C-1-01

A random-forest based method that can predict detailed enzyme functions and also identify specificity determining residues

Chioko Nagao[1] chio@nibio.go.jp

Nozomi Nagano[2] n.nagano@aist.go.jp

Kenji Mizuguchi[1] kenji@nibio.go.jp

[1] National Institute of Biomedical Innovation, 7-6-8 Saito-Asagi, Ibaraki, Osaka 567-0085, Japan

[2] Computational Biology Research Center, AIST, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

Keywords: Enzyme function, EC number, Random forests, Specificity determining residues

Determining enzyme functions is essential for understanding chemical reactions occurring in living cells. Although many prediction methods have been developed, it remains a significant challenge to predict enzyme functions at the fourth-digit level of the Enzyme Commission (EC) numbers [1]. A small number of mutations can often drastically change functional specificity of enzymes. Therefore, information about these specificity determining residues (SDRs) can potentially help discriminate detailed functions. However, because these residues must be identified by mutagenesis experiments, the available information is limited, and the lack of experimentally verified SDRs has hindered the development of detailed function prediction methods and computational identification of SDRs.

In this study, we developed EFPrf, a novel method for predicting enzyme functions at the fourth-digit level of EC numbers and identified a set of putative SDRs (rf-SDRs) by using a machine-learning technique known as random forests [2]. For each enzyme in each CATH homologous superfamily [3], binary predictors were constructed by random forests with full-length sequence similarities and the residue similarities for active sites, ligand binding sites and conserved sites as input attributes. From the most highly contributing attributes, we obtained the rf-SDRs. In a cross-validated benchmark assessment, EFPrf showed a prediction performance comparable to that of a related method currently available (precision=0.98, recall=0.89). The rf-SDRs included many residues, whose importance for specificity had been validated experimentally. The analysis of the rf-SDRs revealed both a general tendency that functionally diverged superfamilies tend to include more active site residues in their rf-SDRs than in less diverged suprefamilies, and more complicated relationships that the rf-SDRs strongly depend on the mechanisms of functional diversification in each superfamily.

[1] Webb E. C., NC-IUBMB, *Enzyme Nomenclature 1992, Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, 1992.

[2] Breiman L., Random Forests, *Machine Learning Journal*: 5-32, 2001.

[3] Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B., Thornton J. M., CATH -a hierarchic classification of protein domain structures-, *Structure* 5: 1093-1108, 1997.