

In-silico Prediction of Surface Residue Clusters for Enzyme-Substrate Specificity

Gong-Xin Yu[‡], Byung-Hoon Park[‡], Praveen Chandramohan, Al Geist, Nagiza F. Samatova
Computational Biology Institute, Oak Ridge National Laboratory

Corresponding Authors: {yug, samatovan}@ornl.gov

[‡] Both authors have contributed equally to this work

Abstract

One of the most remarkable properties of enzyme-substrate binding is the high substrate specificity among homologous enzymes. Identification of regions in enzymes that play an important role in substrate recognition presents an opportunity to understand their basic molecular mechanisms. Current methods are limited to identifying conserved residues, ignoring potential contributions of non-conserved residues. Our method overcomes this limitation. In case studies, we investigated several highly homologous enzymatic protein pairs such as guanylyl vs. adenylyl cyclases and lactate vs. malate dehydrogenases, and applied our method on plant and cyano-bacterial RuBisCos. We identified several critical mono-residue and multi-residue clusters that were consistent with experimental results. Some of the identified clusters, primarily the mono-residue ones, represent residues that are directly involved in enzyme-substrate interactions. Others, mostly the multi-residue ones, represent residues vital for domain-domain and regulator-enzyme interactions, indicating their complementary roles in specificity determination.

1. Introduction

Homologous enzymes exhibit high specificity when binding to their substrates. How do these enzymes achieve such exquisite substrate specificity? Several possible mechanisms have been suggested for these delicate substrate specificities such as substrate-binding in the catalytic centers of enzymes [1], loop-based hinge-motion [2] and cofactor binding and intra- or inter- molecule (domain-domain) interactions [3]. The specificity-determinant regions are mostly small clusters of critical amino acids on the surface of the protein. Mutations in these

residues often force conformational changes, thus, having an immense effect on the substrate specificity. Accurately identifying these regions and residues will have immediate implications for drug design, protein engineering, elucidating molecular pathways through site-directed mutagenesis, and detailed functional annotation.

Current computational approaches, either based on the evolutionary history [4] or HMMs and relative entropy [5], identify only conserved residues but largely ignore non-conserved residues and their potential contributions. In our work, we present a surface patch method to overcome these limitations. In this algorithm, we focus on identifying clusters of spatially co-located surface residues. Our understanding is that critical amino acids responsible for the specificity often cluster in small regions on protein surfaces. An important strength of our approach is its enhanced sensitivity to predict clusters of residues with a low degree of conservation in addition to those that are conserved. Our method also exhaustively exploits residue clusters by focusing on surface patches of varying sizes and ignoring internal residues (more likely linked to the structural integrity of the protein). The method places the contribution of an entire residue cluster at the core of the analysis as opposed to **ET**-like approaches that first evaluate the importance of individual residues and then filter those that are spatially clustered.

2. Method

The key idea of this method is to identify a minimum set of spatially co-located surface residues, named as **Specificity-Determining surface Residue Clusters (SDRC)**, which can discriminate between two classes of functional sub-types with respect to certain enzyme substrate specificity. In “minimum”, we mean that every residue in the cluster contributes to the substrate specificity, either directly (a **SDRC** of single

highly conserved residue) or complementarily (a *SDRC* of multiple non-conserved residues).

There are three major components in this method. First, we specify a search space, in which we confine ourselves to clusters of spatially co-located surface residues, the most likely functional regions. Second, we define a scoring function in terms of classification accuracy provided by multivariate discriminant methods such as SVMs, NNs, or decision trees. The purpose of the scoring function is to determine how well these residue groups can discriminate between different functional sub-types. Finally, we define the statistical significance of the discrimination to select groups of residues with significantly better scores than the others. A high score alone cannot define a group of specificity-determining residues because it strongly depends on the overall amino acid composition in the alignment.

3. Results and conclusions

We applied the method on three benchmark enzyme pairs (function sub-types): Lactate dehydrogenase (LDH) vs. malate dehydrogenase (MDH), Guanylyl cyclase (GC) vs. adenylyl cyclase (AC), and Trypsin (Tr) vs. Chymotrypsin (Ch). By this approach, we identified mono-residue *SDRCs* as well as multi-residue ones, both of which provided equally strong capabilities in the classification of these functional subtypes. We compared our predictions with experimental and structure data and obtained a considerable agreement with them. Specifically, we discovered that some of the *SDRCs*, primarily the mono-residue *SDRCs*, can cover residues that are directly involved in substrate-enzyme interactions, whereas, others, mainly multi-residue *SDRCs*, cover residues vital for domain-domain, and regulator-enzyme interactions.

We extended this method to study RuBisCo enzymes in plants and cyano-bacteria, which differ dramatically in the CO₂/O₂ specificity. RuBisCo is an important enzyme for carbon fixation in photosynthesis and carbon oxidation in photorespiration. The latter, a reversing reaction to photosynthesis, results in net carbon loss, making it the primary limitation of carbon biomass productivity. A better understanding of the biochemical and genetic mechanisms of Rubisco-related CO₂/O₂ specificity would greatly boost our ability to make a great progress in agricultural productions and environmental managements. In this analysis, we focused our

study on large subunits of the RuBisCos since they have been identified as the major specificity-determining resource [6]. By this analysis, we identified *SDRCs* that are strongly associated with residues and surface region critical to the CO₂/O₂ specificity. Interestingly, most of the residues occur in multi-residue *SDRCs*, indicating the potential roles of functionally non-specific residues in specificity determination. These analyses demonstrate that our method can accurately identify residue clusters key to the determination of substrate specificity, and thus help select target residues for mutagenesis experiments focusing on rational protein design and engineering. It can also help in functional improvements of RuBisCo, and other medically, agriculturally and environmentally important enzymes.

4. Reference

- [1] Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. Structural Clusters of Evolutionary Trace Residues are Statistically Significant and Common in Proteins. *Journal of Molecular Biology*, 2001 316: 139-154.
- [2] Miller DW, Agard DA. Enzyme specificity under dynamic control: a normal mode analysis of alpha-lytic protease. *J Mol Biol*. 1999 Feb 12;286: 267-78.
- [3] Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol*. 2002 Feb; 12:21-27.
- [4] Lichtarge O, Bourne HR, Cohen FE. Evolutionarily conserved Galphabeta gamma binding surfaces support a model of the G protein-receptor complex. *Proc Natl Acad Sci U S A*. 1996 Jul 23;93:7507-7511.
- [5] Hannehalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol*. 2000 Oct 13;303(1):61-76.
- [6] Spreitzer RJ, Salvucci ME. Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annu Rev Plant Biol*. 2002;53:449-475.