34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

An adaptation of Pfam profiles to predict protein sub-cellular localization in Gram positive bacteria

G. A. Arango-Argoty[1], J. F. Ruiz-Muñoz[1], J. A. Jaramillo-Garzón[1,2], and C. G. Castellanos-Domínguez[1]

Abstract— **Predicting the sub-cellular localization of a protein can provide useful information to uncover its molecular functions. In this sense, numerous prediction techniques have been developed, which usually have been focused on global information of the protein or sequence alignments. However, several studies have shown that the functional nature of proteins is ruled by conserved sub-sequence patterns known as *domains*. In this paper, an alternative methodology (*PfamFeat*) for gram-positive bacterial sub-cellular localization was developed. *PfamFeat* is based on information provided by *Pfam* database, which stores a series of HMM-profiles describing common protein domains. The likelihood of a sequence, to be generated by a given HMM-profile, can be used to characterize sequences in order to use pattern recognition techniques. Success rates obtained with a simple one-nearest neighbor classifier demonstrate that this method is competitive with popular sub-cellular prediction algorithms and it constitutes a promising research trend.**

I. INTRODUCTION

Prediction of sub-cellular localizations of proteins is one of the greatest issues in bioinformatics. When protein localization is known, it is possible to get information about its role in the cell and also it can help in the design of protein isolation experiments, and in the identification of contaminants in proteomic analysis [1]. In particular, computational predictions for bacterial sequence proteins can be useful for searching novel vaccines and drug targets [2].

Predictors based on machine learning and pattern recognition techniques offer the possibility to analyze massive data and therefore, they have became increasingly used in the last years; e.g., *CELLO* [3] is a multi-class classification system based on support vector machines (SVM) that maps each protein sequence into a feature space by using n-peptide compositions as features. *SubcellPredict* [4] also characterizes proteins by amino acid compositions, but it uses AdaBoost algorithm instead of SVMs to predict cytoplasmic, periplasmic and extracellular localizations sites in prokaryotic and eukariotyc organisms. *Gpos-mPloc* is part of *Cell-PLoc 2.0* [5], a package of web servers designed for prediction of cellular localizations in different organisms. In particular, *Gpos-mPloc* predicts subcellular localization of gram-positive bacterial proteins by using a top-down approach based on BLAST alignments. *Psortb* [6], in turn, is a collection of different modules to decide the sub-cellular localization of a query protein. Although those methods have shown certain success on predicting sub-cellular localizations on bacteria, BLAST alignments used by *Gpos-mPloc* are prone to fail on identifying homologous proteins at significant E-values [7]. Characterizations of *CELLO* and *SubcellPredict* use global properties of the sequences. However, the functional nature of proteins is given by conserved regions known as protein domains [8]. Information of protein domains is commonly stored in databases like *Pfam* [8], a large collection of 13672 conserved protein domains and families, where, each one is represented by a Hidden Markov Model (HMM) [9]. Thus, these profiles should be used in order to identify sub-cellular localizations. This paper presents a methodology for sub-cellular localization prediction on gram-positve bacterial proteins by means of features corresponding to likelihood scores from Hidden Markov Models (HMM); latter belong to representative profiles in *Pfam*. Results were compared with *GPos-mPloc* [10] and *CELLO* version 2.5 [3], and these ones showed that *PfamFeat* can reach similar performances even when it was used a low complexity classifier such as the rule of the nearest neighbor.

[1] Signal Processing and Recognition Group, Universidad Nacional de Colombia, s. Manizales, Campus La Nubia, km 7 vía al Magdalena, Colombia. {gaarangoa, jfruizmu, jajaramillog, cgcastellanosd}@unal.edu.co

[2] The research center of the Instituto Tecnológico Metropolitano, Calle 73 No 76A-354, Medellín, Colombia. {jorgejaramillo }@itm.edu.co

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

II. Materials and Methods

The present methodology is based on constructing a feature space in which proteins are represented by n-dimensional vectors of likelihood scores. The number of components n, corresponding to the dimensions of the feature space is determined by the number of HMM profiles retrieved by *Pfam* for the training dataset. This feature space can thus be used to induce a decision rule through conventional pattern recognition classifiers and predict sub-cellular localization sites based on the distribution of the training data.

A. Database

Proteins for the training set were chosen according to the data set described in [11]. Such data set is comprised by 1206 proteins: 62 from cell wall, 500 cytoplasmic proteins, 500 from the cytoplasmic membrane and 144 extracellular proteins. On the other hand, the data set reported by *Gpos-mPLoc* [10] was used to test the method. This data set contains 523 sequences distributed as follows: 18 to the cell wall, 208 to the cytoplasm, 174 belonging to cytoplasmic membrane and 123 extracellular proteins. In order to eliminate train-test redundancy, an 80% identity cutoff between train and test sets was done with CD-HIT software [12].

B. Local feature descriptor

The procedure to map protein sequences into the feature space is described as follows:

Step 1. The training protein sequences are matched by each sub-cellular localization on the *Pfam sequence search* server [13]. Then a list of related HMM profiles $P = [k_C, k_{CM}, k_P, k_E]$ are identified per localization.

Step 2. The HMMER software package is used to compute the profile-sequence pairwise distance. The reported average posterior probability by HMMER was used for this purpose. Essentially, this measure is the expected alignment accuracy and decides if the alignment is well-determined or not [9]. Thus, the query protein $\mathbf{Q}$ can be formulated in the local feature space P as:

$$\mathbf{Q}_P = [dk_{(SCL,1)} ... dk_{(SCL,l)} ... dk_{(SCL,m)}] \qquad (1)$$

where, $dk_{(SCL,i)}$ is the dissimilarity between the sub-cellular localization domain $SCL\ i$ (i.e., cytoplasmic C) and the query sequence $\mathbf{Q}$.

C. Classification process

Prediction was carried out using the 1-nearest-neighbor classifier (1-nn) following the one-against-all strategy. This strategy produces a strong class imbalance, so, the Synthetic Minority Over-sampling Technique (SMOTE) was employed [14]. Moreover, the 1-nn was trained using all pairwise-distances among the train data set and the selected profiles *HMM* from *Pfam*. Thus, in order to remove redundant local features, the *Fast Correlation-Based Filter* algorithm ([15]) was used.

III. Results and discussion

A total of 1507 HMM profiles were found to be correlated to the training dataset. For cytoplasmic proteins there were detected a total of 780 *Pfam* terms, of whom, 72.69% were domains, 27% families, 0% motifs and 0.3% repeats. In cytoplasmic membrane localization a total of 580 terms were found as follows: 34.51% domains, 64.4% families, 0.6% motifs and 0.4% repeats. It is worth to note that there are more domains for cytoplasm than for cytoplasmic membrane. This could be due to the fact that bacterial proteins can either remain in the cytoplasm or these ones can be targeted to one or more sites through one of several different transport sistems, i.e., type I, which carries proteins into the extracellular space directly from the cytoplasm [16]; type II, which involves insertion into, or translocation across, the cytoplasmic membrane [17]; type III and type IV secretion systems, which directly inject products into the cytoplasm of a neighboring cell [18] and type V which self-transports a passenger *domain* using a C-terminal pore domain [19]. Cytoplasmic membrane is important in cell communication, it is contained from a variety of families such as transporters participating in the secretion of proteins, complex carbohydrates, and lipids into and beyond the cytoplasmic membrane ([20],[21]). Thus, the results of mapping cytoplasmic membrane proteins into pfam shown a high concentration of families than other kind of sub-structure. Domains were the most frequents on extracellular and cell wall proteins with a 57.7% and a 39% respectively.

The cumulative distribution among domains and localization allows to identify frequent probabilities (Figure 2). The *Pfam* terms were found on training data set and then, these ones were mapped into test sequences. Then, the distribution above test sub-cellular localizations are from 0.6 to 0.9 probability interval. So, these distributions demonstrates that the Pfam terms

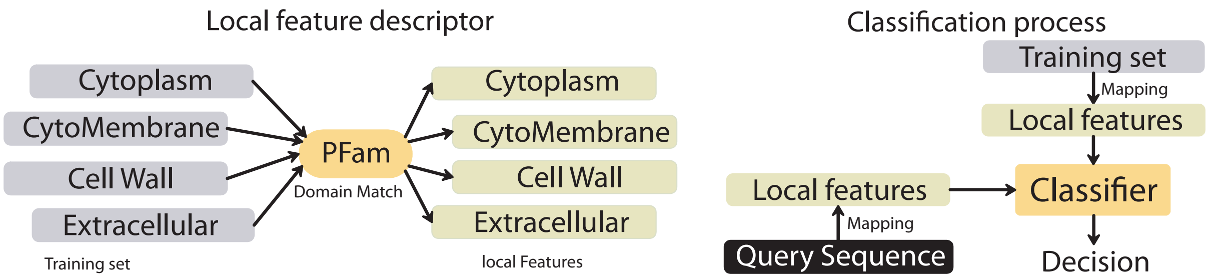

Fig. 1. Feature descriptor based on Pfam terms. The training data set was uploaded to *Pfam* data base to find correlated domains to each sub-cellular localization. These domains are interpreted as *Local Features* and the training data set is mapped into a feature space constructed by the HMM profile-sequence pairwise score. Finally, a *1-NN* classifier was trained to predict sub-cellular localizations.

are a good representation for both, training and testing sequences.

It was determined by a series of trial and error tests that those profiles related with seven or less protein sequences could be discarded without severely affecting the performance of the system. Thus, a set of 323 profiles was finally chosen: 99 correlated to cytoplasm, 86 to cytoplasmic membrane, 49 to cell wall and 98 belonging to extracellular localization. This way, testing dataset comprised a feature matrix of 523 elements with 323 features. In order to validate the procedure, a ten-fold cross validation was carried out and the success rate metric, defined as $TP/(TP+FN)$ (TP: true positives; FN: false negatives), was used to evaluate the performance.

The following approaches were used to compare subcellular localization prediction: **1)** *GPos-mPloc* reports an average success rate of 82.2% whereas the *PfamFeat* proposed method showed an average of 83.5%. **2)** CELLO version 2.5 [3] showed an average success rate of 76.6%, it is about 7% smaller than *PfamFeat*. Predictions of the whole localizations are shown in Table I.

TABLE I
SUCCESS RATES OF THE PROPOSED METHODOLOGY AND CELLO

Subcellular location	Success rate (%)	
	PfamFeat	CELLO
Cell membrane	152/174 = 87.4	151/174 = 86.8
Cell wall	12/18 = 66.7	7/18 = 38.9
Cytoplasm	190/208 = 91.3	200/208 = 96.2
Extracell	109/123 = 88.6	104/123 = 84.5
Overall	**463/523 = 83.5**	**462/523 = 76.6**

IV. Conclusions

In this study, a methodology for prediction of subcellular localizations based on sequences representation according to *HMM*-profiles from *Pfam*, was proposed. Experiments showed that this type of features are discriminant for the classification problem of gram-positive sub-cellular localizations. Performances obtained with *PfamFeat*, where it was used a simple *1-nnc* classifier, were comparable with those obtained by CELLO, even when this last one uses a highly complex classifier such as a support vector machine. In particular, a remarkable result was the obtained for the *cell wall* location, where *PfamFeat* achieved a considerable higher performance in comparison with CELLO. This is possibly due to the fact that global features are not as relevant as local features for this particular class, but further studies would be required to explain this results. The *1-nn* classifier was used, mainly because the goal of this work was to show that by using the proposed methodology, it is possible to obtain a robust representation so that classification could be carried out by simple classifiers. As future work it should be useful to test the proposed methodology over several other datasets and design of a web server for protein prediction. The proposed methodology can be extended to deal with terms that are not included in the *Pfam* data base, whenever some dissimilarity matrix can be computed for the training and testing sequences.

V. Acknowledgments

This work is within the framework of the Dirección de Investigaciones de Manizales (DIMA) of the Universidad Nacional de Colombia and the Centro de Investigación of the Instituto Tecnológico Metropolitano. The work has been partially founded by Colciencias grant 111952128388 and by Jóvenes Investigadores e Innovadores 2010, Convenio Interadministrativo Especial de Cooperación No. 146 de enero 24 de 2011 between COLCIENCIAS and Universidad Nacional de Colombia Sede Manizales

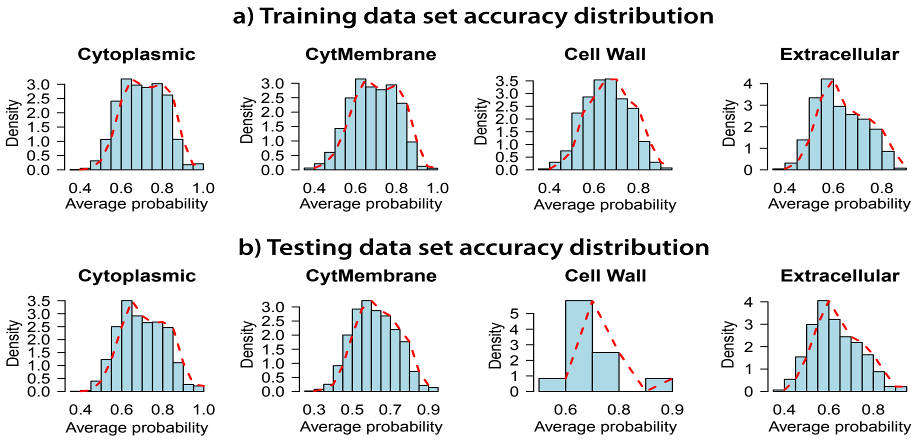

Fig. 2. Training and testing distributions from Pfam founded terms. The x axis is the average probability given by the pairwise alignment based on hidden Markov models and y axes is the density distribution. These histograms show the range between the pfam terms are related for each localization.

REFERENCES

[1] J. Gardy and F. Brinkman, "Methods for predicting bacterial protein subcellular localization," *Nature Reviews Microbiology*, vol. 4, no. 1, pp. 741–751, 2006.

[2] G. Schneider, "How many potentially secreted proteins are contained in a bacterial genome?" *Gene*, vol. 237, no. 1, pp. 113–121, 1999.

[3] C. Yu, C. Lin, and J. Hwang, "Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Science*, vol. 13, no. 5, pp. 1402–1406, 2004.

[4] B. Niu, Y. Jin, K. Feng, W. Lu, Y. Cai, and G. Li, "Using adaboost for the prediction of subcellular location of prokaryotic and eukaryotic proteins," *Molecular diversity*, vol. 12, no. 1, pp. 41–45, 2008.

[5] K.-C. Chou and H.-b. Shen, "Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Science*, vol. 02, no. 10, pp. 1090–1103, 2010.

[6] Y. Nancy, J. Wagner, M. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. Sahinalp, M. Ester, L. Foster *et al.*, "Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, vol. 26, no. 13, pp. 1608–1615, 2010.

[7] T. Hawkins, M. Chitale, S. Luban, and D. Kihara, "PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data." *Proteins*, vol. 74, no. 3, pp. 566–582, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18655063

[8] M. Punta, P. Coggill, R. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements *et al.*, "The pfam protein families database," *Nucleic acids research*, vol. 40, no. D1, pp. D290–D301, 2012.

[9] S. Eddy, "Hmmer3: a new generation of sequence homology search software. url: http://hmmer. janelia. org," *Accessed*, vol. 7, no. 25, p. 2010, 2010.

[10] H. Shen and K. Chou, "Gpos-mploc: A top-down approach to improve the quality of predicting subcellular localization of gram-positive bacterial proteins," *Protein and Peptide Letters*, vol. 16, no. 12, pp. 1478–1484, 2009.

[11] N. Yu, M. Laird, C. Spencer, and F. Brinkman, "Psortdban expanded, auto-updated, user-friendly protein subcellular localization database for bacteria and archaea," *Nucleic acids research*, vol. 39, no. suppl 1, p. D241, 2011.

[12] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "Cd-hit suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[13] R. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. Pollington, O. Gavin, P. Gunasekaran, G. Ceric, K. Forslund *et al.*, "The pfam protein families database," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D211–D222, 2010.

[14] K. Bowyer, N. Chawla, L. Hall, and W. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Arxiv preprint arXiv:1106.1813*, 2011.

[15] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, vol. 20, no. 2, 2003, p. 856.

[16] I. Holland, L. Schmitt, and J. Young, "Type 1 protein secretion in bacteria, the abc-transporter dependent pathway (review)," *Molecular membrane biology*, vol. 22, no. 1-2, pp. 29–39, 2005.

[17] M. Müller and R. Bernd Klösgen, "The tat pathway in bacteria and chloroplasts (review)," *Molecular membrane biology*, vol. 22, no. 1-2, pp. 113–121, 2005.

[18] P. Christie and E. Cascales, "Structural and dynamic properties of bacterial type iv secretion systems (review)," *Molecular membrane biology*, vol. 22, no. 1-2, pp. 51–61, 2005.

[19] D. Thanassi, C. Stathopoulos, A. Karkal, and H. Li, "Protein secretion in the absence of atp: the autotransporter, two-partner secretion and chaperone/usher pathways of gram-negative bacteria (review)," *Molecular membrane biology*, vol. 22, no. 1-2, pp. 63–72, 2005.

[20] A. Davidson, E. Dassa, C. Orelle, and J. Chen, "Structure, function, and evolution of bacterial atp-binding cassette systems," *Microbiology and Molecular Biology Reviews*, vol. 72, no. 2, p. 317, 2008.

[21] M. Saier Jr, "A functional-phylogenetic classification system for transmembrane solute transporters," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 2, p. 354, 2000.