

Computing Statistical Profiles of Active Sites in Proteins

Chang Zhao Jalal Mahmud I.V. Ramakrishnan
Computer Science Dept.,
Stony Brook University,
Stony Brook, NY, 11794, USA
{changz.,jmahmud,ram}@cs.sunysb.edu

Subramanyam Swaminathan
Biology Department
Brookhaven National Laboratory
Upton, NY11973-5000, USA
swami@bnl.gov

Abstract

Active sites in proteins are three dimensional substructures that cause them to perform their function. The problem of finding substructures in a protein that are “similar” to the active sites of another protein has several important applications in biological sciences such as drug design, genetic engineering, and diagnostic tools for analysis of genetically engineered pathogens. Active sites can be grouped into families whose members are related by similarity of their functions. Since similar sites exhibit variability in their physico-chemical and structural features, statistical profiling methods capture the shared features robustly in the presence of such variations. In this paper, we adapt Profile Hidden Markov Models (PHMMs) that have been successfully used for analyzing biological sequences, to statistically profile active site families. Since PHMMs can only profile one dimensional sequences, we develop a serialization of the three dimensional active sites that captures certain shared physico-chemical and geometric features of the family. PHMM parameters are learnt using these serialized sequences. While traditional PHMM learning algorithms deal with discrete physico-chemical feature only, we expand it to include geometric features drawn from a continuous probability distribution. Experimental results with our PHMM based method for profiling active sites suggest that it is effective in practice.

1 Introduction

Proteins are essential to the structure and function of all living cells and viruses. Understanding the function of a protein is fundamental for gaining insight into many biological processes. Technically, proteins are amino acid chains that fold into unique three-dimensional structures that cause them to function. In particular, within the protein structure are key areas called *active sites* and biochemical reactions at these sites with other proteins or other chemical substances cause the protein to perform a function of one type or another.

A problem of significant importance in computational biology is this: *Are active sites of different proteins similar?* i.e., do they share similar physico-chemical and geometric properties. Active sites with such shared properties may perform similar functions. Answer to the aforementioned similarity question drives a number of important biological applications. For instance it can be used to predict the function of a protein with a substructure similar to the active site of another protein whose function is known. Another important application is toxicology tools such as the Toxin Knowledge Base (TKB) system that we have developed [10, 17], for automated diagnosis of bioengineered pathogens. In such pathogens the virulent domains of toxins can be hidden in otherwise non-toxic proteins. Specifically, the active site of a non-toxic protein that is similar to that of a toxin, has the potential to become toxic by suitably altering the *residues*¹ in the site.

State-of-the-art techniques for determining active site similarity are exemplified by the SPASM tool [11, 15]. Its inputs include a protein’s structure; the 3-D coordinates of the residues in the active site of another protein whose function is known, substitutions for these residues and a RMSD (root mean square distance) cutoff value. SPASM attempts to identify 3-D substructure(s) of the former protein that are isomorphic to the active site within the specified RMSD cutoff.

There are two problems with the pairwise similarity testing approach embodied in SPASM. Firstly, although there are general guidelines for choosing RMSD values such as “If you use only a few residues (3-5), an RMSD less than one Å tends to be obtained for similar arrangements of residues,”² in general it is a laborious trial and error process. However, the more serious problem is that similarity tests are done separately with one active site at a time. Consequently, it does not

¹Informally, the residues are the elements joined together in the amino acid sequence.

² Å denotes an angstrom which is the distance measure between atoms. One angstrom is 1.0×10^{-10} meters.

exploit the common physico-chemical and structural features that can exist amongst the *family* of active sites of proteins. A family here means that the active sites of all of its members exhibit similar functionality and can also include evolutionarily unrelated proteins that share no overall sequence or fold similarities. Pairwise comparisons may use features that may not be common to all the family members and hence can fail to identify family members, especially “remote”³ members. For instance, SPASM fails to find the similarity between the active sites of UREASE (PDB ID: 2KAU)⁴ and PHOSPHOTRIESTERASE (PDB ID: 1PTA) (both of which belong to the Amidohydrolase super-family and their active sites are shown to be similar in [7]) for reasonable RMSD cutoffs because atoms not directly related to the protein’s function differ a lot in these two structures. Note however that a “*profile*” of the common features in a collection of active sites belonging to a family would have revealed the irrelevance of such atoms and hence would have been excluded as a shared feature. So a principal benefit of profile based methods is that they capture the essential features shared by all of the family members thereby making it possible to determine the similarity of remote members.

Automated construction of active site family profiles to discern common features is a fairly unexplored problem. In this paper we formulate a solution to this problem inspired by the successful profile-based search methods for homologous protein sequences⁵ [18].

Note that physico-chemical⁶ and structural features⁷ similar active sites may exhibit some degree of variability. So similarity notions rooted in statistics can serve as a robust framework for profiling the active sites of a family.

Profile Hidden Markov Model (PHMM), a statistical learning technique used in profile based sequence homology search methods, has been shown to be very effective for capturing sequence similarity [5]. Several software tools based on PHMM have also been developed [6, 8, 9].

We adapt PHMM for profiling the three dimensional active sites in proteins. Since PHMMs can only profile one dimensional sequences, we first develop a se-

³These are active sites that have few features in common with the other family members.

⁴PDB –<http://www.rcsb.org> – is the Protein Data Bank of 3-D protein structures uniquely indexed by an ID

⁵A protein sequence is simply a linear string of amino acids that constitute the primary structure of a protein. Sequences that are similar are referred to as homologues.

⁶In this paper, we use the word physico-chemical and chemical interchangeably.

⁷In this paper, we use the word structural and geometric interchangeably.

rialization of the three dimensional active sites. The next step is to choose a representative set of active site features. Whereas only residue types (such as Histidine, Glutamate, etc) are used as features in PHMMs for protein sequences we will now have to contend with the structural (i.e., geometric) features of active sites also. So in addition to using the atoms’ types in the active site residues we also use their distances from their center of mass as the structural features. Furthermore these distances are assumed to be drawn from a probability distribution. To handle the joint probability of discrete atom type feature and continuous distance feature, we adapt the training phase of PHMM to learn the parameters of this distribution and finally modify the scoring phase to assign a similarity score to the input data.

Summary of Contribution The main contributions of our work include:

- We have developed a novel serialization of three dimensional active sites in proteins.
- We have expanded traditional PHMMs designed for profiling one dimensional sequences of residues to accommodate both physico-chemical and three dimensional geometric features;
- In contrast to manually building templates for searching similar active sites as described in [22, 7], our profile-based method is fully automatic without any loss in accuracy.
- Our profile-based method enables detection of remote members of active site families. In contrast, pair-wise comparison based methods (e.g. SPASM) are unable to find such members. For example, our approach can determine that the active sites of UREASE (PDB ID: 2KAU) and PHOSPHOTRIESTERASE (PDB ID: 1PTA) are similar so they are members of the same family which SPASM fails to do as was mentioned earlier.
- Finally and most importantly, in our approach there is no need to either figure out RMSD cutoff or specify residue substitutions manually as is required in SPASM. See Section 4.4 for details.

The rest of the paper is organized as follows. Section 2 presents an overview of protein active sites and PHMMs to set the context for understanding the rest of the paper. Section 3 provides details of our adaptation of PHMM for active site profiling. Section 4 presents experimental results of our approach. Related work appears in Section 5 and conclusions in Section 6.

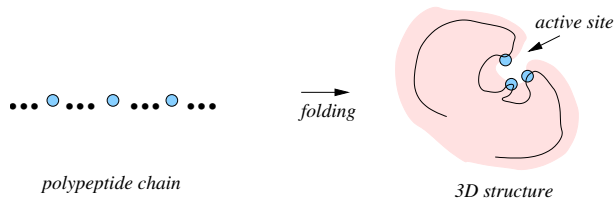


Figure 1: Formation of an Active Site

2 Technical Preliminaries

In this section we review the technical background needed to understand our technical approach. In particular the review focuses on active sites in proteins and PHMM.

2.1 Protein Active Site The building blocks of proteins are twenty amino acids. Examples of these include Alanine, Valine, Histidine, Glycine, etc. They are usually referred to by their symbolic (3-letter and 1-letter) abbreviations e.g., the 3-letter ALA or the 1 letter A for Alanine, VAL or V for Valine and so on. All of the twenty amino acids have in common a central carbon atom (C_α) to which are attached a hydrogen atom, an amino group (NH_2), and a carboxyl group ($COOH$). The rest of an amino acid, which is called the *side chain*, is different for different amino acids. Amino acids are joined end-to-end to form a polypeptide chain when the carboxyl group of one amino acid condenses with the amino group of the next to eliminate water. The remaining part of an amino acid in a polypeptide chain is called a *residue*.

The polypeptide chain of a protein folds in space to form the three-dimensional structure of the protein. The folding of the polypeptide chain typically creates a crevice or cavity on the protein surface. This crevice, called an *active site*, contains a set of residue side chains which might be far apart in the polypeptide chain. They are brought together in the 3-D structure and are disposed in such a way that they can make noncovalent bonds only with certain partners, which can be a protein, DNA, metal ion, etc. The 3-D structure of a protein, especially the localized structure of its active site, determines the functional properties of the protein. Figure 1 sketches the formation of an active site. Note that a protein can have several active sites.

2.2 Profile Hidden Markov Model A PHMM is a statistical learning-based technique for modeling DNA and protein sequences families [5]. The underlying principles of PHMMs are based upon the mathematics of Hidden Markov Models [19] which have found wide applicability in sequence analysis tasks.

HBA_HUMAN	...VGA--HAGEY...
HBB_HUMAN	...V----NVDEV...
MYG_PHYCA	...VEA--DVAGH...
GLB3_CHITP	...VKG-----D...
GLB5_PETMA	...VYS--TYETS...
LGB2_LUPLU	...FNA--NIPKH...
GLB1_GLYDI	...IAGADNGAGV...
	*** *****

Figure 2: A Segment from the Multiple Alignment of 7 Globin Protein Sequences

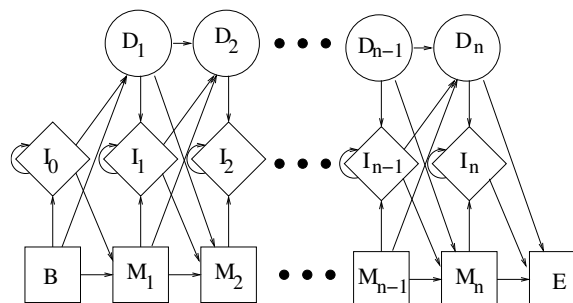


Figure 3: PHMM structure

Protein sequences typically come in families. Members of a family have a common ancestor and normally maintain the same or related function. Although they have diverged during evolution through insertions and deletions, their functional residues are usually conserved which is revealed by a multiple alignment of family members (See Figure 2). In Figure 2, the columns marked with stars are conserved since more than half of the sequences have a residue instead of a dash present in the column. The two non-starred residues in GLB1_GLYDI correspond to insertions. If a sequence has a dash in a conserved column, then it has undergone a deletion.

PHMMs structures, as shown in figure 3, are specialized to capture such conserved residues as well as insertions and deletions in sequence families. Each column, from 1 to n , has three states - a *match*, *insert*, and *delete* state. Intuitively, match states correspond to conserved residues among sequences while insert and delete states correspond to divergence in sequences from a common ancestor due to insertions and deletions respectively. For protein sequences, the emission symbols are the twenty amino acids. Match and insert states emit residues while delete states are non-emitting silent states.

3 Profiling Active Sites with PHMM

Note that active sites with similar functions can exhibit variability in their physico-chemical and geometric configurations. So any technique for profiling active sites should factor in such variations. PHMMs have been used for biological sequence analysis with a high degree of success since they can statistically capture commonalities and variations among sequences that have evolved from a common ancestor. Thus they have the potential to serve as a robust framework for profiling active sites also.

However, adapting PHMMs for this problem is not entirely straightforward. Let us examine the underlying issues. Firstly, observe that PHMM is a sequential model in the sense that it was developed to handle protein sequences which are simply 1-D strings of amino acids. On the other hand active sites are 3-D structures. So the immediate problem is one of serializing these 3-D structures in such a way that salient aspects of their physico-chemical and geometric properties are still retained. Secondly, each state in a traditional PHMM emits only one discrete symbol (i.e., an amino acid) at a time. For active sites these emissions must include both physico-chemical features such as the discrete valued residue types as well as geometric features. So emissions are tuples ranging over the physico-chemical and geometric feature set. A robust description of geometric configurations of active sites is best done using continuous measures. Hence in contrast to traditional PHMMs where only discrete probabilities of emission symbols are estimated we need to estimate the joint distribution of physico-chemical and geometric features. In the rest of this section we describe how we address these issues.

3.1 Serializing Active Sites Since PHMM is a sequential model the task now is to identify a set of 3-D features and serialize them. This serialization will represent the observation sequence corresponding to an active site.

The primary issue in serialization is inventing an ordering for the sequence. For primary protein sequences of amino acid chains this is simply the position of the residue in the chain. For 3-D active sites there is no such obvious ordering. Let us first examine the desiderata for such an ordering. Ideally, if a and b are two conserved atoms in one active site, a' and b' are atoms in another active site corresponding to a and b , respectively, then the order of a and b in the serialized sequence derived from the former active site should be consistent with that of a' and b' in the sequence derived from the latter. A candidate for such an ordering is the distance of the atoms in the active site from their

Atom Name	X-Cord	Y-Cord	Z-Cord	Distance
N	36.729	107.613	20.276	2.427
CA	35.813	107.722	21.395	1.133
C	36.031	109.051	22.149	1.732
O	37.157	109.446	22.496	2.519
CB	35.875	106.405	22.220	1.016
CG	34.949	106.290	23.394	1.622
OD1	33.858	106.833	23.442	2.169
OD2	35.341	105.659	24.387	2.602

Table 1: Example Active Site Atoms

center of mass. Given a set of n atoms with coordinates $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_n, y_n, z_n)$, their center of mass is the expression:

$$\left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n y_i, \frac{1}{n} \sum_{i=1}^n z_i\right)$$

In other words the center of mass is the average over each of the coordinate positions of the atoms. For illustration, suppose an active site contains only one residue D260 with atoms whose coordinates are listed in the first four columns of Table 1. The 3-D coordinate of their center of mass is $(35.719, 107.377, 22.470)$. Distances of each atom from the center of mass are shown in the last column of Table 1. The ordering of atoms arranged in ascending order of their distances from the center of mass is: $\langle \text{CB}, \text{CA}, \text{CG}, \text{C}, \text{OD1}, \text{N}, \text{O}, \text{OD2} \rangle$.

To capture physico-chemical feature, we adopt the atom classification in [16] which classifies all non-hydrogen atoms in proteins into 40 classes according to the atom location (side-chain or backbone), connectivity, and chemical nature. We denote the atom type by *ResidueName.AtomName*, which can be unambiguously mapped to an atom type in [16]. For example, the type of the first atom in Table 1 is represented by D.N.

As far as geometric feature is concerned, an obvious idea is to use an atom’s 3-D coordinate. However, the coordinates of atoms from two active sites are comparable only after those active sites are superposed. Typically, superposing algorithms take two point sets with each point represented by its (x, y, z) coordinate, and perform rigid transformations such as translation and rotation to minimize the RMSD of these two point sets. Since these points are assumed to be typeless, any two points are always superposable. But the problem here is that superposed positions may not be compatible with the atom types at those positions (e.g., in general nitrogen and oxygen atoms cannot be superposed). There are tools such as SPASM [11, 15] that allow users to define superposable atom types. The main problem with this is that knowledge about

2DHC	<D.OD2,2.91>	-	-
1CHO	<D.OD2,3.08>	-	<H.CB ,3.17>
1ACE	<E.OE1,2.32><H.CD2,2.32><H.CB ,2.45>		
	*		*
2DHC	-	<H.CA ,3.10><H.O ,3.20>	
1CHO	<D.CG ,3.22><H.CA ,3.63>	-	
1ACE	-	<H.CA ,3.31><E.CD ,3.46>	
		*	*
2DHC	<H.C ,3.46><D.CG ,3.47><D.OD1,3.50>		
1CHO	<H.C ,4.06>	-	-
1ACE	<H.C ,3.94>	-	<E.OE2,4.06>
	*		*

Figure 4: A Segment of a Multiple Alignment

what are superposable atom types varies from family to family. A desiderata of geometric feature is that it be preserved under serialization. Features that use relative instead of absolute positions can satisfy such a requirement. Observe that distances of atoms to their center of mass are relative quantities and hence can serve as a geometric feature.

In summary, our feature set is the pair $\langle AtomType, Distance_To_CenterOfMass \rangle$, where the first element is the physico-chemical feature and the second is the geometric feature. The general form of an observation sequence corresponding to an active site following serialization using our feature set will be: $\langle t_1, d_1 \rangle, \langle t_2, d_2 \rangle, \dots, \langle t_n, d_n \rangle$ where n is the number of atoms in the active site, t_i is the atom type and d_i is the distance to the center of mass for $i = 1, \dots, n$, and $d_i < d_{i+1}$ for $i=1, \dots, n-1$. For our example active site, it is $\langle D.CB, 1.016 \rangle, \langle D.CA, 1.133 \rangle, \langle D.CG, 1.622 \rangle, \langle D.C, 1.732 \rangle, \langle D.OD1, 2.169 \rangle, \langle D.N, 2.427 \rangle, \langle D.O, 2.519 \rangle, \langle D.OD2, 2.602 \rangle$.

3.2 PHMM for Active Sites When observation sequences of multiple active sites with similar function are put together, one can identify which atoms are conserved by aligning them. Figure 4 shows a segment of the alignment of three similar active sites⁸, namely, acetylcholinesterase (PDB ID: 1ACE) with residues S200, E327, and H440 ; chymotrypsin (PDB ID: 1CHO) with residues H57, D102, and S195; haloalkane dehalogenase (PDB ID: 2DHC) with residues D124, D260, and H289. Although their constituting residues are different, all of them perform similar catalytic function.

This alignment reveals that the consensus sequence has six atoms (see columns marked by '*' in the figure). The atoms appearing in the non-starred columns are insertions. Observe also that the sequence of 2DHC

goes through a deletion between the first match and the third match; 1CHO goes through two deletions: one between the third match and the fifth match and the other after the fifth match.

We can learn the PHMM parameters (transition and emission probabilities) from such multiple alignments by smoothed maximum-likelihood parameter estimation using the frequency counts of transition and emission events. However, it is labor intensive to come up with such a multiple alignment. One can also learn these parameters from unaligned sequences. First, the number of match states (i.e. the length of the PHMM) is estimated by taking the average length of the training sequences. Then the Baum-Welch algorithm [4] is applied to estimate the transition probabilities and emission probabilities.

We adapt this process for learning PHMM parameters from training data consisting of unaligned serialized active site sequences belonging to a family. First, we estimate the length of the PHMM from the training sequences. This is the average length of the sequences. For example, the average length for the sequences in Figure 4 without the dashes is six.

To learn the other two PHMM parameters, we modify the Baum-Welch algorithm. Since emission symbols are pairs $\langle atomtype, distance \rangle$, we will need to compute the joint distribution of these pairs for each state. Making the standard independence assumption done in HMMs, namely, that the random variables in the joint distribution are independent, the probabilities of the atom types and their distances are computed separately. Let us define the probability of atom type t in a state as $P(t)$ and the probability of the distance d from center of mass as $P(d)$. We calculate the emission probability $P(b)$ of the emission symbol $b = \langle t, d \rangle$ to be $P(t) \times P(d)$.

The distance from the center of mass is a continuous feature. We assume that its probability distribution is generated by a multivariate Gaussian distribution whose probability density function is:

$$\frac{e^{-\left(\frac{d-\mu}{2\sigma^2}\right)^2}}{\sigma\sqrt{2\pi}}$$

where d is the distance, μ is the mean and σ is the standard deviation of distances to the center of mass. Suppose the distances to the center of mass from atoms that are emitted by a state are d_1, \dots, d_m . We compute μ and σ at this state using the expressions:

$$\mu = \frac{1}{n} \sum_{i=1}^n d_i$$

⁸Because of width constraints the sequences in the figure run over to multiple lines.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (d_i - \mu)^2 + \epsilon}$$

The small constant ϵ is added so that σ is always positive even when $n = 1$.

Recall that we need 42 parameters to describe the emission distribution for each state. Forty of these parameters correspond to the emission probabilities of the 40 atom types and they must sum up to 1. The remaining two are μ and σ that represent the distribution of the distances of atoms emitted from the state to their center of mass.

For a set of unaligned sequences, Baum-Welch algorithm iteratively updates the parameters of the model to increase the overall probability of the set of training sequences to be generated by the model. We modify the Baum-Welch algorithm to take into account the new emission parameter set and the joint emission probability. At each step of iteration, we calculate the individual probabilities of atom type and distance from center of mass and multiply these probabilities to get the joint probability. For a family of observation sequences of active sites, this modified Baum-Welch algorithm is used to estimate the parameters of the PHMM that profiles this family.

Armed with a PHMM M trained on a family S of serialized active site sequences we can now answer questions about similarity of active sites. To determine if a protein has substructures similar to the active sites in S we proceed as follows: First we find candidate substructures in the protein structure. This can be done with tools such as MOE Active Site Finder [13] and Q-SiteFinder [14]. The advantage of using such tools is that the accuracy of detecting similar active sites can potentially be improved. This is because a method such as ours and SPASM that represents an active site as a set of atoms can only capture the geometric and physico-chemical features of such atoms. It is not known whether these atoms are located in a cavity on protein’s surface. We have learned from ?? that active sites are usually cavities on protein’s surface. Therefore with tools such as Active Site Finder, we can remove false matches which do not have such a property.

Then a serialized observation sequence is generated for each candidate substructure. Those are the candidate observation sequences for the protein.

For each such observation sequence x , we apply the Viterbi algorithm [19] to compute the the probability of its most likely path (i.e., state sequence) in the PHMM model M . In particular, the Viterbi algorithm efficiently computes a state sequence y' that maximize the conditional joint probability $P(x, y|M)$, i.e., $y' = \arg \max_y P(x, y|M)$.

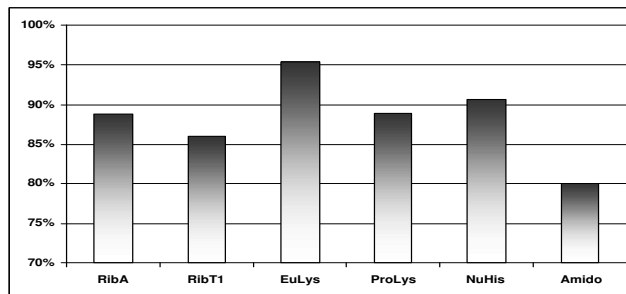


Figure 5: Precision Performance of Protein Families

We modify Viterbi algorithm to compute the probability of observing a pair $\langle t, d \rangle$ at a state. Specifically, we compute the probabilities of observing atom type t and distance d separately using the emission distribution parameters of that state, and then multiply them to get the emission probability of the pair.

Next, we compute $p(x, y'|M')$ where M' is a random model that is identical to M in length and transition probabilities.

The emission parameters are assumed to be uniform for all the insert and match states. These state-independent parameters are computed as follows:

1. The emission probability for atom type a is $\sum_r \text{where } a \in r \frac{q(r)}{\text{num of atoms in } r}$, where r is a residue and $q(r)$ is the frequency of r (see Section 2.2).
2. Randomly sample substructures from PDB, each of which contains the same number of residues as the training examples.
3. For each such substructure, compute the center of mass and the distances of the atoms to this center.
4. Compute the mean μ and the standard deviation σ over all distances and over all substructures.

Finally we compute the base 2 log-odds ratio $\log(\frac{P(x, y|M)}{P(x, y|M')})$ called the *bit score*. If this score falls above a threshold then x is said to be a member of S . The threshold is a global value therefore we do not need to choose such a value for each family. For example, a threshold 0 means that x is taken as a member of S if it fits the model of S better than the random model.

4 Evaluation

We implemented our PHMM-based profiling of active sites. In this section we report its experimental results on six protein families/super-families. Since active site profiling is not well studied and therefore no

Table 2: Data Statistics for Different Protein Families

Protein Families	No of Mem- bers	Average No of Active Site Atoms	Size of Training Set	Size of Test Set	No of Pos- itive Exam- ples in the Test Set	No of Neg- ative Exam- ples in the Test Set
Ribonuclease A	27	34	25	20	9	11
Ribonuclease T1	19	28	12	20	7	13
Eukaryotic Lysozyme	107	17	84	35	23	12
Prokaryotic Lysozyme	94	17	75	42	19	23
Nu:-His-Elec catalytic triad	272	25	176	169	96	73
Amidohydrolase	48	17	8	24	9	15

Table 3: Data Statistics for Subfamilies of Nu:-His-Elec Catalytic Triad

Protein Sub Families	No of Mem- bers	Average No of Active Site Atoms	Size of Training Set	Size of Test Set	No of Pos- itive Exam- ples in the Test Set	No of Neg- ative Exam- ples in the Test Set
sub1	226	25	150	110	76	34
sub2	19	25	10	24	9	15
sub3	9	26	5	15	4	11
sub4	15	24	8	20	7	13

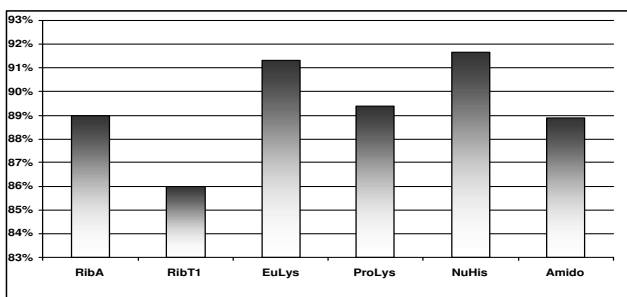


Figure 6: Recall Performance of Protein Families

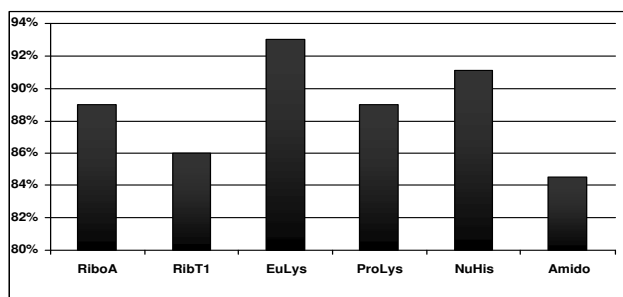


Figure 7: F-Measure Performance of Protein Families

benchmark dataset is available, we extracted our experiment dataset of these six families/super-families from [22] and [7], in which profiles of their active sites are manually built. Among them members of the NuHis-Elec/Amidohydrolase super-family do not share significant sequence similarity and therefore sequence homology tools such as BLAST [2, 3] can not detect family members with high accuracy. We evaluated the performance of our prototype implementation, studied the impact of geometric features and physico-chemical features, and compared with SPASM which is a well-known tool for active site similarity search. Our experiments show that:

- The active site profiles that are automatically trained from examples with our PHMM adaption have similar accuracy as compared to manually built profiles.
- Both geometric/structural and physico-chemical features are important for the accuracy of active

site profiles. However, geometric features have relatively higher impact on the accuracy.

- When using SPASM for active site similarity search, there is no uniform RMSD cut-off or allowed substitution and therefore such parameters have to be chosen manually per family. In contrast, there is no such need in our approach.

This section is organized into the following subsections: The experimental setup for the evaluation; the performance metrics measured; the experimental results of our approach and SPASM; and discussion of the results.

4.1 The Experimental Setup The evaluation was conducted over different sets of protein families detailed below.

Protein Families We developed PHMM profiles for six different protein families, namely, *Ribonuclease A*, *Ribonuclease T1*, *Eukaryotic Lysozyme*, *Prokaryotic*

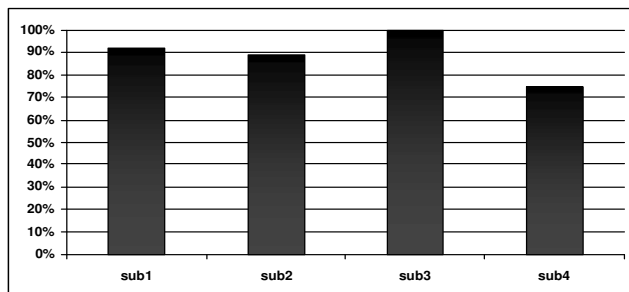


Figure 8: Precision Performance of Protein Sub Families of Nu:-His-Elec Catalytic Triad

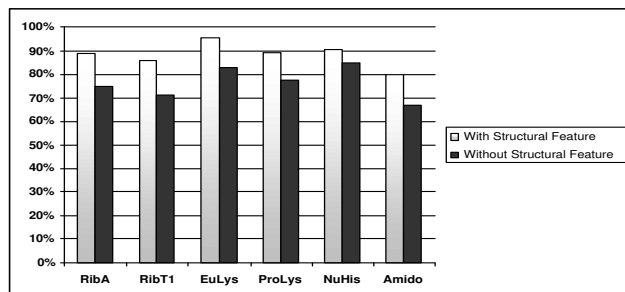


Figure 11: Effect of Structural Feature on Precision Performance

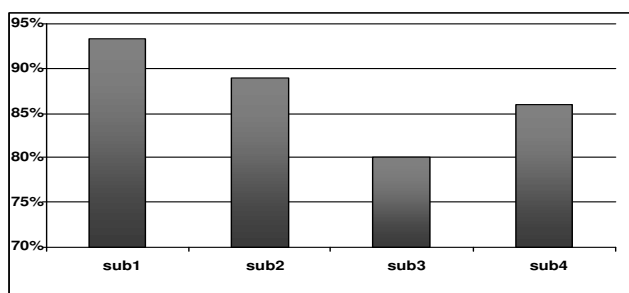


Figure 9: Recall Performance of Protein Sub Families of Nu:-His-Elec Catalytic Triad

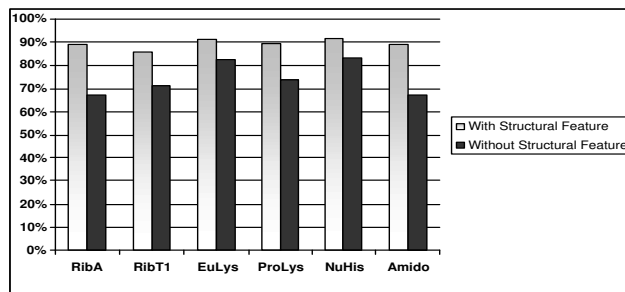


Figure 12: Effect of Structural Feature on Recall Performance

Lysozyme, *Nu:-His-Elec catalytic triad*, *Amidohydroxylase*. We denote them as RibA, RibT1, EuLys, ProLys, NuHis, Amido.

The Nu:-His-Elec family is further divided into five subfamilies according to the residues that comprise the catalytic triads, which are Ser-His-Asp, Ser-His-Glu, Asp-His-Asp, Ser-His-Trp, and Cys-His-Asn. They are denoted by sub1, sub2, sub3, sub4, and sub5, respectively. We built profiles for the first four subfamilies. The fifth subfamily has only 3 members; so we did not build profile for this family.

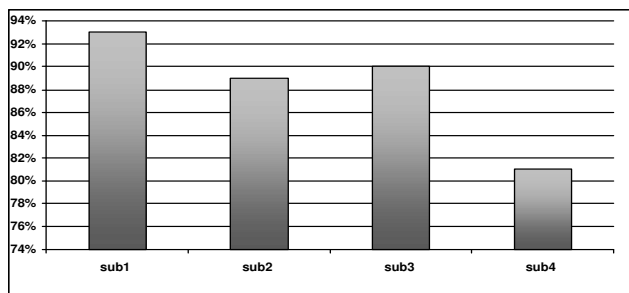


Figure 10: F-Measure Performance of Protein Sub Families of Nu:-His-Elec Catalytic Triad

Training and Test Data The active sites per family were divided into two mutually exclusive training and test sets. The active sites of a family included in the test set associated with the family were labeled as positive test examples. For each family, we augmented its test set with a subset of active sites belonging to other four families. These augmented active sites were labeled as negative test examples. Statistics associated with the experimental data used are listed in table 2 for the six families and in table 3 for the subfamilies of Nu:-His-Elec catalytic triad.

We built a separate PHMM per family. The parameters were learnt using the training set associated with the family. The global threshold for the log-odds ratio was set to 0. For these profiles we used both structural (distance from center of mass) and chemical (atom type) features as emission symbol of the PHMM. **Feature Variation** We also built PHMM per family by using only structural feature as emission symbol. This helps us to observe the effect of structural feature on the performance of the PHMM. Similarly we built PHMM per family by using only chemical feature as emission symbol to observe the effect of this feature.

4.2 Performance Metrics We evaluated the PHMM with respect to three performance metrics:

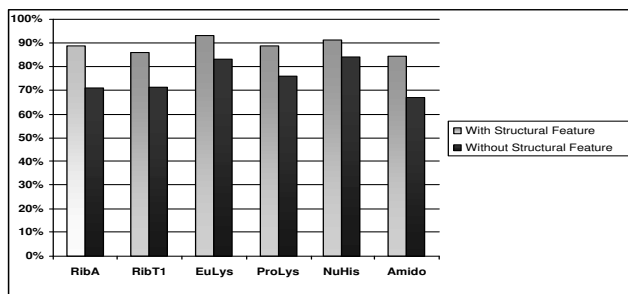


Figure 13: Effect of Structural Feature on F-Measure Performance

recall, precision and f-measure⁹ using the test data set constructed above.

Observe that an active site in the test data for a family was uniquely labeled as a positive or negative test example. These labels are used to classify the similarity results produced by PHMM on the test data into true positives, false positives, true negatives and false negatives. Based on these classifications the recall/precision/F-measures are directly computed from their definitions.

4.3 Experimental Results Here we report the experimental results in terms of precision, recall and f-measure. Figure 5 shows the precision performance for each of the protein families. They range from 86% (for RibT1) to 95% (for EuLys). Figure 8 shows the precision performance for each sub-families of NuHis. Figure 6 shows the recall performance for each of the protein families. They range from 88% (for RibT1 and NuHis) to 91% (for NuHis). Figure 9 shows the recall performance for the subfamilies of NuHis. We also calculated f-measure for each of the families. Figure 7 shows the f-measure for each of the families. It ranges from 86% (for RibT1) to 93% (for EuLys). Figure 10 shows the f-measure performance for the sub-families of NuHis.

Effect of Feature Variation Figure 11, 12 and 13, shows the effect of structural feature (distance from center of mass) on precision, recall and f-measure performance for each of the protein families. When structural features are not used as emission symbol, precision reduces by 5% (for NuHis) to 15% (for RibT1), recall reduces by 8% (for EuLys) to 22% (for RibA) and f-

⁹Recall value for a protein family is defined as the ratio of correctly identified proteins (which are members of the family) over the total number of family members present in the test set. For precision, the denominator is taken as the total number of proteins (both positive as well as negative test examples) present in the test which are identified as members of the family. F-measure is defined as the harmonic mean of recall and precision

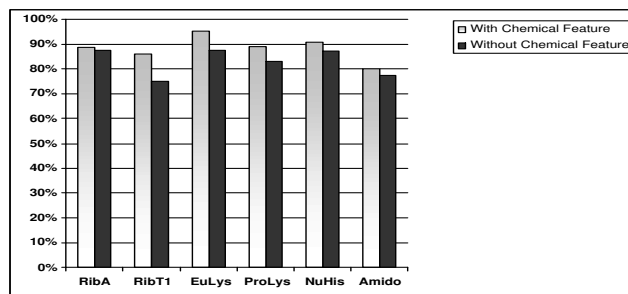


Figure 14: Effect of Chemical Feature on Precision Performance

measure decreases by 7% (for NuHis) to 18% (for RibA). Figure 14, 15 and 16, shows the effect of chemical feature (Atom Type) on precision, recall and f-measure performance for each of the protein families. As we remove chemical feature from emission symbol, precision reduces by 1% (for RibA) to 13% (for EuLys). Recall is unchanged for RibT1 and EuLys but decreases by as high as 12% for RibA. F-measure reduces by 4.8% (for NuHis) to 8% (for ProLys).

The experimental result of feature variation suggests that both structural and chemical features effect the performance of the PHMM. But structural feature has higher impact on the performance than chemical feature.

4.4 SPASM Performance We also conducted experiments on using SPASM to identify similar active sites in the six families/super-families listed in Table 2. For RibA, RibT1, EuLys and ProLys, we randomly picked one member and its active site was given as the input to SPASM. For NuHis and Amido, one member from the largest family of the super-family was randomly taken and its active site was the input to SPASM. For each input active site, we used SPASM to search for similar active sites in the database which consists of the three-dimensional structures of all members of the corresponding family/super-family. Table 4 and 5 show the experimental result of SPASM when we varied RMSD cut-offs and allowed substitutions in terms of BLOSUM45 cut-offs¹⁰. A star (*) in an entry denotes that SPASM aborts because the number of hits has exceeded a pre-set value(1,000).

Table 4 shows the number of true hits and the number of total hits for each family when no substitution is allowed and the RMSD cut-offs increases from 1.5Å to

¹⁰BLOSUM45 is a matrix defining a score for each pair of residue types. If the score of a pair of residues is greater than or equal to the cut-off, then they are considered as substitutable by each other

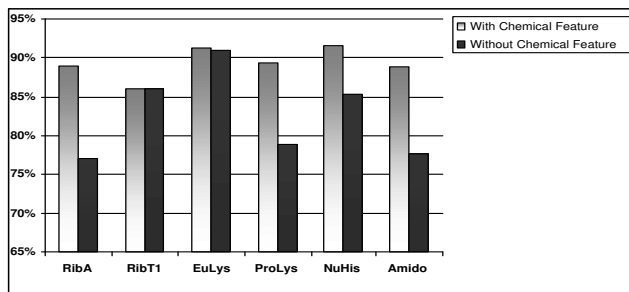


Figure 15: Effect of Chemical Feature on Recall Performance

4.5Å. In most cases the cut-off 1.5Å works fine. But with this cut-off, the recall is only about 87% for NuHis and 41% for Amido. When increasing the RMSD cut-off, the recall increases and the precision decreases. The sensitiveness to the change of RMSD cut-off is quite different from family to family. For example, RibT1 is much more stable than Prolys.

Table 5 shows the number of true hits and the number of total hits for each family when RMSD cut-off is fixed to 1.5Å and the allowed substitution defined by BLOSUM45 cut-off is varied. The most distant (in terms of BLOSUM45 score) pair of corresponding residues in our dataset is ASP and TRP in sub1 and sub4 of NuHis, respectively. The score is -4. That is to say, the BLOSUM45 should be set to -4 to find all similar active sites. However, we can see from Table 5 that the precision is already too low for some families when the BLOSUM45 cut-off is set to 0. It might look that the performance of SPASM is good enough when no substitution is allowed. But the recall can be as low as 1% when a different input active site is taken for NuHis.

Our experiments with SPASM show that there is no uniform way of deciding the RMSD cutoff and allowed substitutions needed for such tools. It is basically a painful trial-and-error process.

4.5 Discussion The experimental performance suggests that PHMM-based methods described in this paper for determining similarity of active sites works well in practice. The performance of our method is comparable to the template-based manual approaches as described in [22, 7]. The PHMM constructed for each family exhibits reasonably high recall, precision and f-measure values. A high degree of shared features by family members results in higher performance metrics. For instance the active sites of Eukaryotic Lysozyme shares many atom types along with their geometric configuration. This is reflected by its high recall, precision

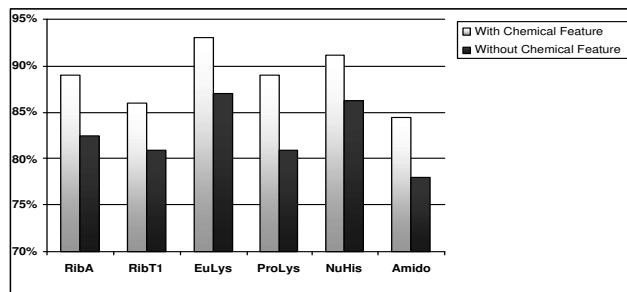


Figure 16: Effect of Chemical Feature on F-Measure Performance

and f-measures (95%, 91%, 93%). On the other hand the low degree of shared features observed in Ribonuclease T1 has translated into low recall, precision and f-measure values (86%, 86%, 86%). Our experimental results also demonstrate the relatively higher impact of structural/geometric feature than chemical feature on the performance of PHMM.

5 Related Work

We review here computational tools and techniques related to the problem of determining similarity of active sites.

On the tool front the best known system is SPASM [11, 15]. It takes the pair (protein structure, target active site) as the input and finds substructures in the protein that are similar to the active site. As we had discussed earlier comparing a substructure to an active site independently of other members of the active site’s family fails to exploit the commonality amongst them. Consequently, it can fail to establish similarity with some family members, especially remote ones. A profile based approach as is done in the paper addresses this problem since profiles can capture common features of family members.

The idea of profiling active sites was first explored in the context of building the PROCAT database [22, 23], in which the term “functional template” was used for what we refer to as the active site profile in this paper. In PROCAT, functional templates are manually defined for several enzyme families. For example, it includes templates for Ribonuclease A and the five subfamilies of Histidine-based catalytic triad (see Tables 2 and 3). These templates consist of only a subset of atoms in the active site residues. For instance, only the O^γ atom is included in the template for the Ser-His-Asp subfamily. The decision of which atoms to include is done manually through close inspection of the structures and functional mechanisms of all the proteins in the

RMSD cut-off	1.5		2.5		3.5		4.5	
	true hits	total hits	true hits	total hits	true hits	total hits	true hits	total hits
RibA	28	28	30	30	34	119	34	158
RibT1	18	18	18	18	18	18	18	19
EuLys	95	104	98	125	103	209	106	310
ProLys	93	198	93	485	93	529	88	1010*
NuHis	192	194	202	256	221	728	164	1004*
Amido	7	7	7	7	10	17	11	99

Table 4: Varying RMSD Cut-off

BLOSUM45 cut-off	no substitution		3		0		-2	
	true hits	total hits	true hits	total hits	true hits	total hits	true hits	total hits
RibA	28	28	28	28	28	57	28	914
RibT1	18	18	18	18	18	79	18	280
EuLys	95	104	95	104	15	1034*		
ProLys	93	198	93	217	7	1015*		
NuHis	192	194	192	194	82	1004*		
Amido	7	7	7	7	7	8	7	15

Table 5: Varying Allowed Substitution

family. The template so constructed captures the features shared by the family members. The problem here is that template construction is a manual process thereby limiting scalability. In contrast our approach to “learn the templates” is highly automated.

A more recent work is Catalytic Site Atlas (CSA) [21], a database documenting enzyme active sites and catalytic residues present in enzymes with 3-D structures. The active sites are labeled either original or derived. The former are extracted from scientific literature while the latter are associated with proteins whose primary sequences are homologous to the primary sequences of proteins containing the original active sites. An original active site and all of its derived sites constitutes a family. Templates with shared features are again constructed manually for each family.

MultiBind is yet another recent work that takes a set of active sites and automatically aligns all of them [20]. The multiple alignment reveals what are the subset of atoms that are conserved among all the active sites in the set. Firstly, this approach is not statistical unlike ours. But the more important difference is that multiple alignment alone does not provide any quantitative measure of how close an active site is to the aligned sites. Without such measures it is not possible to algorithmically deduce similarity.

PHMMs were used for profiling entire protein structures in [1]. The 3-D structure is serialised into a sequence of 3-D coordinates. In other words this work uses only one geometric feature. Such an approach is

useful for determining similarity of entire protein structures whose superposition has the lowest RMSD value. As we had discussed earlier (see Section 3.1) 3-D coordinates alone may not adequately capture the salient shared features of the family. Good superpositions in terms of low RMSD values may produce incompatible atom types at the superposed positions. Factoring in both physico-chemical and geometric features as is done in our approach can result in more accurate determinations of similarity and our experimental results seem to validate this hypothesis.

Finally, we remark that protein functions can also be predicted based on sequence homology or overall structure similarity. However it has been observed that there is no significant correlation between conservation of sequences, structures and active sites. Hence function prediction by detection of substructures in proteins that are similar to active sites of proteins with known functions complement those based on sequence homology and structural similarity methods.

6 Conclusion

In this paper we described computational techniques for statistically profiling active sites in proteins. Specifically we adapted the successful PHMM based approach for analysis of linear sequences to encode the profiles of 3-D active sites belonging to a family. Our preliminary experience with a prototype implementation of our approach indicates that it is effective in practice.

There are several avenues for future work along the

lines pursued in this paper. In our experimentation, we only utilized one geometric feature, namely distance to the center of mass. This is a relatively coarse measure. It should be possible to incorporate other geometric features as well, such as pair-wise distances between atoms. We can also incorporate additional physico-chemical features such as stereochemical and charge constraints of the active sites. Adding these features will yield richer profiles and may further improve the accuracy of prediction.

Another major idea is to depart from the linear structure of HMMs. Transitions in HMMs depend only on the previous state. While HMMs are appropriate for modeling primary structures of proteins, active sites are 3-D structures and a state transition is necessarily influenced by a set of neighboring states. Linearizing 3-D structures fails to capture such dependencies between neighboring states. Hidden Markov Random Fields (HMRF) [12] relax this limitation. HMRFs operate over undirected state graphs. The probability distribution of a random variable associated with a state in a HMRF is a function of the states in its neighborhood as defined by the graph structure. It appears that HMRFs offer a natural computing model for profiling active sites. Estimating HMRF parameters for this problem is a promising research direction.

References

- [1] V. Alexandrov and M. Gerstein. Using 3d hidden markov models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformatics*, 5, 2004.
- [2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- [4] L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972.
- [5] S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [6] S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.
- [7] L. Holm and C. Sander. An evolutionary treasure: Unification of a broad set of amidohydroloases related to urease. *Proteins: Structure, and Genetics*, 28:72–82, 1997.
- [8] R. Hughey and A. Krogh. Hidden markov models for sequence analysis: Extension and analysis of the basic method. *Comput. Appl. Biosci.*, 12:95–107, 1996.
- [9] K. Karplus, C. Barrett, and R. Hugher. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14:846–856, 1999.
- [10] M. Kifer, I. Ramakrishnan, A. Ramanathan, C. Zhao, S. Jayaraman, and S. Swaminathan. Tkb: Toxin knowledge base for discovering bio-engineered threats. In *ISMB 2005*, 2005. Tool Demo and Poster.
- [11] G. Kleywegt. Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, 285:1887–1897, 1999.
- [12] H. Kuensch, S. Geman, and A. Kehagias. Hidden markov random fields. *Annals of Applied Probability*, 5:577–602, 1995.
- [13] P. Labute and M. Santavy. Locating binding sites in protein structures. <http://www.chemcomp.com/journal/sitefind.htm>.
- [14] A. Laurie and R. Jackson. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.
- [15] D. Madsen and G. Kleywegt. Interactive motif and fold recognition in protein structures. *J. Appl. Cryst.*, 35:137–139, 2002.
- [16] F. Melo and E. Feytmans. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, 267:207–222, 1997.
- [17] S. Mukherjee, C. Zhao, and I. Ramakrishnan. Profiling protein families from partially aligned sequences. In *SIAM Conference on Data Mining*, 2006.
- [18] J. Park, K. Karplus, C. Barrett, R. Hugher, D. Hausler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, 284:1201–1210, 1998.
- [19] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2), 1989.
- [20] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H. Wolfson. recognition of binding patterns common to a set of protein structures. In *RECOMB*, pages 440–455, 2005.
- [21] J. Torrance, G. Bartlett, C. Porter, and J. Thornton. Using a library of structural templates to recognize catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.*, 347:565–581, 2005.
- [22] A. Wallace, N. Borkakoti, and J. Thornton. Tess: A geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Science*, 6:2308–2323, 1997.
- [23] A. Wallace, R. Laskowski, and J. Thornton. Derivation of 3d coordinate templates for searching structural databases: Application to the ser-his-asp catalytic triads of the serine proteinases and lipases. *Protein Science*, 5:1001–1013, 1996.