Prediction of subplastidial localization of chloroplast proteins from spectral count data - Comparison of machine learning algorithms

Thomas Burger[(1)], Samuel Wieczorek[(1)], Christophe Masselon[(1)], Daniel Salvi[(2)], Norbert Rolland[(2)], Myriam Ferro[(1)]

[(1)] CEA/Grenoble, iRTSV, Biologie à Grande Echelle (équipe EDyP), CNRS FR 3425, INSERM U1038, Université Joseph Fourier, F-38054 Grenoble, France.
[(2)] CEA/Grenoble, iRTSV, Physiologie Cellulaire Végétale, CNRS UMR5168, INRA UMR1200, Université Joseph Fourier, F-38054, Grenoble, France.

Context: In order to study chloroplast metabolism and functions, subplastidial localization is a prerequisite to achieve protein functional characterization. As the accurate localization of many chloroplast proteins often remains hypothetical, we set up a proteomics strategy in order to assign the subplastidial localization of chloroplast proteins.

State-of-the-art: A comprehensive study of *Arabidopsis thaliana* chloroplast proteome has been carried out in our group [1], involving high performance mass spectrometry analyses of highly fractionated chloroplasts. In particular, spectral count data were acquired for the three major chloroplast sub-fractions (stroma, thylakoids and envelope) obtained by sucrose gradient purification. As the distribution of spectral counts over compartments is a fair predicator of relative abundance of proteins [2], it was justified to propose a prime statistical model [1] relating spectral counts to subplastidial localization. This predictive model was based on a logistic regression, and demonstrated an accuracy rate of 84% for chloroplast proteins.

Contribution and results: In the present work, we conducted a comparative study of various machine learning techniques to generate a predictive model of subplastidial localization of chloroplast proteins based on spectral count data. To do so, we trained on the same dataset containing spectral count information for 555 proteins, various classification algorithms:

1. **Support Vector Machines**, **Random Forest**: the state-of-the-art in terms of performances.
2. ***k*-nearest neighbors**: a baseline reference, the performances of which, when compared to the state-of-the-art, provide interesting clues on the computational complexity of the problem.
3. **PerTurbo**: A new classification algorithm [3] based on kernel tricks and matrix perturbation theory, which provides results similar to SVM, while providing several qualitative advantages (fewer parameter to tune, efficient with high number of classes, no risk of over-fitting, etc.)

From this comparison, it appears that the most efficient predictive models provide accurate subplastidial localization for 91% of the proteins. Thus, compared to the original model based on logistic regression, it corresponds to an improvement of 7 points the accuracy rate, and an avoidance of ~40% of the misclassifications.. In addition, we also focused on more qualitative elements: The coverage of the training set, the processing of mislabeled data, and the influences of the parameters of the algorithm. These essential elements will be of prime importance in subsequent work, aimed at developing more accurate models based on comprehensive datasets, and leading to accurate prediction even in the case of multi-localized proteins.

References:

1. **Ferro, M., et al. (2010).** AT_CHLORO: comprehensive chloroplast proteome database with subplastidial localization and information for functional genomics using quantitative label-free analyses, *Mol. Cell. Proteomics,* 9(6): 1063-1084.
2. **Gilchrist A., et al. (2006).** Quantitative Proteomics Analysis of the Secretory Pathway, *Cell 127:1265–1281.*
3. **N. Courty, T. Burger, J. Laurent (2011).** "*PerTurbo: a new classification algorithm based on the spectrum perturbations of the Laplace-Beltrami operator*", *ECMLPKDD 2011.*