Prediction of Enzyme Catalytic Sites on Protein Using a Graph Kernel Method

Benaragama V.M.V. Sanjaka
Department of Computer Science
North Dakota State University
Fargo, ND, USA
malinda.sanjaka@my.ndsu.edu

Changhui Yan
Department of Computer Science
North Dakota State University
Fargo, ND, USA
changhui.yan@ndsu.edu

***Abstract*—Structural Genomics projects are producing structural data for proteins at an unprecedented speed. The functions of many of these protein structures are still unknown. To decipher the functions of these proteins and identify functional sites on their structures have become an urgent task. In this study, we developed an innovative graph method to represent protein surface based on how amino acid residues contact with each other. Then, we implemented a shortest-path graph kernel method to measure the similarities between graphs. We tried three variants of the nearest neighbor method to predict enzyme catalytic sites using the similarity measurement given by the shortest-path graph kernel. The prediction methods were evaluated using the leave-one-out cross validation. The methods achieved accuracy as high as 77.1%. We sorted all examples in the order of decreasing prediction scores. The results revealed that the positive examples (catalytic site residues) were associated with higher prediction scores and they were enriched in the region of top 10 percentile. Our results showed that the proposed methods were able to capture the structural similarity between enzyme catalytic sites and would provide a useful tool for catalytic site prediction.**

Keywords—graph kernel; nearest neighbor method; enzyme catalytic sites; prediction

I. INTRODUCTION

Over the past decade various structural genomics projects [1] have produced structural data for over 75,000 proteins. But the functions of many of them are still unknown. To determine the functions of these proteins using traditional laboratory approaches is laborious and time-consuming. Computational methods play an important role to address this problem. Many different methods have been proposed for predicting protein functions or identifying functional sites on proteins [2-5]. Among them are a group of methods based on graph theory [6-8]. In these methods, graphs are used to describe and analyze the geometric and physicochemical properties of protein structures. Then, various methods are used to compare graphs to identify predictive patterns that are correlated with protein function or functional sites. A key step in graph-based structure analysis is to measure the similarities between graphs. Graph kernels have become a favorite solution to this problem. In a simple way, a kernel function is a positive definite matrix that measures the similarities between all pairs of input data. Originally, kernel methods only took vectors as input. Later, researchers developed graph kernel methods that took 2-dimensional (2D) or 3-dimensional (3D) structures as input [9]. These graph kernels varied in what components of the graphs they compared, how they searched the components in a graph and how they compared each pair of components. The most prominent of them was the marginalized kernel [10] that used a probability function to model the distribution of labeled walks, and calculated the similarities between all pairs of labeled walks from different graphs, and then summed them up to get the overall similarity between two graphs. To circumvent the computational difficulties associated with the marginalized kernel, researchers used different ways to approximate the marginalized kernel, which resulted in a group of new kernels called spectrum kernels [11]. For example, the Tanimoto kernel was a spectrum kernel that only considered whether a walk existed in a graph and the MinMax Tanimoto kernel took into account the frequency of a walk in a graph [9]. Instead of comparing labeled walks in graphs, other graph kernels methods compared trees [12-13]. Although these kernel methods varied in many details, they all produced a kernel matrix that showed the similarity between all pairs of instances. The kernel matrix could be embedded into kernel-based machine-learning methods like Support Vector Machine (SVM) [14] to build predictors. It can also be interpreted as a similarity matrix and embedded with other machine learning methods like nearest neighbor method.

It the currently study, we developed an innovative graph method to represent protein surface based on how amino acid residues contact with each other. Then, we implemented a shortest-path graph kernel that was originally developed by Borgwart and Kriegel [15] to compare the similarity between labeled graphs. The shortest-path graph kernel compared all pairs shortest-paths between two graphs. It took into account edge and vertex labels that were real numbers. The method was faster than other graph kernel methods. We embedded the resulting kernel matrix into three variants of nearest neighbor methods to build predictors. We applied the proposed approach to predict enzyme catalytic sites on protein structures. The results showed that that the proposed methods were able to capture the similarity between enzyme catalytic sites and would provide a useful tool for catalytic site prediction.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

II. MATERIALS AND METHODS

Protein dataset and catalytic site residues

Enzymes and their catalytic sites were downloaded from the Catalytic Site Atlas (CSA) [16]. In CSA, enzymes were hierarchically organized based on the Enzyme Commission (EC) number [17]. There are six groups at the first level of the hierarchy, which are EC1 through EC6. We examined the number of proteins in each group of the second level. Group EC3.4 had the most proteins at the second level. Thus, we chose EC3.4 as the dataset to test our method. We used program blastclust from the BLAST [18] to remove redundancy so that pairwise similarity between proteins was less than 30%. In the end, 73 proteins were left. There were a total of 201 active catalytic site residues (positive examples) and 20,398 non-catalytic site residues (negative examples) in these proteins. Position-specific scoring matrix (PSSM) of a protein was built by running 4 iterations of PSI-BLAST [18] against the NCBI non-redundant (nr) database. In the PSSM, each residue position was associated with 20 values.

Graph representation

Each example was represented using a graph, which included the amino acid residue corresponding to the example and the residues that it contacted. Two residues were considered contacting if the shortest distance between their atoms was less than the sum of the radii of the corresponding atoms plus 0.5 Å. In the graph representation, each amino acid residue was represented using a node labeled with the 20 PSSM values of the residue. An edge was added between two nodes if the corresponding residues were contacting.

Graph kernel

A shortest-path graph kernel was used to calculate the similarity between graphs as in Alvarez et al. [19]. Briefly, the first step of the shortest-path kernel was to transform original graphs into shortest-path graphs. A shortest-path graph had the same nodes as its original graph, and between each pair of nodes, there was an edge labeled with the shortest distance between the two nodes in the original graph. Then, the shortest-path graph kernel compared all pairs of walks of length 1 from different shortest-path graphs. The comparison of a pair of walks included the comparisons of the involved edges and vertices. Two vertices were compared using a Gaussian kernel as in (1)

$$k_{vertex}(v,w) = \exp(-\frac{\| labels(v) - labels(w) \|^2}{2\delta^2}) \qquad (1)$$

where $1/2\delta^2$ was set to 72, v and w were two vertices, and function *labels ()* returned the labels of a vertex. Two edges were compared using a Brownian kernel as in (2)

$$k_{edge}(e_1,e_2) = \max(0, c - |weight(e_1) - weight(e_2)|) \quad (2)$$

where c was set to 2, e_1 and e_2 were edges, and function *weight()* returned the weight (or length) of an edge.

Classification methods and leave-one-out cross validation

We tested three variants of the nearest neighbor method (NNM), namely NNM_AVE, NNM_MAX, and NNM_TOP10, to build predictors for enzyme catalytic site prediction. For a test example, its pairwise similarities to all examples in the training set were calculated using the shortest-path graph kernel. The three NNMs were defined as follows: (1) Let Ave_pos be the average similarity between the test example and all positive examples, and Ave_neg be the average similarity between the test example and all negative examples. Then, the NNM_AVE method predicted the test example to be a catalytic site if Ave_pos ≥ Ave_neg, and non-catalytic site otherwise. The prediction score for the test example was defined as Ave_pos-Ave_neg; (2) Let Max_pos be the maximum similarity between the test example and all positive examples, and Max_neg be the maximum similarity between the test example and all negative examples. The NNM_MAX method predicted the test example to be a catalytic site if Max_pos ≥ Max_neg, and non-catalytic site otherwise. The prediction score for the test example was defined as Max_pos-Max_neg; (3) Let Top10_pos be the average of the 10 highest similarities between the test example and all positive examples, and Top10_neg be the average of the 10 highest similarities between the test example and all negative examples. In the NNM_TOP10 method, the test example was predicted to be a catalytic site if Top10_pos ≥ Top10_neg, and non-catalytic site otherwise. The prediction score for the test example was defined as Top10_pos-Top10_neg. All the predictors were evaluated using leave-one-out cross-validation at protein level, so that when an example was used as the test example, examples from the same proteins were removed from the training set.

III. RESULTS AND DISCUSSION

Classification performance

We extracted catalytic site residues and non-catalytic site residues from the proteins and represented each of them using a graph. The dataset was extremely unbalanced, with 201 positive examples and 20,398 negative. To make a better evaluation of the methods, we randomly selected 201 non-catalytic site residues from the negative examples and put them with the 201 positive examples to form a balanced dataset. Then, we used the shortest-path graph kernel to calculate pair-wise similarities between graphs. We used three variants of the NNM to build predictors for catalytic site prediction. The predictors were evaluated using leave-one-out cross validation using the balanced dataset. The results (TABLE I) show that the three methods achieved comparable accuracy, with NNM_AVE being slightly better than the others. The accuracy of the NNM_AVE was 77.1%. The NNM_AVE predicted a total of 46 false positives. We analyzed the locations of the false positives on the protein structures and found that 5 of them are contacting with known catalytic site residues. Some methods have been published for predicting enzyme catalytic sites. The accuracy reported varies over a big range depending on the dataset used and other parameters. We have not been able to make a direct comparison between our method and other methods due to difference in the datasets used.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

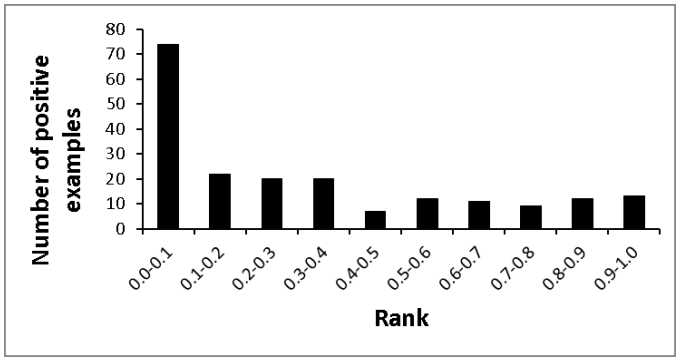

Fig. 1. The rank distribution of positive examples. NNM_AVE was used to make prediction. Lower values of rank correspond to higher prediction scores. The histogram shows a clear trend that positive examples (i.e., catalytic site residues) were enriched in the regions with higher prediction scores.

TABLE I. PERFORMANCE FOR PREDICTING ENZYME CATALYTIC SITES

	TP[a]	FP	TN	FN	Accuracy
NNM_AVE	155	46	155	46	77.1%
NNM_MAX	150	64	137	51	71.3%
NNM_TOP10	156	51	150	45	76.1%

[a] TP: true positive; FP: false positive; TN: true negative; FN: false negative

Enrichment of positive examples in the high scoring region

We repeated the leave-one-out cross validation procedure on the whole dataset, which included 201 positive and 20,398 negative examples. Then, for each protein, we sorted the examples in the order of decreasing prediction scores. We then looked at the ranks of positive examples. The rank of an example was defined as the percentage of examples from the same protein that had higher scores than it. For example, for a given example, if 5% of examples from the same protein had higher prediction scores than it, then its rank was 0.05. Good predictors should assign higher scores to positive examples than to negatives, thus positive examples should have higher ranks (which correspond to smaller values for ranks) than negative ones. Fig. 1 shows the rank distribution for positive examples when NNM_AVE was used. The results revealed a clear trend that positive examples are enriched in the regions corresponding to high prediction scores. Analysis of results of NNM_MAX and NNM_TOP10 revealed the same trend.

IV. CONCLUSIONS

In this work, we developed an innovative graph method to represent protein surface based on how amino acid residues contact with each other. Then, we implemented a shortest-path graph kernel method and used it to compute the similarity between graphs. We developed three nearest neighbor methods to predict enzyme catalytic sites based on the similarity matrix that the graph kernel method produced. The predictors were able to predict catalytic sites with accuracy up to 77.1%. Analysis of the prediction scores showed that positive examples had a clear bias towards the high prediction score regions. This work showed that the proposed methods were able to capture the similarity between enzyme catalytic sites and would provide a useful tool for catalytic site prediction.

ACKNOWLEDGMENT

The project described was partially supported by NIH Grant Number P20 RR016471 from the INBRE Program of the National Institute of General Medical Sciences.

REFERENCES

[1] S. K. Burley, An overview of structural genomics. Nat. Struct. Biol., 2000, Struc. Genomic Supplement, 932-934.

[2] G. Casari, et al., A method to predict functional residues in proteins. Nat. Struct. Biol., 1995.2, 171–178.

[3] M. J. Ondrechen, et al., THEMATICS: a simple computational predictor of enzyme function from structure. Proc. Natl Acad. Sci. USA, 2001, 98, 12473–12478.

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

[5] O. Lichtarge, H. Bourne, F. Cohen, Evolutionary Trace method defines binding surfaces common to protein families. J. Mol. Biol., 1996, 257, pp. 342–358.

[6] H. Deng, G. Chen, W. Yang, J. J. Yang, Predicting calcium binding sites in proteins - a graph theory and geometry approach. Proteins, 2006, 64 (1), 34-42.

[7] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, A. Tropsha, Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. J. Comput. Biol., 2005, 12 (6), 657-671.

[8] V. Vacic, L. M. Iakoucheva, S. Lonardi, P. Radivojac, Graphlet kernels for prediction of functional residues in protein structures. J. Comput. Biol. (2010) 17(1): 55-72.

[9] L. S. Ralaivola, et al., Graph kernels for chemical informatics. Neural Networks 2005, 18, 1093-1110.

[10] P. Mahe, et al., Graph Kernels for Molecular Structure-Activity Relationship Analysis with Support Vector Machines. J Chem Inf Model 2005, 45: 939-951.

[11] C. Leslie, et al., The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput, 2002, 564-575.

[12] J. Ramon, T. Gartner, Expressivity versus efficiency of graph kernels. In Proceedings of First International Workshop on Mining Graphs, Trees, and Sequences, 2003, 65-74.

[13] P. L. Mahe, J. P. Vert, Graph kernels based on tree patterns for molecules. Machine Learning, 2009, 75: 3-35.

[14] C. Cortes, V. N. Vapnik, Support-Vector Networks, Machine Learning, 1995, 20, 273-297.

[15] K. M. Borgwardt, H. P. Kriegel, Shortest-path kernels on graphs. In Proceedings of The fifth IEEE International Conference on Data Minning (ICDM'05), 2005, 8.

[16] C. T. Porter, G. J. Bartlett, J. M. Thornton, The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, Nucleic Acids Res ,2004, 32(Database issue), D129-D133.

[17] E. C. Webb, Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes, 1992, San Diego, Academic Press, ISBN 0-12-227164-5.

[18] S. Altschul, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. ,1997, 25(17): 3389-3402.

[19] M. Alvarez, X. Qi, C. Yan, A shortest-path graph kernel for estimating gene product semantic similarity. J Biomed Semantics, 2011, 2:3.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE