Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Prediction of Enzyme Catalytic Sites from Sequence Using Neural Networks

Swati Pande
Department of Biological Sciences
California State Polytechnic University, Pomona
3801 W. Temple Ave.
Pomona, CA 91768
sspande@csupomona.edu

Amar Raheja
Department of Computer Science
California State Polytechnic University, Pomona
3801 W. Temple Ave.
Pomona, CA 91768
raheja@csupomona.edu

Dennis R. Livesay
Bioinformatics Research Center
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC 28223
drlivesa@uncc.edu

The accurate prediction of enzyme catalytic sites remains an open problem in bioinformatics. Recently, several structure-based methods have become popular; however, few robust sequence-only methods have been developed. In this report, we demonstrate that three different feed forward neural networks, trained on a variety of sequence-based properties, can reliably predict enzyme catalytic sites. To the best of our knowledge, this is only the second report using neural networks to predict catalytic sites, and is the first relying solely on sequence-derived information. Scaled conjugate gradient is used during training of the models. The simplest of the models uses only sequence conservation, diversity of position and residue identity within the input. Surprisingly, model accuracy is largely unaffected when sequence-based predictions of structural properties (i.e. solvent accessibility and secondary structure) are added to the input. A similar lack of improvement is observed when evolutionary information in the form of phylogenetic motifs is included. These results are noteworthy because they indicate that routine neural network architectures can accurately predict catalytic using only residue identity and conservation inputs. However, applying these methods on a per protein basis still produces a significant number of false positives, which significantly reduces the model's utility to experimentalists.

I. Introduction

The identification of protein functional and/or catalytic sites is an especially important bioinformatic problem. This information is important because it can aid understanding of catalysis, identify new drug targets and predict the effects of nonsynonymous single nucleotide polymorphisms. There are currently several widely used protein functional site prediction strategies (see [1] and [2] for two recent reviews); however, most are (at least in part) based on protein structure. Due to the fact that the number of solved protein structures continues to lag behind sequence coverage by at least three orders of magnitude, the utility of such methods is restricted. Moreover, the gap between sequence and structural coverage will continue to grow as more and more high-throughput genome sequencing efforts are completed. Consequently, the importance of being able to accurately predict functional sites from sequence will continue to be important post-genomic task.

Previously [3], we have demonstrated that phylogenetic motifs (PMs), which are sequence alignment fragments that approximate the overall familial phylogeny, are very good predictions of regions surrounding enzyme active sites. PMs are similar in spirit to the evolutionary trace approach, yet they are explicitly designed to not rely on structural information. Our previous investigations across a wide range of protein architectures and functional classes demonstrate that PMs consistently cluster around protein active sites and/or substrate binding epitopes [3,4]. Moreover, through comparison to Poisson-Boltzmann electrostatic calculations, we have elucidated the functional role of many PM residues [5]. Frequently, these residues makeup evolutionarily conserved electrostatic networks at the enzyme active site that serve to fine-tune the activity of the catalytic sites.

Catalytic sites are defined as residues that are directly involved in the enzyme-mediated reaction pathway, meaning that catalytic sites represent a small subset of all *functional sites*. In this investigation, we attempt to predict enzyme catalytic sites from sequence information alone. Toward this goal, several different neural networks are trained on residue, sequence alignment and sequence-derived information. The inspiration for this work is from the recent report from Gutteridge et al. [6], where a neural network was successfully trained on a combination of sequence and structure inputs. While neural networks are frequently applied to a wide variety of bioinformatic problems (i.e., secondary structure, contact map and splice site prediction), the Gutteridge et al. effort is, to the best of our knowledge, the only other report applying neural networks to the catalytic site problem. The accuracy of the earlier effort is a consequence of both the sequence and the structural inputs.

In order to partially compensate for the loss of structural information in our sequence-only effort, we test the ability of sequence-derived predictions of structural quantities (i.e., solvent accessibility and secondary structure) to improve learning. Surprisingly, our results indicate that model accuracy is largely unaffected, if not slightly reduced, when including such predictions. A similar performance decrease is observed when PMs are encoded within the input. These results are noteworthy because they demonstrate that accurate

1-4244-0710-9/07/$20.00 ©2007 IEEE

catalytic site prediction can be achieved using only residue identity and familial conservation within routine neural network architectures.

II. Methods

A. Neural network

In this investigation, we use the multilayer perceptron, which is a feed forward neural network, implemented within MATLAB [7]. A scaled conjugate gradient search algorithm is used during training. Our dataset is composed of 132 proteins taken from the Catalytic Site Atlas (CSA) [8]. The CSA is a well-curated database of experimentally validated and predicted (from homologous entries) enzyme catalytic sites. Each site within the database is annotated as catalytic or not, making supervised learning straightforward. In this report, we only utilize the experimentally validated sites. For training and assessment, the dataset is randomly divided into nine similarly sized groups. Training is performed on eight of the datasets, and network performance is assessed by predictions on the ninth. Thus, nine different data permutations are investigated.

Training on the complete CSA is problematic due to the limited number of catalytic sites within each protein. On average, each enzyme has only 3-6 catalytic sites, whereas the average enzyme length is greater than 200 residues. Consequently, successful learning on such sparse data is an exceedingly difficult computational problem. In order to circumvent this, we employ the same solution adopted by Gutteridge et al. All catalytic sites are included in the training set; however, only a randomly selected fraction of the non-catalytic sites are utilized. By testing several different ratios, Gutteridge et al. found that a ratio of 1 to 6 (catalytic to non-catalytic) resulted in the best network performance. We use the same ratios here.

Systematic testing of model performance vis-à-vis the number of neurons within the single hidden layer reveals twenty to be ideal (results not provided). Twenty neurons are used within all of the results reported here. In order to facilitate comparisons between different inputs, all networks are trained to a fairly consistent error threshold. The number of epochs required to reach the threshold is provided in Table 1. The approximate amount of compute time within in Table 1 is ~1.8 seconds/epoch. In all cases, the log-sigmoid transfer function is used between the input and hidden layers, as well as between the hidden and output layers.

B. Input encoding

Our inputs closely resemble those from Gutteridge et al. A unique vector describes each site within an input protein sequence. An example input vector is provided in Figure 1. Within each vector, amino acid identity is orthogonally encoded over twenty different elements. In order to provide sequence conservation information, each input sequence is PSI-BLASTed [9] against the NCBI Non-Redundant Protein Database [10]. All homologs from four PSI-BLAST iterations are collected. A Diversity of Position Score (DOPS) and a Conservation Score (CS) are calculated from the PSI-BLAST alignment using the program SCORECONS [11]. The DOPS and CS values are rescaled as floating-point numbers (0.0-1.0) and added to the input vector. Inclusion of these three properties only constitutes our sequence only neural network (SeqNN).

In an attempt to compensate for the lack of explicit structural information, sequence-based predictions of residue solvent accessibility (RSA) and secondary structure (SS) are included. Structurally derived RSA and SS were included in the input by Gutteridge et al. RSA and SS predictions are calculated using third-party one-dimensional recurrent neural network programs. RSA is predicted using ACCPro [12], whereas SS is predicted using SSPro [13]. Both are from the SCRATCH suite of programs [14] from the University of California at Irvine. As done previously by Gutteridge et al., we orthogonally encode three SS states (helix, sheet and coil) within the input vector. For example, {0 0 1} corresponds to coil. The RSA predictions from ACCPro are simply binary values (exposed or buried); consequently, RSA is orthogonally encoded in the same way directly from the ACCPro predictions.

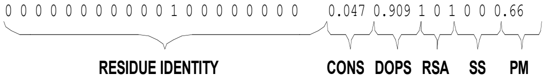

Fig. 1. An example of the neural network encoding. A unique vector describes the attributes of each residue within the dataset. The vector shown corresponds to the CompNN; the NoPMM and SeqNN exclude the PM and the RSA + SS + PM elements, respectively.

C. Phylogenetic motif identification

Not surprisingly, one of the most heavily weighted structural metrics within the Gutteridge et al. investigation was cleft information. It is well known that catalytic sites generally reside in the depths of active sites clefts within the protein structure [15]. Using Surfnet [16], Gutteridge et al. identified all clefts within the input structure before constructing the input. Cleft information was orthogonally divided into four categories: no cleft, largest cleft, second or third largest cleft and fourth to ninth largest cleft. Unfortunately, reliable predictions of protein clefts from sequence remain elusive. Therefore, we alternatively encode PM information into the input in an attempt to compensate for the lack of cleft information.

As stated above, PMs consistently correspond to enzyme active site regions, which should, at least partially, offset the loss of cleft information. Across a structurally and functionally diverse protein family dataset, PMs consistently correspond to known functional sites defined by surface loops, active site clefts, and partially buried regions interacting with prosthetic groups. This point is exemplified in Figure 2. In each, it is clear that the identified PMs cluster around the active site.

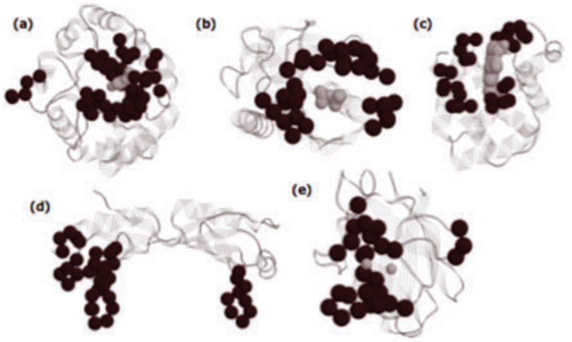

Fig.2. PMs consistently correspond to known functional sites. This figure shows a sampling of the structurally diverse examples previously investigated. Identified PMs are structurally clustered and correspond to functional sites in: (a) TIM, (b) inorganic pyrophosphatase, (c) myoglobin, (d) TATA-box binding protein and (e) CuZn superoxide dismutase. Dark spheres represent PM α-carbons; light spheres represent the substrate analog, pyrophosphate, heme, and copper/zinc ions in (a), (b), (c) and (e), respectively. The PMs in (d) correspond to the DNA binding site.

PMs are identified using MINER [17], which is our automated PM identification software. MINER begins by sliding a window across an input multiple sequence alignment. A phylogenetic tree is constructed for each alignment fragment, which is then compared to the complete familial phylogenetic tree using a modified bipartition metric [18]. All overlapping windows that score past some tree similarity threshold are deemed to be a PM. In our early implementation of the approach, this threshold had to be set manually for each trial. Threshold values vary significantly between different examples, meaning that no single threshold is robust enough for large-scale application of the method. Threshold determination was subsequently automated using a novel algorithmic approach in order to determine ideal thresholds on a case-by-case basis [4].

Based on the way PMs are defined, encoding PM information into the input vector poses two challenges. The first challenge arises from the fact that a single PM is defined to be the union of several overlapping sequence windows, each with its own bipartition score. Furthermore, these constituent bipartition scores can vary significantly. Simply averaging these scores to define a global score might be attractive due to ease; however, there is no theoretical rational to support such an approach. In order to circumvent this global score problem, we simply coarse-grain each site into a binary (1, 0) depending on whether it (does, does not) correspond to a PM. A second problem with encoding PM information is due to the fact that the automated threshold identification algorithm within MINER is explicitly designed to be overly stringent, meaning that it is biased towards false negatives versus false positives. However, a qualitative analysis of the PM results vis-à-vis the Catalytic Site Atlas [8] indicates it to be overly stringent when predicting catalytic sites (unpublished results). Therefore, we have added a second, less stringent, threshold to the PM results. The second threshold is exactly one standard deviation away from the automated threshold. If a site scores past the second threshold, but not the automated threshold, then a value of 0.66 is added to the input. If the site scores past the automated threshold, then a 1.0 is input, else the site receives a 0.0.

The neural network based on the complete set of input metrics, which is described in Figure 1, is denoted CompNN. In order to test the ability of PMs to add information to the network, a third neural network is constructed. The NoPMNN includes all elements of the CompNN except the final element encoding the MINER results is omitted.

III. Results and Discussion

A. CompNN

The ability of CompNN to successfully learn on the training data is described in Table 1. The dataset designation corresponds to predictions on that portion of the dataset based on training on the remainder. Predictive abilities are assessed by Receiver Operating Characteristic (ROC) graphs. ROC plots sensitivity values (true positive rate) versus 1-specificity values (false positive rate). Overall model performance is based on the area under the ROC curve. A method is considered better than random if its curve is over the y = x curve, whereas a method is worse than random if it is beneath the y = x curve. An example ROC plot is provided in Figure 3.

As can be seen from the area under the ROC curves (Figure 3 and Table 1), the quality of the CompNN predictions is much better than random. In all cases, the plots are shifted significantly to the left of the y = x line. An ROC curve can be thought of as the tradeoff between true and false positive rates at every possible threshold. In this example, it is clear that the tradeoff between the two is very good, especially at low false positive rates. In catalytic site prediction, it is common to consider false positive rates below 20% acceptable [19]. Figure 3 demonstrates that at a false positive rate of 20% (indicated by the vertical dashed line), the true positive rate is 80%. These results are very good, especially considering they lack any structural information. For example, these results are inline with those reported in Amitai et al. [19] and Thibert et al. [20], both of which use an immerging structure-based catalytic site prediction scheme. In both, the protein structure is recast as a topological network, and residues with the highest centrality values are put forth as catalytic site predictions.

The above ROC results are reinforced by examining the data in a second way. Figure 4 plots the neural network score for all sites within dataset 7, which, based on ROC performance, is rank-ordered only fifth of nine. Using a prediction threshold of 0.75 (meaning network scores greater than 0.75 are considered to be catalytic site predictions), there are 16 false negatives and only 5 false positives. This translates to a total error of 8.5%; total error is calculated by: (# false negatives + # false positives) / total sites * 100. The overall accuracy is simple 1 – error. Interestingly, the error in dataset 7 is the best of the nine CompNN examples, despite the fact that its performance is judged to only be average by the ROC analysis. Similar contradictions between accuracy and ROC areas are observed within the other data permutations.

TABLE 1
SUMMARY OF TRAINING AND ROC RESULTS

CompNN			
Dataset	Error	Epochs	ROC Area
1	0.0189	1,803	0.599
2	0.0190	3,324	0.518
3	0.0190	2,012	0.622
4	0.0189	1,985	0.696
5	0.0189	1,549	0.559
6	0.0189	5,016	0.650
7	0.0194	15,000	0.635
8	0.0190	15,000	0.757
9	0.0190	15,000	0.676
SeqNN			
Dataset	Error	Epochs	ROC Area
1	0.0241	35,000	0.547
2	0.0240	77,002	0.589
3	0.0228	80,000	0.639
4	0.0233	80,000	0.665
5	0.0228	80,000	0.741
6	0.0251	104,852	0.886
7	0.0243	80,000	0.631
8	0.0270	85,076	0.661
9	0.0265	100,751	0.633
NoPMNN			
Dataset	Error	Epochs	ROC Area
1	0.0199	2,336	0.595
2	0.0199	3,217	0.658
3	0.0190	3,537	0.598
4	0.0199	4,956	0.670
5	0.0199	8,764	0.552
6	0.0199	2,895	0.634
7	0.0200	7,771	0.694
8	0.0199	3,331	0.800
9	0.0199	4,059	0.713

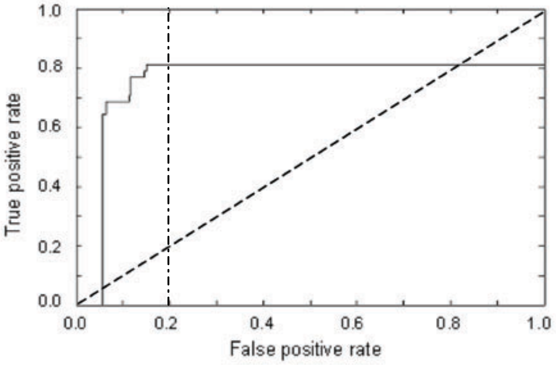

Fig. 3. Example ROC plot. Sensitivity (TP rate) is plotted against 1-specificity (FP rate). Dataset number 8 is shown. The area under the ROC curve is 0.757. The dashed diagonal line indicates the random expectation. The vertical dashed line demonstrates that at false positive rates below 20%, the true positive rate is 80%. ROC curves are calculated using the Matlab ROC toolkit (http://theoval.sys.uea.ac.uk/matlab/default.html).

Table 2 highlights the false negative, false positive and total errors for all nine datasets. The average error over all nine datasets is an impressive 12.0% (the standard deviation is 2.6%). Equally encouraging is the fact that the CompNN results are very specific. Overall, the total number of false positives (161) is less than the total number of false negatives (174). This global result is observed in all but two of the individual datasets.

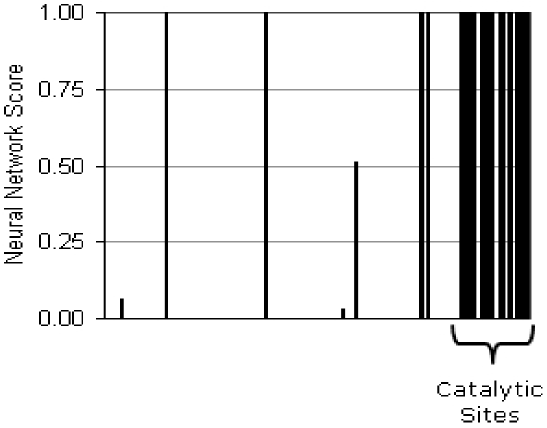

Fig. 4. Neural network scores for all sites within dataset 7. The sites have been sorted such that all non-catalytic sites are to the left, whereas, as indicated, the catalytic sites are to the far right. In this example, there are five false positive and 16 false negatives, which corresponds to an error of 8.5%.

B. SeqNN and NoPMNN

In general, it is much more difficult to train SeqNN compared to CompNN. In fact, 80,000 epochs of training are unable to match the training errors of CompNN (which average 6,743 epochs of training). Moreover, extensive training, over 200,000 epochs in one case (data not shown), the error never approached those achieved by CompNN. Nevertheless, the overall prediction accuracy of SeqNN is similar to the CompNN results. In fact, the average error and ROC areas of the SeqNN results (see Figure 5) are slightly better than those from CompNN. Based on the conclusions drawn within the Gutteridge et al. report, we naively assumed that, despite the innate problems associated with secondary structure and solvent accessibility predictions, the additional descriptors would improve prediction accuracy. This is clearly not the case, which is exciting because it demonstrates that residue identity and alignment conservation are all that is required to accurately predict catalytic sites using current neural network techniques.

When an error estimate of one standard deviation is added, the differences between CompNN and SeqNN are shown to be statistically equivalent. (Future work will apply more sophisticated statistical analyses to the two methods.) A case-by-case comparison of the ROC areas further highlights how similar the accuracies of the two methods actually are. CompNN outperforms SeqNN five times (out of the nine total), whereas SeqNN outperforms CompNN five times when considering total errors. As before, the total number of false positive (142) is again less than the total number of false negatives (179).

TABLE 2
FALSE POSITIVES (FP) VS. FALSE NEGATIVES (FN)[1]

CompNN				
Dataset	#FP	#FN	Total	Error
1	42	33	461	16.3%
2	14	26	294	13.6%
3	5	19	236	10.2%
4	9	14	230	10.0%
5	19	23	294	14.3%
6	14	15	251	11.6%
7	5	16	248	8.5%
8	31	15	341	13.5%
9	22	13	351	10.0%
SeqNN				
Dataset	#FP	#FN	Total	Error
1	50	32	461	17.8%
2	13	23	294	12.2%
3	9	18	236	11.4%
4	5	15	230	8.7%
5	9	15	294	8.2%
6	11	8	251	7.6%
7	5	23	248	11.3%
8	20	25	341	13.2%
9	20	20	351	11.4%
NoPMNN				
Dataset	#FP	#FN	Total	Error
1	62	31	461	20.2%
2	6	26	294	10.9%
3	4	21	236	10.6%
4	4	13	230	7.4%
5	11	22	294	11.2%
6	23	19	251	16.7%
7	3	16	248	7.7%
8	15	25	341	11.7%
9	24	20	351	12.5%

[1] In all cases, a prediction threshold of 0.75 is used.

Removing PM information from the CompNN (called NoPMNN) does not change the results appreciably. Based on both performance measures used above, the NoPMNN results are statistically indistinguishable from the CompNN and SeqNN results (Figure 5). As before, the uniformity of method accuracy is demonstrated by case-by-case comparisons of the three networks (Table 3). Like the previous two models, the total number of NoPMNN false positive (152) is again less than the number of false negatives (193).

While we were initially surprised by the uniformity of the results from the three networks, careful consideration reveals that we should have expected it. This is because within the Gutteridge et al. paper, they surprisingly demonstrate that the input element with the largest relative weight is simply whether or not the amino acid in question is histidine. The second most important input element is sequence conservation, followed by whether or not the amino acid is lysine, cysteine, aspartate, glutamate and arginine --- which are all included within the SeqNN. Relative solvent accessibility is the most important of the structural characteristics within the Gutteridge et al. results, which is only the eighth most heavily weighted across their whole input. The diversity of position score, which is a "chemical" conservation metric is the ninth most heavily weighted. The 10th-15th most important input elements in Gutteridge et al. are whether or not the residue is serine, tyrosine, glutamine, threonine, asparagines, methionine and glycine, meaning that 14 of the 15 most important input elements are included within the simple SeqNN. The fact that 13 of the top 15 input elements corresponds to a specific residue identity highlights that only certain amino acids have the appropriate physiochemical properties to be catalytic. The observation that alignment conservation is so important highlights the evolutionary pressures acting on an enzyme family to conserve function.

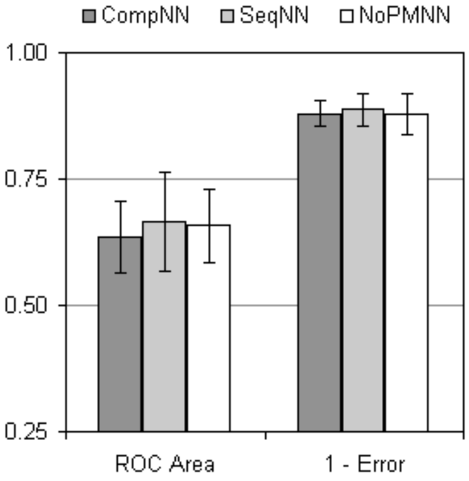

Fig. 5. Performance of the three different models employed here. Reported values are the average value over all nine datasets; the standard deviation of the distribution is used as an error estimate. While the total errors reported in Table 2 are percentages, the 1-Error (i.e., accuracy) values reported here are simply one minus those values (scaled 0 to 1).

The inability of the predicted solvent accessibility information to improve catalytic site prediction accuracy demonstrates again how limited the predictions are. While the authors of ACCpro report an accuracy >77%, all predictions are returned as a binary --- exposed or not. The explicit structure determined accessibilities used by Gutteridge et al. were floating-point descriptions of the percent burial. It is a well-known phenomenon [15] that catalytic sites are frequently partially exposed; for example, they are commonly found *partially* buried within the depths of the active site cleft. Presumably, coarse graining the predicted accessibilities, coupled with their innate inaccuracies, is the reason that the CompNN fails to improve upon the SeqNN results. SSpro is consistently among the best of the benchmarked secondary structure prediction algorithms. However, even if the predictions are 100% accurate, the point is largely moot because as Gutteridge et al. demonstrate, they are the least important of all elements within the input vector.

TABLE 3
CASE-BY-CASE ACCURACY COMPARISONS

	ROC Area	1 – Error
CompNN / SeqNN	5 / 4	4 / 5
CompNN / NoPMNN	5 / 4	4 / 5
SeqNN / NoPMNN	6 / 3	4 / 5

The fact that the CompNN and NoPMNN results are so similar is again surprising to us. Based on our earlier results (see Figure 2), we expected the PMs to act as surrogate for the cleft data in Gutteridge et al. The lack of improvement within CompNN is likely due to two factors. First, we have utilized here the simplest possible encoding of the PM data. Perhaps it would be better to scale the PM results as a floating-point number to more finely discriminate the phylogenetic data. Future work will investigate other methods of encoding PM information. Second, it follows from the discussion above that the most important inputs within Gutteridge et al. are all, except for solvent accessibility, sequence-based. This result suggests that even if the PMs results did corresponded one-to-one with cleft data (which they do not), the model accuracy would not be changed significantly.

C. Specific enzyme examples

Figure 6 plots the CompNN scores against residue number for the enzyme phosphoenolpyruvate mutase (PEPM). The results provided are after training on dataset 7, which has the lowest error of the nine data permutations. The SeqNN and NoPMNN plots look qualitatively similar. According to the CSA, this enzyme has four catalytic sites: Gly47, Leu48, Asp58 and Lys120. As can clearly be seen from Figure 7, the network significantly over-predicts catalytic sites within PEPM. This result should not be confused with the false positive vs. false negative discussions above. In the results above, the dataset includes only a partial set of non-catalytic residues (1:6); however, in this example, the network is applied to the entire protein sequence. Figure 6 is inline with the per protein predictions by Gutteridge et al. In both cases, too many false positives occur to provide real utility to experimentalists wishing to use the approach to guide large-scale mutation efforts. In fact, Gutteridge et al. applied a structural clustering algorithm to post-process the network predictions in order to improve model accuracy. Since we are overtly investigating sequence-only methods, we are unable to apply such an approach.

The four PEPM catalytic sites are structurally highlighted in Figure 6 (PDBid: 1PYM). Two of the sites (Leu48 and Asp58) are strongly predicted to be catalytic. A third (Gly47) is weakly predicted, whereas Lys120 is not predicted at all. The observation that the three correct predictions are clustered in sequence is tantalizing because it is possible that this could be exploited to improve results. A cursory analysis of several other proteins reveals similar overall performance. However, sequence clustering is not observed elsewhere, indicating that it is just a coincidence in the PEPM case. In fact, there is a large body of biochemical evidence that establishes that catalytic sites are generally nonlocal in sequence. The total error for the complete protein sequence ranges between 15 and 20%; the error in the PEPM example is 15.4%. It should be pointed out that if structural clustering within Figure 7 were used (as done by Gutteridge et al. to post-process the neural network predictions), the prediction accuracy would improve due to the close proximity of Gly47, Leu48 and Asp58. Additionally, note that the catalytic site predictions are occurring uniformly across solvent exposed and inaccessible regions, which is likely due to the poor quality of sequence-based RSA predictions.

IV. CONCLUSIONS

This report clearly demonstrates that prediction of enzyme catalytic sites from sequence is viable using standard neural network techniques. In all cases, ROC and total error analyses demonstrate the predictions to be much better than random; in fact, prediction accuracies average ~88%. Moreover, the balance between false positives and false negatives is satisfactory upon the 6:1 (non-catalytic:catalytic site) datasets. Surprisingly, the addition of phylogenetic motif and sequence-derived predictions of structural properties to the input provides no appreciable accuracy increase. This result is mostly due to the importance of residue type and familial conservation on the ability to predict enzyme catalytic sites; nevertheless, it is likely to also reflect the limited quality of such predictions. We are currently exploring this result in more detail using more detailed PM inputs. Future work will utilize other methods of predicting RSA and SS, additional input permutations and alternate neural network architectures. Applying the method on a per protein basis is less exciting due to false positives. Future work will also attempt to figure out ways of improving the per protein results.

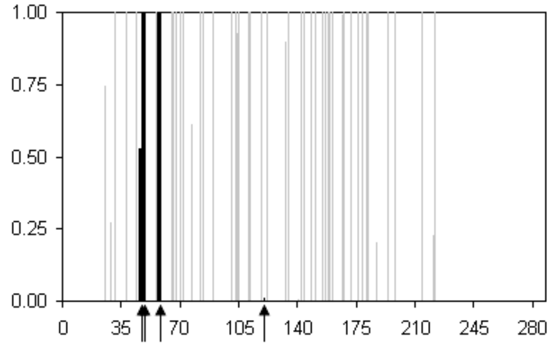

Fig. 6. Neural network score vs. residue number for phosphoenolpyruvate mutase. The four catalytic sites are indicated by arrows and boldface. The total error in this example is 15.4%. Qualitatively similar plots are observed in other protein examples.

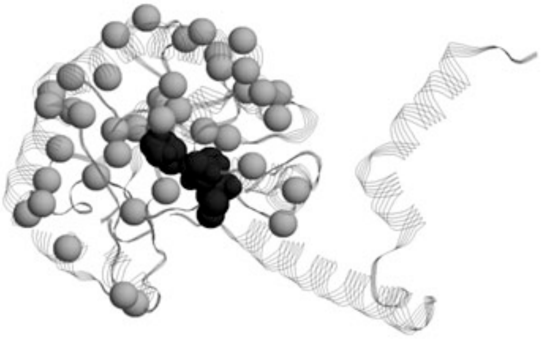

Fig. 7. Structure of phosphoenolpyruvate mutase. Alpha-carbons of the CompNN predictions of catalytic sites (using the 0.75 threshold) are shown in grey. The four catalytic sites are shown (dark grey) in spacefill.

ACKNOWLEDGEMENT

We thank members of Janet Thornton's research group for their kind help with this project. Specifically, James Torrance is acknowledged for help with the Catalytic Site Atlas, William Valdar for providing SCORECONS and Alex Gutteridge for clarifying details concerning reference [6]. This work is partially supported by a California State University Program for Education and Research in Biotechnology-Joint Venture Grant to DRL. Correspondence should be addressed to DRL.

REFERENCES

[1] Watson JD, Laskowski RA, Thornton JM (2005). Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* 15:275-284.

[2] Pazos F, Bang JW (2006). Computational Prediction of Functionally Important Regions in Proteins. *Curr Bioinformatics* 1:15-23.

[3] La D, Sutch B, Livesay DR (2005). Predicting protein functional sites with phylogenetic motifs. *Proteins* 58:309-320.

[4] La D, Livesay DR (2005). Predicting functional sites with an automated algorithm suitable for heterogeneous datasets. *BMC Bioinformatics* 6:116.

[5] Livesay DR, La D (2005). The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins. *Protein Sci* 14:1158-1170.

[6] Gutteridge A, Bartlett GJ, Thornton JM (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* 330:719-734.

[7] MATLAB, Release 14. The MathWorks, 3 Apple Hill Drive, Natick, MA 01760-2098 USA.

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

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

[10] Pruitt KD, Tatusova T, Maglott DR (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33:D501-D504.

[11] Valdar WS (2002). Scoring residue conservation. *Proteins* 48: 227-241.

[12] Pollastri G, Baldi P, Fariselli P, Casadio R (2002). Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* 47:142-153.

[13] Pollastri G, Przybylski D, Rost B, Baldi P (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 47:228-235.

[14] Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 33:W72-W76.

[15] Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 324:105-121.

[16] Laskowski RA (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13:323-328.

[17] La D, Livesay DR (2005). MINER: Software for phylogenetic motif identification. *Nucleic Acids Res* 33:W267-W270.

[18] Roshan U, Livesay DR, La D (2005). Improved phylogenetic motif detection using parsimony. *Proceedings of the IEEE BIBE Meeting* BIBE05:19-26.

[19] Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S (2004). Network analysis of protein structures identifies functional residues. *J Mol Biol* 344:1135-1146.

[20] Thibert B, Bredesen DE, del Rio G (2005). Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics* 6:213.