Predicting regulatory motifs of enzymes using machine learning

Wenfa Ng

Department of Chemical and Biomolecular Engineering, National University of Singapore,
Email: ngwenfa771@hotmail.com
ORCID ID: https://orcid.org/0000-0001-7629-2140

Abstract

Enzymes perform critical biochemical reactions in the cell, but they are not given a free ride to roam in the cellular milieu. Specifically, a variety of regulatory factors such as substrate, product, inducer, or activator interact with different domains and motifs on the enzyme to enact regulatory action that tunes, in broad strokes, enzyme activity. Much of such understanding comes from detailed biochemical assays augmented with insights from structural biology. Progress in elucidation of regulatory actions on enzymes is slow and defined by generalizing characterized phenomena into broad categories. But, much remains unknown. Recent advances in fast computational protein structure prediction by machine learning tools generates a larger set of structures for which structural biologists could link amino acid sequence to enzyme structure, and therefore predict putative binding sites for regulatory factors on proteins. Such insights could be further augmented by convolution neural network (CNN) exploration of protein-ligand binding affinity that annotated particular structural folds from structure prediction with biological function. While nascent, the combination of the above machine learning tools could revolutionize the speed and accuracy at which regulatory motifs on new enzymes could be predicted, which narrows the search space for biochemical experimental verification.

Keywords: regulatory motifs, enzymes, computational structure prediction, protein-ligand affinity predictions,

Subject areas: biochemistry, structural biology, computational biology, biotechnology,

Enzymes in the cellular milieu do not exist in isolation; rather, they are plugged into the cellular communication network where small molecules could influence enzyme activity through binding to allosteric site of a multi-domain enzyme. Such regulatory action manifests as conformational changes elicited by binding of a regulatory molecule on an enzyme motif that results in either enhanced or dampened enzymatic activity at the active site. Prediction of such regulatory action prior to selecting an enzyme for pathway construction holds importance given that metabolic engineering involves tapping on endogenous metabolism and signaling cascades for performing biotransformations. If such endogenous mediated regulatory actions are not predicted a priori, constructed pathways may not perform in an optimal fashion, and which may not yield to further optimization without knowledge of the stated enzyme regulatory action.
Predicting such regulatory action through machine learning methodologies would necessitate the use of artificial intelligence software for predicting enzyme structure from amino acid sequence information or predicting protein-ligand binding affinity or interactions. In contrast to enzyme activity prediction, structure prediction does not suffer from a severe lack of data given the increasing number of annotated protein structures deposited in Protein Data Bank. More importantly, current demonstrations of machine learning based protein structure prediction has shown that the approach is much faster compared to force-field based molecular dynamics approaches. But, as a data-driven approach, machine learning structure prediction critically requires structures whose regulatory motifs have been well-characterized by experimental approaches. Currently, application of machine learning tools to predict enzyme regulatory motifs remain in its infancy with few papers on the topic. However, the field is expected to emerge strongly as machine learning algorithms gain maturity in predicting protein structures, particularly those with multi-domains.

Convolutional neural network (CNN) has been used for predicting ligand binding sites in proteins based on learning of specific molecular interaction data, hydrogen bond, aromatic and hydrophobic contacts. Although trained to detect nucleotide and heme binding sites, the method has been generalized to steroid-binding proteins and peptidase enzymes with structures modelled using interaction energy based attributes. In another study, CNN was also used in a sequence-based predictor of ligand binding site residues of a protein. The method achieves higher prediction accuracy compared to the state-of-the-art COACH approach. In general, CNN has shown utility in uncovering protein residues capable of binding a particular ligand through surveying the relative distance between protein residues and ligand atoms and identifying regions of closest approach which highlights a potential binding site. Such algorithms for identifying potential ligand binding sites in proteins could possibly be applied to determine enzyme regulatory motifs in a further extension of the work.

Besides predicting enzyme regulatory motifs, possibility also exists in using the propensity of ligand in binding a protein in understanding potential regulatory action. This inevitably necessitate the use of protein-ligand binding affinity data, which has been shown to be capable of being predicted by a combination of Random forest and Gaussian process regression algorithms with molecular properties of ligand and protein as inputs.

Collectively, much of our current understanding of protein regulatory mechanism comes from detailed and laborious biochemical assays augmented with structural biology elucidation. But, given the diversity of proteins available, the current approach could not appreciably sample a sizeable fraction of the regulatory space of enzymes. Machine learning tools such as computational structure prediction and neural network-based prediction of protein-ligand binding affinity offer tools that could help augment the biochemist’s toolkit. While not definitive, these approaches do suggest possible regulatory motif of enzymes that could be verified experimentally. Such approaches are likely to gain importance in the future, and may be the first step to machine learning based predictions of regulatory motifs of enzymes.
Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work.

References