

# Computing Statistical Profiles of Active Sites in Proteins\*

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

Subramanyam Swaminathan  
Biology Department  
Brookhaven National Laboratory  
Upton, NY 11973-5000, 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. In this paper, we adapt Profile Hidden Markov Models (PHMMs) to statistically profile active site families. We develop a serialization of the three dimensional active sites that captures certain shared physico-chemical and geometric features of the family. Experimental results with our PHMM based method for profiling active sites suggest that it is effective in practice.

## 1 Introduction

Active sites of proteins are key areas within proteins’ three-dimensional structures where biochemical reactions with other proteins or other chemical substances happen. 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. 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 [3], for automated diagnosis of bioengineered pathogens, which are non-toxic proteins containing an active site that are similar to that of a toxin and thus have the potential to become toxic by suitably altering the amino acids in the site.

State-of-the-art techniques for determining active site similarity are exemplified by the SPASM tool [4].

Given the structure of a protein and the structure of an active site, SPASM attempts to identify 3-D substructure(s) of the former protein that are isomorphic to the active site within user-specified RMSD (Root Mean Square Distance) cutoff.

There are two problems with the pairwise similarity testing approach embodied in SPASM. First, choosing the RMSD cutoff is a laborious trial and error process. The second and the more serious problem is that similarity tests are done separately with one active site at a time. Consequently, it does not exploit the common physico-chemical and structural features that can exist amongst the *family* of active sites of proteins. Pairwise comparisons may use features that may not be common to all the family members and hence can fail to identify the similarity between family members. For instance, SPASM fails to detect the active site similarity of UREASE (PDB ID: 2KAU)<sup>1</sup> and PHOSPHOTRIESTERASE (PDB ID: 1PTA), both of which were shown to be members of the Amidohydrolase family in [2]. This is because atoms in the active sites of these two proteins that are not directly related to the proteins’ function exhibit considerable differences. 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 make it possible to determine the similarity.

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 based on Profile Hidden Markov Models (PHMMs). PHMM is a statistical learning technique which has been shown to be very effective for capturing protein sequence similarity [1]. We adapt PHMM for profiling the three dimensional active sites in proteins. Since PHMMs can only profile one dimensional sequences, we first develop a serialization of the three dimensional active sites. The next step is to choose a representative set of active site features. Whereas only amino acid types (such as Histidine, Glutamate, etc) are used as fea-

\*This research is supported by U.S. Army Medical Research Acquisition Activity Contract DAMD17-03-1-0520 and New York State Department of Health Contract C020593.

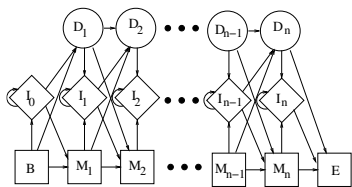
<sup>1</sup>PDB –<http://www.rcsb.org> – is the Protein Data Bank of 3-D protein structures uniquely indexed by an ID.

tures 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 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.

The rest of the paper is organized as follows. Section 2 provides details of our adaptation of PHMM for active site profiling. Section 3 presents experimental results and Section 4 discusses related work.

## 2 Profiling Active Sites with PHMM

Profile Hidden Markov Model (PHMM) is a statistical learning-based technique for modeling protein sequences families [1]. 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 amino acids are usually conserved. PHMMs structures, as shown in Figure 1, are specialized to capture such conserved amino acids 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 amino acids among sequences while insert and delete states correspond to divergence in sequences from a common ancestor due to insertions and deletions respectively. The emission symbols of match and insert states are the 20 amino acids while the delete states are non-emitting silent states. As in the case of general Hidden Markov Models [7], the Baum-Welch algorithm is usually used to learn the transition and emission probabilities and the Viterbi algorithm is used to decide the best state sequence for a given protein sequence.



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). Tools such as SPASM allow users to define superposable atom types. The main problem with this is that knowledge about what are superposable atom types varies from family to family. For these reasons, a desiderata of geometric feature is that it be preserved under transformations. 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  $i=1, \dots, n-1$ . Note that our serialization framework is not restricted to the feature pair  $\langle AtomType, Distance\_To\_CenterOfMass \rangle$  used in this paper. It can easily incorporate more geometric and physico-chemical features.

**2.2 PHMM for Active Sites** Now we have the serialization of the 3-D active sites, an immediate problem is to learn the PHMM parameters from training data, i.e., to decide the number of columns in the PHMM and to estimate the transition and emission probabilities. For the number of columns, we follow the heuristics in [1] and take the average length of all training sequences as the number of columns of the PHMM.

To learn the transition probabilities and emission probabilities, 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:

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-(d-\mu)^2/2\sigma^2}$$

where  $d$  is the distance,  $\mu$  is the mean and  $\sigma$  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  $\mu$  and  $\sigma$  at this state using the expressions:

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

The small constant  $\epsilon$  is added so that  $\sigma$  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  $\mu$  and  $\sigma$  that represent the distribution of the distances of atoms emitted from the state to their center of mass.

For a set of training 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 Q-SiteFinder [5]. 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 possible to decide whether these atoms are located in some cavity on a protein’s surface, which is a characteristic of active sites. With Q-SiteFinder, we select candidate substructures which are cavities on protein’s surfaces.

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 [7] to compute 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.,  $y' = \arg \max_y P(x, y|M)$ .

We modify Viterbi algorithm to compute the probability of observing a pair  $(t, d)$  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 an amino acid and  $q(r)$  is the frequency of  $r$  in the protein sequence database PROSITE.
2. Randomly sample substructures from PDB, each of which contains the same number of amino acids 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  $\mu$  and the standard deviation  $\sigma$  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 the observation sequence  $x$  is said to be a member of the family  $S$  which is modeled by  $M$ . The threshold is a global value. Therefore we do not need to choose such a value for each family. One such global value is zero. If the bit score for  $x$  is greater than zero, then it means that  $x$  is more likely to be generated by  $M$  than by the random model  $M'$ . We use zero as the threshold in our work.

### 3 Evaluation

We implemented our PHMM-based profiling of active sites and experimentally evaluated its performance. Herein we report on the experimental results.

Active site profiling is a relatively unexplored topic. Consequently, there are no standard data sets available for benchmarking 3D profiling of proteins. Rather than creating our own arbitrary data set, we experimented with those used in [2, 9]. In these two works, profiles of active site families were constructed manually.

**3.1 The Experimental Setup** The evaluation was conducted over different protein families detailed below.

**Protein Families:** The families in [2, 9] are: *Ribonuclease A*, *Ribonuclease T1*, *Eukaryotic Lysozyme*, *Prokaryotic Lysozyme*, *Nu:-His-Elec catalytic triad*, *Amidohydrolase*. We denote them as RibA, RibT1, EuLys, ProLys, NuHis, Amido. We developed PHMM profiles for the active sites of these six families.

**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 five families. These augmented active sites were labeled as negative test examples. Statistics associated with the experimental data used are listed in table 1.

Protein Family	No of Members	Size of Training Set	No of Positive Test Examples	No of Negative Test Examples
RibA	34	25	9	11
RibT1	19	12	7	13
EuLys	107	84	23	12
ProLys	94	75	19	23
NuHis	272	176	96	73
Amido	17	8	9	15

Table 1: Data Statistics for Different Protein Families

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 the structural feature as emission symbols. This lets us observe the effect of structural features on the performance of the PHMM. Similarly we built PHMM per family using only the chemical feature as emission symbols to observe their effect.

**3.2 Experimental Results** Here we report the experimental results in terms of precision and recall of our prototype implementation in recognizing members of each of the six families from their test sets. Figure 2 shows these measures for each family. The precisions range from 80% (for Amido) to 95% (for EuLys). The recalls range from 88% (for RibT1 and NuHis) to 91% (for NuHis).

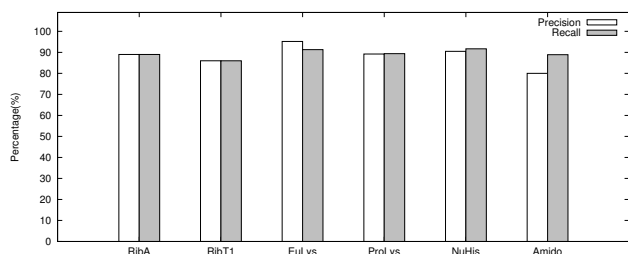


Figure 2: Performance of Protein Families

**Effect of Feature Variation:** Figure 3 shows the effect of structural and chemical feature on precision and Figure 4 shows the effect of features on recall.

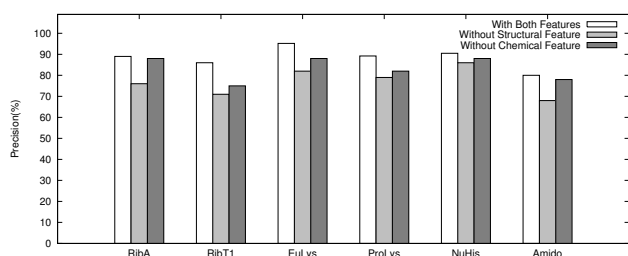


Figure 3: Effect of Structural and Chemical Features on Precision

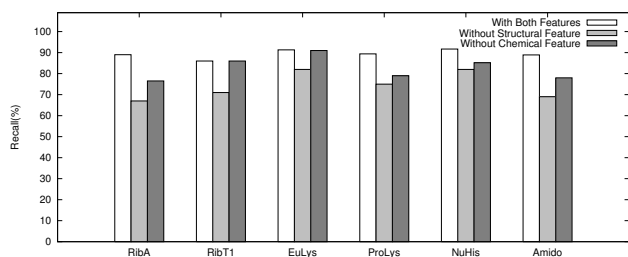


Figure 4: Effect of Structural and Chemical Features on Recall

The reduction of precision ranges from 5% (for NuHis) to 15% (for RibT1) when structural feature is not used as emission symbols. It ranges from 1% (for

RibA) to 11% (for RibT1) when chemical feature is removed from emission. Recall reduction ranges from 8% (for EuLys) to 22% (for RibA) when structural feature is not used and from 0% (for RibT1 and EuLys) to 12% (for RibA) when chemical feature is not used.

**3.3 SPASM Performance** We also conducted experiments on using SPASM to identify similar active sites in the six families listed in Table 1. 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 subfamily 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 a database consisting of the three-dimensional structures of all the other members of the corresponding family. Tables 2 and 3 show the experimental results of SPASM when we varied RMSD cut-offs and allowed substitutions in terms of BLOSUM45 cut-offs<sup>2</sup>. A star (\*) in an entry denotes that SPASM aborts because the number of hits has exceeded a pre-set value(1,000).

Family #members	1.5		2.5		3.5	
	true	total	true	total	true	total
RibA(34)	28	28	30	30	34	119
RibT1(19)	18	18	18	18	18	18
EuLys(107)	95	104	98	125	103	209
ProLys(94)	93	198	93	485	93	529
NuHis(192)	272	194	202	256	272	728
Amido(17)	7	7	7	7	10	17

Table 2: Varying RMSD Cut-off

Family (#members)	3		0		-2	
	true	total	true	total	true	total
RibA(34)	28	28	28	57	28	914
RibT1(19)	18	18	18	79	18	280
EuLys(107)	95	104	15	1034*		
ProLys(94)	93	217	7	1015*		
NuHis(272)	192	194	82	1004*		
Amido(17)	7	7	7	8	7	15

Table 3: Varying Amino Acid Substitutions

Table 2 shows the number of true hits and the number of total hits for each family when no substitution is allowed and the RMSD cut-off increases from 1.5Å to 3.5Å. In most cases the cut-off 1.5Å is satisfactory. But with this cut-off, the recall is only about 71% for NuHis

<sup>2</sup>BLOSUM45 is a matrix defining a score for each pair of amino types. If the score of a pair of amino acids is greater than or equal to the cut-off, then they are considered as substitutable by each other

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 3 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 amino acids in our dataset have a score of -4. That is to say, the BLOSUM45 should be set to -4 to find all similar active sites. However, we can see from Table 3 that the precision is already too low for some families when the BLOSUM45 cut-off is set to 0. When the cut-off is 3, the precision is only 43% for ProLys while the recall is only 41% for Amido.

**3.4 Discussion** The experimental performance suggests that PHMM-based methods described in this paper for determining similarity of active sites works well in practice.

The PHMM constructed for each family exhibits reasonably high precision and recall. 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 precision and recall (95% and 91%). On the other hand the low degree of shared features observed in Ribonuclease T1 has translated into low precision and recall (86% and 86%).

Our experimental results also demonstrate that both geometric and physico-chemical features are needed to achieve high accuracy. Furthermore, we observe that geometric features have relatively higher impact on the performance than physico-chemical features.

Last but not least, our experiments with SPASM indicate that there is no uniform way to decide the RMSD cutoff and amino acid substitution and thus they have to be chosen manually per family. In contrast, there is no such need in our approach.

## 4 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 [4]. It can be used to find substructures in proteins that are similar to an input 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 [9], 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. These templates consist of only a subset of atoms in the active site amino acids. 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 family. In contrast our approach to “learn the templates” is highly automated.

A more recent work is Catalytic Site Atlas (CSA) [8], a database documenting enzyme active sites and catalytic amino acids present in enzymes with 3-D structures. Active site templates are again constructed manually for each family.

## References

- [1] S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [2] 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.
- [3] 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.
- [4] G. Kleywegt. Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, 285:1887–1897, 1999.
- [5] 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.
- [6] F. Melo and E. Feytmans. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, 267:207–222, 1997.
- [7] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2), 1989.
- [8] 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.
- [9] 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.