35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Contact potentials via wavelet transform for prediction of subcellular localizations in gram negative bacterial proteins

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

Abstract—**Predicting the localization of a protein has become a useful practice for inferring its function. Most of the reported methods to predict subcellular localizations in Gram-negative bacterial proteins have shown a low false positive rate. However, some subcellular compartmens like "periplasm" and "extracellular medium" are difficult to predict and remain high false negative rates. In this paper, a method based on representation from statistical contact potentials and wavelet transform is presented. The wavelet-based method achieves an overall high performance holding low false and negative rates particularly on periplasm and extracellular medium. Results suggest the contact potentials as an useful alternative to characterize protein sequences.**

I. Introduction

Protein subcellular localizations can indicate how and what kind of cellular environments the proteins interact, helping to elucidate its function and role in biological process [1]. Experimental techniques such as immunolocalization, fluorescent tagged, and isotopes could be accurate, but they are slow and labor-intensive [2]. To cope with this drawback, several computational approaches have been developed as an alternative to predict subcellular localizations, among others: PSORTb v.3 [3], CELLO [4], PSLpred [5], LOCtree [6], P-CLASSIFIER [7], and GNeg-mPLoc [1], which cover different types of algorithms such as support vector machines (SVM), amino-acid composition, Bayesian networks, signal peptides, motif matching, homology based prediction, hidden Markov models (HMM), and text labeling among others. In general terms, they all report adequate performance, but, in spite of the low false positive rate in most of them, a high false negative rate remains.

In this work, a method to predict five distinct subcellular localizations in Gram negative bacteria is developed. The method uses local features patterns distributed along the protein sequence. The identification of such patterns is done by using the continuous wavelet transform, which has shown to be a powerful tool for the characterization of motifs [8,9]. However, the most important aspect, in the wavelet analysis, is the protein representation; here, we introduce the use of pairwise protein contact potentials in conjunction with the wavelet transform, in order to identify strongly conserved local features correlated with a specific cellular compartment. Thus, the Aaindex database, which comprises 47 protein pairwise contact potentials, is used [10]. These potentials are obtained from statistical analysis and have been extensively used to predict protein structures [11]. After all proteins are decomposed in their own local features, a clustering and modeling module based on HMM allows to compress all local features in a set of profiles [12] that can be used further as features to train an SVM and make a prediction.

Comparisons are made with three of the currently active services for subcellular localization prediction in Gram-negative bacteria: Psortb, CELLO and SOSUIGramN. In this work, the statistical contact potentials have shown to be a useful representation of the proteins. Then, if a set of proteins has similar interactions among adjacent amino acids at any position in the proteins, the wavelet transform can efficiently detect those interactions. Unlike SOSUIGramN, CELLO and Psortb in which several types of protein representations had been proposed (amino acid composition, partitioned amino acid composition, local amino acid composition, SCL-blast, signal peptides, N and C-terminal composition, profile motifs among others) the wavelet-based method involves just the local feature representation. Thoroughly, Psortb uses a set of known profile motifs per subcellular localization in contrast to the proposed method which generates its own set of profiles. SOSUIGramN consists of a set of filters in which proteins are divided into ten segments and compute average values of physicochemical properties. CELLO divides the sequence into k subsequences of equal length and each partition is encoded by a particular amino acid composition. On the other hand, the proposed method uses core local features encoded by the amino acid sequence, thus making use of the main protein information contained in the amino acid distribution. Results show the potential-wavelet method as a reliable and efficient alternative to improve the performance in the prediction of protein subcellular localization in Gram negative bacteria.

II. Materials and Methods

The method is divided into two principal stages as follows: **1) a local feature descriptor** that represents a set of sequences belonging to a determined subcellular location by a set of profiles based on HMMs. This descriptor assumes the proteins to be folded in many structural elements (motifs) which are conserved among related proteins. Particularly, these motifs are more frequent to the function that they develop than any other. **2) A classification framework** makes use of those profiles (local features) to build a representation space, in which, a query protein is

[1]Signal Processing and Recognition Group, Universidad Nacional de Colombia, s. Manizales, Campus La Nubia, km 7 via al Magdalena, Colombia. gaarangoa,jajaramillog, cgcastellanosdg@unal.edu.co
[2]The research center of the InstitutoTecnologicoMetropolitano, Calle 73 No 76A-354, Medellín, Colombia. jorgejaramillo@itm.edu.co

depicted as a vector of protein-profile distances. If there are similarities among profiles and query sequences, they would produce similar distributions. Thus, classifiers such as SVM may then identify these distributions and hence make an appropriate prediction. A general scheme of the method is shown in Figure 1

A. *Protein numerical representation*

A protein sequence $S = \{s_1, \dots s_i \dots s_t\}$ of length t can be represented in terms of the numerical signal $f = \{f_1, \dots f_i \dots f_t\}$ by the contact potential Y (see Figure 2-left), where, $f_i = Y[s_i, s_{i+1}]$ is the pairwise contact potential between the ith and $i+1$th amino acid, e.g., for the sequence $S=\{ARGNG\}$, the numerical representation is given by the pairwise contact potential between the adjacent amino acids as follows:

$$f=\{Y[A,R], Y[R,G], Y[G,N], Y[N,G], Y[G,N]\}$$

B. *Local feature detection*

Given a numerical signal $f(t)$, the Continuous Wavelet Transform (CWT) allows the identification of patterns located, simultaneously, in both scaling and spatial information. It provides the localization of conserved and variable length sub-sequences along the protein sequence S (Figure 2b).

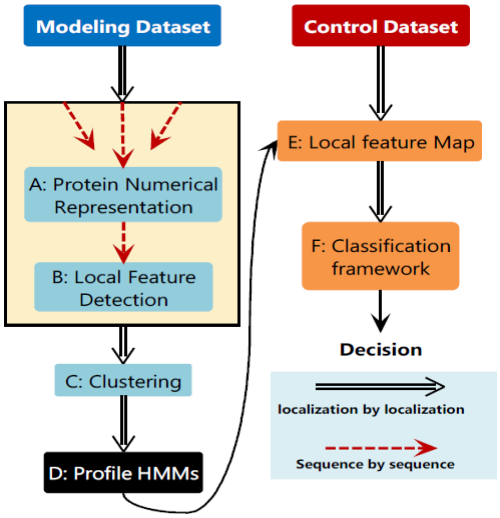

Figure 1: General workflow of the method. **Local feature descriptor (left):** The modeling data set is depicted as a set of HMMs as: A-B) Sequences from a specific subcellular localization are decomposed into a set of subsequences by the wavelet transform. Then, C) subsequences in each cellular compartment are clustered, thus, D) each group is modeled by a profile HMM. **Classification framework (right)** E) Protein-profile distances are computed over the control data set. Then, a feature space based on this distances is used to test the validity of the profiles. F) A SVM with 10 fold cross validation is carried out following the one-against-all strategy. Parameters on the SVM are tunned by the particle swarm optimization.

The CWT is defined as the projection of a function or the signal f (t) onto the wavelet function:

$$W_f(a,b) = \left(\frac{1}{\sqrt{|a|}}\right) \int_{-\infty}^{\infty} f(t)\phi\left(\frac{t-b}{a}\right)dt, \quad (1)$$

$$\phi_{a,b}(t) = \left(\frac{1}{\sqrt{|a|}}\right)\phi\left(\frac{t-b}{a}\right), \quad (2)$$

where $\phi_{a,b}(t)$ is the basis wavelet function at a particular scale a and a translation b, $a,b \in R$, $a \neq 0$. In order to identify the γ_k conserved regions throughout the sequence, the W_f matrix is decomposed into binary matrices W_f^+ and W_f^- defined as follows:

$$W_f^+ = \begin{cases} 1 & if \quad W_f > thr \\ 0 & other \end{cases} \quad (3)$$

$$W_f^- = \begin{cases} 1 & if \quad W_f < thr \\ 0 & other \end{cases} \quad (4)$$

$$thr = \frac{1}{N_a * t} \sum_{i=1}^{N_a} \sum_{j=1}^{t} W_f(i,j), \quad (5)$$

where N_a is the total number of scales and t is the length of the sequence.

The general mean value of W_f is used as a threshold *thr*. This value defines the boundary of the conserved regions over the protein sequence. Thus, the amino acid sub-sequence x_j related to each one of the region γ_j given W_f^+ and W_f^- is found. Therefore, the protein sequence S can be represented as a set of k variable length sub-sequences $x_s = \{x_1, \dots, x_k\}$

Wavelet coefficients W_f represent the adjacent and non-adjacent amino acid interactions of variable-length, depending on the scale. These interactions are given by the maximal and minimal patterns. This is the reason why the matrix W_f should be decomposed into the W_f^+ and W_f^- matrices.

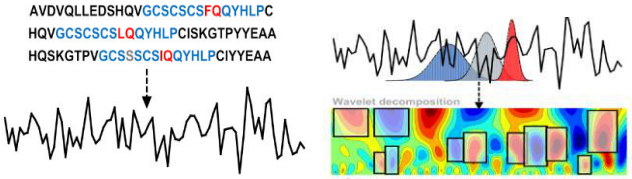

Figure 2: (left) Numerical representation of the amino acid sequence by some protein contact potential. (right) The continuous wavelet transform decompose the series *f(t)* into a set of coefficients through the sequence allowing the identification of specific patterns contained by the protein.

C. Clustering

Let the complete set of sub-sequences τ defined as the collection of all *k* founded fragments by the wavelet transform (i.e., all objects given by the wavelet decomposition of all cytoplasmic proteins). Objects in τ are grouped using the software package Clustal Ω [13]. Thus, the resulting dendrogram preserves the structural relationship among subsequences and localization. In order to identify the optimal number of clusters, the *DynamicCutTree,* which is a fast and accurate method for cutting tree, is used [14]. So, when the clustering step is done, a set of related sequences is now expressed as a set of clusters $\zeta=\{\zeta_1,...,\zeta_l\}$ preserving the core features of the sequences.

D. Profile HMMs

Consider a set of sequences ζ_i with a similar amino acid distribution. The profile HMM h_i is a statistical model for these sequences, in that for any query protein, it defines a probability whether protein belongs or does not to the set ζ_i. as introduced to model protein families and domains [12]. Several software packages implement profile HMMs with an important difference in the architecture they adopt. These methods are based on the original profile HMMs proposed in [12]. HMMER3, which uses a robust model architecture to deal with multiple domains, sequence fragments and local alignments is used to build a profile HMM for each cluster. Thus, each subcellular localization is depicted as a set of profiles $H=\{h_1,...,h_k\}$

E. Local feature map

The local feature space can be viewed as the distribution of the protein sequences over the profiles HMMs, in which the profile-protein relationship $P(S|h_i)$ is the probability that the sequence *S* belongs to profile h_i. Since sequences may contain the same domain multiple times, the value of the *i* profile is set to $\phi(i)=max_{1..n}\{P(S|h_i)\}$, where $\phi(i)$ is the probability of the highest scoring of protein *S* on the profile HMM h_i.

	Wavelet*		**Psortb**		**SOSUIGram**		**CELLO**	
	%Sn	*%Sp*	*%Sn*	*%Sp*	*%Sn*	*%Sp*	*%Sn*	*%Sp*
C	**0.92**	**0.92**	0.82	0.97	0.90	0.88	0.93	0.83
CM	0.78	0.99	**0.82**	**0.99**	0.72	0.99	0.62	0.99
P	**0.93**	**0.98**	0.79	0.99	0.59	0.98	0.50	0.97
OM	0.86	0.98	0.81	1.00	**0.92**	**0.99**	0.55	0.96
E	**0.83**	**0.99**	0.77	0.99	0.50	0.99	0.44	0.96
Av	**0.88**	**0.98**	**0.82**	**0.99**	**0.81**	**0.98**	**0.73**	**0.95**

Table1 The overall performance of the wavelet-based method with a 10-fold cross validation

F. Classification framework

Once the protein sequences from the control data set are mapped onto the profiles HMMs, the SVM is used as predictor. A 10-fold cross validation procedure is used to obtain performance results. Redundant information on the profile space is removed by means of the *fast correlation-based filter* algorithm [15]. In addition, Principal Component Analysis (PCA) is applied to this space, after which the first five principal components are selected. SVMs are designed following the one-against-all strategy that produces a strong class imbalance, and thus, the Synthetic Minority Over-sampling Technique SMOTE is employed [17]. Parameters of the SVM are tuned using the Particle Swarm Optimization algorithm [18].

In order to find the best representation per class, all 47 statistical contact potentials from aaindex are used to decompose the proteins, so, each cellular compartment comprises 47 classifiers. Then, the best one with the highest performance score is selected. This selection process is out of the classification framework avoiding bias and overtraining.

III. Results and Discussion

A. Database

In order to build the local feature descriptor 500 cytoplasmic proteins, 500 inner membrane proteins, 359 periplasmic proteins, 349 outer membrane proteins and 288 extracellular proteins were selected from ePSORTdb [19], omitting sequences with an identity percent superior to 60%. This dataset is called the modeling dataset. A control dataset reported in [20] had been used to test all methods. This control dataset comprises 299 protein sequences distributed as follows: 145 cytoplasmic proteins, 69 cytoplasmic membrane proteins, 29 periplasmic proteins, 38 outer membrane proteins and 18 extracellular proteins. In addition, any proteins sharing >60% identity of modeling data set with respect to the control data set were removed. All identity filters were carried out using the software cdHit [16].

To evaluate and compare the performance of the methods, both measures, sensitivity $Sn=\frac{TP}{TP+FN}$, and specificity $Sp=\frac{TN}{FP+TN}$ are used, where *TP, FP, TN*, and *FN* denotes true positive, false positive, true negative and false negative, respectively. The following web servers are used to predict the subcellular localizations in addition to the standalone version of Psortb V3.0.2, CELLO version 2.5 and SOSUIGramN. In order to ensure that psortb classifications are not biased, the modeling data set is used in the blast module, so, the predictions are carried out on the control data set. Also, it is necessary to clarify that is not possible to verify whether test sequences are or are not in the training set of CELLO and SOSUIGramN servers. Performance prediction of the wavelet-based method and the corresponding comparisons are shown in Table 1.

The performance of the individual methods reveals that the wavelet approach achieves the highest overall sensitivity. This fact can be highlighted in "periplasmic" and "extracellular" localizations, in which the proposed method has a significant increase above 10% over psort, SOSUIGramN and CELLO. Also, the specificity for these classes are basically the same in all methods showing that the wavelet approach can improve the true positive rate holding a low false positive rate (Table 1). For cytoplasmic proteins, CELLO shows the highest sensitivity (0.93) followed by the proposed approach (0.92), SOSUIGramN (0.89) and psort (0.82). However, CELLO and SOSUIGramN have a low specificity (0.83 and 0.88 respectively) which is interpreted as a high false positive rate.

A protein can remain in the cytoplasm or be targeted into different sites by a transport system, and thus, proteins associated to the "cytoplasm" localization are highly diverse and comprise a big variety of domains. It is also the case of transmembrane proteins, which are simultaneously located on both sides of the membrane and transport molecules from one side to the other, making difficult to characterize this kind of proteins through local features (we use this term to refer to domains, motifs and sites). Accordingly, both "cytoplasmic" and "outer membrane" are the classes with the lowest performances of sensitivity in comparison to psortb and SOSUIGramN, respectively. Psortb shows an upper sensitivity of 5% respect to the proposed method, while the specificity remains equal. For "outer membrane" proteins, SOSUIGramN achieved the best sensitivity of 92% followed by our method and psortb with a 6% and 11% upper, respectively.

IV. CONCLUSIONS

For the five major subcellular localizations in Gram-negative bacterial proteins, the wavelet method showed the best performance prediction decreasing the false negative rate and holding the false positive rate. One of the main advantages of the method is its capability to find correlated and variable length local features, followed by a precise representation by HMMs. Thus, the proposed contact-potential characterization is an alternative to the classic models based on the amino acid composition and physicochemical properties. This method, unlike Psortb, CELLO and SOSUIGramN, uses just one protein characterization. As future work, the implementation of other representations such as physicochemical properties, amino acid composition, or homology modules like blast can be implemented to improve even more the final result.

Acknowledgments: This work is within the framework of the DirecciondeInvestigaciones de Manizales (DIMA) of the Universidad Nacional de Colombia and the Centro de Investigacion of the InstitutoTecnologico Metropolitano. The work has been partially funded by Colciencias grant 111952128388 and by Jovenes Investigadores e Innovadores 2010, Convenio Interadministrativo Especial de Cooperacion No. 146 de enero 24 de 2011 between COLCIENCIAS and Universidad Nacional de Colombia Sede Manizales

V. REFERENCES

[1] Chou, Kuo-Chen, and Hong-Bin Shen. "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms." Nature protocols 3.2 (2008): 153-162.

[2] Dunkley, Tom PJ et al. "Localization of organelle proteins by isotope tagging (LOPIT)." Molecular & Cellular Proteomics 3.11 (2004): 1128-1134.

[3] Yu, Nancy Y et al. "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes." Bioinformatics 26.13 (2010): 1608-1615

[4] Yu, Chin- Sheng et al. "Prediction of protein subcellular localization." Proteins: Structure, Function, and Bioinformatics 64.3 (2006): 643-651.

[5] Bhasin, Manoj, AartiGarg, and GPS Raghava. "PSLpred: prediction of subcellular localization of bacterial proteins." Bioinformatics 21.10 (2005): 2522-2524.

[6] Nair, Rajesh, and BurkhardRost. "Mimicking cellular sorting improves prediction of subcellular localization." Journal of molecular biology 348.1 (2005): 85-100.

[7] Wang, Jiren et al. "Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines." BMC bioinformatics 6.1 (2005): 174.

[8] Murray, Kevin B, Denise Gorse, and Janet M Thornton. "Wavelet transforms for the characterization and detection of repeating motifs." Journal of molecular biology 316.2 (2002): 341-363

[9] Arango-Argoty, GA et al. "Prediction of protein subcellular localization based on variable-length motifs detection and dissimilarity based classification." Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE 30 Aug. 2011: 945-948.

[10] Kawashima, Shuichi, and Minoru Kanehisa. "AAindex: amino acid index database." Nucleic acids research 28.1 (2000): 374-374.

[11] Shen, Min- yi, and Andrej Sali. "Statistical potential for assessment and prediction of protein structures." Protein Science 15.11 (2009): 2507-2524

[12] Finn, Robert D, Jody Clements, and Sean R Eddy. "HMMER web server: interactive sequence similarity searching." Nucleic acids research 39.suppl 2 (2011): W29-W37.

[13] Sievers, Fabian et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Molecular systems biology 7.1 (2011).

[14] Langfeldera, Peter, Bin Zhangb, and Steve Horvatha. "Dynamic Tree Cut: in-depth description, tests and applications." (2007).

[15] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE- 21 Aug. 2003: 856

[16] Li, Weizhong, and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics 22.13 (2006): 1658-1659.

[17] Chawla, Nitesh V. et al. "SMOTE: synthetic minority over-sampling technique." arXiv preprint arXiv:1106.1813 (2011).

[18] Clerc, Maurice. Particle swarm optimization. Wiley-ISTE, 2010.

[19] Nancy, Y Yu et al. "PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea." Nucleic acids research 39.suppl 1 (2011): D241-D244.

[20] Gardy, Jennifer L, and Fiona SL Brinkman. "Methods for predicting bacterial protein subcellular localization." Nature Reviews Microbiology 4.1 (2006): 741-751.