Hierarchical Motif Vectors for Prediction of Functional Sites in Amino Acid Sequences Using Quasi-Supervised Learning

Bilge Karaçali

Abstract—We propose hierarchical motif vectors to represent local amino acid sequence configurations for predicting the functional attributes of amino acid sites on a global scale in a quasi-supervised learning framework. The motif vectors are constructed via wavelet decomposition on the variations of physico-chemical amino acid properties along the sequences. We then formulate a prediction scheme for the functional attributes of amino acid sites in terms of the respective motif vectors using the quasi-supervised learning algorithm that carries out predictions for all sites in consideration using only the experimentally verified sites. We have carried out comparative performance evaluation of the proposed method on the prediction of N-glycosylation of 55,184 sites possessing the consensus N-glycosylation sequon identified over 15,104 human proteins, out of which only 1,939 were experimentally verified N-glycosylation sites. In the experiments, the proposed method achieved better predictive performance than the alternative strategies from the literature. In addition, the predicted N-glycosylation sites showed good agreement with existing potential annotations, while the novel predictions belonged to proteins known to be modified by glycosylation.

Index Terms—Functional attribute prediction, hierarchical motif vectors, protein sequence analysis, quasi-supervised learning.

1 Introduction

Understanding structural and functional properties of proteins from large amounts of quantitative data with statistical reliability requires techniques from statistical learning theory just as understanding any other phenomenon bearing uncertainty [1]. In spite of this close association of bioinformatics with statistical learning, a wide scale inclusion of general-purpose learning algorithms into bioinformatics and computational biology has been hampered due to a critical shortcoming of the sequence data: while the sequence data are digital and can be stored using alpha-numeric codes in a computer environment, it is not numeric in and of itself. As statistical learning algorithms model data instances, or samples, as vectors in a vector space of observations, sequence data cannot be readily subjected to the variety of algorithms in store in the statistical learning literature.

Several studies have addressed this problem by calculating numeric features on the composition and the configuration of nucleic acid or amino acid sequences for protein classification. One strategy is to compute histograms of nucleic acid bases or amino acid residues occupying a succession of sites, as in amino acid and dipeptide compositions [2], [3], [4]. The frequency of critical tripeptides, tetrapeptides, and pentapeptides have also been considered for protein characterization via n-gram analysis [5], [6], [7]. Several issues limit the efficacy of the frequency histograms in protein understanding, however, such as the combinatorial nature of the letter groups and their inability to characterize long-range dependencies that are important especially in secondary structure prediction [8], [9].

An alternate strategy toward the same end computes scalar attributes for amino acid sequences by converting them into numeric sequences using physico-chemical properties of amino acids [10], [11], [12]. While these attributes can provide very rich descriptions of amino acid sequences, their global nature introduces a risk of overlooking the presence of critical subsequences, whose effects can be concealed, dominated by the remaining parts of the amino acid sequence. In addition, even when accurate predictions are obtained, the molecular cues exploited by the predictors remain obscure as the features are derived from complex applied mathematics formulas.

Determination of amino acid subsequences that may bear functional significance has been addressed in the literature in terms of sequence motifs. A large number of computational algorithms have been used to trace the repeated amino acid combinations within protein subgroups for sequence motif discovery. Among the most notable of these algorithms is the MEME method [13], [14] that searches for contiguous blocks repeated across the amino acid sequences of a given set of proteins using a probabilistic motif model and the Expectation-Maximization algorithm [15]. The motifs discovered by such methods are collected into public databases such as MnM [16], ELM [17], Prosite [18], [19], [20], CDART [21], RPSBLAST [22], and BLIMPS [23]. Several databases keep and continuously update detailed records of protein families identified through conserved amino acid patterns such as InterPro [24], PFAM [25], SMART [26], [27], and PROSITE [28].

Supervised classification methods can be employed for functional site identification by defining distance measures

• *The author is with the Department of Electrical and Electronics Engineering, Izmir Institute of Technology, Gülbahçe, 35430 Urla Izmir, Turkey. E-mail: bilgekaracali@iyte.edu.tr.*

Manuscript received 26 Aug. 2011; revised 14 Mar. 2012; accepted 27 Apr. 2012; published online 8 May 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-08-0219. Digital Object Identifier no. 10.1109/TCBB.2012.68.

between local amino acid patterns, such as sequence identity [29]. These methods construct decision rules to separate the patterns that possess a specific characteristic from the others, using a collection of patterns for which the correct decisions are already known [30]. In the prediction of glycosylation sites, for instance, online protein databases provide up-to-date lists of sites that are experimentally verified to be glycosylated [31]. The data sets, however, do not provide a complementary list of sites that are experimentally verified to lack glycosylation.

One option, then, is to gather the sites in proteins that are not modified by glycosylation, such as mouse interleukin-3 (P01586) [32], [33]. Another is to collect the nonglycosylated sites in glycosylated proteins [29], [34]. The essential problem in both cases is poor characterization of the sites lacking glycosylation: the small number of sites collected in this manner cannot be expected to adequately characterize the actual nonglycosylated sites across large protein sets. Finally, it is also possible to improvise true-negative data sets from the sites that are not annotated to be modified by glycosylation [35], albeit at an even higher risk of misrepresentation: these sites may in fact be positive glycosylation sites yet in line to be identified as such.

To complicate the matter further, the true positive data sets are not exempt from errors either. Protein databases continuously evolve with new findings that occasionally alter the existing annotations. For instance, the site at position 271 on alpha1-antitrypsin (P01009), used previously as a true negative site for glycosylation [36], [37], is now annotated to be a positive glycosylation site in the UniProt Knowledgebase database. In another instance, 4F2 cell-surface antigen heavy chain (P08195) has glycosylation annotations at six sites in the UniProt Knowledgebase database release July 2010, revised down to 4 in release January 2011. These examples illustrate that the molecular biology data are inherently noisy and subject to change in time, and statistical learning algorithms that require absolute examples to train on are bound to suffer in structural or functional prediction tasks.

In this paper, we propose a novel computational method for predicting functional attributes of sites along amino acid sequences on a global scale using only the sites that have been experimentally verified to possess the functional attribute of interest. To this end, we first compute hierarchical motif vectors as novel descriptors of local amino acid configurations that encode the short, mid, and long-range variations in physico-chemical composition around each site along amino acid sequences using the wavelet decomposition [38], [39]. We then carry out statistical learning over the motif vectors of all sites under consideration for the specific functional attribute using the quasi-supervised learning algorithm that identifies the sites that are likely to possess the functional attribute among all prospective sites based only on the motif vectors of the experimentally verified sites [40]. In experiments, the quasi-supervised learning algorithm identified the likely N-glycosylation sites among all candidates possessing the consensus sequon across all human proteins based on their motif vectors, and achieved a higher predictive performance than existing N-glycosylation prediction methods from the literature. This paper represents the first application of the quasi-supervised learning strategy in the field of computational biology and bioinformatics.

TABLE 1
Units Statistics of the Protein Sequence Data Used in the Study

Number of entries in the original dataset	524420
Number of human proteins	20252
Number of proteins with no unidentified sites (working set)	20187
Total number of sites in the working set	11227834
Average sequence length per protein in the working set	556.1913

The sequence data was downloaded from the Uniprot/Swiss-Prot Knowledgebase, release January 2011.

Details of motif vector construction from amino acid sequences are provided in the next section along with a brief summary of the quasi-supervised learning paradigm as well as its application to N-glycosylation prediction. The experiment results are shown in Section 3, followed by concluding remarks in Section 4.

2 Methods

In this section, we first describe the amino acid sequence data used in this study and elaborate on the construction of hierarchical motif vectors. Then, we briefly summarize the quasi-supervised learning paradigm for pattern recognition and its application to the recognition of N-glycosylation sites using motif vectors. Finally, we elucidate the procedure we have used to carry out the comparative performance evaluation of the proposed functional prediction method against possible alternatives from the existing literature.

2.1 Sequence Data

The sequence data of all known human proteins were obtained from the UniProt/Swiss-Prot Knowledgebase (http://www.uniprot.org/), release of January 2011. The data set was then parsed twice, first to extract the human protein data, and a second time to load the data into the Matlab mathematical analysis environment (The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098, USA) for subsequent processing. The statistics of the resulting sequence data is summarized in Table 1.

2.2 Construction of Hierarchical Motif Vectors

The hierarchical motif vectors were computed from amino acid sequences of human proteins using the physicochemical properties associated with each amino acid. These properties were obtained from the publicly available APDbase data set [41] (accessible at the web address http://www.rfdn.org/bioinfo/APDbase/) providing a collection of 243 properties from the literature for the 20 naturally occurring amino acids. The computation of the hierarchical motif vectors relied on constructing numeric property sequences from amino acid sequences by replacing the amino acids in a sequence with a select physicochemical property, and subjecting it to a wavelet decomposition. The wavelet decomposition effectively separated the variations in the select property along the amino acid sequence at varying ranges or scales.

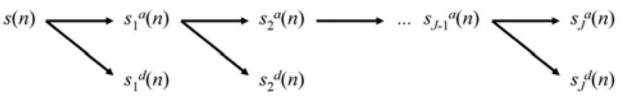

Fig. 1. Computation of the approximation and detail property sequences via wavelet decomposition.

Details of the wavelet decomposition can be found in the vast literature on the topic (see [38] and the references contained therein). Briefly, given a numeric sequence $s(n)$ of a select property along a given amino acid sequence, the wavelet decomposition computes the coefficients $a_{i,j}$ and $d_{i,j}$ that satisfy the equality

$$s(n) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} d_{i,j}\phi_{i,j}(n) + \sum_{i=1}^{n_J} a_{i,J}\psi_{i,J}(n) \tag{1}$$

for scales $j = 1, 2, \ldots, J$, where J denotes the maximum scale of the decomposition, i indexes the wavelet coefficients at the corresponding scale, and n denotes the site position along the sequence. The functions φ and ψ represent the so-called mother and father wavelets. The mother wavelet, φ, allows generating all the subsequent wavelets $\varphi_{i,j}$ by scaling and shifting along the sequence. The father wavelet, ψ, accompanies the mother wavelet and ensures that the slow varying components in the numeric sequence discarded by the wavelets $\varphi_{i,j}$ are preserved in the representation. The coarsest approximation property sequence $s_J^a(n)$ given by

$$s_J^a(n) = \sum_{i=1}^{n_J} a_{i,J}\psi_{i,J}(n) \tag{2}$$

then represents the average value of the select property along the sequence across a range of about 2^J sites, while the detail property sequence $s_j^d(n)$

$$s_j^d(n) = \sum_{i=1}^{n_j} d_{i,j}\phi_{i,j}(n) \tag{3}$$

represents the variations of the property along the sequence across approximately 2^j sites wide neighborhoods (Fig. 1). Note that in the multiscale organization of the approximation and detail property sequences, each level of decomposition corresponds to a distinct level of a representation hierarchy. In other words, the hierarchy in the representation in (1) draws directly from the wavelet decomposition that separates the short, mid, and long-range variations along the numeric property sequences.

The notion of a hierarchical motif vector rests on the understanding that the vector

$$w_s(n) = \begin{bmatrix} s_J^a(n) \\ s_J^d(n) \\ \vdots \\ s_1^d(n) \end{bmatrix} \tag{4}$$

captures the configuration of the select property around site n, and thus, is informative on the functional/structural properties associated with the site n of the corresponding amino acid sequence. The concatenation of all such vectors for each of the 243 amino acid properties from the APDbase as

$$\omega(n) = \begin{bmatrix} w_{s_1}(n) \\ w_{s_2}(n) \\ \vdots \\ w_{s_{243}}(n) \end{bmatrix} \tag{5}$$

then represents the most detailed description of the physico-chemical organization around all sites across the given amino acid sequence over a representation hierarchy, hence the term *hierarchical motif vectors*. Similarly, the approximations $s_1^a(n), s_2^a(n), \ldots, s_J^a(n)$ can also be collected into a column vector $\omega_s^a(n)$, combined with those of other property sequences into an approximation motif vector $\omega^a(n)$ providing an alternative perspective on the same physico-chemical organization around the site n. To distinguish between $\omega(n)$ and $\omega^a(n)$, we termed the former the *decomposition motif vector* and the latter the *approximation motif vector* at site n.

Prior to the computation of the motif vectors, the amino acid properties obtained from the APDbase were subjected to a gamma normalization to have their values across the 20 amino acids cover the unit interval as uniformly as possible, and thus, be comparable to each other in terms of their respective dynamic ranges. The gamma normalization entailed first scaling the observed property values p_1, $p_2, \ldots, p_{20}$ linearly so that

$$\min_i p_i = 1/21$$

and

$$\max_i p_i = 20/21,$$

and then finding the coefficient γ that minimizes the functional

$$E(\gamma) = \frac{1}{2}\sum_{i=1}^{20} \left(p_{(i)}^{\gamma} - \frac{i}{21} \right)^2 \tag{6}$$

that assesses the discrepancy between the distribution of the ordered property values $p_{(i)}$ taken to the power γ, $\{p_{(i)}^{\gamma}\}$, and the uniform distribution within the unit interval. The ordered property value $p_{(i)}$ represents the ith smallest value in the set $\{p_i\}$. This normalization allows heavily one-sided data distributions to effectively span their dynamic range.

2.3 Prediction of N-Glycosylation via Quasi-Supervised Learning

The quasi-supervised learning paradigm addresses the difficulties associated with obtaining ground-truth data sets required for supervised learning in biomedical data analysis tasks [40]. Briefly, for patterns $\{\omega_i\}, i = 1, 2, \ldots, P$, in one of two different collections C_0 and C_1, it computes estimates for the posterior probabilities $\Pr\{C_0 \mid \omega_i\}$ and $\Pr\{C_1 \mid \omega_i\}$ that assess the relative likelihoods for a pattern ω_i of belonging to C_0 and C_1, respectively. The framework is especially powerful in cases where the collection C_0 represents patterns that all share a common property while no knowledge is available on which patterns in C_1 also share the same property. The algorithm can then identify the C_1 patterns ω that are likely to share the property in consideration as those for which

$$\Pr\{C_0 \mid \omega\} \approx \Pr\{C_1 \mid \omega\} \approx 0.5,$$

while those that are exclusively specific to C_1 would be characterized by

$$\Pr\{C_0 \mid \omega\} = 1 - \Pr\{C_1 \mid \omega\} \leq \varepsilon,$$

where ε is a small number. Note that this makes quasi-supervised learning fit the functional and/or structural site prediction problem perfectly: motif vectors of the sites that have been verified experimentally to possess a given property, such as glycosylation, would be collected in C_0, while the others in C_1 as a mixed bunch. The sites in C_1 with motif vectors ω for which the probability $\Pr\{C_0 \mid \omega_i\}$ is discernibly higher than $\Pr\{C_1 \mid \omega_i\}$ would then be predicted to share the property of interest [40].

In order to evaluate the potential of motif vectors for functional site prediction, we have addressed the problem of predicting the glycosylation of all sites possessing the consensus N-glycosylation sequon among amino acid sequences of human proteins. The consensus sequon N-X-S/T consists of an asparagine residue followed by any amino acid X other than proline, and either a serine or a threonine residue [42], [43]. A great majority of N-glycosylation sites adhere to the consensus sequon, and the exceptions are relatively scarce [44].

We have parsed the amino acid sequences of all human proteins for which we computed the motif vectors and identified the sites that possessed the consensus N-glycosylation sequon. We then collected the motif vectors of those that were experimentally validated for N-glycosylation in a collection C_0, and the rest in C_1 along with the ones annotated as probable or potential. Next, we have carried out a modified version of the quasi-supervised learning algorithm to compute the probabilities $\Pr\{C_0 \mid \omega\}$ and $\Pr\{C_1 \mid \omega\}$ for all motif vectors ω. The modifications included technical considerations to prevent the large discrepancy between the collection sizes from biasing the results, and are described in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2012.68. A Matlab implementation of the quasi-supervised learning algorithm is distributed at the internet address http://web.iyte.edu.tr/~bilgekaracali/Projects/QSL/. The sites of C_1 vectors ω for which $\Pr\{C_1 \mid \omega\} < P_c$ were then predicted as likely N-glycosylation sites, where P_c represents a decision threshold between 0 and 1.

We have used three measures to evaluate the separation between the known N-glycosylation vectors in C_0 from those in C_1 that are likely to be nonglycosylated. The first measure, the *probability of detection* denoted by P_D, computed the fraction of C_0 vectors satisfying the condition for predicted glycosylation via

$$P_D = \frac{1}{|C_0|} \sum_{\omega \in C_0} 1(\Pr\{C_1 \mid \omega\} < P_c). \tag{7}$$

The second measure, termed the *probability of quasi-detection*, P_{Q-D},

$$P_{Q-D} = \frac{1}{|C_1|} \sum_{\omega \in C_1} 1(\Pr\{C_1 \mid \omega\} < P_c) \tag{8}$$

computed the fraction of C_1 vectors predicted to be N-glycosylated like the C_0 vectors. In the expressions above, the function $1(\cdot)$ returns 1 if its argument holds true and 0 otherwise, and $|C_0|$ and $|C_1|$ denote the number of vectors in the respective collections. A good prediction performance would be characterized by high detection rates P_D accompanied with low quasi-detection rates P_{Q-D}. Note that the quasi-detection rate P_{Q-D} is intrinsically related to the rate P_R at which the N-glycosylation of the C_1 vectors are rejected via

$$P_{Q-D} = 1 - P_R. \tag{9}$$

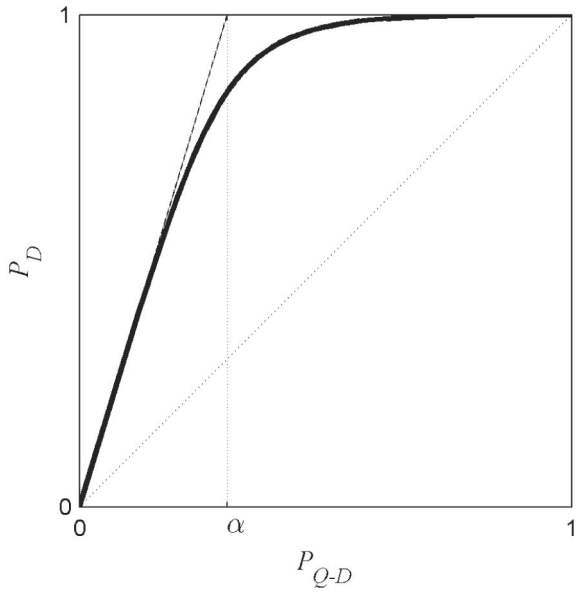

Fig. 2. Illustration of a typical P_D-P_{Q-D} curve. The unknown fraction α of glycosylated sites in C_1 causes the linear rise from the origin. The ideal curve would attain the point $(\alpha, 1)$ and achieve an AUQ-C value of $1 - \alpha/2$.

Finally, we have also computed the area under the P_D-P_{Q-D} curve traced by letting P_c increase gradually from 0 to 1 as a third performance measure, denoted by AUQ-C. A good separation between C_0 and C_1 would result in a large AUQ-C value, while poor separations would produce AUQ-C values around 0.5. Note, however, that since a good portion of the motif vectors in C_1 may in fact be associated with true N-glycosylation sites, the AUQ-C measure is not expected to attain a value of 1. Indeed, if a fraction α of the sites in C_1 is glycosylated, the best AUQ-C measure would reach $1 - \alpha/2$, representing an upper bound on the prediction performance (Fig. 2).

In addition to carrying out the quasi-supervised learning algorithm separately for the decomposition and the approximation motif vectors, we have also subjected the motif vectors to feature selection prior to learning. First, we have computed utility scores u_i via

$$u_i = \frac{\sqrt{|C_0| + |C_1| - 2}\ \ |\mu_{i|0} - \mu_{i|1}|}{\sqrt{(|C_0| - 1)\sigma_{i|0}^2 + (|C_1| - 1)\sigma_{i|1}^2}} \tag{10}$$

for all motif vector components indexed by i, in terms of the sample means $\mu_{i|0}$ and $\mu_{i|1}$, and the sample variances $\sigma_{i|0}^2$ and $\sigma_{i|1}^2$ across C_0 and C_1. The *individual feature selection* strategy then chose the K features with the largest utility scores. Second, we have computed a least squares affine fit

$$L(\omega) \cong \omega^T \omega^{aff} + \omega_0 \tag{11}$$

to the log-likelihood ratio $L(\omega)$ defined by

$$L(\omega) = \log_2 \frac{\Pr\{C_0 \mid \omega\}}{\Pr\{C_1 \mid \omega\}} \tag{12}$$

over the posterior probabilities provided by applying the quasi-supervised learning algorithm on the full motif vectors. The largest components of ω^{aff} in absolute value then identified the features with the greatest contribution to the log-likelihood ratio, through the *affine feature selection*. Finally, we have computed the Fisher's linear discriminant vector ω^{Fld}

$$\omega^{Fld} = (\Sigma_0 + \Sigma_1)^+(\mu_0 - \mu_1), \tag{13}$$

where μ_0 and μ_1 denote the means of the motif vectors, respectively, in C_0 and C_1, and Σ_0 and Σ_1 are the corresponding covariance matrices [30]. The plus sign in the superscript denotes the pseudoinverse operation. Just as before, the greater the component of the vector ω^{Fld} in absolute value, the more significant the feature, and the *Fisher's discriminant feature selection* strategy then collected the K top features.

2.4 Comparative Performance Evaluation

Carrying out a comparative performance evaluation of the proposed method for predicting functional sites is problematic for several reasons. The greatest difficulty is the lack of samples experimentally verified to not possess the functional attribute of interest. However, the prediction task addressed here consists of samples that may potentially exhibit the functional attribute accompanied only by samples that are experimentally verified to do so. In the absence of samples that can be wrong when predicted, the conventional measures of precision rate and the recall rate used in computational biology research to evaluate prediction performance are meaningless. The only option then is to derive measures of separation between the samples predicted to exhibit the functional attribute of interest and those predicted otherwise among all samples lacking experimentally verified functional annotation, as provided by the measures P_D, P_{Q-D} or alternatively P_R, and AUQ-C.

Another difficulty stems from the novelty of the functional site prediction problem without any information of which samples lack the functional attribute. In the specific instance of N-glycosylation prediction, all existing methods improvise a collection of sites that presumably lack glycosylation along with experimentally verified sites into a training data set upon which they base their predictions. Consequently, they are not applicable to the data set analyzed by the proposed method. Conversely, as the proposed method does not anticipate sites specified to lack the functional attribute of interest, it would not receive a fair comparison on the data sets used by the other prediction algorithms.

The only remaining option, then, is to compare the predictions produced by the proposed method against those of the existing N-glycosylation prediction methods on sites possessing the consensus N-glycosylation sequon. We have considered the following methods to carry out this comparison:

1. NetNGlyc available publicly at the internet address http://www.cbs.dtu.dk/services/NetNGlyc/.
2. EnsembleGly available publicly at the internet address http://turing.cs.iastate.edu/EnsembleGly/.
3. GPP available publicly at the internet address http://comp.chem.nottingham.ac.uk/glyco/.

Note that the working data set used here consists of 15,104 human proteins, amounting to a substantial computational load for these publicly available servers. In order to test the predictions, we have therefore selected 100 proteins randomly from the list of proteins that were not included in the training of the prediction servers EnsembleGly and GPP. The training data set of NetNGlyc server was not available for download. We have then computed the $P_D - P_{Q-D}$ graphs for each set of predictions and computed the respective AUQ-C measures for comparative performance evaluation. In order to maintain comparability, we have also reevaluated the predictions obtained by these sites using the proposed method by limiting the computation of the posterior probabilities to use the sites in the remaining 15,004 proteins only.

As a final performance evaluation of the proposed method, we have also subjected the motif vector data presented by the collections C_0 and C_1 to a support vector machine classification presenting the collection C_1 of undetermined sites to the classification algorithm as if they were experimentally verified to lack glycosylation [45], [46]. For the nonlinearity, the Gaussian kernel given by

$$K(\omega_i, \omega_j) = \exp\left(-\frac{\left\|\omega_i - \omega_j\right\|^2}{2\sigma^2}\right), \tag{14}$$

was used where the optimal scale parameter σ was determined to minimize the number of support vectors via line search. The computations were carried out using the SVM^{light} software package available for download at the internet address http://svmlight.joachims.org/. As before, the performance evaluation consisted of computing the AUQ-C measures for the predictions.

3 Results

The hierarchical motif vectors associated with the amino acid sequences of the human proteins included in the study were computed in the Matlab environment (Fig. 3). The wavelet decomposition was carried out up to a maximal decomposition scale $J = 7$ using a Daubechies wavelet of order 4 [47]. Since the decomposition scale governs the amino acid segment sizes over which the variations of the physicochemical properties are evaluated, the calculation of motif vectors was restricted to those proteins with amino acid sequences no shorter than $2^{J+1} = 256\text{aa}$ within the working set of human proteins, also excluding MUC16_HUMAN (Q8WXI7) and TITIN_HUMAN (Q8WZ42) that possessed amino acid sequences longer than 10,000 sites (22,152 and 34,350, respectively) in order to prevent the compositions of these proteins from dominating the analysis. With these exclusions, the number of human proteins included in the analysis reduced to 15,104.

Parsing the amino acid sequences of these proteins identified a total of 55,184 occurrences of the consensus

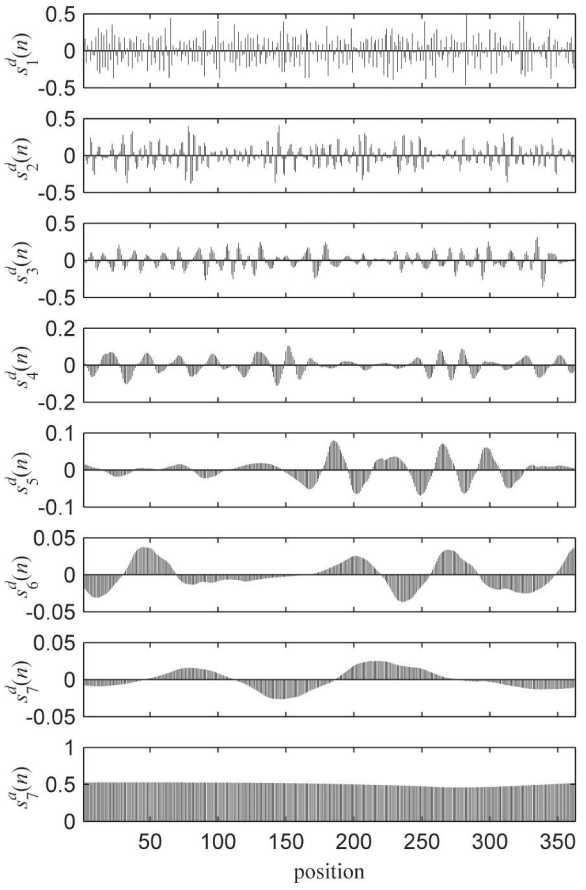

Fig. 3. Wavelet decomposition of the normalized hydrophobicity scale property sequence $s(n)$ for ADA_HUMAN into detail sequences at increasing scales $s_1^d(n), s_2^d(n), \ldots, s_7^d(n)$, ending with the coarsest approximation sequence $s_7^a(n)$, for $n = 1, 2, \ldots, 363$. Note that the short-range variations are captured by the lower level detail sequences, and the longer range variations by the higher level detail sequences. This illustrates the hierarchy in the representation for the select property via the wavelet decomposition.

N-glycosylation motif. Among these, 1,939 were experimentally verified glycosylation sites in the UniProt Knowledgebase data set excluding the potential and probable glycosylation annotations. The motif vectors of these sites constituted the C_0 collection. The motif vectors of the remaining 53,245 consensus sites were pooled into the C_1 collection. The motif vector data were then linearly normalized so that each feature exhibited unit standard deviation across all motif vectors.

The quasi-supervised learning algorithm was carried out on the motif vector data using the configurations described in Section 2. The AUQ-C measures obtained for each configuration revealed that the best separation of the motif vectors associated with experimentally verified consensus N-glycosylation sites in C_0 and the unknown prospects in C_1 is achieved using the Fisher's discriminant features selection on approximation motif vectors, at a value of 0.7708 using $K = 150$ features (Table 2). The corresponding P_D-P_{Q-D} curve indicates that about 60 percent of the consensus sites in C_1 can be rejected while maintaining accurate prediction of 80 percent of the experimentally verified N-glycosylation sites in C_0 (Fig. 4).

The list of 100 proteins used in the comparative performance evaluation experiments contained 368 sites possessing the consensus N-glycosylation sequon, out of which only 19 were experimentally verified for glycosylation. For comparison purposes, AUQ-C measures were computed on this limited data set from the predictions provided by the NetNGlyc and the EnsembleGly servers as

TABLE 2
AUQ-C Measures Associated with Different Predictor Configurations

Decomposition motif vectors				
no selection	*K*= 1944	0.6901		
feature selection		individual	affine	Fisher's disc.
	K = 5	0.6926	0.6999	0.7064
	K = 10	0.7056	0.7229	0.7285
	K = 20	0.7181	0.7387	0.7585
	K = 50	0.7372	0.7484	0.7637
	K = 100	0.7380	0.7420	0.7588
	K = 150	0.7427	0.7382	0.7579
	K = 170	0.7393	0.7365	0.7594
	K = 200	0.7347	0.7336	0.7561
	K = 500	0.7238	0.7262	0.7374
Approximation motif vectors				
no selection	*K*= 1701	0.7297		
feature selection		individual	affine	Fisher's disc.
	K = 5	0.6926	0.7077	0.6971
	K = 10	0.7056	0.7205	0.7179
	K = 20	0.7193	0.7366	0.7564
	K = 50	0.7399	0.7593	0.7678
	K = 100	0.7419	0.7681	0.7642
	K = 150	0.7374	0.7695	0.7708
	K = 170	0.7400	0.7690	0.7694
	K = 200	0.7384	0.7670	0.7693
	K = 500	0.7335	0.7526	0.7621

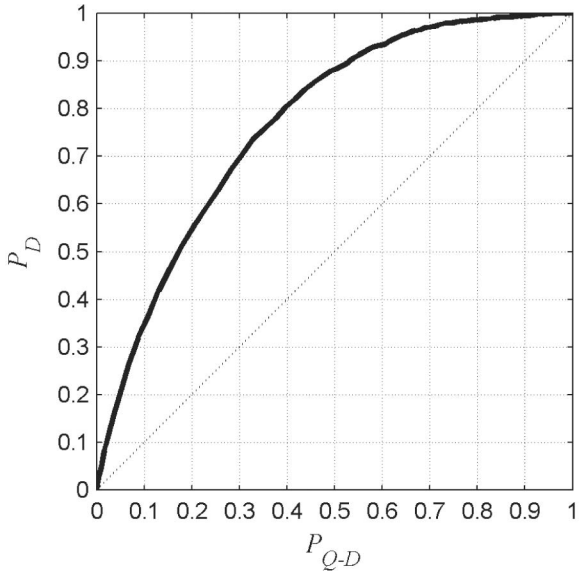

Fig. 4. The P_D-P_{Q-D} curve associated with the best performing N-glycosylation predictor configuration. The corresponding AUQ-C measure was 0.7708.

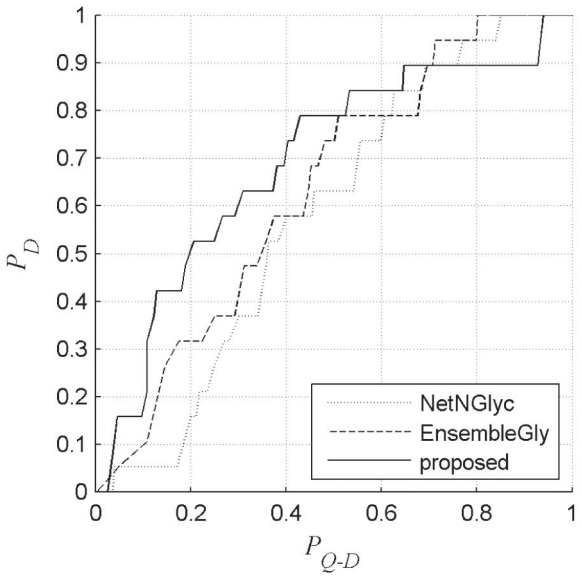

Fig. 5. The P_D-P_{Q-D} curves associated with the predictions by the NetNGlyc and EnsembleGly servers along with the proposed method on a limited data set of 100 proteins. The corresponding AUQ-C measures were 0.5814, 0.6340, and 0.6918, respectively.

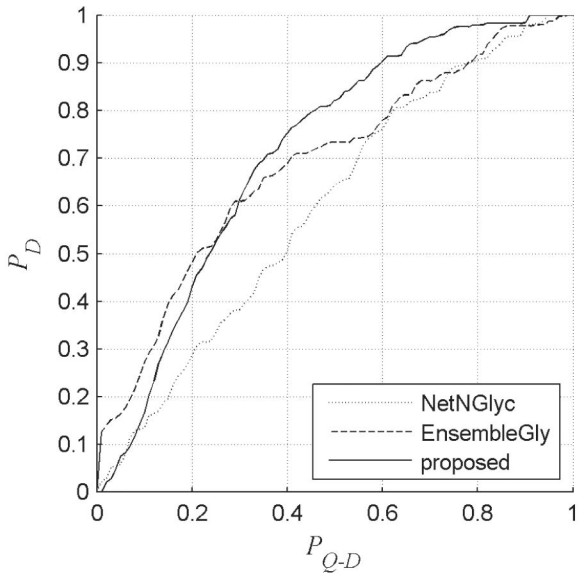

Fig. 6. The average P_D-P_{Q-D} curves associated with the predictions by the NetNGlyc and EnsembleGly servers along with the proposed method on a total of 10 randomized prediction experiments over 100 proteins selected independently for each case.

well as the proposed method, that were 0.5814, 0.6340, and 0.6918, respectively. In computing the predictions by the proposed method, the quasi-supervised algorithm was operated on the sites on the test proteins using those corresponding to the remaining 15,004 proteins only.

The corresponding P_D-P_{Q-D} curves in Fig. 5 indicate that at the default thresholds for glycosylation scores, the proposed method attained a much better separation of the sites that are glycosylated from the undetermined sites, correctly predicting 78.95 percent of the known glycosylation sites while rejecting 52.44 percent among the 349 prospective sites using $P_c = 0.5$ as the prediction threshold. These rates were 89.47 percent correct prediction at 28.65 percent rejection for the NetNGlyc predictions and 100.00 percent correct prediction at 2.58 percent rejection for the EnsembleGly predictions. The GPP server provided predictions only on 31 proteins out of the 100 submitted, containing 122 consensus N-glycosylation sites 19 of which were experimentally validated. The GPP predictions did not possess a predictive score and neither a P_D-P_{Q-D} curve nor an AUQ-C measure could be derived, but all 122 sites were predicted to be glycosylated, amounting to 100.00 percent correct prediction with 0.00 percent rejection.

In order to achieve a more general comparison of the prediction performance achieved by the proposed method and the predictions by the NetNGlyc and EnsembleGly webservers, we have repeated the above experiment 10 more times, each time selecting an independent set of 100 random proteins. The GPP webserver was excluded from this analysis as it did not allow computation of P_D-P_{Q-D} curves. Out of these 10 experiments, the NetNGlyc and EnsembleGly webservers produced predictions only for eight cases, rejecting the remaining 2 due possibly to larger than allowed sequence lengths. The average P_D-P_{Q-D} curves in Fig. 6 show good agreement with those in Fig. 5: the P_D-P_{Q-D} curve of the proposed method rises faster to the $P_D = 1$ level and covers a greater area than those obtained for the NetNGlyc and EnsembleGly predictions. Note that the initial quick rise observed in the P_D-P_{Q-D} curve of the EnsembleGly predictions falls outside of the theoretically expected linear behavior illustrated in Fig. 2, and can thus be attributed to chance factors. The average AUQ-C measures also sustain these observations, with the proposed method leading the others at 0.7125, with the NetNGlyc and EnsembleGly servers trailing at 0.5979 and 0.6873, respectively. The EnsembleGly predictions were obtained using an ensemble of support vector machine classifiers. The averages for the NetNGlyc and EnsembleGly predictions were calculated using only the eight cases for which these webservers produced predictions, while the averages for the proposed method were calculated using predictions obtained for all 10 cases.

The comparative evaluation of the proposed method against support vector classification could not be carried out in an exhaustive manner for all motif vector configurations summarized in Table 2 due to the colossal computational expense associated with the algorithm that performs the quadratic optimization in support vector machine training. Instead, the predictive performance of the support vector classification was assessed on the configuration that maximized the performance of the proposed method; using 150 features determined by the Fisher's linear discriminant feature selection method on the approximation motif vectors, and via 10-fold cross validation. At each cross-validation experiment, a true positive data set of 1,843 experimentally verified glycosylation sites was paired with 13,157 undetermined sites in a presumed true negative data set, both selected randomly. Note that this amounts to a collection of 15,000 points to be used for training, leaving the remaining 40,184 sites including 96 experimentally verified to be modified by glycosylation. The classifier construction was carried out on this training set, and the resulting classifier acted upon the sites that were not included in the training set. In each case, a separate P_D-P_{Q-D} curve was computed, and the corresponding AUQ-C measures were calculated for a final assessment. In the experiments, the

TABLE 3
List of Top 15 Sites Most Likely to be Modified by N-Glycosylation

$\Pr\{C_0 \mid \omega\}$	Protein name	AC code	Position	Existing annotation
0.8467	Methionyl-tRNA synthetase, cytoplasmic	P56192	574	
0.8294	Protein HEG homolog 1	Q9ULI3	1317	
0.8287	5,6-dihydroxyindole-2-carboxylic acid oxidase	P17643	385	N-linked (GlcNAc...) (Potential)
0.8281	Low-density lipoprotein receptor-related protein 2	P98164	3355	N-linked (GlcNAc...) (Potential)
0.8230	Phlorizin hydrolase	P09848	1340	
0.8221	Lysosomal thioesterase PPT2	Q9UMR5	206	N-linked (GlcNAc...) (Potential)
0.8217	Methionyl-tRNA synthetase, cytoplasmic	P56192	531	
0.8206	LIM domain-binding protein 1	Q86U70	65	
0.8203	Interleukin-31 receptor subunit alpha	Q8NI17	67	N-linked (GlcNAc...) (Potential)
0.8189	Extracellular sulfatase Sulf-1	Q8IWU6	148	N-linked (GlcNAc...) (Potential)
0.8176	Interleukin-13 receptor subunit alpha-2	Q14627	115	N-linked (GlcNAc...) (Potential)
0.8159	Anoctamin-2	Q9NQ90	856	N-linked (GlcNAc...) (Potential)
0.8159	PDZ domain-containing protein 8	Q8NEN9	99	
0.8156	Mucin-2	Q02817	1154	N-linked (GlcNAc...) (Potential)
0.8128	Nucleotide pyrophosphatase	P22413	578	

The predictions show good agreement with existing potential annotations in the UniProt/Swiss-Prot Knowledgebase while identifying novel potential N-glycosylation sites.

average AUQ-C measure was 0.7466, indicating that the predictions by the quasi-supervised learning algorithm provided higher rejection rates of the sites not verified for glycosylation for the same rates of correctly predicting the experimentally verified sites.

The superior performance of the proposed method against the alternatives can be attributed to several factors. First and foremost, the quasi-supervised strategy is attuned by design to the nature of the recognition problem that offers only the experimentally verified N-linked glycosylation sites along with sites that can potentially be also glycosylated, but no sites that have been verified to lack N-linked glycosylation. The alternative strategies are based on supervised learning, and require a presumed data set for negative N-linked glycosylation. As discussed previously, in the absence of such data sets, an improvised negative data set is to be formed from among the potential N-linked glycosylation sites that may very well be glycosylated and only waiting to be identified as such. Without reliable positive and negative data sets, supervised learning algorithms cannot produce reliable predictions.

In addition, note also that the proposed method forms predictions based on all evidence available in the sequence data. The alternative strategies, on the other hand, operate on limited data sets aggravated further by the lack of a bona fide negative data set. An additional limitation impeded the application of a support vector machine classification strategy to the motif vector data as it was not possible to train such a classifier on all 55,184 motif vectors due to the colossal computational expense involved in solving the underlying quadratic optimization problem.

On a final performance evaluation, the list of top C_1 sites most likely to be N-glycosylated in Table 3 showed good agreement with the existing probable and potential glycosylation annotations in the UniProt Knowledgebase data set. The sites lacking any annotation are therefore identified as novel N-glycosylation sites. At the present time, however, there is no way of knowing whether these predictions are indeed true as there are no published results on their glycosylation status, and the true performance test of any prediction method invariably has to come from experimental validation. Yet, the existing literature does provide some support for these findings. For instance, the phlorizin hydrolase (P09848) is known to be modified by glycosylation, and the consensus N-glycosylation motif at position 1,340 provides a likely candidate site [48], [49]. The case of the cytoplasmic methionyl-tRNA synthetase (P56192) is intriguing, as there is evidence that the activity of its close homologue in mice is altered in the presence of the N-glycosylation inhibitor tunicamycin in HeLa cells [50]. Consequently, these sites provide prime candidates for investigating the glycosylation of the corresponding proteins.

4 Discussions and Conclusion

In this paper, we have introduced the hierarchical motif vectors for functional characterization of sites along amino acid sequences. Following the initial use of the motif vector approach to amino acid alignment [39], this represents the first application of the hierarchical motif vectors to protein characterization. As the motif vectors encode the variation of physico-chemical properties at short, mid, and long ranges along both directions in an amino acid sequence, they capture the structural and functional characteristics at and around their respective sites. Similarity of physico-chemical configurations around sites then translates into the similarity of the corresponding motif vectors. Furthermore, the motif vector perspective of sequence analysis formulates the problem of sequence characterization in a vector space organization very much suited to quantitative evaluation via statistical learning algorithms.

Another critical component of the proposed method for site characterization is the quasi-supervised learning algorithm that allowed a global evaluation of all 53,245 candidate sites possessing the consensus N-glycosylation sequon

identified across 15,104 human proteins against the backdrop of only 1,939 experimentally verified sites, without requiring any specification on negative glycosylation sites. This is a substantial benefit of the proposed method as it eliminates all the risks associated with limited or improvised true-negative data sets that afflict the existing prediction methods based on conventional supervised statistical learning algorithms. While the best AUQ-C measure of prediction performance in Table 2 stands noticeably lower than 1.0000 at 0.7708, it should be pointed out that if 40 percent of the consensus sites populating C_1 is in fact glycosylated, the AUQ-C of the best possible predictor would be bound from above by 0.8000. Nevertheless, the smoothness of the P_D-P_{Q-D} curve in Fig. 4 suggests a considerable overlap between the motif vector distributions of the experimentally verified N-glycosylation sites and the others, indicating that the motif vector representation of local amino acid configurations can be improved further for an even higher prediction performance.

As an additional benefit, the quasi-supervised learning algorithm did not require cross-validation tests for performance evaluation as the posterior probabilities $\Pr\{C_0 \mid \omega\}$ and $\Pr\{C_1 \mid \omega\}$ are computed for each motif vector ω with no regard to the collection to which it belongs [40]. This saves substantial amounts of time in the analysis as well as limiting the susceptibility of the technique to issues related with data overfitting. The same, however, is not true for conventional classification methods such as support vector machines that require extensive cross-validation experiments in order to derive a P_D-P_{Q-D} curve.

On a final note, it should be pointed out that the results presented here demonstrate the prediction power of motif vector representation coupled with the quasi-supervised learning algorithm on large-scale N-glycosylation data at a combined collection size of 55,184 vectors. On the other hand, in order for this strategy to be applicable to arbitrary structural or functional prediction tasks, the learning algorithm has to be adapted to very large scales allowing joint evaluation of all 11,227,834 sites in the working set of amino acid sequences of human proteins and possibly more from other species. While this is a daunting task even for the simplest of the statistical learning methods, efforts are currently under way to advance the technique to such an analytical capacity.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the European Commission (PIRG03-GA-2008-230903).

REFERENCES

[1] L. Parida, *Pattern Discovery in Bioinformatics: Theory & Algorithms.* Chapman and Hall/CRC, 2008.

[2] M. Reczko and H. Bohr, "The Def Data-Base of Sequence Based Protein Fold Class Predictions," *Nucleic Acids Research,* vol. 22, pp. 3616-3619, Sept. 1994.

[3] M. Bhasin and G.P.S. Raghava, "Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition," *J. Biological Chemistry,* vol. 279, pp. 23262-23266, May 2004.

[4] S.J. Hua and Z.R. Sun, "Support Vector Machine Approach for Protein Subcellular Localization Prediction," *Bioinformatics,* vol. 17, pp. 721-728, Aug. 2001.

[5] J.K. Vries, X. Liu, and I. Bahar, "The Relationship between N-Gram Patterns and Protein Secondary Structure," *Proteins-Structure Function and Bioinformatics,* vol. 68, pp. 830-838, Sept. 2007.

[6] A.M. Facchiano and S. Costantini, "Prediction of the Protein Structural Class by Specific Peptide Frequencies," *Biochimie,* vol. 91, pp. 226-229, Feb. 2009.

[7] S. Anishetty, R. Anishetty, and G. Pennathur, "Understanding Mutations and Protein Stability through Tripeptides," *FEBS Letters,* vol. 580, pp. 2071-2080, Apr. 2006.

[8] A. Ceroni and P. Frasconi, "On the Role of Long-Range Dependencies in Learning Protein Secondary Structure," *Proc. IEEE Int'l Joint Conf. Neural Networks,* vol. 3, pp. 1899-1904, 2004.

[9] D. Kihara, "The Effect of Long-Range Interactions on the Secondary Structure Formation of Proteins," *Protein Science,* vol. 14, pp. 1955-1963, Aug. 2005.

[10] Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, and Y.Z. Chen, "PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence," *Nucleic Acids Research,* vol. 34, pp. W32-W37, 2006.

[11] Z.R. Li, H.B. Rao, F. Zhu, G.B. Yang, and Y.Z. Chen, "Update of PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence," *Nucleic Acids Research,* vol. 39, pp. W385-W390, July 2011.

[12] C. Chen, L.X. Chen, X.Y. Zou, and P.X. Cai, "Predicting Protein Structural Class Based on Multi-Features Fusion," *J. Theoretical Biology,* vol. 253, pp. 388-392, July 2008.

[13] T.L. Bailey and C. Elkan, "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization," *Machine Learning,* vol. 21, pp. 51-80, Oct./Nov. 1995.

[14] T.L. Bailey, N. Williams, C. Misleh, and W.W. Li, "MEME: Discovering and Analyzing DNA and Protein Sequence Motifs," *Nucleic Acids Research,* vol. 34, pp. W369-W373, July 2006.

[15] C.E. Lawrence and A.A. Reilly, "An Expectation Maximization (Em) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences," *Proteins-Structure Function and Genetics,* vol. 7, pp. 41-51, 1990.

[16] S. Balla, V. Thapar, S. Verma, T. Luong, T. Faghri, C.H. Huang, S. Rajasekaran, J.J. del Campo, J.H. Shinn, W.A. Mohler, M.W. Maciejewski, M.R. Gryk, B. Piccirillo, S.R. Schiller, and M.R. Schiller, "Minimotif Miner: A Tool for Investigating Protein Function," *Nature Methods,* vol. 3, pp. 175-177, Mar. 2006.

[17] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D.M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W.N. Hunter, R. Aasland, and T.J. Gibson, "ELM Server: A New Resource for Investigating Short Functional Sites in Modular Eukaryotic Proteins," *Nucleic Acids Research,* vol. 31, pp. 3625-3630, July 2003.

[18] A. Bairoch, "PROSITE: A Dictionary of Sites and Patterns in Proteins," *Nucleic Acids Research,* vol. 19, no. Suppl, pp. 2241-2245, Apr. 1991.

[19] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B.A. Cuche, E. de Castro, C. Lachaize, P.S. Langendijk-Genevaux, and C.J. Sigrist, "The 20 Years of PROSITE," *Nucleic Acids Research,* vol. 36, pp. D245-D249, Jan. 2008.

[20] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, M. Pagni, and C.J. Sigrist, "The PROSITE Database," *Nucleic Acids Research,* vol. 34, pp. D227-D230, Jan. 2006.

[21] L.Y. Geer, M. Domrachev, D.J. Lipman, and S.H. Bryant, "CDART: Protein Homology by Domain Architecture," *Genome Research,* vol. 12, pp. 1619-1623, Oct. 2002.

[22] N.C.W. Goonesekere and B. Lee, "Context-Specific Amino Acid Substitution Matrices and Their Use in the Detection of Protein Homologs," *Proteins-Structure Function and Bioinformatics,* vol. 71, pp. 910-919, May 2008.

[23] J.G. Henikoff, S. Pietrokovski, C.M. McCallum, and S. Henikoff, "Blocks-Based Methods for Detecting Protein Homology," *Electrophoresis,* vol. 21, pp. 1700-1706, May 2000.

[24] S. Hunter, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R.D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A.F. Quinn, J.D. Selengut, C.J.A. Sigrist, M. Thimma, P.D. Thomas, F. Valentin, D. Wilson, C.H. Wu, and C. Yeats, "InterPro: The Integrative Protein Signature Database," *Nucleic Acids Research,* vol. 37, pp. D211-D215, Jan. 2009.

[25] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, and A. Bateman, "The Pfam Protein Families Database," *Nucleic Acids Research,* vol. 38, pp. D211-D222, Jan. 2010.

[26] I. Letunic, T. Doerks, and P. Bork, "SMART 6: Recent Updates and New Developments," *Nucleic Acids Research,* vol. 37, pp. D229-D232, Jan. 2009.

[27] J. Schultz, F. Milpetz, P. Bork, and C.P. Ponting, "SMART, a Simple Modular Architecture Research Tool: Identification of Signaling Domains," *Proc. Nat'l Academy Sciences USA,* vol. 95, pp. 5857-5864, May 1998.

[28] C.J. Sigrist, L. Cerutti, E. de Castro, P.S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo, "PROSITE, a Protein Domain Database for Functional Characterization and Annotation," *Nucleic Acids Research,* vol. 38, pp. D161-D166, Jan. 2010.

[29] C. Caragea, J. Sinapov, A. Silvescu, D. Dobbs, and V. Honavar, "Glycosylation Site Prediction Using Ensembles of Support Vector Machine Classifiers," *BMC Bioinformatics,* vol. 8, article 438, 2007.

[30] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* second ed. Wiley-Interscience, 2000.

[31] N. Blom, T. Sicheritz-Ponten, R. Gupta, S. Gammeltoft, and S. Brunak, "Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence," *Proteomics,* vol. 4, pp. 1633-1649, June 2004.

[32] K. Julenius, A. Molgaard, R. Gupta, and S. Brunak, "Prediction, Conservation Analysis, and Structural Characterization of Mammalian Mucin-Type O-Glycosylation Sites," *Glycobiology,* vol. 15, pp. 153-164, Feb. 2005.

[33] T.P. Knepper, B. Arbogast, J. Schreurs, and M.L. Deinzer, "Determination of the Glycosylation Patterns, Disulfide Linkages, and Protein Heterogeneities of Baculovirus-Expressed Mouse Interleukin-3 by Mass Spectrometry," *Biochemistry,* vol. 31, pp. 11651-11659, Nov. 1992.

[34] S.E. Hamby and J.D. Hirst, "Prediction of Glycosylation Sites Using Random Forests," *BMC Bioinformatics,* vol. 9, article 500, 2008.

[35] S. Li, B. Liu, R. Zeng, Y. Cai, and Y. Li, "Predicting O-Glycosylation Sites in Mammalian Proteins by Using SVMs," *Computational Biology and Chemistry,* vol. 30, pp. 203-238, June 2006.

[36] Y. Gavel and G. von Heijne, "Sequence Differences between Glycosylated and Non-Glycosylated Asn-X-Thr/Ser Acceptor Sites: Implications for Protein Engineering," *Protein Eng.,* vol. 3, pp. 433-442, Apr. 1990.

[37] R.W. Carrell, J.O. Jeppsson, L. Vaughan, S.O. Brennan, M.C. Owen, and D.R. Boswell, "Human Alpha 1-antitrypsin: Carbohydrate Attachment and Sequence Homology," *FEBS Letters,* vol. 135, pp. 301-303, Dec. 1981.

[38] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[39] B. Karaçalı, "Hierarchical Motif Vectors for Amino Acid Sequence Alignment," *Proc. Ninth IASTED Int'l Conf. Biomedical Eng.,* 2012.

[40] B. Karaçalı, "Quasi-Supervised Learning for Biomedical Data Analysis," *Pattern Recognition,* vol. 43, pp. 3674-3682, 2010.

[41] V.S. Mathura and D. Kolippakkam, "APDbase: Amino Acid Physico-Chemical Properties Database," *Bioinformation,* vol. 1, pp. 2-4, 2005.

[42] A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, G.W. Hart, and M.E. Etzler, *Essentials of Glycobiology,* second ed. Cold Spring Harbor Laboratory Press, 2008.

[43] E. Weerapana and B. Imperiali, "Asparagine-Linked Protein Glycosylation: From Eukaryotic to Prokaryotic Systems," *Glycobiology,* vol. 16, pp. 91R-101R, June 2006.

[44] J.P. Miletich and G.J. Broze Jr., "Beta Protein C is Not Glycosylated at Asparagine 329, The Rate of Translation may Influence the Frequency of Usage at Asparagine-X-Cysteine Sites," *J. Biological Chemistry,* vol. 265, pp. 11397-11404, July 1990.

[45] V.N. Vapnik, *The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science),* second ed. Springer-Verlag, 1999.

[46] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, pp. 273-297, Sept. 1995.

[47] I. Daubechies, *Ten Lectures on Wavelets.* SIAM, 1992.

[48] E.M. Danielsen, H. Skovbjerg, O. Noren, and H. Sjostrom, "Biosynthesis of Intestinal Microvillar Proteins, Intracellular Processing of Lactase-Phlorizin Hydrolase," *Biochemical and Biophysical Research Comm.,* vol. 122, pp. 82-90, July 1984.

[49] H.Y. Naim, E.E. Sterchi, and M.J. Lentze, "Biosynthesis and Maturation of Lactase-Phlorizin Hydrolase in the Human Small Intestinal Epithelial Cells," *Biochemical J.,* vol. 241, pp. 427-434, Jan. 1987.

[50] N. Netzer, J.M. Goodenbour, A. David, K.A. Dittmar, R.B. Jones, J.R. Schneider, D. Boone, E.M. Eves, M.R. Rosner, J.S. Gibbs, A. Embry, B. Dolan, S. Das, H.D. Hickman, P. Berglund, J.R. Bennink, J.W. Yewdell, and T. Pan, "Innate Immune and Chemically Triggered Oxidative Stress Modifies Translational Fidelity," *Nature,* vol. 462, pp. 522-526, Nov. 2009.

Bilge Karaçali received the master's degree in 1999, and then the doctor of philosophy degree in 2002, both in electrical engineering, with a minor in mathematics, from the Electrical and Computer Engineering Department, North Carolina State University. He has worked in the field of biomedical image analysis as a postdoctoral research fellow in the Radiology Department of the School of Medicine, University of Pennsylvania, before joining the School of Biomedical Engineering, Science and Health Systems at Drexel University in 2005 as a research assistant professor and the assistant director of bioimaging of the Center for Integrated Bioinformatics. Since 2008, he has been with the Electrical and Electronics Engineering Department of Izmir Institute of Technology, where he serves as an associate professor and director of the Biomedical Information Processing Laboratory.