Integration of structural bioinformatics and deep-learning protein structure prediction: application to human protein kinases

Bulat Faezov and Roland Dunbrack
Fox Chase Cancer Center
Philadelphia, PA 19111 USA

The human proteome encodes 438 catalytically competent kinases that take on the typical protein kinase domain. Less than half of these kinases are represented in the Protein Data Bank in active conformations. In order to utilize deep-learning structure prediction methods to obtain active conformations of these domains, it is important to define what makes a kinase structure "active." We have used a rigorous structural bioinformatics approach, starting with the structures of substrate-bound kinase-substrate complexes and ATP-bound kinase structures. We find that in addition to the placement of the Phe residue of the DFG motif (DFGin) that kinases must have a certain pattern of dihedral angles of the whole DFG motif. From contacts with substrate peptides, it is clear that they also must position the N-terminal six amino acids of the activation loop and the C-terminal 10 amino acids of the activation loop in specific positions to allow substrate binding. By applying these criteria to potential templates for AlphaFold2 (including both experimental structures and AlphaFold2 predicted structures, which we call "distillation templates") and utilizing shallow multiple sequence alignments and heavy sampling, we are able to obtain active substrate-binding-capable models of all 438 catalytically active human protein kinases.