the conventional classification of MMPs, with the MMP class labeled for each clade. MMP-3 formed two separate clusters (MMP-3A and MMP-3B). From left to right: MMP-3A, MMP-3B, MMP-12 & 13, MMP-11 & 7, MMP-9, MMP-8, and MMP-1

Figure 3.

The signature pockets of metzincin metalloendopeptidase enzymes as determined by the shaded clusters in the hierarchical tree shown in Figure 2. Each signature pocket is shown in two coloring schemes. The first is by preservation ratio ρ , and the second is by location variance v .

Figure 4.
The signature pockets of the MMP class of enzymes corresponding to the shaded clusters in the hierarchical tree of Figure 2B. The color scheme is the same as Fig 3.

Figure 5.

The two signature pockets of the MMP-3 sub-group accommodate different configurations of the side chains of the bound inhibitors. The signature pocket for MMP-3A (Fig 4A) mapped on to a representative structure A) with the inhibitor shown (PDB 1hfs), and B) with inhibitor removed for visualization (PDB 1c8t). C). Four inhibitors taken from structures of the MMP-3 A cluster. All MMP-3 A bound inhibitors have extended side chains (shown in square). The MMP-3 B signature pocket 4B mapped onto a representative structure (PDB 1g05) D) with the inhibitor shown, and E) with inhibitor removed for visualization (1d8f). F). Four inhibitors taken from representative structures of the MMP-3B cluster. All MMP-3 B bound inhibitors have short side chains.

Figure 6.

Structure based multiple surface sequence fragment alignment (sbMSFA) of a group of the MMP subfamilies. A) The topological tree of the subtree of the binding pockets for these MMP subgroups. The residue at corresponding positions from each member structure enters the alignment. For clarification we show only the sequence fragment for one representative structure for each subclass of MMPs. All residues in the alignment occur in the signature pocket, namely, their preservation ratio $p > 50\%$. Residues that occur in the signature pocket with $> 99\%$ identical in residue type across all members are colored in blue, those between 75–99% identical are colored green, and those with $< 75\%$ identical residues are colored in gray. A special column colored in red represents a position that occurs in all subfamilies, but with a different residue type for each subfamily. This column corresponds to the X residue in the Pro221-X222-Tyr223 motif. For MMP-1 subfamily, there is in addition a signature residue (colored in red) that does not appear in other subfamilies. This corresponds to the key residue Arg198 that determines the depth of the S1' pocket in MMP1 and hence is an important determinant of MMP1 specificity. B–D). Visualization of the underlying surfaces where the colored residues in A) are mapped onto the surface of the representative structure for each subfamily. B): a representative structure of MMP-9, C) and D): a representative structure of MMP-1 with and without the key Arg residues for visualization of the effects of this Arg on the depth of the S1' subsite.

Figure 7.

Classification of NAD binding proteins and their signature bind pockets. A.) The hierarchical tree resulting from clustering of NAD-binding pockets. B–J.) The signature pockets of the NAD binding proteins, along with the NAD cofactors bound in the member structures of the respective clusters superimposed. The NAD cofactor has two distinct conformations. Those in an extended conformation are marked with an “X”, and those in a compact conformation are marked with a “C”.

Figure 8.

The binding pockets from the catalytic domains of two different stromelysin (pocket 29 from PDB 1hv5 chain A and pocket 19 from 1qic chain D). The two pockets align well with an cRMSD of 0.76Å for 29 atoms from 10 different residues. The structurally aligned residues are placed in the order given by primary sequence of 1hv5. A sequence order dependent structural alignment method would not be able to generate this alignment as aligned atoms from both proteins do not follow the order given by their respective primary sequences.

Table 1

The sensitivity and specificity in predicting functional surfaces of metalloendopeptidase enzymes. The prediction sensitivity is calculated as $\frac{TP}{TP+FN}$, and specificity as $\frac{TN}{TN+FP}$, where *TP* represents true positive, *FN* false negative, *TN* true negative, and *FP* false positive.

Class	Sensitivity	Specificity
Bontoxilysin	0.96	0.94
Ser-protease	0.89	0.92
Astacin	0.91	0.87
Thermolysin	0.94	0.90
Pep-lys-mep	0.92	0.89
Reprolysin	0.94	0.91
MMP	0.95	0.89
Overall	0.93	0.89

Table 2

The sensitivity and specificity in predicting the functional class of 100 metzincin structures (in all 4 Enzyme Commission digits), after discriminating each from 300 structures of other enzymes and structures of metzincins of other subfamilies with different E.C. digits. These predictions require locating the correct binding pockets from 900 surface pockets of non-metzincin enzymes and 100 surface pockets on metzincins. “Signature”: Using similarity to individual signature pockets, with MMPs excluded; “Basis Set”: Using similarity to each basis set, with MMPs included.

Class	Signature		Basis Set	
	Sensitivity	Specificity	Sensitivity	Specificity
Bontoxilysin	0.97	0.91	0.97	0.91
Ser-protease	0.86	0.83	0.87	0.84
Astacin	0.89	0.85	0.89	0.85
Thermolysin	0.92	0.88	0.92	0.88
Pep-lys-mep	0.91	0.90	0.91	0.90
Reprolysin	0.92	0.91	0.93	0.91
MMP			0.90	0.89
Overall	0.85	0.81	0.95	0.85