Poster I-13

Predicting Enzyme Class From Protein Structural Parameters Using STING_DB parameters and Bagging Predictors

Authors:
Michel E. B. Yamagishi (*Embrapa Informática*)
Stanley R. M. Oliveira (*Embrapa Informática*)
Luiz C. Borro (*Embrapa Informática*)
Edgard H. Santos (*Embrapa Informática*)
José G. Jardine (*Embrapa Informática*)
Fábio D. Vieira (*Embrapa Informática*)
Ivan Mazoni (*Embrapa Informática*)
Marcelo G. Narciso (*Embrapa Informática*)
Paulo R. Kuser-Falcão (*Embrapa Informática*)
Goran Neshich (*Embrapa Informática*)

Short Abstract: In this work we present a new method to classify enzymes that uses the STING_DB physical-chemical parameters and Bagging predictors. By building models based on "decision tree" and "neural network", we obtained an accuracy of 74% on average. These results outperform the similar models proposed in literature.

Long Abstract:

Every year, the number of protein structures available in the Protein Data Bank, PDB, (Berman et al., 2000) is increasing. As a side effect, the number of structures lacking functional annotation is also growing very fast because some of the new structures have no sequence or structural homolog.

Most of the annotation methods rely on the similarity of sequence or structure between a functionally annotated protein and the query protein, and there are well known tools for sequence similarity detection (Altschul et al., 1990) or structural similarity identification (Shindyalov & Bourne, 1998). However, it is also desirable to have alternative methods to classify these new structures with no similar structure or sequence. Many approaches have emerged lately. One of them (Dobson & Doig, 2005) has called our attention because the authors established a connection between structure and function through structural parameters. Using the 6 protein enzyme classes and their respective members, according to Enzyme Classification (EC) number, from ASTRAL SCOP (Chandonia et al., 2002), for each structure, they created a matrix where the rows were assigned to one of the 20 amino acids and the columns, structural attributes. In that work the authors combined the prediction of one-class versus one-class support vector machine models to make overall assignments of EC number to an accuracy of 35% with the top-ranked prediction. In the same direction, but with a different approach, our method predicts the enzyme class using the STING_DB structural parameters. More than 300 attributes are available in the database, which makes the attribute selection step particularly difficult. Moreover, the number of residues varies from structure to structure which means that we have to create matrices with different number of columns. In order to build the model, all the matrices should have the same dimensions. This criteria was accomplished by using a modified 2D auto-correlation function (Broto, Moreau &

Vandycke , 1984). Instead of the original topological distance in the correlation function, the Euclidean distance was used in order to take in account the three-dimensional amino acids' environment. Another important contribution of our work is to apply the bagging predictors (Breiman, 1996) to improve the accuracy of our models. Bagging is an acronym to "bootstrap aggregating" procedure. It is a strategy borrowed from machine learning field and has proved to be very effective. By building models based on "decision tree" and "neural network", we obtained an accuracy of 74% on average. These results outperform the similar models proposed in literature.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E (2000). The Protein Data Bank. Nucl. Acids Res. 28, 235-242.
3. Breiman, L. (1996). Bagging predictors. Machine Learning 26, pp. 123-140.
4. Broto, P., Moreau, G. & Vandycke,C.,
Eur. J. Med. Chem. - Chim. Ther. 19, 61-65
5. Chandonia, J. M., Walker, N. S., Conte, L. L., Koehl P., Levitt, M. & Brenner, S. E. (2002) ASTRAL compendium enhancements. Nucl. Acids Res. 30, 260-263.
6. Dobson, P. D. & Doig A. J. (2005) Predicting enzyme class from protein structure without alignment J. Mol. Biol. 345, 187-199.
7. Shindyalov, I. N. & Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739-747