Predicting Sub-cellular Location of Proteins Based on Hierarchical Clustering and Hidden Markov Models

Jorge Alberto Jaramillo-Garzón[1,2], Jacobo Castro-Ceballos[1], and Germán Castellanos-Dominguez[1]

[1] Universidad Nacional de Colombia, Sede Manizales, Colombia
[2] Institute Tecnológico Metropolitano, Medellín, Colombia
{jcastroc,cgcastellanosd}@unal.edu.co,jorgejaramillo@itm.edu.co

Abstract. Sub-cellular localization prediction is an important step for inferring protein functions. Several strategies have been developed in the recent years to solve this problem, from alignment-based solutions to feature-based solutions. However, under some identity thesholds, these kind of approaches fail to detect homologous sequences, achieving predictions with low specificity and sensitivity. Here, a novel methodology is proposed for classifying proteins with low identity levels. This approach implements a simple, yet powerful assumption that employs hierarchical clustering and hidden Markov models, obtaining high performance on the prediction of four different sub-cellular localizations.

1 Introduction

The information derived from sequenced genomes has grown with an exponential behaviour, and the number of protein sequences with missing annotation increases rapidly. Therefore, the functional annotation of proteins has become a theme of great importance in molecular biology. Nonetheless, this task posses a big challenge due to the huge amount of available data.

The localization of a given protein can indicate how and what kind of cellular environments the proteins interact, and thus, it can help to elucidate its function [1]. A commonly used experimental method for determining the localization of a given protein is to fuse the sequence encoding a green fluorescent protein (GFP) to one end of the gene sequence for the query protein, and then use its intrinsic fluorescence to monitor where the protein is in the cell [2]. However, as this approach must be focused on specific proteins, it turns very expensive and time consuming, especially when considering the current size of unannotated protein sequences [3].

Several computational predictors of protein sub-cellular localizations have been proposed in the past few years. The most common methods among biologists are the alignment-based methods, which consist on searching query proteins against public databases of annotated proteins by using local alignment search tools such as BLAST or PSI-BLAST [4]. These methods, however, tend to attain low sensitivity for databases of proteins with low identity levels, due to the inability of the method for identifying homologous proteins at significant E-values [5].

A second category of methods are based on machine learning strategies. In this kind of methods, a set of numerical features from protein sequences is computed and a classifier is trained to label query protein sequences according to one of the classes from the

F. Ortuño and I. Rojas (Eds.): IWBBIO 2015, Part II, LNCS 9044, pp. 256–263, 2015.

training dataset. Among the classifiers used in this kind of methods are support vector machines [6–8, 3], neural networks [9], or ensembles of multiple specialized classifiers [10–13].

A more biologically driven alternative are the subsequence-based methods, which explore the fact that the functionality of proteins is mainly due to functional domains that may reside in different portions of the proteins. These methods employ stochastic models for describing protein families. Large collections of protein families and domains can be found in public databases [14] and methods based on Hidden Markov Models (HMM) can efficiently represent family profiles [15]. However, both machine learning based methods and subsequece-based methods have shown low sensitivities for categories with high diversity among its samples, as it is the case for several sub-cellular localizations. Since these methods try to represent the whole category by a single trained model, potentially useful information is necessarily discarded and a big amount of false-negatives appear.

Considering those precedents, this work proposes a subsequence-based methodology, that aims to improve the prediction of sub-cellular localizations with low identity levels, by using HMM models. The proposed methodology assumes that there is not a single cluster of samples belonging to a given category and that each cluster may have its own distinctive profiles. The proposed methodology is compared with two other subsequence-based methods in the literature PfamFeat [15] and Plant-mPloc [16] for the prediction of four sub-cellular localizations in a set of *Embryophyta* proteins. The results show that implementing this simple, yet powerful assumption, the classification model is enhaced and the sensitivity of the system rises, thus increasing the overall performance of the system.

2 Background

The proposed methodology involves two software packages that have been extensively used in the literature. First sequences are clustered together with the CD-HIT software package [17] and then, each cluster is modelled into an HMM profile using the HMMer software [18]. These packages will be described in the present section.

Hierarchical Clustering of Protein Sequences Using CD-HIT: CD-HIT [17] uses a method based on greedy incremental clustering for detecting clusters of similar sequences in the data. Briefly, sequences are first sorted in order of decreasing length and the longest one becomes the representative of the first cluster. Then, each remaining sequence is compared to the representatives of existing clusters. If the similarity with any representative is above a given threshold, it is grouped into that cluster. Otherwise, a new cluster is defined with that sequence as the representative. The process continues until all sequences have been assigned to a cluster.

This algorithm has been extensively used for a large variety of applications ranging from non-redundant dataset creation [19], protein family classifications [20], metagenomics annotation [21], among others.

HMM-Based Modelling of Sequence Clusters Using HMMER: HMMs are stochastic model, which assume the system can be modelled as a Markov process, with unknown

parameters. In discrete cases, these parameters are represented by a set of Q states $\boldsymbol{\theta}\in\mathbb{R}^Q$ that has to be computed by an underlying optimization process. Each state is associated to one of K possible observation values. The model is then composed by three parameters: an initial state probability $\boldsymbol{\pi}$ with elements $\{\pi(\theta_i)\in\mathbb{R}[0,1]\}$, that describes the distribution over the initial state set; a transition matrix $\boldsymbol{A}\in\mathbb{R}^{Q\times Q}$ where each $a_{ij}\in\mathbb{R}^+, i,j\in[1,Q]$ represents the transition probability from state i to state j; and an observation matrix $\boldsymbol{B}\in\mathbb{R}^{Q\times K}$ with the elements $\{b_{i,k}\in[0,1]\}$ representing the probability of each observed symbol $k\in[1,K]$, given that the system remains at state i [22].

The HMMER software package [18] uses the (MSV) score for target sequence $\boldsymbol{x}$. it is a log likelihood ratio score of singles optimal (maximally likely) alignment: the ratio of the probability of the optimal alignment $\boldsymbol{\Omega}$ for $\boldsymbol{x}$ given the MSV model $\mathcal{M}$ and the probability of the sequence given a null hypothesis model $\mathcal{R}$:

$$S^{MSV}(\boldsymbol{x}) = \log_2 \frac{Prob(\boldsymbol{x},\boldsymbol{\Omega}|\mathcal{M})}{Prob(\boldsymbol{x}|\mathcal{R})} \tag{1}$$

For a query of length m positions, the MSV Profile has Km match emission parameters (where K is the alphabet size, 4 nucleotide of 20 amino acids), plus $m+8$ additional state transition parameters involving the flanking and N, B, E, C and J states that account for non-homologous residues. Other state transitions in the original profile are ignored, which means implicitly in the original profile are ignored, which means implicitly treating match-match transitions as 1.0.

The null model R is assumed to be an HMM with a single state R emitting residues a with background frequencies $f(a)$ (i.e. a standard i.i.d null model: independent, identically distributed residues), with a geometric length distribution specified by a transition parameter t_{RR}.

The mK positions-specific match scores $\sigma_k(a)$ are precomputed as log-odss ratios for a residues a emitted from match state M_k with emission probability $e_k(a)$, compared to the null model background frequencies f_a:

$$\sigma_k(a) = \log_2 = e_k(a)/f_a \tag{2}$$

These match scores (as well as the emission probabilities and background frequencies) are the same as in the original profile. The only state transition parameters in the MSV model are those that control target sequences length modelling, the uniform local alignment fragment length distribution, and the number of hits to the core homology model per target sequence [23]. These too are identical to the parametrization of the original profile.

3 Experimental Set-Up

The workflow of the experimental set-up has five main components: the input *database* that is labeled with the corresponding sub-cellular localizations; the *preprocessing* stage, where sequences are grouped into clusters with the CD-HIT software; the alignment stage, where a sequence alignment is obtained from the sequences in each cluster using

the Clustal Omega software package [24]; and the *HMM construction* stage, where the HMMER software package is used to generate profiles from each alignment. Finally a statistical *validation* is performed in order to test the performance of the designed predictor and obtain performance measures. Figure 1 shows the experimental set-up workflow.

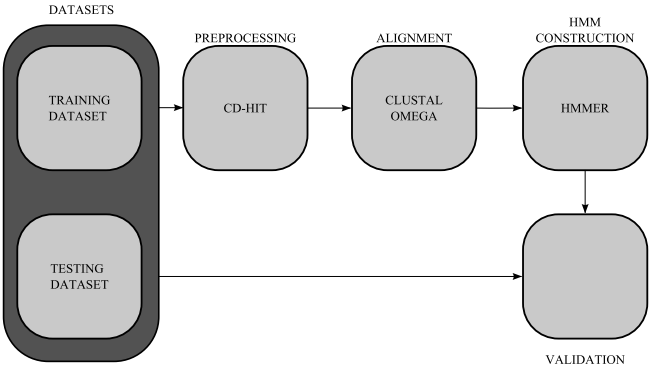

Fig. 1. Scheme of the baseline classifiers

3.1 Database

The dataset used in this work is a subset of the database published in [19]. It is conformed from all the available *Embryophyta* proteins at UniProtKB/Swiss-Prot database ([25] file version: 10/01/2013), with at least one annotation in the Cellular Component Ontology according to *Gene Ontology Annotation* (GOA) project ([26], file version: 7/01/2013). Proteins with unknown evidence of existence or resulting from computational predictions were discarded.

This dataset is composed by 2643 proteins, associated to four Gene Ontology terms, as shown in Table Table 1.

Table 1. Number of protein sequences associated to each sub-cellular localization

Localization	GO ID	Samples
Cytoplasm	GO_0005737	572
Cell Wall	GO_0005618	412
Extreacellular Region	GO_000556	374
Membrane	GO_0016020	1285

Preprocessing: In order to identify clusters of sequences with similar primary structures, the CD-HIT software package is used. The software searches for sequences with identities under a predefined cut-off. This identity cut-off was set at 30% in order to test

the accuracy of the proposed methodology at a low identity level. CD-HIT performs a hierarchical clustering and retrieves the clustered proteins sequences. Then, a multiple sequence alignment is performed with the Clustal Omega software package in order to obtain an alignment model for being used as input of HMMER.

HMM Training: With the aligned protein sequences grouped into clusters, the HMMs can be generated from the alignments. The obtained profiles are associated with the GO terms which in turn are associated with the sequences from which the models came, in order to to find the relationship between them, and probability of the dataset belonging to any model.

Validation: Three performance measures are used to analyse the generalization capability of the predictor: sensitivity (s_n) describes the capacity of the algorithm to recognize as positives the sequences that are indeed associated to a given sub-cellular component; specificiy (s_p) describes how the algorithm is able to reject sequences that are not associated to the sub-cellular component; and the geometric mean (g_m) between those measures as a global performance measure.

$$s_n = \frac{n_{TP}}{n_{TP} + n_{FN}} \quad s_p = \frac{n_{TN}}{n_{TN} + n_{FP}}$$

$$g_m = \sqrt{\frac{n_{TP} n_{TN}}{(n_{TP} + n_{FN})(n_{TN} + n_{FP})}}$$

where n_{TP}, n_{TN}, n_{FP} and n_{FN} are true positive, true negative, false positive, and false negative, respectively.

4 Results and Discussion

Figure 2 show the results obtained for the four sub-cellular components. It is important to note how *Cytoplasm* and *Membrane*, which are subcellular localizations containing protein sequences with a high variety, reached very high performance results. In contrast, *Extracellular region* and *Cell Wall*, which are more specific categiories, attained high specificities but remained with poor sensitivities.

In order to compare the results of the proposed methodology against other proposed strategies, Table 2 contrasts the results with the ones reported by the PfamFeat algorithm [15]. Also, since the database used for the experiments contains only *Embryophyta* (land plants) proteins, the Plant-mPloc [16] server was also tested for comparison purposes.

As can be observed on Table 2, the proposed methodology outperforms the results of PfamFeat on three out of four sub-cellular localizations, and the results from Plant-mPloc in all cases. The most important result relies on the fact that the specificity of the system was significantively improved for all the subcellular localizations, while keeping an acceptable sensitivity. Although Plamnt-mPloc obtained the highest values of sensitivity for three localizations, it also obtained the lowest specificities, thus achieving poor global performances.

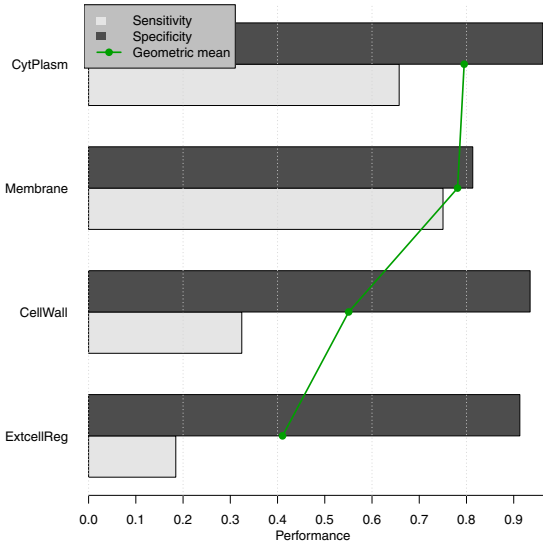

Fig. 2. Main results obtained for each sub-cellular localization

Table 2. Comparison of methods

Class	HMM			Pfam			Plant-mPloc		
	$s_p\%$	$s_n\%$	$g_m\%$	$s_p\%$	$s_n\%$	$g_m\%$	$s_p\%$	$s_n\%$	$g_m\%$
Cytoplasm	**96.15**	**65.75**	**79.51**	72.5	37.7	52.28	1.43	41.70	7.74
Cell Wall	**93.47**	32.43	**55.06**	88.01	19.7	41.63	2.32	**88.08**	14.31
Extcell. Region	**91.30**	18.46	41.05	78.5	40.7	**56.52**	1.25	**77.73**	9.88
Membrane	**81.33**	75.03	**78.10**	72.2	37.1	51.75	0.00	**77.30**	0.00

This results demonstrate that the inclusion of the hierarchical clustering stage, in conjunction with the HMM profile representations, provides a methodology that efficiently describes diversity of protein sequences associated to each localization, even when sequences have low identity levels among them.

5 Conclusions and Future Work

This paper presented a novel methodology for the prediction of sub-cellular localization of proteins. It proved to have better overall performances than other similar methods recently proposed in the literature, enhacing the sensitivity of the predictor for categories with high diversity. As a future work, it is important to test the methodology with the other GO ontologies (Molecular Function and Biological Process). Also it would be interesting to test several other clustering algorithms and semi-supervised strategies in order to improve even further the classification performances.

Acknowledgment. This work has been supported by the research project "*Metodología de clasificación multi-etiqueta mediante técnicas de optimización aplicado a la clasificación funcional de proteínas*" (*"Jóvenes Investigadores e Innovadores 2012"* program by COLCIENCIAS), and 20101007497 Universidad Nacional de Colombia.

References

1. Chou, K.-C., Shen, H.-B.: Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3(2), 153–162 (2008)
2. Baldi, P., Brunak, S.: Bioinformatics: the machine learning approach. The MIT Press (2001)
3. Jaramillo-Garzón, J., Perera-Lluna, A., Castellanos-Domiínguez, C.: Predictability of protein subcellular locations by pattern recognition techniques. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5512–5515. IEEE (2010)
4. Conesa, A., Götz, S.: Blast2go: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics 2008 (2008)
5. Hawkins, T., Chitale, M., Luban, S., Kihara, D.: PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 74(3), 566–582 (2009)
6. Yu, C., Lin, C., Hwang, J.: Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science 13(5), 1402–1406 (2004)
7. Shi, J., Zhang, S., Pan, Q., Cheng, Y., Xie, J.: Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33(1), 69–74 (2007)
8. Nanni, L., Lumini, A.: An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence. Amino Acids 35(3), 573–580 (2008)
9. Ma, J., Liu, W., Gu, H.: Predicting protein subcellular locations for Gram-negative bacteria using neural networks ensemble. In: Proceedings of the 6th Annual IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, pp. 114–120. The Institute of Electrical and Electronics Engineers Inc. (2009)
10. Shen, Y., Burger, G.: 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. BMC Bioinformatics 8(1), 420 (2007)
11. Shen, H., Yang, J., Chou, K.: Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33(1), 57–67 (2007)
12. Niu, B., Jin, Y., Feng, K., Lu, W., Cai, Y., Li, G.: Using adaboost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Molecular Diversity 12(1), 41–45 (2008)
13. Khan, A., Majid, A., Choi, T.: Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers. Amino Acids 38(1), 347–350 (2010)
14. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al.: The pfam protein families database. Nucleic Acids Research 40(D1), D290–D301 (2012)
15. Arango-Argoty, G., Ruiz-Munoz, J., Jaramillo-Garzon, J., Castellanos-Dominguez, C.: An adaptation of pfam profiles to predict protein sub-cellular localization in gram positive bacteria. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5554–5557. IEEE (2012)

16. Chou, K.-C., Shen, H.-B.: Plant-mploc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PloS One 5(6), e11335 (2010)
17. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23), 3150–3152 (2012)
18. Finn, R.D., Clements, J., Eddy, S.R.: Hmmer web server: interactive sequence similarity searching. Nucleic Acids Research 39(suppl. 2), W29–W37 (2011)
19. Jaramillo-Garzón, J.A., Gallardo-Chacón, J.J., Castellanos-Domínguez, C.G., Perera-Lluna, A.: Predictability of gene ontology slim-terms from primary structure information in embryophyta plant proteins. BMC Bioinformatics 14(1), 68 (2013)
20. Yooseph, S., Li, W., Sutton, G.: Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. BMC Bioinformatics 9(1), 182 (2008)
21. Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E.E., Ellisman, M., Grethe, J., et al.: Community cyberinfrastructure for advanced microbial ecology research and analysis: the camera resource. Nucleic Acids Research 39(suppl. 1), D546–D551 (2011)
22. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
23. Freyhult, E.K., Bollback, J.P., Gardner, P.P.: Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding rna. Genome Research 17(1), 117–125 (2007)
24. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Molecular Systems Biology 7(1) (2011)
25. Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B., Martin, M., McGarvey, P., Gasteiger, E.: Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10(1), 136 (2009)
26. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., Apweiler, R.: The goa database in 2009-an integrated gene ontology annotation resource. Nucleic Acids Research 37(suppl. 1), D396–D403 (2009)