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

To cite this version:

Thomas Burger, Samuel Wieczorek, Christophe Masselon, Daniel Salvi, Norbert Rolland, et al.. Prediction of subplastidial localization of chloroplast proteins from spectral count data - Comparison of machine learning algorithms. RECOMB sat. conf. on proteomics 2012, 2012, San Diego La Joya, United States. hal-00683257

HAL Id: hal-00683257
https://hal.science/hal-00683257
Submitted on 28 Mar 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Prediction of sub plastidial 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 UMR1250, Université Joseph Fourier, F-38054, Grenoble, France.

Introduction

To study chloroplast metabolism and functions, sub plastidial 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 sub plastidial 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 main chloroplast sub-fractions (stroma, thylakoids and envelope) obtained by sucrose gradient purification. As the distribution of spectral counts over compartments is a fair major chloroplast sub-fractions ([1] fractionated chloroplasts. In particular, spectral count data were acquired for the three major chloroplast sub-fractions (stoma, thylakoids and envelope) obtained by sucrose gradient purification. As the distribution of spectral counts over compartments is a fair major chloroplast sub-fractions ([1] fractionated chloroplasts. In particular, spectral count data were acquired for the three major chloroplast sub-fractions (stoma, thylakoids and envelope) obtained by sucrose gradient purification. As the distribution of spectral counts over compartments is a fair

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 understood 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

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 reduncancy 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:

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Parameters</th>
<th>Performances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>--</td>
<td>84% (=)</td>
</tr>
<tr>
<td>SVM</td>
<td>σ=10, C=10</td>
<td>92.2% (1.2)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Ntree = 200, Deep=1</td>
<td>91.7% (1.4)</td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>k = 3</td>
<td>89.4% (1.3)</td>
</tr>
<tr>
<td>PerTurbo</td>
<td>σ = 0.25, λ=10^-5</td>
<td>91.9% (1.1)</td>
</tr>
</tbody>
</table>

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 set: 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 σ, λ and C. The kernel involved in SVM and PerTurbo was the Gaussian one.

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

References: