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

Thomas Burger, Samuel Wieczorek, Christophe Masselon, Daniel Salvi, Norbert Rolland, Myriam Ferro

# 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].

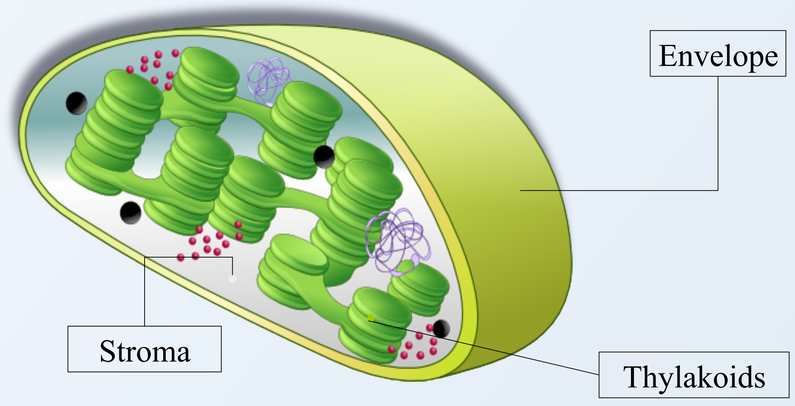
[1] CEA/Grenoble, iRTSV, Biologie à Grande Echelle (équipe EDyP), CNRS FR 3425, INSERM U1038, Université Joseph Fourier, F-38054 Grenoble, France.
[2] CEA/Grenoble, iRTSV, Physiologie Cellulaire Végétale, CNRS UMR5168, INRA UMR1200, Université Joseph Fourier, F-38054, Grenoble, France.

CEA / DSV / iRTSV

## Introduction

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 accurate subplastidial localization.

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.




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.

## Comparison of various machine learning algorithms

1. **Support Vector Machines** (SVM) and **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.). It can be understand as both a subspace classifier and a class-wise kernel Principal Component Analysis.

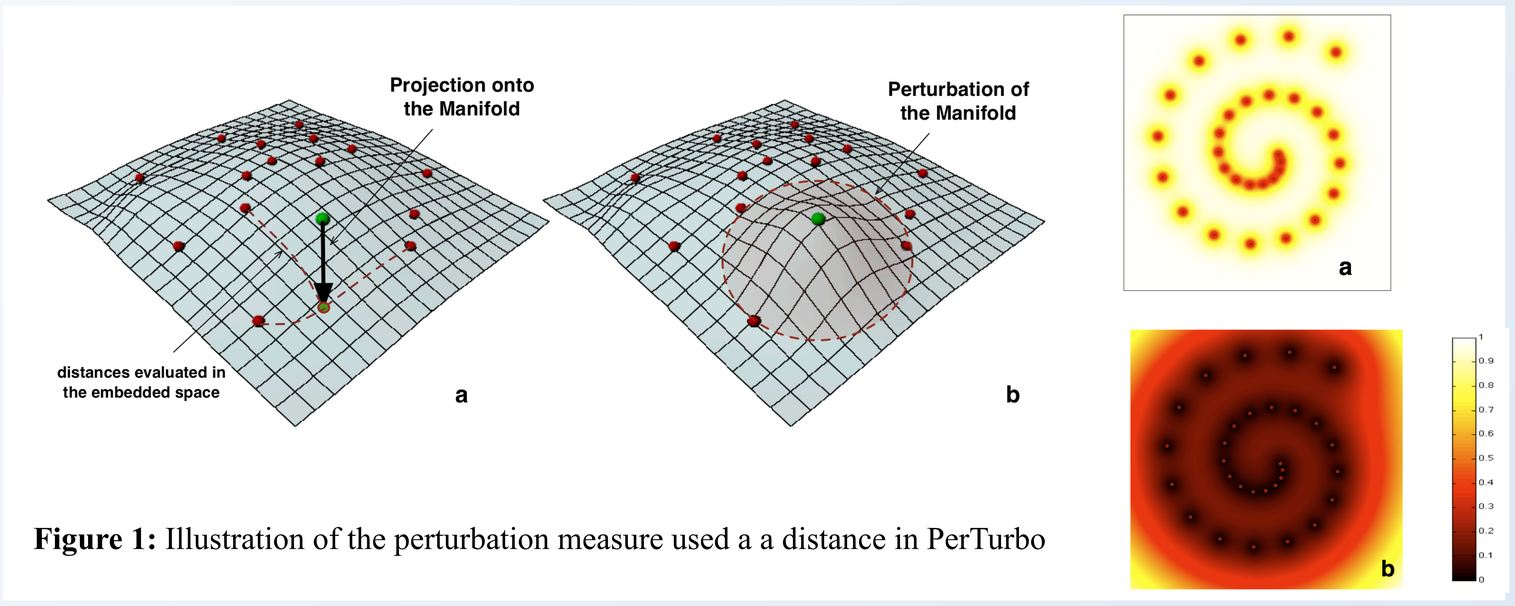



**Figure 1:** Illustration of the perturbation measure used a a distance in PerTurbo

| Algorithms | Parameters | Performances |
|---|---|---|
| Logistic regression | -- | 84% (--) |
| SVM | $\sigma$ =10, C =$10^5$ | 92.2% (1.2) |
| Random Forest | Nbtree = 200, Deep= 1 | 91.7% (1.4) |
| *k*-nearest neighbors | $k$ = 3 | 89.4% (1.3) |
| PerTurbo | $\sigma$ = 0.25, $\lambda$=$10^{-8}$ | 91.9% (1.1) |

**Labeling the dataset:** the same dataset contains spectral count information for 555 proteins. Among them, 216 were labeled according to the expertise of biologists of the team, while the 339 others were labeled according to PPDB (curated location).

**Composition of the training/test sets:** With logistic regression, the set of 216 proteins was used for training, and the other for testing. For the new experiments, the two datasets were merged together, and 50% was randomly sample for training (the remaining is the test set). The operation was repeated 50 times to compute a mean accuracy rate and a standard deviation.

**Parameter optimization:** The single or two parameter(s) of each algorithm were optimized by cross-validation on grid search. The grid search was logarithmic for $\sigma$, $\lambda$ and C. The kernel involved in SVM and PerTurbo was the Gaussian one.

## Beyond the performances

### Coverage of the training / Generalization power

Considering the evolution of the classification accuracy rate when the training/testing ratio varies, it appears that the variation of the performances are really low. This indicates that:

- There is little margin between baseline and state of the art performances indicating the problem is simple
- There is an important redundancy in the training set
- The remaining uncertainty of the model is not captured: Relevant hidden variables are still to discover

### Label Correction

As some of the labels appear as conflictive with respect to the various spectral data count, the whole dataset was manually curated by biologists. After 10 corrections and 43 deletions (as uncertainty remains), a new training/testing phase was performed (on less noisy datasets) and the following results are found :
*SVM: 97.8% (0.75), PerTurbo : 98.3% (0.85), k-NN: 98.0% (0.74), random forest: 98.0 (0.84)*

### Future works: Accounting for multiple localizations

The main interest of a generative classifier such as PerTurbo is to provide a similarity measure between each protein and each localisation. Hence, by comparing the similarities between a protein and all the locations, it is possible to infer possible multiple localizations. As is, there is not enough data to conduct a complete study of multiple locations. However, it will be considered in furure works. In parallel, additional experiments are conducted within the framework of Prime-XS project (FP7), in order to investigate more precise subplastidial localizations (6 classes or sub-classes).

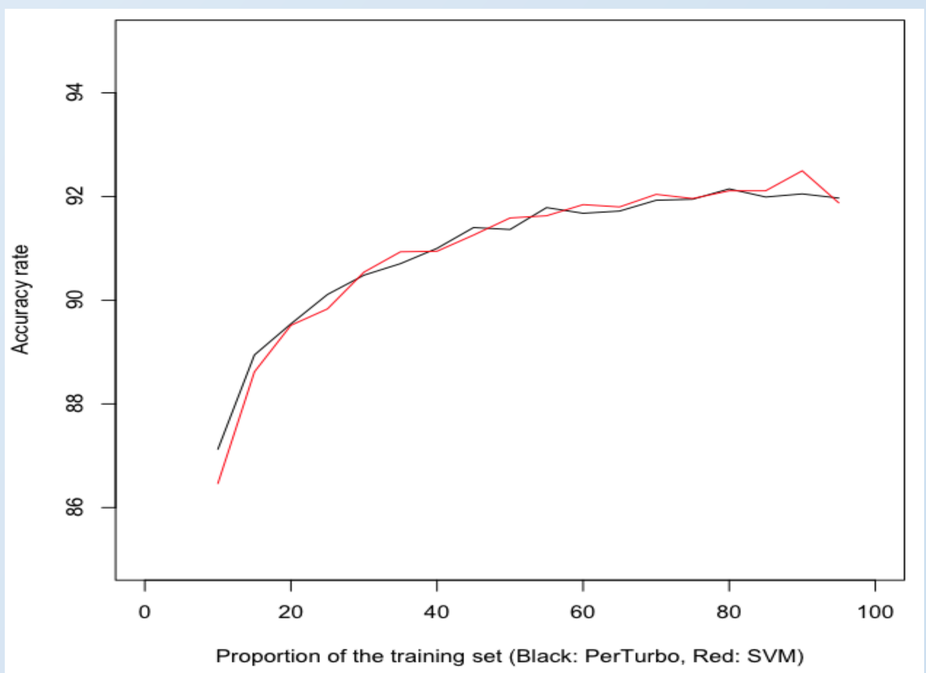



**Figure 2:** Evolution of the accuracy rate (in %) when the size of the training set increases (in % of the total dataset)

## 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.*

Exploring the Dynamics of Proteomes