Prediction of enzyme-catalyzed reactions based on atomic and molecular properties of metabolites

Fangping Mu[1], Robert F. Williams[2], Pat J. Unkefer[2], Clifford J. Unkefer[2] and William S. Hlavacek[1]

***Short Abstract* — Our knowledge of the metabolites in cells and their reactions is far from complete, as revealed by metabolomic measurements that detect many more small molecules than are documented in databases. Here, we developed an approach for predicting the reactivity of metabolites in enzyme-catalyzed reactions that combines expert knowledge and machine learning. Metabolic reactions are represented as 80 functional groups transformations. Based on atomic properties of atoms in the reaction center, which models the local environment of the functional group, and molecular properties, classifiers have been trained to discriminate the functional groups to be reactive and non-reactive. On average, 80% accuracy is obtained with respect to both sensitivity and specificity.**

***Keywords* — Metabolomics, reactivity, machine learning**

METABOLOMICS, the study of a cell's complete complement of small-molecule metabolites, has recently been recognized as an important part of post-genomics science(1). Identification of the structures of metabolites and reconstruction of the network of enzyme-catalyzed reactions in which metabolites participate are two of the important challenges for metabolomic analysis of a cell. Large-scale metabolic network reconstructions currently available have largely been inferred from genome annotation(2), which identifies genes encoding enzymes, and literature data. The sequence-based approach to network reconstruction can only infer reactions that have been characterized experimentally in some system, but large gaps in our knowledge of metabolism seem likely. Unknown metabolic reactions cannot be discovered using automatic reconstruction methods.

The gaps in our knowledge of metabolomics are being revealed by metabolomic detection of small-molecules not previously known to exist in cells. An important challenge is to determine the reactions in which these compounds participate, which can lead to the identification of gene products responsible for novel metabolic pathways. To address this challenge, we investigate how machine learning can be used to predict potential substrates and products of enzyme-catalyzed reactions(3). Based on 6577 reactions downloaded in the KEGG database(4), we classify the vast majority of these reactions into 80 groups, each of which is marked by a characteristic functional group biotransformation. The functional group patterns of the reaction centers of the 82 substrate(s) types and 88 product(s) types of reactions are represented using SMARTs. These patterns are not unique to these reactions but are widely distributed among KEGG metabolites. Based on atomic properties of atoms in the reaction center, which models the local environment of the functional group, and molecular properties, classifiers have been trained to discriminate the functional groups to be reactive and non-reactive. On average, 80% accuracy is obtained with respect to both sensitivity and specificity. Our results suggest that metabolic reactivity in enzyme-catalyzed reactions can be predicted with reasonable accuracy based on functional groups and its local environmental in the molecule, which is characterized by the atomic properties and molecular properties.

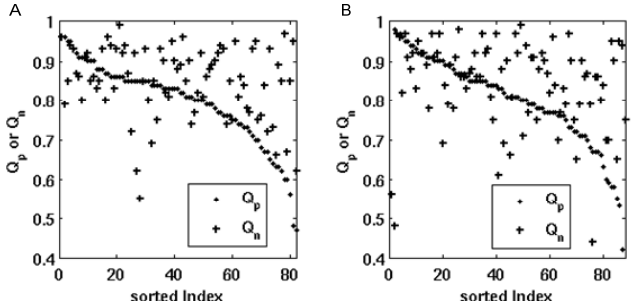

Fig. 1. Sensitivity (Q_p) and specificity (Q_n) of classifiers for 82 substrates (A) and 88 products types (B). X-axis are sorted according to Q_p. Q_p and Q_n are plotted as dots and plus sign, respectively.

[1]Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, NM 87545, USA E-mail: fmu@lanl.gov

[2]National Stable Isotope Resource, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

REFERENCES

1. Fiehn O (2002) *Plant Mol Biol* **48,** 155-171.
2. Price ND, Reed JL, & Palsson BØ (2004) *Nature Reviews Microbiology* **2,** 886-897.
3. Mu F, Unkefer PJ, Unkefer CJ, & Hlavacek WS (2006) *Bioinformatics* **22,** 3082-3088.
4. Goto S, Okuno Y, Hattori M, Nishioka T, & Kanehisa M (2002) *Nucleic Acids Res* **30,** 402-404.