33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Prediction of protein subcellular localization based on variable-length motifs detection and dissimilarity based classification

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

***Abstract*— Predict the function of unknown proteins is one of the principal goals in computational biology. The subcellular localization of a protein allows further understanding its structure and molecular function. Numerous prediction techniques have been developed, usually focusing on global information of the protein. But, predictions can be done through the identification of functional sub-sequence patterns known as *motifs*. For *motifs* discovery problem, many methods requires a predefined fixed window size in advance and aligned sequences. To confront these problems we proposed a method based on variable length motifs characterization and detection using the continuous wavelet transform (CWT) and a dissimilarity space representation. For analyzing the motifs results generated by our approach, we divide the entire dataset into training (60%) and validation (40%). A Support Vector Machine (SVM) classifier is used as predictor for validation set. The highest $Sn = 82.58\%$ and $Sp = 92.86\%$, across 10-fold cross validation, is obtained for *endosome* proteins. Average results $Sn = 74\%$ and $Sp = 75.58\%$ are comparable to current state of the art. For data sets whose identity is low ($< 40\%$), the motifs characterization and localization based on CWT shows a good performance and the interpretability of the subsequences in each subcellular localization.**

***Index Terms*— Motifs, wavelet transform, hydrophathy scale, subcellular localization, support vector machine.**

I. Introduction

ONE of the main goals of genomic projects is to provide reliable functional annotations for gene products. In particular, protein subcellular localizaton provides useful clues for revealing protein functions and for understanding the intricate pathways that regulate biological processes at cellular level [5]. The location of specific proteins can be determined through experimental approaches such as the attachment of green fluorescent protein coding sequences to one end of the sequence encoding the protein of interest in order to monitor its intrinsic fluorescence and subsequently locate it within the cell [1]. However, such procedures are expensive and highly time consuming, leading to the development of computational predictors wich are able to identify the subcellular localization of newly found proteins based on their primary sequence information alone [4].

A vast number of predictors based on pattern recognition methods have been designed in the last few years. Their main difference lies in the attributes they extract to characterize protein sequences: statistical and physical-chemical properties of amino acids ([10],[16]), energy concentrations form time-frequency representations, distance measures, word statistics, information theory and others [17]. However, most of them only describe global attributes of the whole protein sequence, ignoring the fact that functional domains may reside in different portions of proteins within the same family. For instance, the same functional domain may reside at the beginning of one protein and at the end of another. Moreover, amino acids that have an important role in protein function and structure cannot mutate without an important effect on protein activity, but change very slow in a given protein family during evolution [13]. Thus, for a set of sequences that stretch a great evolutionary distance it is possible to identify regions of amino acids that are highly conserved even if they greatly differ from a global perspective.

Such recurring patterns are called *motifs* and can be used to identify representative regions of the proteins, revealing their potential location within the cell. As an example, transmembrane proteins have three principal regions called cytoplasmic, transmembrane and extracellular domains, each one of them constituted by amino acids with specific hydrophobic properties that give rise to a distinctive motif. However, common algorithms for motif identification are limited by several restrictions such as the need for a predefined motif size or globally aligned sequences [13]. Additionally, most of them are limited to the discovery of motifs but do not apply this information to the prediction of protein subcellular localization.

Regarding subcellular location prediction, few methods have been proposed. Most of them for bacterial organisms as *PSORT I, PSORT b, proteome analyst, amino acid composition SVM based methods, CELLO and P-Classifier*, See the review [6] for details of each one of them. These methods use local alignment, amino acid composition, signal peptides among others global features. Plant-Ploc [5] and TestLoc [16] are methods based on the amino acid composition and ensemble classifiers for deducing the subellular localization of uncharacterized proteins in plants.

In this paper, a methodology for protein subcellular localization predicton is proposed, based on the identification of variable-length motifs by using the continuous wavelet transform. Query proteins are mapped into these prototype motifs resulting in a dissimilarity space in which a support vector machine classifier is built to predict the location to which a specific protein belong. The remainder of this paper is organized as follows. Section II describes methods for

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

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

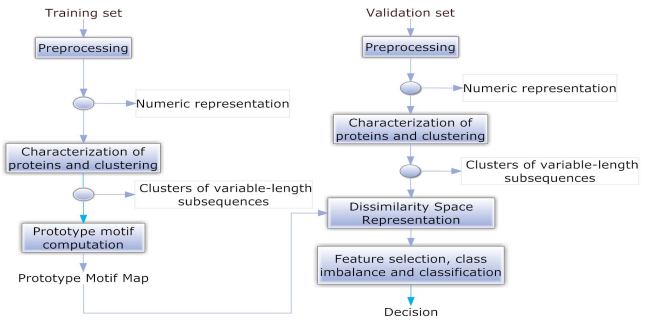

Fig. 1. Prototye motif representation and dissimilarity based classification flow diagram.

prototype motif identification and subcellular localization prediction. Section III presents the results and performance of the classification. Conclusions are presented in section IV.

II. Materials and Methods

The proposed method is described in Fig. 1. First, training proteins of each class (subcellular location) are preprocessed to extract short sub-sequences of variable length, determined by high energy concentrations on the scalogram. The scalogram communicate the time frequency localization property of the wavelet transform. Then, sub-sequences are clustered to extract prototype motifs, defined as the cosensus of all sub-sequences belonging to one cluster. The number of prototype motifs describing a subcellular location is thus equal to the number of clusters obtained by the clustering algorithm. Query proteins are mapped into a dissimilarity space comprised of alignment-score distances to prototype motifs and a support vector machine is trained to decide whether or not the query protein belongs to that specific location. All experiments where carried out on a plant (embryophyta) protein database, belonging to eight different subcellular locations.

A. Database

The database used is comprised of 1097 sequences extracted from the public resource Uniprot [9]. Sequences belonging to the taxonomic class *embryophyta* (land plants) were selected, with at least one annotation in the Gene Ontology Annotation (GOA) project [2]. Sequences predicted by computational tools and with no real experimental evidence were discarded. The subcellular localization data set is comprised of eight different locations: vacuole, peroxisome, golgi apparatus, ribosome, nucleoplasm, endosome, endoplasmatic reticulum and cytoplasm. The dataset does not contain protein sequences with a sequence similarity superior to 40% in order to avoid bias due to the presence of protein families in the database.

B. Preprocessing

A protein can be represented as a signal in function of its length, by substituting each amino acid by its equivalent value of a given physico-chemical property. Let the sequence of interest be represented as the discrete symbolic signal $X = \{x_1, x_2, ..., x_i, ..., x_n\}$, for $1 \leq i \leq n$, n denotes the length of the sequence and $x_i \in S$, where $S = \{s_1, s_2, ..., s_j, ..., s_M\}$ is the set of possible symbols and $H = \{h_1, h_2, ..., h_j, ..., h_M\}$ is the set of values associated to the physicochemical property. The relationship between the symbolic and numeric representation is $H\{s_j\} = h_j$. Then, the discrete "position" signal Y can be represented as $Y = \{y_1, y_2, ..., y_i, ..., y_n\}$ where $y_i = H\{x_i\}$. For proteins the set of symbols corresponds to the number of amino acids found in nature, in other words $M = 20$.

As protein folding is driven by hydrophobic forces, the hydropathy distribution along the protein sequence has been recognized as a useful feature for characterizing protein structures. In this study, the Kyte Doolittle hydropathy scale is used. It is based on the free energy transfer of each amino acid between organic solvent and water [12]. This scale aims to group the amino acids into three categories: Strongly hydrophilic, Strongly hydrophobic and Weakly hydrophilic or weakly hydrophobic (see Fig. 2).

C. Extraction of variable length sub-sequences

With the proteins converted into numerical signals, it is possible to treat them with statistical methods. The continuous wavelet transform (CWT) allows the identification of patterns located simultaneously in both spectral and spatial information within the sequences. The wavelet transform was proposed in [15] for known proteins that contain repeating motifs.

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

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

$$\psi_{a,b}(t) = \left(\frac{1}{\sqrt{|a|}}\right)\psi\left(\frac{t-b}{a}\right) \tag{1b}$$

where $\psi_{a,b}(t)$ is the basis function at a particular scale a and a translation b, $a, b \in R$, $a \neq 0$.

The mother wavelet used in our experiments is the *Bior6.8* because the decomposition wavelet function and scaling are very rugged and have abrupt changes, allowing an adequate

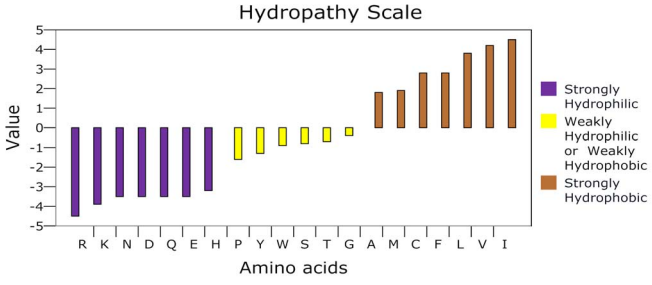

Fig. 2. The Kyte Doolittle hydropathy scale for each one of the amino acids and the three classifications of the amino acids. Hydrophobic amino acids tend to be internal (with regard to the protein's 3 dimensional shape) while hydrophilic amino acids are more commonly found towards the protein surface.

representation of protein signals with high variability. For regions with maximal concentrations of energy through the sequence it is possible to locate the coordinate point of the centroid in the scale-position space. Then, this point grows in sequence position axis both towards the left and the rigth until the value of the actual position is less than the minimal value of the region, thus determining the number of motifs in a sequence. This process is applied to each sequence in the training set.

D. Prototype motifs and dissimilarity space

The *Iterative Self Organizing Data Analysis Technique - Isodata* was used to cluster subsequences in order to obtain representative motifs. The normalized score of the sequence local alignment was used as the metric for clustering with alignment-score distance, defined as:

$$d(x,y) = (\frac{1-s(x,y)}{s(x,x)}) * (\frac{1-s(x,y)}{s(y,y)}) \quad (2)$$

where s is the similarity between two sequences x and y computed by:

$$s(x,y) = \sum_{i=1}^{l} M(x(i), y(i)) \quad (3)$$

being l the length of the subsequences and $M(x(i), y(i))$ the value of the similarity matrix for the $i-th$ elements of x and y. For M, we used the Point Accepted Mutation (PAM150) scoring matrix [18].

A *prototype motif* is generated for each resulting cluster as the consensus sequence for locally aligned sub-sequences. Given a cluster Ck, the consensus for this cluster is:

$$Q'_k = M * P_k \quad (4a)$$

$$P_k(i,j) = \log(\frac{f_k(i,j)+k}{|C_k|}) \quad (4b)$$

where P_k is the profile of the cluster C_k, $f_k(i,j)$ represents the count of amino acid j at position i of the subsequences in C_k. The consensus $Q_k(i)$ is the indexing amino acids with highest values on Q'_k for each column.

Once the set of prototype motifs are generated, a new sequence S is represented as the distribution of their sub-sequences s_i. The *Dissimilarity Vector Construction* aims to take the minimum alignment-score distance between the set of subsequences and a prototype motif. Each subsequence s_i is compared with each consensus Q_k. The value for the $k-th$ dimension of the dissimilarity space F is set to:

$$F(k) = \min_{s_i \in S} \{d(s_i, Q_k)\} \quad (5)$$

Conceptually, this quantity is a measure of the extent at which the prototype motif Q_x is present in the sequence S.

E. SVM-based predictor

The entire database was divided into a training and a validation set. For each class, 60% of the sequences were selected for training and 40% for validation. The *Fast Correlation-Based Filter* [19] was used for feature selection, providing a reduced matrix F_r of distances to non redundant motifs. Since support vector machines are designed only for two-class problems, classification was implemented following the one-against-all strategy. This method produces a strong class imbalance, so *Synthetic Minority Over-sampling Technique* (SMOTE) was employed to overcome it, in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement [3]. Parameters of the SVM were tuned with a *Particle Swarm Optimization* algorithm [11]. Validation of results were obtained by 10-fold cross-validation. Sensitivity (Sn), specificity (Sp) and geometric mean (gm) were used as classification performance measures.

III. Results and Discussion

Detection of motifs can be illustrated with the wavelet representation shown in Fig. 3. This scalogram belongs to the $A8R7K9$ protein of *Arabidopsis thaliana*, located within the cell in the endosome and is responsible for directing the movement of substances from endosomes to lysosomes [8]. For this protein, the wavelet representation found a total of 71 motifs. Some of these motifs were also found by the web tools ScanProsite [7] and InterPro scan [14], demonstrating the validity of the results:

- **N-P-[ST]-P**, called N-glycosylation site pattern, found at positions 6-9, 244-247, 375-378, 444-447. In the scalogram they all correspond to the motifs: 1, 21, 31 and 41.
- **[ST]-x(2)-[DE]**, called Casein kinase II phosphorylation site pattern is found in most of the known physiological substrates. The positions are 84-87, 160-163, 444-447 and 753-756, wich correspond with the scalogram motifs 8, 16, 35 and 68.
- **L-x(6)-L-x(6)-L-x(6)-L**, known as the Leucine zipper pattern, It consists of a periodic repetition of leucine residues at every seventh position over a distance covering eight helical turns. The segments containing these periodic arrays of leucine residues seems to exist in an α-helical conformation. The positions are 346-367 and 353-374. Given the nature of the amino acids, these sequences are found in the union of motifs 29, 30 and 31 in the scalogram.

Prediction of protein subcellular localization was compared with TESTLoc [16] and Plant-PLoc [5]. TESTLoc is to the best of our knowledge, the closest method to ours, since

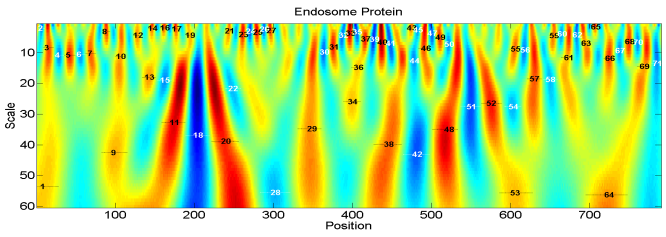

Fig. 3. Wavelet transform location of sub-sequences in an Endosome protein

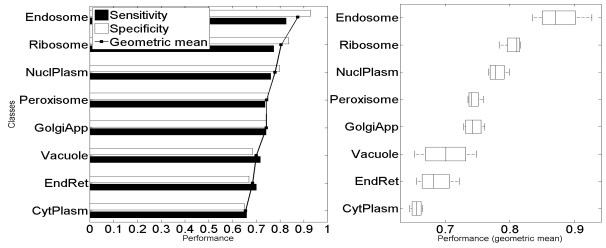

Fig. 4. Prediction performance of subcellular localizaton in $Embryophyta$ proteins. Left plot shows mean performance statistics, while right plot depicts performance variation through ten repetitions.

it also perform predictions on proteins of $embryophyta$. In contrast, they characterize protein sequences according to physicochemical properties of amino acids, six different types of amino acid composition, grouped amino acid composition according to their properties and gapped amino acid composition. None of these properties take into account the spatial distribution of the sequences in which the feature space comprises high dimensionality. Additionally, their dataset includes sequences with up to 60% of identity. For endoplasmic reticulum, peroxisomes and vacuoles they reported average sensitivities for all sequence feature of 20%, 40% and 63.3%, respectively. For three top-performing features the average sensitivities were 20%, 50% and 63.3% respectively. Whereas our average sensitivities were 70.12%, 73.78% and 71.72%. The entire results for our system are shown in Fig. 4. Plant-PLoc is a method for protein subcellular localization in plants. The prediction quality was examined employing two datasets, the first for training and the second for validation using 406 and 265 proteins, respecivelly. Where none of the proteins had $\leq 25\%$ sequence identity to any other in the same subcellular localization. The feature mapping is given by hybridation of GO and amphipilic pseudo amino acid composition and the classification is made by the hybridation of ensemble classifiers. Plant-PLoc report an average sensitivity of $S_n = 78.9\%$ whereas our method for the same database showed an average sensitivity of $S_n = 76.6\%$. With the advantage that we found discriminating sites (prototype motifs) for each subcellular localization which are not found with conventional methods for sequence alignment.

IV. Conclusions

In this paper, we proposed a characterization and localization of variable-length motifs based on continuous wavelet transform for subcellular location problem. For a set of related sequences, is likely to find short patterns distribuited throughout the sequences, and the nature of this patterns is given by its amino acid interactions. The scalogram is a representation of these interactions, then, for training sequences, we found a set of prototype motifs that represented a specific subcellular location by isodata algorithm. The endosome and ribosome proteins showed high discriminability with regard to remaining classes. Therefore, the results showed that the motifs characterization based on continuous wavelet transform and the posterior dissimilarity space representation allowed discrimination between different subcellular localizations. Our method proved to be competitiveness with respect to other developed methods in the state of the art for data sets with a low identity.

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 Centro de Investigaciones ITM grants P10240 and P09225.

References

[1] P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach*. The MIT Press, 2001.

[2] D. Barrell, E. Dimmer, R.P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic acids research*, 37(Database issue):D396, 2009.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[4] K.C. Chou and H.B. Shen. Recent progress in protein subcellular location prediction. *Analytical Biochemistry*, 370(1):1–16, 2007.

[5] K.C. Chou, H.B. Shen, and E. Newbigin. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PloS one*, 5(6):259–270, 2010.

[6] J.L. Gardy and F.S.L. Brinkman. Methods for predicting bacterial protein subcellular localization. *Nature Reviews Microbiology*, 4(1):741–751, 2006.

[7] A. Gattiker, E. Gasteiger, and A. Bairoch. ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*, 1(2):107–108, 2002.

[8] Y. Jaillais, I. Fobis-Loisy, C. Miège, and T. Gaude. Evidence for a sorting endosome in Arabidopsis root cells. *The Plant Journal*, 53(2):237–247, 2008.

[9] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B.E. Suzek, M.J. Martin, P. McGarvey, and E. Gasteiger. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC bioinformatics*, 10(1):136, 2009.

[10] J.A. Jaramillo-Garzón and, A. Perera-Lluna, and C.G. Castellanos-Domónguez. Predictability of protein subcellular locations by pattern recognition techniques. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5512 –5515, 31 2010-sept. 4 2010.

[11] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[12] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein* 1. *Journal of molecular biology*, 157(1):105–132, 1982.

[13] X.I. Liu, N. Korde, U. Jakob, and L.I. Leichert. CoSMoS: conserved sequence motif search in the proteome. *BMC bioinformatics*, 7(1):37, 2006.

[14] N. Mulder and R. Apweiler. InterPro and InterProScan. *Comparative genomics*, page 59, 2007.

[15] K.B. Murray, D. Gorse, and J.M. Thornton. Wavelet transforms for the characterization and detection of repeating motifs1. *Journal of molecular biology*, 316(2):341–363, 2002.

[16] Y.Q. Shen and G. Burger. TESTLoc: protein subcellular localization prediction from EST data. *BMC bioinformatics*, 11(1):563, 2010.

[17] S. Vinga and J. Almeida. Alignment-free sequence comparisona review. *Bioinformatics*, 19(4):513, 2003.

[18] D. Wheeler. Selecting the Right Protein-Scoring Matrix.

[19] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, volume 20, page 856, 2003.