Understanding Performance of Protein Surface Classifiers

Alper Sarikaya
sarikaya@cs.wisc.edu

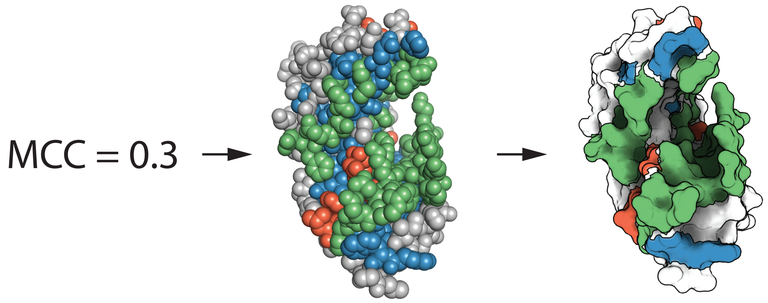
Danielle Albers
dalbers@cs.wisc.edu

Michael Gleicher
gleicher@cs.wisc.edu

Department of Computer Sciences
University of Wisconsin–Madison

The Problem

Surface classifiers make predictions as to which regions of the molecules have target properties. These classifiers can help determine the functionality of previously uncharacterized proteins and help elucidate the structural features that lead to a particular function. Current strategies for assessing the performance of these structural classifiers depend on aggregate summary statistics, which are insufficient and limit analysis.

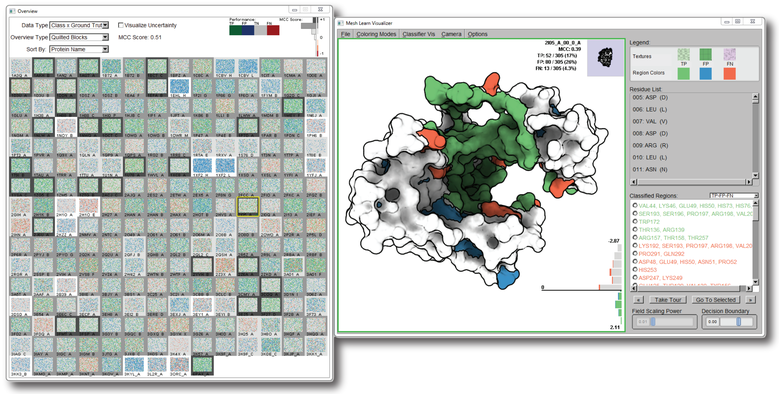


Visualizing a Solution

To understand how the classifier behaves, we must be able to see performance over the **corpus** and on the **molecule**. This suggests an overview+detail visualization. Special considerations need to be made to display spatial data in a way that decreases unnecessary visual complexity and allows inspection of all data.

To investigate performance across the test corpus, the **corpus overview** uses a re-orderable small multiples framework with glyphs designed for aggregate judgment.

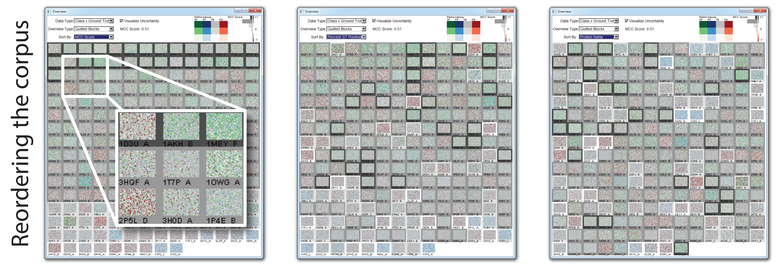
To see data on individual proteins, the molecular **detail view** uses the protein surface as a scaffold to display spatial classifications and input features in three dimensions. This representation affords bivariate layering using textures and automatic camera touring of classifications.



The visualization prototype shows a test corpus of proteins with the output of a DNA-binding classifier. The overview (left) shows aggregate performance across the corpus, while the detail view (right) shows an individual protein and its classifications.

Corpus Overview

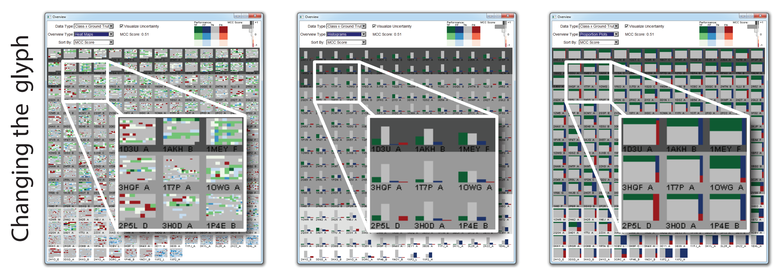
Each protein is represented by a small multiple. The visualization in the glyphs can be changed to afford different views of the data and the corpus can be reordered to discover patterns in performance.



Ordering by MCC score, best for finding outliers; *using* quilted blocks, best for extracting mean distributions

Ordering by percentage of positive labels, best for discovering the role of initial bias

Ordering by PDB ID, best to regain bearings over the corpus



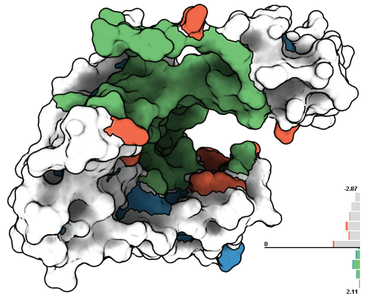
Using the heatmap glyph, best for high-fidelity numerosity judgments

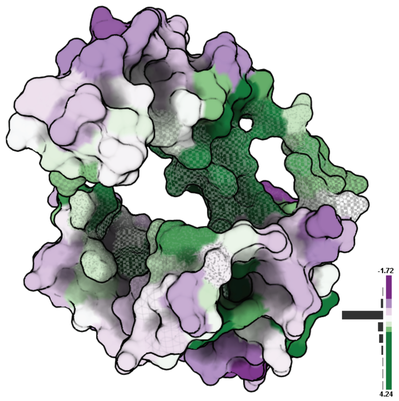
Using the histogram glyph, best for explicit classification distributions

Using the proportion plot glyph, best for determining classification correctness

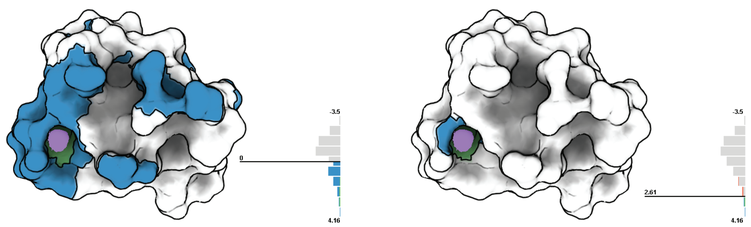
Detail View

Clustering classifications on the protein surface helps simplify the localization of classifications and transforms a large amount of tiny spatial decisions to a much smaller number of decisions grouped together. This simplifies spatial analysis and allows for automated camera touring.





Using textures to layer classifications over an input variable affords a separable bivariate encoding of features and classification. The user has control over the transfer function for the color.



Changing the decision boundary of the classifications allows the user to reclassify classifier predictions and see the resulting distribution of classifications in the histogram.

Acknowledgments

We thank our domain collaborators, Julie Mitchell and Spencer Ericksen, for discussions of the tool and providing the DNA-binding classifier data. This project was supported in part by NSF awards CMMI-0941013 and IIS-1162037.

References are available in the abstract submission.