Understanding Performance of Protein Structural Classifiers

Alper Sarikaya*
University of Wisconsin – Madison

Danielle Albers†
University of Wisconsin – Madison

Michael Gleicher‡
University of Wisconsin – Madison

ABSTRACT

Many bioinformatics applications utilize machine learning techniques to create models for predicting which parts of proteins will bind to targets. Understanding the results of these protein surface binding classifiers is challenging, as the individual answers are embedded spatially on the surface of the molecules, yet the performance needs to be understood over an entire corpus of molecules. In this project, we introduce a multi-scale approach for assessing the performance of these structural classifiers, providing coordinated views for both corpus level overviews as well as spatially-embedded results on the three-dimensional structures of proteins.

Keywords: J.3.1 [Computer Applications]: Life and Medical Sciences — Biology and Genetics;

1 INTRODUCTION

Developing classifiers for proteins is a common task in bioinformatics. The goals of these classifiers are diverse, predicting functionality like binding affinity, druggable sites, and flexible regions. Generally, these structural classifiers make predictions as to which regions on the protein have a set of target properties. While a vast literature provides example models for a particular structural task (see Irsoy *et al.* for a survey [5]), tools that analyze the output of these predictive models are typically simple and rely on tabular data.

Currently, analyzing classifier performance relies on aggregated summary statistics and individual examination of specific proteins using standard molecular graphics tools, providing an incomplete picture of performance and an inability to localize errors. With this project, we describe an approach to visually understand the results of a structural classifier over a test corpus. Using visual techniques, classifier performance can be presented in a concise manner that affords better error localization and pattern detection.

We created a prototype tool to view the generality of the predictive model across the corpus and to identify patterns of strong and weak performance. Understanding the predictive performance of these classifiers requires analyzing both the correctness and confidence for each classification value at both the protein and corpus level. While reporting predictions leads to a simple representation, our tool must handle the size of the data as well as the spatial organization of data on molecular surfaces. Displaying classification results in this manner must also allow the user to augment descriptive statistics and discover where strong and weak performance occur.

Trends of classifier performance can manifest themselves over different groups of proteins and over different spatial regions on individual proteins, suggesting our multi-scale approach. These two views are designed to address error identification and localization at their respective scales, and coordinate to provide a multi-scale workflow. Our overview visualization leverages small multiples to provide a summary of individual performance over the collection of proteins. The individual molecular detail view shows a surface-abstracted three-dimensional view of the protein, providing a scaffold to display data mapped to the surface in a manner similar to canonical molecular graphics applications such as PyMol [3]. However, we group spatially-congruent results on the molecular surface to help collapse the data into a small set of regions that supports interactive and guided exploration. Our analytical approach is realized by a prototype that we have applied to various structural binding classifiers.

*e-mail: sarikaya@cs.wisc.edu
†e-mail: dalbers@cs.wisc.edu
‡e-mail: gleicher@cs.wisc.edu

2 TASK ANALYSIS

Our problem domain focuses on proteins: large macro-molecules that are composed of 20 naturally-occurring amino acids (residues) chained together in sequence. The residues, in turn, fold over one another to form the conformation that governs the protein's biological and chemical function. Structural classifiers leverage chemical and physical features of these residues in a three-dimensional spatially-aware fashion, using chemical and physical features such as charge, hydrophobicity, and surface curvature.

Protein surface classifiers are tested on tens to hundreds of proteins, each with 40 to 600 residues and larger. These testing corpora have classification prediction, prediction confidences, and ground truth labels for each residue for every molecule. The hierarchical nature of this data lends support for our multi-scale approach. Identifying individual molecules or classes of molecules where the predictive model has weak or strong performance of true positives (TP) helps to understand the generality of the model. False positive (FP) classifications may identify decoys that may indicate over-generalization. Patterns of false negatives (FN) can signify biochemical motifs not captured in training and can indicate that the model is specific to a particular subproblem.

Several aspects of the problem create requirements for visualization tools. First, we need to cope with the size and density of our data, both in terms of the number of proteins in the corpus and the number of residues in each protein. Second, residue-based data is inherently three-dimensional, which makes it largely incompatible with standard overview techniques. Third, the ground truth distribution is skewed toward negative labels.

3 OUR APPROACH

Figure 1 shows our prototype implementation using the output of a DNA-binding classifier. On the left-hand side, the overview window provides a visual encoding of the performance over the entire test corpus. The overview provides a small multiples display where each protein's classification performance is summarized. A protein is shown on the right-hand side in the molecular viewer, accessed by clicking on a small multiple. The surface of the protein displays a spatial clustering of classification values (TP, FP, FN), conveying clusters of classification values directly on the protein structure. These clusters are enumerated in the list to the right of the molecule.

Performance classes are encoded by color. Four three-step sequential ColorBrewer ramps [4] (green for TP, blue for FP, grey for TN, and red for FN) use luminance and saturation variance to encode confidence. The molecular detail window also utilizes these colors. The overview provides summaries of each molecule's performance to support localizing different performance trends in the corpus. To help localize patterns of performance on the molecule, residues on the surface are grouped based on their classifications.

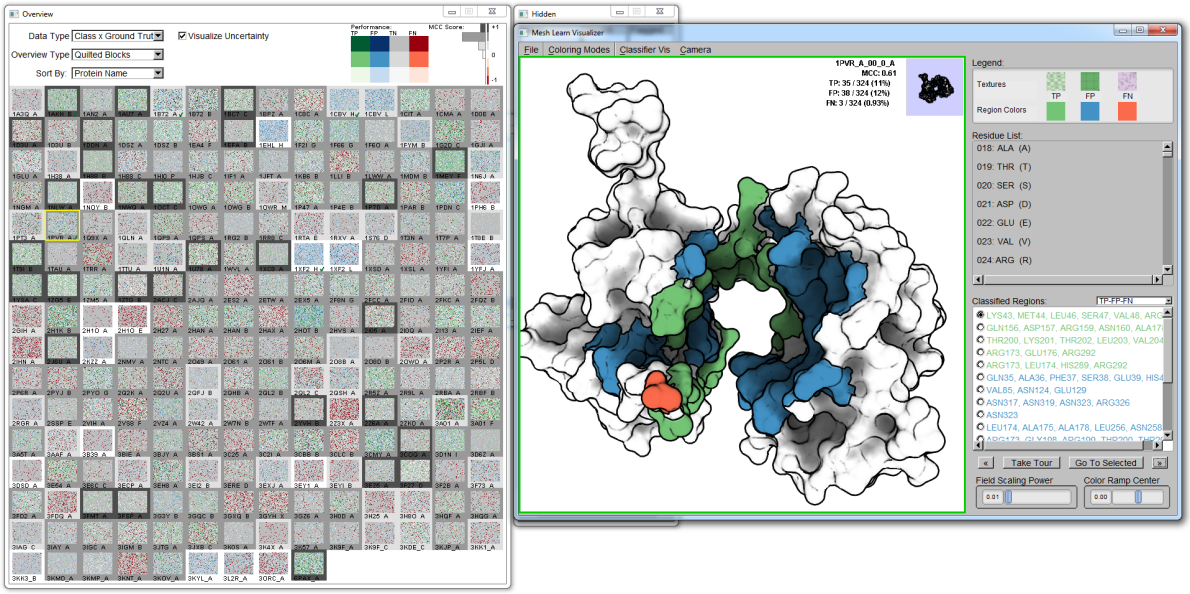
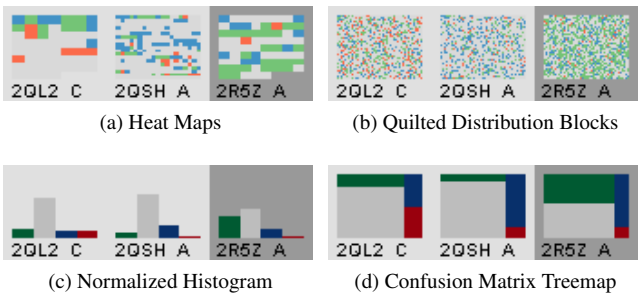


Figure 1: An overall view of our predictive model visualization for a surface DNA-binding predictive model. The overview window (left) displays the corpus rendered as quilted blocks, showing performance distributions per protein. The legend in the top-right of both windows provides a color key for the data. The detail window (right) shows the classification regions for PDB: 1PVR_A [1].

3.1 The Corpus-Level View

Figure 2 shows four different types of summaries that our system can use in the overview display. Heat maps (Fig. 2a) provide a way of viewing the classification performance of each residue of a protein sequentially, supporting judgements of numerosity in sequence order. Quilted blocks (Fig. 2b) use an adaptive color weaving technique to display distributional data by permuting pixel order, allowing for perceptual summarization of classification performance [2]. Histograms (Fig. 2c) provide a way to perform relative comparisons between classification categories. Confusion matrix treemaps (Fig. 2d) use a space-filling representation to support the distributional analysis of correctly and incorrectly classified residues.

The small multiples that make up the overview can be ordered using performance (e.g. accuracy) or molecular statistics (e.g. number of residues) in order to see trends in performance in relation to metadata. This dynamic querying takes advantage of perceptual strengths in these different views of performance. Localization of trends in strong or weak performance helps to identify those elements that warrant closer inspection with the molecular-level view.



(a) Heat Maps (b) Quilted Distribution Blocks

(c) Normalized Histogram (d) Confusion Matrix Treemap

Figure 2: Four overview encodings provide visual summaries of per-protein classifier output.

3.2 The Molecule-Level View

Displaying the protein as a solvent-excluded surface [6] allows for multivariate data to be displayed in the context of the three-dimensional structure of the folded residues. Collapsing residues into *classification regions* enables user iteration through the classification data. Residues that are spatial neighbors and have the same classification value can be grouped together using connected components. Using these regions, residues can be itemized (see right scroll in Figure 1) and toured via automatic viewpoint selection.

4 Conclusion

This multi-level visualization allows for performance insight to be obtained from the predictions of protein structural classifiers. We intend to generalize our approach to support performance judgments across multiple scales in order to help determine the strengths and weaknesses of domain-specific predictive models. Initial feedback from collaborators shows promise for the insights supported by our approach.

Acknowledgments

Many thanks to Julie Mitchell and Spencer Ericksen for providing the DNA-binding classifier data. This project was supported in part by NSF awards CMMI-0941013 and IIS-1162037.

References

[1] H. M. Berman et al. The protein data bank. *Nucleic Acids Res*, 28:235–242, 2000.

[2] M. Correll et al. Comparing averages in time series data. In *Proc 2012 ACM Human Factors in Computing Systems*, pages 1095–1104, 2012.

[3] W. L. DeLano. The PyMOL molecular graphics system. 2002.

[4] M. Harrower and C. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *Cartogr J*, 40(1):27–37, 2003.

[5] O. Irsoy et al. Design and analysis of classifier learning experiments in bioinformatics: Survey and case studies. *IEEE/ACM Trans Comput Biol Bioinf*, 9(6):1663–1675, Nov. 2012.

[6] F. M. Richards. Areas, volumes, packing, and protein structure. *Annu Rev Biophys Bio*, 6(1):151–176, 1977. PMID: 326146.