Leveraging Protein Structure and Dynamics for Variant Interpretation in Coding Regions

Sushant Kumar, Declan Clarke and Mark Gerstein

Program in Computational Biology and Bioinformatics, Department of Molecular Biophysics and Biochemistry, Yale University

The pace of data generation by next-generation sequencing is presenting considerable challenges in terms of variant interpretation. Though deep sequencing is unearthing large numbers of rare single-nucleotide variants (SNVs), the rarity of these variants makes it difficult to evaluate their potential deleteriousness with conventional phenotype-genotype associations. Furthermore, many disease-associated SNVs act through mechanisms that remain poorly understood. 3D protein structures may provide valuable substrates for addressing these challenges algorithmically. We present two general frameworks for doing so. In our first, we employ models of conformational change to identify key allosteric residues by predicting essential surface pockets and information-flow bottlenecks (a new software tool that enables this analysis is also described). In our second approach, we use localized frustration, which quantifies unfavorable residue interactions, as a metric to investigate the local effects of SNVs. In contrast to this metric, previous efforts have quantified the global impacts of SNVs on protein stability, despite the fact that local effects may impact functionality without disrupting global stability (e.g. in relation to catalysis or allostery). Importantly, although these two frameworks are fundamentally structural in nature, they are algorithmically extremely fast, thereby making analyses on large datasets accessible. We detail how these database-scale analyses shed light on signatures of conservation, as well as known disease-associated variants, including those involved in cancer.