BaCelLo: a Balanced subCellular Localization predictor.

Andrea Pierleoni
University of Bologna - Dept. of Biology - Via Irnerio 42, 40126 Bologna, ITALY
Pier Luigi Martelli
University of Bologna - Dept. of Biology - Via Irnerio 42, 40126 Bologna, ITALY
Piero Fariselli
University of Bologna - Dept. of Biology - Via Irnerio 42, 40126 Bologna, ITALY
Rita Casadio
University of Bologna - Dept. of Biology - Via Irnerio 42, 40126 Bologna, ITALY

Method Article

Keywords: subcellular localization, SVM, bioinformatics, eukaryotic cell, protein sorting

Posted Date: March 13th, 2007

DOI: https://doi.org/10.1038/nprot.2007.165

Abstract

Introduction

Compartmentalization plays a major role in eukaryotic cells by making possible the fine regulation of complex biochemical pathways. Each protein needs the right biochemical context to operate, therefore the knowledge of the subcellular localization of a protein is essential in order to understand its functions and its pattern of interactions in protein networks. BaCelLo is a predictor for the subcellular localization of eukaryotic proteins and it is based on several Support Vector Machines \(SVMs) arranged in a decision tree \(Fig 1). Starting from the residue sequence, BaCelLo discriminates five different localizations: secretory pathway, cytoplasm, nucleus, mitochondrion and chloroplast. The predictor analyzes the protein residue sequence and its evolutionary profile considering information from the whole sequence and from its N- and C-terminal regions. Three different predictors are available for three different eukaryotic kingdoms: Metazoa, Viridiplantae and Fungi. The distinctive features of BaCelLo are: 1. a homology-reduced dataset for training and testing the predictor, in order to avoid redundancy. This dataset was compiled starting from the Swissprot data base \(release 48) and contains proteins whose subcellular localization was experimentally annotated. The dataset was reduced by similarity so that no protein in the dataset share more than 30% identity; 2. the implementation of three kingdom-specific predictors to take into account differences in subcellular localization mechanisms; 3. the evolutionary profile to extract evolutionary information from the residue sequence. 4. a hierarchic tree for the predictions; 5. the introduction of a unique balancing procedure in SVMs that corrects the biases between the different classes due to the disproportions in the training set . BaCelLo proved to outperform all the other state-of-art methods publicly available, when validated on a set of protein sequences independent of the training set[1].

Reagents

Sequences of the proteins to be predicted are required in FASTA format.

Equipment

1. A personal computer with a web browser program \(Internet Explorer 6 and upper, Firefox and Opera 8 and upper were tested and support the prediction server) 2. An internet connection

Procedure

How to predict the subcellular localization for a protein: 1. Go to "http://gpcr.biocomp.unibo.it/bacello/pred.htm":http://gpcr.biocomp.unibo.it/bacello/pred.htm 2. Select the kingdom of the organism expressing your protein\(s) \(choosing between Animals, Fungi or Plants). 3. Paste the sequences \(up to five sequences per time) in the corresponding field. 4. Submit the request and wait for results. How to read results: • The result page will be available for a maximum of 24h • In the

result page you will find, for each protein: a) the prediction of the subcellular localization b) the path along the decision tree \(Figure 1). As shown in table 1, the performance depends on the hierarchy of the tree.

Timing

Approximatively 30 seconds per protein sequence.

Troubleshooting

BaCelLo is able to assign a subcellular localization only for soluble proteins. For membrane proteins other prediction methods have to be considered BaCelLo needs the whole residue sequence of the protein; using a fragment can lead to mispredictions. For bug reports please contact us at: andrea@biocomp.unibo.it

Anticipated Results

A summary of BaCelLo performance for the three kingdoms is shown in Table 1, while in the original paper additional information can be found[1]. At the first level of the prediction tree, BaCelLo discriminates between extracellular and intracellular proteins with a rate of correct prediction that ranges from 91% to 96%, depending on the kingdom. At the second level, intracellular proteins are further discriminated between nucleocytoplasmic and organellar, so that three classes are separated with an overall accuracy ranging from 84% to 89%. At the third level nuclear proteins are discriminated from cytoplasmic ones, with a score of about 75% of correct assignments, when four classes are discriminated. Only in the case of plant protein there is another level that separates mitochondrial proteins from chloroplastic ones with an overall accuracy for five classes of as high as 66.6%.

References

1. A Pierleoni , PL Martelli, P Fariselli and R Casadio. BaCelLo: a Balanced subCellular Localization predictor. _Bioinformatics_ **22** e408-e416 \(2006).

Acknowledgements

RC acknowledges the receipt of the following grants: FIRB 2003 LIBI—International Laboratory of Bioinformatics and the support to the Bologna node of the Biosapiens Network of Excellence project within the European Union's VI Framework Programme \(contract number LSHG-CT-2003-503265). AP is supported by a FIRB 2003-LIBI grant.

Figures

Level	Classes	Plants			Animals			Fungi		
		Cov	*nAcc*	*nQ*	*Cov*	*nAcc*	*nQ*	*Cov*	*nAcc*	*nQ*
1	Secr	85.4	95.3	**90.6**	90.8	95.6	**93.3**	94.3	97.7	**96.0**
	Intr	85.8	86.7		95.8	91.2		97.7	94.5	
2	Nucl/Cyto	80.4	76.0	**84.4**	92.9	82.9	**86.6**	91.2	83.2	**89.0**
	Secr	85.4	89.8		90.8	85.0		94.3	92.8	
	Mito/Chlo	87.5	88.2		76.1	93.6		81.4	91.8	
3	Nucl	71.9	69.1	**74.1**	64.8	67.8	**74.2**	67.1	65.7	**75.8**
	Cyto	51.7	65.5		65.3	60.3		60.2	62.3	
	Secr	85.4	81.3		90.8	83.1		94.3	90.6	
	Mito/Chlo	87.5	78.1		76.1	87.4		81.4	83.8	
4	Nucl	71.9	66.8	**66.6**						
	Cyto	51.7	61.6							
	Secr	85.4	80.0							
	Mito	50.7	77.3							
	Chlo	73.0	53.6							

Figure 1

Table 1 Summary of BaCelLo performances over the three considered kingdoms. Performances were evaluated in a 10-fold cross-validation, so they are indicative of the performance that can be achieved with new sequences unrelated to the training dataset. Cov = Coverage: percentage of correctly predicted proteins of a class. nAcc = Normalized Accuracy: probability of correct predictions in a class. nQ = Normalized Overall Accuracy: estimates of total correct predictions where an equiprobability among the different classes is assumed. For theoretical details see the BaCelLo original paper^1^. (adapted from reference 1)

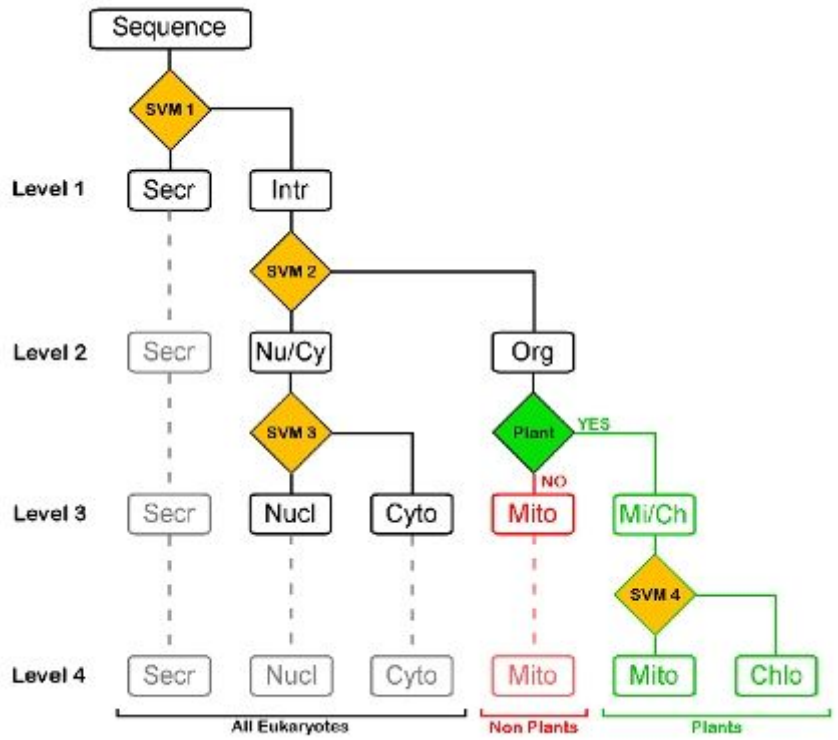

Figure 2

Figure 1 Architecture of BaCelLo decision tree. (originally published in reference 1)