Using Support Vector Machines to predict expression-purification success of recombinant integral membrane proteins in Escherichia coli

Punta M[1,2], Love J[2], Kloss B[2], Bruni R[2], Hillerich B[2], Mancia F[2,3], Shapiro L[2,4], Hendrickson WA[2,4,5], Rost B[1,2,4,6]

Motivation

At the New York Consortium on Membrane Proteins Structure (NYCOMPS), we have performed a retrospective analysis of outcomes in membrane protein expression and purification. We have developed a Support Vector Machine (SVM)-based method that predicts expression-purification success under our current experimental protocols. The predictor is meant to become a new tool for prioritization of experiments at the consortium.

Methods

The data for this analysis came from 2,674 proteins, a subset of those successfully cloned and tested in the protein production pipeline of NYCOMPS. All proteins in the training set were predicted to have >= 2 transmembrane helices. 22% of them were successfully purified and expressed to levels that were considered acceptable for further experimental processing towards the goal of structure determination. For the predictor's development we used the LIBSVM package [1].

Results

We first looked at correlation between expression-purification success and individual protein features. Features that showed significant correlations included, among others, GC content, Codon Adaptation Index, protein termini localization and the protein's organism of origin. Next, we applied SVMs to combine 15 such features into a predictor of protein expression-purification success. The newly developed SVM was used to derive success scores for all proteins currently found in the NYCOMPS list of valid targets [2]. In order to blind-test the predictor's performance, we selected 3 novel valid target subsets: TOP, comprising proteins with the highest SVM scores (i.e. proteins most likely to purify and express);

[1] Department of Informatics, Bioinformatics, TU Muenchen and Institute for Advanced Study, Garching b. Muenchen, Germany [2] New York Consortium on Membrane Protein Structure, New York Structural Biology Center, New York, NY, USA [3] Department of Physiology and Cellular Biophysics, Columbia University, New York, NY, USA [4] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA [5] Howard Hughes Medical Institute, Columbia University, New York, NY, USA [6] Columbia University Center for Computational Biology and Bioinformatics (C2B2) and Northeast Structural Genomics Consortium (NESG), New York, NY, USA

BOTTOM, comprising proteins with the lowest SVM scores; RANDOM, a control set of randomly chosen proteins. Each set comprised 626 targets. All 3 sets were submitted to the NYCOMPS protein production pipeline for testing. Expression-purification success was 7%, 20% and 27% for BOTTOM, RANDOM and TOP datasets, respectively. This data showed that the predictor could indeed separate the bad from the good targets, with TOP success about 4 times higher than BOTTOM. On the other hand, TOP success was lower than expected from cross-validation experiments performed on the training set. While we are currently investigating the reasons for such an outcome, the proved ability of the predictor to identify “stay-away” targets means that it can already be used as a filtering tool in target selection. [1] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm [2] Punta et al. J Struct Funct Genomics 2009, 10:255-268

Contact e-mail

punta@rostlab.org