Proper integration of feature subsets boosts GO subcellular localization predictions

Flavio E. Spetale[1], Elizabeth Tapia[1], Javier Murillo[1],
Flavia Krsticevic[1,2], Sergio Ponce[2], Laura Angelone[1], and Pilar Bulacio[1,2]

[1]*BioInformática,CIFASIS-Conicet-UNR Institute,Ocampo y Esmeralda, S2000EZP Rosario,Argentina.*
[2]*Facultad Regional San Nicolás-UTN, Colón 332, B2900LWH San Nicolás, Argentina.*

Abstract*— Prediction of multiple subcellular localizations in proteins brings relevant information for biological function discovery. The use of computational methods based on knowledge can be a helpful starting point for guiding the costly experimental validation. In this work, we present a multilabel classifier framework to perform Gene Ontology - Cellular Component prediction focused on the improvement of two design aspects: *i*) the protein sequence characterization, regarding biological knowledge with experimental evidence, and *ii*) the error evaluation by considering a noise model inherent in real prediction frameworks. Our proposal is validated against sets of well-known protein sequences of four model organisms *D. rerio*, *A. thaliana*, *S. cerevisiae* and *D. melanogaster

***Keywords*— Cellular Component, Prediction, Multilabel Classification.**

***Resumen*— La predicción de múltiples localizaciones subcelulares en proteínas brinda información relavante para el descubrimiento de funciones biológicas. El uso de métodos computacionales basados en el conocimiento puede ser un buen punto de partida para conducir a las costosas validaciones experimentales. En este trabajo, presentamos un framework de clasificación multi-etiqueta para para realizar la predicción en Gene Ontology - Componente Celular enfocada en la mejora de dos aspectos del diseño: *i*) la caracterización de la secuencia proteica, relacionando el conocimiento biológico con la evidencia experimental; y *ii*) la evaluación de errores al considerar un modelo de ruido inherente a los frameworks de predicción reales. Nuestra propuesta es validada contra un conjunto de secuencias de proteínas de cuatro organismos modelos *D. rerio*, *A. thaliana*, *S. cerevisiae* and *D. melanogaster*.**

***Palabras clave*— Clasificación multi-etiqueta, Machine Learning, Gene Ontology - Componente Celular.**

I. Introduction

Protein subcellular localization (SCL) is a key point to enrich the evidence information of biological function prediction for genome annotation [?]. Although the experimental testing is the quintessential method, it is usually an unaffordable burden in terms of time and money. Alternatively, computational methods arise as a first prediction approximation to guide and focus experimental validations. Under this baseline, machine learning classifiers can be a proper framework, where dataset design for learning and testing stages is a crucial point. Specifically, dataset design focus on the sample (protein) characterization through a set of features. Characterization strategies [?] evaluate different aspects of the protein sequence: peptide composition, structural, physico-chemical, sorting signal and biological knowledge, depending on the specific biological question related to SCL prediction to be answered.

After sequence characterization, a classification method is chosen to generate the predictive model. Initially, classification methods predicted just one SCL from a predefined -organism dependent- set of labels, e.g., 22 labels for eukaryotic [?]. More recent studies [?], [?] noted that some proteins may exist in more than one localization in a cell. Consequently, prediction approaches based on multi-label classification, as iloc-Euk [?] and Fuel-mLoc [?] approaches, would be more appropriate. One step further are Cello2go [?] and FFPred3 [?] methods, which besides predicting SCL in an organism dependent set of labels, do so over Gene Ontology (GO) Cellular Component (CC) terms, i.e., labels predefined in a hierarchical structured (GO-CC) independently of considered organisms. Briefly, Cello2go relies on BLAST for searching homologous proteins already GO-CC annotated, providing null results when homology with other organisms are poor, i.e., cutoff of e-value. FFPred3 is based on the integration of feature groups (14) which are characterized from protein sequences with a fixed number of GO-CC terms (104). FFPred's groups are generated following characterization strategies above-mentioned and the classification is done by binary SVM-Light classifiers, i.e., SVMs with outputs fitting by a sigmoid function. Note that the proposed characterization accomplish near to 300 features which are reduced through recursive feature elimination (SVM-RFE) by FFPred's groups. Finally, predicted GO-CC terms are those that exceed a user-defined threshold.

In this paper, we work on this later prediction strategy: inferring protein SCL directly over GO-CC terms, improving two aspects of the prediction system. The first one, is related to the characterization assuming that the boosting of biological knowledge k -with experimental evidence- hidden in the local-

Contact: Flavio E. Spetale, CIFASIS-Conicet-UNR Institute, Ocampo y Esmeralda, Phone +543414237248, S2000EZP Rosario, Argentina, *spetale@cifasis-conicet.gov.ar*

ization signals of 8 basic SCL, should improve the prediction recall through the robust guide of that key information in the complete inference process. The second improvement works on the uncertainty measures associated with the predictions. Specifically, we take into account the noise n present in every inference process in a similar way to Information Theory. We consider classifiers with real-value outputs which were achieved from an ideal binary outputs corrupted with additive white Gaussian noise.

The proposed prediction framework, hereafter called GO-CC$_{kn}$, is designed in three step: *i)* selection of features applying all characterization strategies; *ii)* designing of SVMs, one for each GO-CC term prediction; and *iii)* building the multilabel classifier system.

This paper is organized as follows. In Section **??**, multilabel classifier system is detailed . Section III discusses the results on *D. rerio*, *A. thaliana*, *S. cerevisiae* and *D. melanogaster* in CC-GO. In the last Section, conclusions are presented.

II. Materials and Methods

The GO-CC$_{kn}$ method consists of three steps: *i*) the design of datasets for learning and validation, i.e., the characterization of proteins in a set of features related to cellular component classes; *ii*) the design of a set of classifiers, i.e., one classifier for each considered CC-GO class (GO-term); and *iii*) the multiple classifier system for CC-GO prediction.

A. Feature Representation

The first step involves the characterization by a set of feature descriptors for protein sequences. The characterization may involve one or more of following issues: amino acid composition, physicochemical properties, secondary structure, sorting signals, and experimental localization information. More specifically, the amount of localization signals to 8 basic SCL in the LocSigDB database [**?**], sorting signals (SS) [**?**], [**?**], [**?**], coiled coils (CCoil) [**?**] and the measurement of 457 physicochemical/secondary structure properties (PC-SSP), 453 of the physicochemical type [**?**] including 20 amino acid compositions, 400 dipeptide compositions, 33 physicochemical properties and 4 of the secondary structure type [**?**], [**?**]. This characterization was divided in four characterization processes: LocSigDB, SS, CCoil, and PC-SSP. LocSigDB is a numerical vector; each element indicates the amount of experimental localization signals in 8 basic SCLs (Nucleus, Mitochondria, Secreted, Lysosome, Peroxisome, Golgi, Plasmatic Membrane, Endoplasmaic Reticulum). CCoil, SS and PC-SSP are real vectors, where elements indicate the sum of coil-forming probabilities in scanning windows of 14, 21 and 28 residues, the value of signals peptides, and the value of physicochemical/secondary structure property respectively. Practically, protein sequence characterization methods were implemented with R-cran.

B. SVM classifiers

The set of classifiers are designed with SVMs: one classifier for each considered GO-CC terms. The SVMs are set with single soft-margin radial: default constant complexity C=1 and gamma. In order to fulfill Gaussian assumption of prediction noise [**?**], real valued predictions are set to the margin of SVM classifier outputs. Practically, SVMs were implemented with e-1071 R package [**?**].

C. Multiple classifier system for GO-CC$_{kn}$ prediction

The complete pipeline for GO-CC prediction, shown in Fig. **??**, comprises m SVM classifiers for m GO-terms. Each GO-term ($GO - CC_i$) is represented by a binary classifiers SVM_i. For each query, a target protein is characterized by the set of features (input of predictor), and a set of predicted probabilities corresponding to certainties values of protein/GO-term association are returned by classifiers. To determinate predicted GO-terms (positive GO-terms), we consider SVMs with a level decision higher than a cut-off. In the paper, we consider probability value cut-off= 0.5.

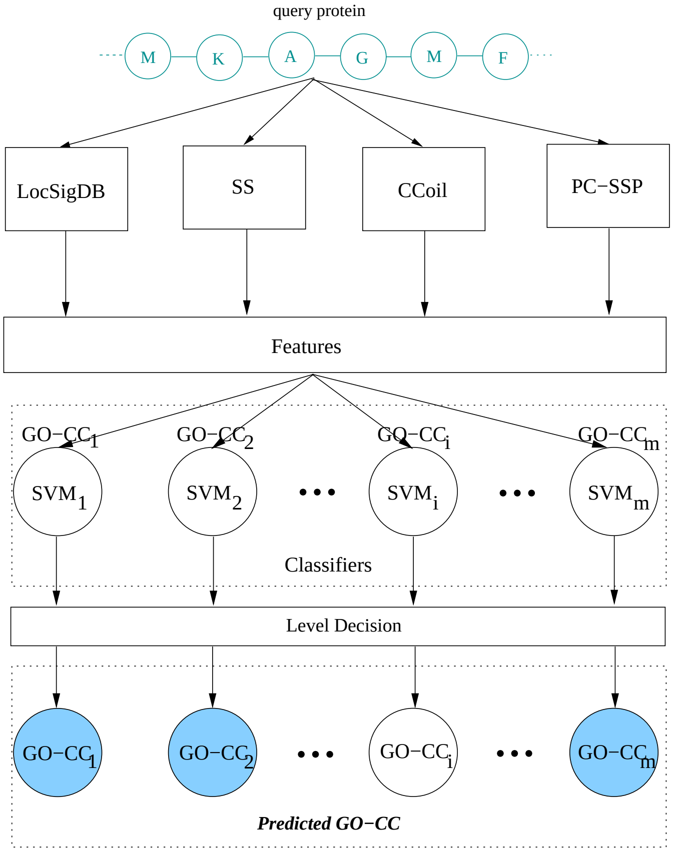

Fig. 1. Multiple classifier system for CC-GO prediction. LocSigDB: localization signals database, SS: sorting signals, CCoil: coiled coils, PC-SSP: physicochemical/secondary structure properties. Features represent the group of selected characterization to our model. Level decision represents the cut-off. In cyan, the predicted GO-CC.

D. Experimental protocol

Four models organisms, *D. rerio* [**?**], *A. thaliana* [**?**], *S. cerevisiae* [**?**] and *D. melanogaster* [**?**] were considered. For each of them, the annotation datasets were built from protein sequences considering experimental and computational analysis evidence codes[1] of GO. For experimental codes are considered: inferred from Experiment (EXP), inferred from Direct Assay (IDA), inferred from Physical Interaction (IPI), inferred from Mutant Phenotype (IMP), inferred from Genetic

[1] http://geneontology.org/page/guide-go-evidence-codes

TABLE I
DATASETS IN THE GO-CC

Organism	# GO-terms	# Samples
D. rerio	52	1243
A. thaliana	144	22788
D. melanogaster	165	6176
S. cerevisiae	174	5134

Interaction (IGI), inferred from Expression Pattern (IEP). For computational analysis, the following evidence codes are considered: inferred from Sequence or structural Similarity (ISS), inferred from Sequence Orthology (ISO), inferred from Sequence Alignment (ISA), inferred from Sequence Model (ISM). The considered GO subgraph was restricted to have leaves (GO-term) with a minimum of 50 positively annotated protein sequences. To assemble conveniently balanced binary training datasets [**?**], positive annotated protein sequences to individual GO-terms were complemented with negative annotated instances using the *inclusive* separation policy [**?**] (see Table **??**).

In order to evaluate which sets of features behaves better for GO-CC$_{kn}$ prediction, we evaluated the performance characterization groups individually and jointly in *D. rerio*.

The evaluation of the predictive performance of our approach was made with 5-fold cross-validation test. The values of recall, precision and F1 were calculated for the classifier.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

Where TP indicates the total numbers of true positives, FP indicates the total numbers of false positives and FN indicates the total numbers of false negatives.

III. RESULTS AND DISCUSSION

A. Comparing different group features

To select the set of features that improve the predictions, we compare precision/recall measures. able **??** shows the results of applying the four characterizations with the model organism *D. rerio* to SVM classifiers. The results show that four group of features together have the best performance. PC-SSP and SS may be associated with precision due to the information determining the subcellular localization site of a protein is encoded in its amino acid sequence and share similarity across certain physicochemical properties and sorting signals [**?**], [**?**]. On the other hand, LogSigDB may be associated with recall due to enclose the biological knowledge with experimental validation is related with specific functions [**?**]. So, their inclusion entails similar (or slight less) precision levels but with much higher recall levels, i.e., the system has less noise.

TABLE II
AVERAGE PRECISION (P), RECALL (R) OF THE OUR APPROACH IN THE GO-CC. THE SELECTED ORGANISM MODEL IS *D. rerio*. CHARACTERIZATIONS ARE LOCSIGDB DATABASE, SS: SORTING SIGNALS, CCOIL: COILED COILS AND PC-SSP: PHYSICOCHEMICAL/SECONDARY STRUCTURE PROPERTIES.

Characterization	P	R
LocSigDB	0.32	0.58
SS	0.36	0.69
CCoil	0.28	0.39
PC-SSP	0.41	0.67
SS + CCoil	0.37	0.67
SS + CCoil + LocSigDB	0.38	0.64
SS + PC-SSP	0.41	0.73
SS + CCoil + LocSigDB + PC-SSP	0.41	0.75

B. Prediction performances of GO-CC$_{kn}$

Our method is evaluated with feature descriptors selected in the previous stage, reporting the average measure of precision, recall and F1 for four model organisms. The results are shown in Table **??**.

TABLE III
AVERAGE PRECISION (P), AVERAGE RECALL (R) AND AVERAGE F_1 SCORE OF THE EACH METHOD IN THE GO-CC IS SHOWN.

Organism	P	R	F1
D. rerio	0.41	0.75	0.51
A. thaliana	0.35	0.79	0.39
D. melanogaster	0.45	0.72	0.52
S. cerevisiae	0.32	0.82	0.42

C. Comparison with other methods

Table **??** compares the performance of GO-CC$_{kn}$ method against Cello2go and FFPred3. The results show that our approach achieve higher recall and F1 values than Cello2go and FFPred3. On the other hand, Cello2go has the highest precision due to it is working with a model organism. Note that Well-known methods based on homology provide good results in model organism such as *D. rerio*. However, their performance falls when predictions are far from model organisms [**?**], e. g., *S. lycopersicum* (Solyc06g076520). These results motivate us to develop alternative computational methods to homology-based ones, i.e., methods based on machine learning techniques to infer a consistent ontological structure (GO).

TABLE IV
AVERAGE PRECISION (P), AVERAGE RECALL (R) AND AVERAGE F_1 SCORE OF THE EACH METHOD IN THE GO-CC IS SHOWN. THE SELECTED ORGANISM MODEL IS *D. rerio*.

Method	P	R	F1
GO-CC$_{kn}$	0.41	0.75	0.51
FFPred3	0.40	0.29	0.32
Cello2go	0.64	0.28	0.37

IV. CONCLUSION

The computational methods based on machine learning strategy improve their performance when features associated with physicochemical issues and secondary structure properties, which are both associated with subcellular localization. In addition, the inclusion of experimentally validated biological knowledge boost the recall of predictors, i.e., the model is more exhaustive. GO-CC$_{kn}$ can predict any protein sequence that exceeds 10 amino acids.

As further work, an interesting point is to study the consistency checking strategies related to the hierarchical structure of Gene Ontology.

Acknowledgment

The authors would like to acknowledge support received from the following projects: STIC-AmSud, "SNPs-PHE - Computational models for predicting SNPs-phenotype associations"; and INI 3600TC, "Anotación funcional GO basada en clasificación jerárquica supervisada", Universidad Tecnológica Nacional.

References

[1] L. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, and et al., "Prediction of human protein function from post-translational modifications and localization features," *Journal of Molecular Biology*, vol. 319, no. 5, pp. 1257–1265, 2002.

[2] L. Anna, S. M. B, O. C. A, and J. D. T, "Inferring function using patterns of native disorder in proteins," *PLOS Computational Biology*, vol. 3, no. 8, pp. 1–13, 2007.

[3] Chou Kuo-Chen and Shen Hong-Bin, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nat. Protocols*, vol. 3, no. 2, pp. 153–162, 2008.

[4] A. H. Millar, C. Carrie, B. Pogson, and J. Whelan, "Exploring the function-location nexus: Using multiple lines of evidence in defining the subcellular location of plant proteins," *The Plant Cell*, vol. 21, no. 6, pp. 1625–1631, 2009.

[5] R. F. Murphy, "Communicating subcellular distributions," *Cytometry Part A*, vol. 77A, no. 7, pp. 686–692, 2010.

[6] Kuo-Chen, W. Zhi-Cheng, and X. X. Chou, "iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins," *PLOS ONE*, vol. 6, no. 3, pp. 1–10, 2011.

[7] S. Wan, M.-W. Mak, and S.-Y. Kung, "Fuel-mloc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms," *Bioinformatics*, vol. 33, no. 5, p. 749, 2017.

[8] C.-S. Yu, C.-W. Cheng, W.-C. Su, K.-C. Chang, S.-W. Huang, J.-K. Hwang, and C.-H. Lu, "Cello2go: A web server for protein subcellular localization prediction with functional gene ontology annotation," *PLOS ONE*, vol. 9, no. 6, pp. 1–9, 2014.

[9] D. Cozzetto, F. Minneci, H. Currant, and D. T. Jones, "Ffpred 3: feature-based function prediction for all gene ontology domains," *Scientific reports*, vol. 6, p. 31865, 2016.

[10] N. Simarjeet, P. Sanjit, S. S. M, M. Akram, and G. Chittibabu, "LocSigDB: a database of protein localization signals," *Database: The Journal of Biological Databases and Curation*, vol. 2015, p. bav003, 2015.

[11] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 4.0: discriminating signal peptides from transmembrane regions," *Nat Meth*, vol. 8, no. 10, pp. 785–786, 2011.

[12] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence," *Journal of Molecular Biology*, vol. 300, no. 4, pp. 1005–1016, 2000.

[13] H. Paul, P. Keun-Joon, O. Takeshi, F. Naoya, H. Hajime, A.-C. CJ, and N. Kenta, "WoLF PSORT: protein localization predictor," *Nucleic Acids Research*, vol. 35, no. Web Server issue, pp. W585–W587, 2007.

[14] A. Lupas, M. Van Dyke, and J. Stock, "Predicting coiled coils from protein sequences," *Science*, vol. 252, no. 5009, pp. 1162–1164, 1991.

[15] B. Lee, M. Shin, Y. Oh, H. Oh, and K. Ryu, "Identification of protein functions using a machine-learning approach based on sequence-derived properties," *Proteome Science*, vol. 7, no. 1, p. 27, 2009.

[16] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," *Biochemistry*, vol. 13, no. 2, pp. 222–245, 1974.

[17] ——, "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins," *Biochemistry*, vol. 13, no. 2, pp. 211–222, 1974.

[18] E. Tapia, P. Bulacio, and L. Angelone, "Sparse and stable gene selection with consensus svm-rfe," *Pattern Recognition Letters*, vol. 33, no. 2, pp. 164–172, 2012.

[19] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. (2014) Misc functions of the department of statistics (e1071), tu wien. Version: 1.6-4, Accessed: 2015-09-02. [Online]. Available: http://cran.r-project.org/web/packages/e1071/index.html

[20] M. Carlson. (2016) Genome wide annotation for zebrafish. Version: 3.2.3, Accessed: 2017-04-06. [Online]. Available: http://bioconductor.org/packages/org.Dr.eg.db/

[21] ——. (2016) Genome wide annotation for arabidopsis. Version: 3.2.3, Accessed: 2017-04-06. [Online]. Available: http://bioconductor.org/packages/org.At.tair.db/

[22] ——. (2016) Genome wide annotation for yeast. Version: 3.2.3, Accessed: 2017-04-06. [Online]. Available: http://bioconductor.org/packages/org.Sc.sgd.db/

[23] ——. (2016) Genome wide annotation for fly. Version: 3.2.3, Accessed: 2017-04-06. [Online]. Available: http://bioconductor.org/packages/org.Dm.eg.db

[24] Q. Wei and R. L. Dunbrack, "The role of balanced training and testing data sets for binary classifiers in bioinformatics." *PloS one*, vol. 8, no. 7, 2013.

[25] R. Eisner, B. Poulin, D. Szafron, P. Lu, and R. Greiner, "Improving protein function prediction using the hierarchical structure of the gene ontology," in *Proc. IEEE CIBCB*, 2005, pp. 1–10.

[26] P. Bork, *Analysis of Amino Acid Sequences*, ser. Advances in protein chemistry. Academic Press, 2000.

[27] D. Sarda, G. H. Chua, K.-B. Li, and A. Krishnan, "pslip: Svm based protein subcellular localization prediction using multiple physicochemical properties," *BMC Bioinformatics*, vol. 6, no. 1, p. 152, 2005.

[28] Anton Brian P, Kasif Simon, Roberts Richard J, and Steffen Martin, "Objective: biochemical function," *Frontiers in Genetics*, vol. 5, p. 210, 2014.

[29] W. R. Pearson, *An Introduction to Sequence Similarity ("Homology") Searching*. John Wiley & Sons, Inc., 2002.

Flavio E. Spetale is an Electronic Engineer and PhD in Computer Science. His areas of interest are the automatic inference in ontologies about biological data and the processing of spectroscopic signals. He is a Postdoctoral of the Cifasis-UNR Institute, integrating the BioAgroInformática group and a professor in Computer Science.

Elizabeth Tapia is an Electronic Engineer and PhD in Telematic Systems. Her areas of interest are multi-label classifiers for the processing of biological data and DNA barcodes based on error-correcting codes. She is a researcher at the Cifasis-UNR Institute, managing the BioAgroInformática group, and a tenured professor in the Electronic Engineering career.

Javier Murillo has a Bachelor's Degree in Computer Science and PhD in Computer Science. His areas of interest are fuzzy measures, information integration and aggregation operators. He is a researcher at the Cifasis-UNR Institute, integrating the BioAgroInformática group and a professor in Computer Science.

Pilar Bulacio is an Electronic Engineer and PhD in Telematic Systems. Her areas of interest are the selection of variables through measures of sets for the processing of biological and spectroscopic data. She is a researcher at the Cifasis-UNR Institute, integrating the BioAgroInformática group and a professor in Computer Science.