

A Novel Approach for Prediction of Protein Subcellular Localization from Sequence Using Fourier Analysis and Support Vector Machines

Zhengdeng Lei
Department of Bioengineering
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607, USA
zlei2@uic.edu

Yang Dai^{*}
Department of Bioengineering
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607, USA
yangdai@uic.edu

ABSTRACT

A novel method is presented for the prediction of protein subcellular localization from sequence using Fourier analysis and support vector machines. To extract the features of a protein sequence, each amino acid is replaced by a value representing its scale of hydrophobicity and then a fast Fourier transform is applied to the numerically encoded sequence. The transformed sequence data are then used as the input for the training of support vector machines to predict subcellular localization. The motivation for this method of encoding resides fundamentally on (1) the fact that periodicities are critically important factors in protein structure and (2) the ability of this method to capture information about long-range correlations and global symmetries which are completely missed by approaches based on global amino acid composition. Our method is evaluated against the integrated system PSORT-B for the prediction of subcellular localizations of proteins in Gram-negative bacteria. It is demonstrated that the new method outperforms PSORT-B in prediction for the inner membrane, the outer membrane, and extra cellular localizations in a 5-fold cross-validation. It is expected that integrated systems such as PSORT-B may benefit from inclusion of the advanced individual predictor presented in this paper.

Keywords

Protein Subcellular Localization, Gram-negative bacteria, Fourier Transform, Support Vector Machine.

1. INTRODUCTION

Advances in proteomics and genome sequencing are generating enormous numbers of genes and proteins. The development of automated systems for the annotation of protein structure and function has become extremely important. Since many cellular functions are compartmentalized in specific regions of the cell, subcellular localization of a protein is biologically highlighted as a key element in understanding

its function. Specific knowledge of subcellular localization can inform and direct further experimental studies of proteins.

Several methods and systems have been developed during the last decade for the predictive task of protein localization. Machine learning methods such as Artificial Neural Networks, the k -nearest neighbor method, and Support Vector Machines (SVM) have been utilized in conjunction with various modalities of feature extraction from protein sequences. Most of the early approaches employed the amino acid composition and the di-peptide frequency [7; 13; 26] to represent sequences. This method may miss the information on sequence order and the inter-relationships between the amino acids. In order to overcome this shortcoming, it has been shown that motifs, frequent-subsequences, and functional domains, which are obtained from various databases (SMART, InterPro, PROSITE) or extracted using Hidden Markov Models and data mining techniques, can be used for the representation of protein sequences for the prediction of subcellular localizations [2; 3; 6; 28; 29]. Methods have also been developed based on the use of the N-terminal sorting signals [1; 5; 10; 20; 22; 23; 24] and sequence homology searching [21].

It has become clear that no single method of prediction can achieve high predictive accuracy for all localizations. Therefore, most robust methods adopt an integrative approach by combining several methods, each of which may be a suitable predictor for a specific localization or a generic predictor for all localizations. PSORT is an example of such a successful system. Developed by Nakai and Kanehisa [23], PSORT, recently upgraded to PSORT II [12; 22], is an expert system that can distinguish between different subcellular localizations in eukaryotic cells. It also has a dedicated subsystem PSORT-B for bacterial sequences [9]. Obviously, further improvement of the quality of such an integrated system relies on advances in the individual predictors, namely, improvements that arise from the employ of sophisticated protein encoding schemes and powerful machine learning and data mining techniques.

In this study, we describe a new approach for the prediction of protein subcellular localization from protein sequences using Fourier analysis as the feature extracting tool and sup-

^{*}Corresponding author.

port vector machines as the learning framework. In order to extract the features from a given protein sequence, each amino acid is replaced by a value representing its scale of hydrophobicity and a fast Fourier transform is subsequently applied to the numerically encoded sequence. These transformed data are then trained by support vector machines.

Fourier analysis has been used for (1) the recognition of protein folds [27] and gene-encoding regions of DNA sequences [8; 30] and (2) the detection of periodic patterns and tandem repeats of residues in both DNA and protein sequences [25]. The motivation for this method of encoding resides fundamentally on the observation that periodicities are critically important factors in protein structure [27]. The approach based on the Fourier transform analysis is capable of capturing information about long-range correlations and global symmetries; both are completely missed by approaches based on global amino acid composition. For comparison, we also present another encoding method based on the tri-peptide frequency. This encoding scheme is an extension of the method using the amino acid decomposition and has been used for the prediction of protein folds [18].

Our method is evaluated against PSORT-B for the prediction of subcellular localizations for Gram-negative bacteria [9]. It is demonstrated by the result of a 5-fold cross-validation that the new method outperforms PSORT-B predictions associated with the outer membrane, the inner membrane, and extra cellular localizations. It is expected that PSORT-B may benefit from the integration of this new predictor into the system.

2. METHOD

This section introduces two sequence encoding methods. One is the encoding method based on the Fourier analysis of protein sequences; the other is based on the tri-peptide frequency. The latter approach has been used in protein fold recognition [18], but has never been evaluated for the prediction of subcellular localizations. We also present a short description of support vector machines, the machine learning method used in this study.

2.1 Feature Extraction based on the Fourier Transform

There are many ways to describe amino acids, most of which are correlated to some degree. For example, the AAindex database contains indices representing 434 different physico-chemical and biological properties of amino acids [16]. We concentrate on the amino-acid hydrophobicity in this work, as it is the one of major properties influencing the structure and function of a protein [14]. A simple three-state hydrophobicity scale is used to map hydrophobic residues to 1, hydrophilic residues to -1 , and "neutral" residues to 0 [27]. More precisely,

$$(A, C, F, I, L, M, V) \rightarrow 1,$$

$$(D, E, H, K, N, Q, R) \rightarrow -1,$$

and

$$(G, P, S, T, W, Y) \rightarrow 0.$$

Once a protein sequence has been encoded into the above numerical format, it is converted to a sequence in the frequency

domain with a Fourier transform. A common use of the Fourier transform is the identification of frequency components of a weak time-dependent signal buried in noise. Prior to the application of the Fourier transform, the numerical sequences have to be lengthened by padding with zeros, since the length of the input sequences is required to be a power of two. Let $n = 2^M$ denote the smallest number that is greater than or equal to the length of the longest protein sequence in a given set, where M is some integer. Let $\{x(1), \dots, x(n)\}$ be the numerically encoded sequence of a protein according to the three-state hydrophobicity scale after padding. The Fast Fourier Transform (FFT) will transform the encoded sequence into another sequence $\{X(1), \dots, X(n)\}$ in the frequency domain. The procedure of the FFT used in this research is based on the algorithm of Masters [19], which is an implementation of the discrete Fourier transform (DFT) given by

$$X(f) = \sum_{t=1}^n x(t) \exp[i(2\pi tf/n)] \quad (f = 1, \dots, n),$$

and

$$x(t) = \frac{1}{n} \sum_{f=1}^n X(f) \exp[-i(2\pi tf/n)] \quad (t = 1, \dots, n).$$

Figures 1-3 present the encoded sequences before and after the application of the FFT for two representative proteins from extra cellular, inner membrane, and outer membrane localizations, respectively. The sequences obtained from the FFT display enhanced characteristics for each localization in comparison with the sequences before the use of the FFT.

Another advantage of the FFT based feature extraction is that the number of extracted features is almost the same as the length of the longest protein sequence in the data. This is a compact representation for protein sequences in contrast to the features extracted based on the tri-peptide frequency described below.

2.2 Feature Extraction based on the Tri-peptide Frequency

In order to evaluate the FFT encoding method presented above, an approach based on the tri-peptide frequency for feature extraction has also been considered. This encoding method extends the concept of the amino acid composition and di-peptide frequency encoding methods. These have been used intensively for the representation of protein sequences in numerous applications. These are, for example, the prediction of (1) protein secondary structures, (2) protein folds, and (3) subcellular localizations, and the efficacy of these encoding methods has been established.

In order to encode a protein sequence with the tri-peptide frequency, a vector of $21^3 = 9261$ dimensions is required. Each entry of the vector is associated with a possible pattern of three amino acids. Since the symbol "X" may appear in some sequences, it is added to the set of the original 20 symbols of the amino acids to give a total of 21. A window with a length of three is moved along the sequence from the first amino acid to the third amino acid from the end. Every 3-letter pattern that appears in the window is recorded with increments of 1 in the corresponding entry of the vector. Upon the termination of this procedure, the vector provides the tri-peptide frequency of the sequence.

The final vector is normalized by dividing the number of window positions associated with that sequence. Note that the resulting vector is sparse, as only a small collection of the possible 3-letter patterns will appear in each protein sequence.

2.3 Support Vector Machine

Suppose that we are given a set of m points \mathbf{x}_i ($1 \leq i \leq m$) in an n -dimensional space. Each point \mathbf{x}_i is labeled by $y_i \in \{1, -1\}$ denoting the membership of the point. An SVM is a learning method for binary classification. Using a nonlinear transformation ϕ , it maps the data to a high dimensional feature space in which a linear classification is performed. It is equivalent to solving the quadratic optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_1, \dots, \xi_m} \quad & \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y_i (\phi(\mathbf{x}_i) \cdot \mathbf{w} + b) \geq 1 - \xi_i \quad (i = 1, \dots, m), \\ & \xi_i \geq 0 \quad (i = 1, \dots, m), \end{aligned}$$

where C is a parameter. The decision function is defined as $f(\mathbf{x}) = \phi(\mathbf{x}) \cdot \mathbf{w} + b$, where $\mathbf{w} = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i)$ and α_i ($i = 1, \dots, m$) are nonnegative constants determined by the dual problem of the optimization defined above. Therefore, the function is

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b$$

through the definition of the appropriate kernel function K . For details of SVMs refer to Cristianini and Shawe-Taylor [4].

3. RESULTS AND DISCUSSION

We employed the SVMs in conjunction with the features extracted by the methods described above for training and testing. The evaluation of the methods was conducted on the following dataset.

3.1 Dataset

The set of proteins from Gram-negative bacteria used in the evaluation of PSORT-B [9] was considered (available at <http://www.psort.org/>) in this experiment. It consists of 1443 proteins with experimentally determined localizations. The dataset comprises 1302 proteins resident at a single localization site: 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane, and 190 extra cellular; it additionally contains a set of 141 proteins resident at multiple localization sites: 14 cytoplasmic/inner membrane, 50 inner membrane/periplasmic, and 77 outer membrane/extracellular. In our experiment, we considered only the 1302 proteins possessing a single localization. The longest protein sequence in this dataset is about 4000 amino acids, so the length of the final FFT encoded sequences is approximately, 2000.

3.2 Experiment and Results

We have compared the performance of our new methods with that of PSORT-B, a powerful tool for the prediction of protein subcellular localization for Gram-negative bacteria.

The system PSORT-B was designed to seek precision other than recall to allow for confident predictions, and prevents

the propagation of erroneous predictions. It utilizes six modules for the generation of an overall prediction of a localization site:

- (1) BLAST search based predictor *SCL-BLAST* for all localizations [21];
- (2) Motif based predictor *Motif* for all localization sites [23];
- (3) Hidden Markov Model based predictor *HMMTOP* for the inner membrane localization [28; 29];
- (4) Motif based predictor *OPT Motif* for the outer membrane localization [9];
- (5) Amino acid composition based predictor *SubLocC* for the cytoplasmic localization [13];
- (6) Signal peptide based predictor *Signal peptides* for the non-cytoplasmic localization [9; 24].

Based on the output from each module, the system uses a Bayesian network to generate a final probability value for each localization site. The system achieved an overall prediction accuracy of 75% for all localizations, a significant improvement over the previous results of PSORT I.

Besides the tri-peptide and the FFT based methods, we also implemented the method based on the amino acid composition. The experiment was carried out using a 5-fold cross-validation for each specific localization. Each time, the relevant dataset consisting of the proteins with the specific localizations was designated as the positive set; the remainder of the proteins was denoted as the negative set. The radial basis function was chosen as the kernel function for the SVM, since a preliminary experiment indicated this kernel exhibited better performance.

As the sizes of the positive and negative sets are substantially different, the performance of SVM was evaluated for precision (or sensitivity):

$$\text{precision} = \frac{tp}{tp + fp},$$

and recall (or positive prediction value):

$$\text{recall} = \frac{tp}{tp + fn},$$

where tp (resp. tn) is the number of the predicted positive (resp. negative) proteins which are true positive (resp. negative), and fp (resp. fn) is the number of the predicted positive (resp. negative) proteins which are true negative (resp. positive). The precision and recall of the 5-fold cross-validation were computed as the averages of the values from 5 folds.

The generalization performance of an SVM is controlled by the following parameters:

- (1) the trade-off C between the training error and the class separation;
- (2) the parameter g in the radial basis function, i.e., $\exp(-g\|\mathbf{x}_i - \mathbf{x}_j\|^2)$;
- (3) the biased penalty J for error from positive and negative training points.

Table 1: Results obtained from four different methods for the proteins from Germ-negative bacteria. (The numbers represent percentages).

Method	Composition		tri-peptide		FFT		PSORT-B	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Cytoplasmic	83.38	69.22	83.43	50.53	61.20	68.00	97.6	69.4
Inner membrane	98.65	83.57	99.52	80.75	96.12	87.30	96.7	78.7
Periplasmic	91.36	54.56	90.37	50.34	50.00	54.20	91.9	57.6
Outer membrane	87.21	84.12	95.28	83.66	95.70	94.30	98.8	90.3
Extra cellular	88.38	53.68	92.57	50.53	92.10	80.70	94.4	70.0

Composition : the method using SVM with the features from the amino acid composition ;
tri-peptide : the method using SVM with the features from the tri-peptide frequency ;
FFT : the method using SVM with the features from the FFT of hydrophobicity encoding ;
PSORT-B : the integrated predictor in [9]. The results are from Gardy *et al.* [9].

The values of precision and recall of a 5-fold cross-validation were computed for each triplet (C, g, J) . The choices of the parameters in the experiment for the composition and tri-peptide encoding sequences are given as follows:

C : from 1 to 150 with an incremental size of 10;

g : 1 to 100 with an incremental size of 10;

J : from 0.1 to 3.0 with an incremental size of 0.2.

The FFT encoded sequences are dense, therefore, they demand an intensive training time. Accordingly, a search over the full range of parameters would be prohibited. In order to deal with this problem, a two-step strategy for searching was employed. In the first round, the procedure scanned through all triplets (C, g, J) determined as follows.

C : from 2^{-8} to 2^7 with $c = 2 * c$ for each step;

g : from 2^{-8} to 2^7 with $g = 2 * g$ for each step;

J : from 0.1 to 3.0 with $j = j + 0.2$ for each step.

After identifying the best g value g^* from the first round, a more intensive search localized around g^* was performed. More precisely, it searched all triplets determined as follows.

C : from 1 to 21 with $C = C + 3$ for each step;

g : from $2^{g^*} - 1$ to $2^{g^*} + 1$ with $g = g + 0.003$ for each step;

J : from 0.1 to 3.0 with $J = J + 0.2$ for each step.

The SVMLight package was used as the SVM solver [15]. The best values of precision and recall for each method are given in Table 1, where the results for PSORT-B are taken from Gardy *et al.* [9]. Note that we compare the performance of the single predictor against the integrated predictive results from PSORT-B.

The FFT based method demonstrated superior performance over that of PSORT-B for the prediction of all three localizations: the inner membrane, the outer membrane, and the extra cellular case. While maintaining similar levels of precision, the improvement on the corresponding recall is from 78.7 to 87.3 for the inner membrane localization, from 90.3 to 94.3 for the outer membrane localization, and from 70.0 to 80.7 for the extra cellular localization. The FFT based method achieved substantial improvement in recall for the inner membrane and extra cellular localizations as compared with the remaining three methods. However, the FFT based

approach provided inferior findings for the cytoplasmic and periplasmic localizations.

On the other hand, the tri-peptide based method demonstrated good predictive power for the inner membrane localization as compared with PSORT-B. However, its ability for the other localizations did not surpass that of PSORT-B. Notably, the prediction of the periplasmic localization seems to be the hardest for all methods.

Although the FFT encoding method generates a compact set of features, we experienced longer times for training and testing in comparison with the tri-peptide encoding method, even though the tri-peptide frequency approach has a significantly larger number of features. We propose the following interpretation of this behavior. The FFT encoded sequences have a full dense structure while the tri-peptide encoded sequences are very sparse, although the lengths are longer. A feature selection scheme using a cut-off value to discard lower frequency features in the FFT encoded sequences may be able to achieve a similar level of predictive quality.

4. CONCLUSIONS

This work has introduced a novel Fast Fourier Transform based method for the feature extraction of protein sequences in conjunction with the use of support vector machines for the prediction of subcellular localizations. In addition, a tri-peptide based encoding method was considered in parallel.

The performances of these methods were empirically evaluated on a set of proteins with experimentally determined localizations from Germ-negative bacteria. Compared with the integrated system PSORT-B, the experimental results demonstrated that the SVM learned from the FFT encoded sequences exhibited superior performance for the prediction of the inner membrane, the outer membrane, and the extra cellular localizations, but was inferior for the prediction of cytoplasmic and periplasmic localizations. This implies that the hydrophobicity alone can not properly represent the sequence information which characterizes these two localizations. Combination with the tri-peptide based method may improve the predictive performance. This can be realized by using a kernel that combines the information from the FFT encoded sequences and the tri-peptide encoded sequences. The use of a different hydrophobicity index of amino acids, for example, the index shown in Table 2 [17], may also improve the quality of prediction.

Table 2: Hydrophobicity index of amino acids in Kyte and Doolittle.

amino acid	I	V	L	F	C	M	A	Z	T	S
index	4.5	4.2	3.8	2.8	2.5	1.9	1.8	-0.4	-0.7	-0.8
amino acid	W	Y	P	H	E	Q	D	N	K	R
index	-0.9	-1.3	-1.6	-3.2	-3.5	-3.5	-3.5	-3.5	-3.9	-4.5

5. ACKNOWLEDGMENTS

This research is partially supported by National Science Foundation (EIA-022-0301) and Naval Research Laboratory (N00173-03-1-G016).

6. REFERENCES

- [1] Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. and Miyano, S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18, 298-305.
- [2] Cai, Y.D. and Chou, K. C. (2003) Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics*. 20, 1151-1156.
- [3] Chou, K.C. and Cai, Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, 277, 45765-4576.
- [4] Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*, Cambridge University Press.
- [5] Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, 300, 1005-1016.
- [6] Emanuelsson, O. (2002) Predicting protein subcellular localisation from amino acid sequence information. *Brief. Bioinform.*, 3, 361-376.
- [7] Feng, Z.P. (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, 58, 491 - 499.
- [8] Fickett, J.W. and Tung, C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, 20, 6441-50.
- [9] Gardy, J.L. et al. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, 31, 3613-3617.
- [10] von Heijne, G. (1994) Signals for protein targeting into and across membranes. *Subcell. Biochem.*, 22, 1-19.
- [11] Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, 27, 215-219.
- [12] Horton, P. and Nakai, K. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, 24, 34-36.
- [13] Hua, S. and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17, 721-728.
- [14] Irbäck, A. and Sandelin, E. (2000) On hydrophobic correlations in protein chains. *Biophysical Journal*, 79, 2252-2258.
- [15] Joachims, T. (1999) *Making Large Scale SVM Learning Practical. Advances in Kernel Methods-Support vector learning*. MIT Press, Cambridge.
- [16] Kawashima, S. and Kanehisa, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, 28, 374.
- [17] Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, 157, 105.
- [18] Leslie, C., Eskin, E., Cohen, A., Weston, J. and Noble, W. (2002) Mismatch String Kernels for Discriminative Protein Classification. [Journal version of NIPS 2002 paper.] To appear in *Bioinformatics*.
- [19] Master, T. (1994) *Signal and Image Processing with Neural Networks : a C++ Sourcebook*. New York : John Wiley & Sons.
- [20] Menne, K.M.L., Hermjakob, H. and Apweiler, R. (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, 16, 741-742.
- [21] Nair, R. and Rost, B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, 11, 2836-2847.
- [22] Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein. Chem.*, 54, 277-344.
- [23] Nakai, K. and Kanehisa, M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, 11, 95-110.
- [24] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, 8, 581-599.
- [25] Pasquier, C.M., Promponas, V.I., Varvayannis, N.J. and Hamodrakas, S.J. (1998) A web server to locate periodicities in a sequence *Bioinformatics*, 14, 749-704.
- [26] Reinhardt, A. and Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, 26, 2230-2236.

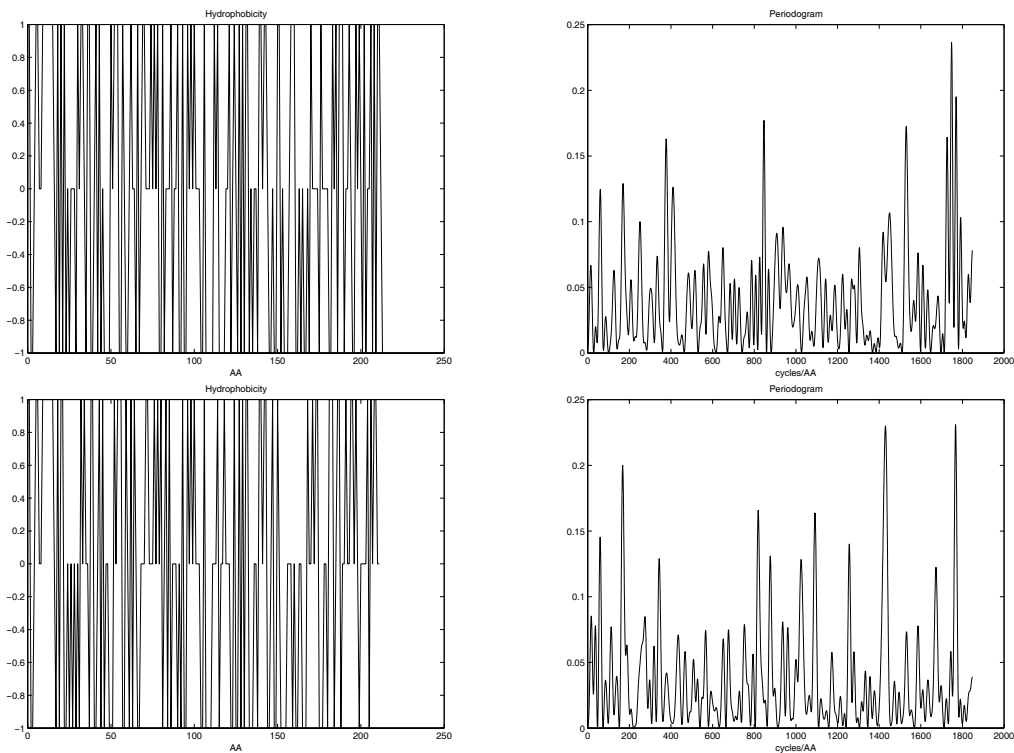


Figure 1: The encoded sequences for two extra cellular proteins before (left) and after (right) the fast Fourier transform.

- [27] Shepherd, A.J., Gorse, D. and Thornton, J.M. (2003) A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks. *PROTEINS: Structure, Function, and Genetics*, 50, 290-302.
- [28] Tusnady, G.E. and Simon, I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, 283, 489-506.
- [29] Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17, 849-850.
- [30] Yan, M., Lin, Z.S. and Zhang, C.T. (1988) A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, 14, 685-690.

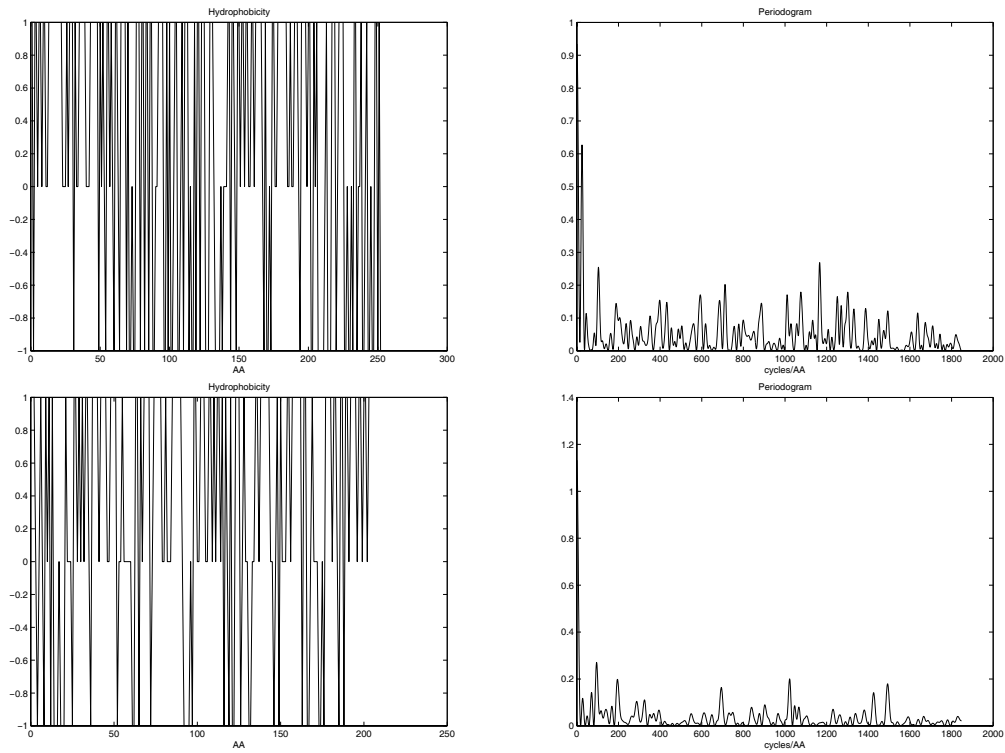


Figure 2: The encoded sequences for two inner membrane proteins before (left) and after (right) the fast Fourier transform.

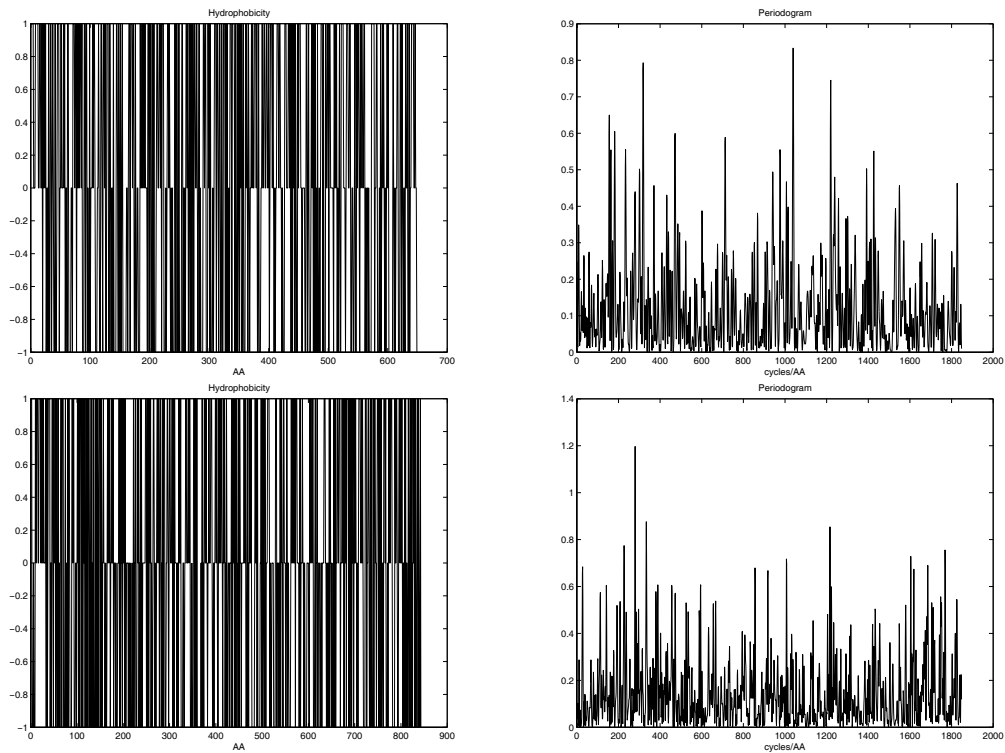


Figure 3: The encoded sequences for two outer membrane proteins before (left) and after (right) the fast Fourier transform.