CAVITY-AWARE MOTIFS REDUCE FALSE POSITIVES IN PROTEIN FUNCTION PREDICTION

Brian Y. Chen[a], Drew H. Bryant[b], Viacheslav Y. Fofanov[c], David M. Kristensen[d]
Amanda E. Cruess[a], Marek Kimmel[c], Olivier Lichtarge[d,e], Lydia E. Kavraki[a,b,e]
[a]Department of Computer Science,
[b]Department of Bioengineering,
[c]Department of Statistics,
Rice University
Houston, TX 77005, USA

[d]Program in Structural Computational
Biology and Molecular Biophysics,
[e]Department of Molecular and Human Genetics,
Baylor College of Medicine
Houston, TX 77030, USA

Determining the function of proteins is a problem with immense practical impact on the identification of inhibition targets and the causes of side effects. Unfortunately, experimental determination of protein function is expensive and time consuming. For this reason, algorithms for computational function prediction have been developed to focus and accelerate this effort. These algorithms are comparison techniques which identify matches of geometric and chemical similarity between motifs, representing known functional sites, and substructures of functionally uncharacterized proteins (targets). Matches of statistically significant geometric and chemical similarity can identify targets with active sites cognate to the matching motif. Unfortunately, statistically significant matches can include false positive matches to functionally unrelated proteins. We target this problem by presenting Cavity Aware Match Augmentation (CAMA), a technique which uses C–spheres to represent active clefts which must remain vacant for ligand binding. CAMA rejects matches to targets without similar binding volumes. On 18 sample motifs, we observed that introducing C–spheres eliminated 80% of false positive matches and maintained 87% of true positive matches found with identical motifs lacking C–spheres. Analyzing a range of C–sphere positions and sizes, we observed that some high-impact C– spheres eliminate more false positive matches than others. High-impact C–spheres can be detected with a geometric analysis we call Cavity Scaling, permitting us to refine our initial cavity-aware motifs to contain only high-impact C–spheres. In the absence of expert knowledge, Cavity Scaling can guide the design of cavity-aware motifs to eliminate many false positive matches.