Available online at www.sciencedirect.com
ScienceDirect
Mathematical Biosciences 207 (2007) 204–218
Mathematical Biosciences
www.elsevier.com/locate/mbs

Application of complex demodulation on bZIP and bHLH-PAS protein domains

Zhi Wang [a], Charles E. Smith [a,*], William R. Atchley [a,b]

[a] *Graduate Program in Biomathematics, North Carolina State University, Raleigh, NC 27695-8203, USA*
[b] *Department of Genetics and Center for Computational Biology, North Carolina State University, Raleigh, NC 27695-8203, USA*

Received 21 July 2006; received in revised form 29 December 2006; accepted 10 January 2007
Available online 3 February 2007

Abstract

Proteins are built with molecular modular building blocks such as an α-helix, β-sheet, loop region and other structures. This is an economical way of constructing complex molecules. Periodicity analysis of protein sequences has allowed us to obtain meaningful information concerning their structure, function and evolution. In this work, complex demodulation (CDM) is introduced to detect functional regions in protein sequences data. More specifically, we analyzed bZIP and bHLH-PAS protein domains. Complex demodulation provided insightful information about changing amplitudes of periodic components in protein sequences. Furthermore, it was found that the local amplitude minimum or local amplitude maximum of the 3.6-aa periodic component is associated with protein structural or functional information due to the observation that the extrema are mainly located in the boundary area of two structural or functional regions.

Keywords: Periodicity; Complex demodulation; Functional region

* Corresponding author. Tel.: +1 919 515 1907; fax: +1 919 515 1909.
E-mail address: bmasmith@stat.ncsu.edu (C.E. Smith).

0025-5564/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2007.01.004

1. Introduction

Recent developments in computational methodology have provided mechanisms to statistically transform alphabetic sequence information into biologically meaningfully arrays of numerical values [5,6]. Using a multivariate statistical approach, these authors generated five multidimensional indices (factors) of amino acid attributes that reflect polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge of sequences. This advance will make possible a number of statistical and mathematical analyses that will facilitate our understanding of the structure and function of biological sequence data.

For example, it has been suggested that the periodicity of a sequence can be evaluated by using a Fourier transformation or spectral analysis [21]. Periodicity of biological sequences is an important indicator of protein structure and DNA folding [14,25]. However, the Fourier analysis or spectral analysis is not useful in assessing the changes in cycle parameters, such as the amplitude and phase of the periodic components over the sequence [8,11].

Recently, the technique of complex demodulation (CDM) has been introduced to provide a continuous assessment of the periodic amplitude and thereby identify regions of change in structural and functional aspects of biological sequences. Complex demodulation has been widely used in many fields such as physiology, psychology and oceanography research [7,12,19,24,26]. However, there are no previous applications of CDM procedures in computational biology and bioinformatics according to our literature query results.

This paper intends to illustrate the application of CDM method on protein sequences based on several trials of bZIP and bHLH-PAS protein domains. This paper is also an exploratory work to ascertain if the amplitude of a certain periodic component of a protein sequence contains readily interpretable biological information. The bZIP and bHLH-PAS proteins are selected because of the complexity of their function and structure, which can represent quite complex characteristics of biological signals.

First we show that CDM can describe the changing amplitude of a particular periodic component. The amplitude pattern of the 3.6-aa periodic component is closely associated with the secondary structure of the protein sequences. It is found that the amino acid sites with local amplitude maximums and local amplitude minimums primarily occur at the boundaries of helices and strands. This strongly suggests that the CDM method is a new computational tool to aid us in the understanding of biological sequences. There are several other methods available to predict the regular secondary structure, however, the number of correctly predicted α-helix start positions was not large, namely 38% [27]. This research should trigger increased interest in CDM and more exploratory works to apply the CDM method to analyze additional biological sequences associated with signaling.

2. Methods

2.1. Principle of complex demodulation

Not every ''periodic'' series has a simple representation in terms of cosine or sine functions. A perturbed periodic component may have changing amplitude and changing phase. The goal of complex

demodulation is to quantify the amplitude and phase as a function of time. The amplitude and phase are determined by the data in the neighborhood of t, rather than by the whole series. The principle of complex demodulation has been well documented by [8], it is briefly described here before showing the case study results on the bZIP and bHLH-PAS proteins. Given the fundamental period of a biological sequence, CDM can extract approximations of the changing amplitude and changing phase as a function of the position of nucleotides or amino acid residues. Since this paper deals with protein sequences as a function of amino acid sites, the spatial function x_p is used to describe the numerical series of proteins instead of the time series x_t, where p represents the amino acid site. If the numerical series data x_p of a biological sequence is known to include a component oscillating around a frequency of λ (the amplitude and the phase may vary), then x_p can be written as

$$X_p = A_p \cos(\lambda p + \phi_p) + z_p \tag{1}$$

where A_p and $\emptyset_p$ are the changing amplitude and phase of the periodic component and z_p is residue including all other components and noises. Fig. 2.1 is a good illustration of power spectrum analysis and complex demodulation of simulated data [12]. Similarly, CDM is able to extract approximations of A_p (amplitude) as a function of time. $\emptyset_p$ can be also represented as a function of time, but this phase plot hasn't been shown here. In fact, the real-valued time series (1) can be regarded as complex-valued series and hereby can be easily processed in computation. With the Euler relation $\cos\lambda + i\sin\lambda = \exp(i\lambda)$, the time series X_p in (1) is converted to its complex analogue

$$X_p = \tfrac{1}{2}A_p\{\exp[i(\lambda p + \phi_p)] + \exp[-i(\lambda p + \phi_p)]\} + z_p, \tag{2}$$

where i is the unit complex number and $i^2 = -1$.

We then obtain a new signal y_t by shifting all the frequencies in X_p by $-\lambda$. This procedure is called CDM and y_p is expressed as

$$y_p = 2X_p \exp[-i\lambda p]. \tag{3}$$

Inserting Eq. (2) into (3), Eq. (3) then becomes

$$y_p = A_p \exp(i\phi_p) + A_p \exp[-i(2\lambda p + \phi_p)] + 2z_p \exp(-i\lambda p). \tag{4}$$

The first item of Eq. (4) is smooth (the frequency is around zero), the second term oscillates at a frequency of -2λ and the third item is assumed to contain no component around the zero frequency from the definition of z_p.

Therefore, when we let Y_p be the signal obtained by passing y_p through a low-pass filter, we would obtain Y_p in complex version as

$$Y_p = A_p \exp(i\phi_p). \tag{5}$$

Here Y_p is represented by a set of complex numbers in terms of its magnitude and phase, A_p and $\emptyset_p$. The instantaneous amplitude of the periodic component is defined as

$$A_p = |Y_p| = \sqrt{Y_p Y_p^*}, \tag{6}$$

where Y_p* is the complex conjugate of Y_p. The phase $\emptyset_p$ can be then calculated.

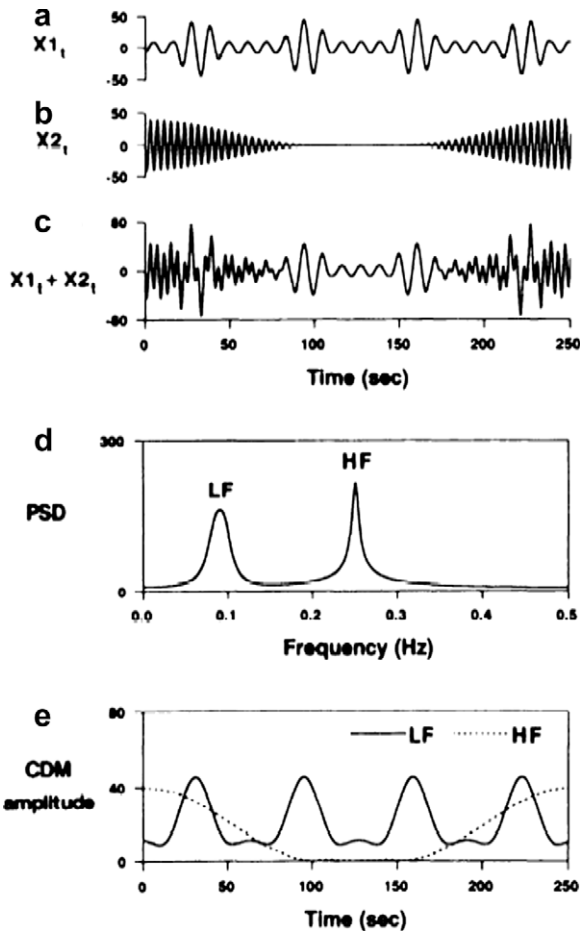

Fig. 2.1. Comparison between autoregressive power spectrum analysis and complex demodulation (CDM) of simulated data containing two periodic components (a) and (b). (a) Simulated low-frequency (LF) component. (b) Simulated high-frequency (HF) component. $X1_t$ and $X2_t$, 0.09 and 0.25 Hz sine functions with a fluctuating amplitude, respectively. (c) Time series generated by adding 2 sine functions ($X1_t + X2_t$). (d) Autoregressive power spectrum density (PSD). (e) Time series of instantaneous amplitude of LF and HF components obtained by CDM (figure from [12]).

The FORTRAN program of CDM was listed by [8]. From a frequency-domain perspective, the power spectrum of x_p has a peak around a frequency of λ. As the result of CDM, the peak is moved leftward to around zero frequency in the power spectrum of Y_p (in the PSD-frequency plot). For example in Fig. 2.1, if CDM is applied to the low-frequency periodic component A with a frequency around 0.09 Hz, then the Low-frequency peak will move leftward to around zero frequency. The peaks of all other components in x_p, if any, are also moved leftward, those at an original frequency above λ do not reach zero frequency, and those below λ move into the negative part of the frequency axis.

Thus, it is desirable for a low-pass filter to exclude all components except the zero-frequency component and then the amplitude can be determined. The low-pass filter is designed according

to the least squares filter design method presented by [8]. The transfer function of the ideal low-pass filter is

$$H(\omega) = \begin{cases} 1 & \text{if } 0 \leqslant \omega \leqslant \omega_c, \\ 0 & \text{if } \omega_c < \omega \leqslant \pi, \end{cases} \tag{7}$$

where ω_c is the cutoff frequency. The Fourier coefficients of $H(\omega)$ are

$$h_u = \frac{\sin u\omega_c}{\pi u} \quad u \geqslant 1 \text{ and } h_0 = \frac{\omega_c}{\pi} \tag{8}$$

We have to construct a smoothing function to approximate the ideal low-pass filter in computation. Convergence factors are used to accelerate the convergence of Fourier series and achieve better approximation of the transfer function $H(\omega)$ in Eq. (7). According to Bloomfield, the smoothed function approximating $H(\omega)$ is

$$\tilde{H}_s(\omega) = h_0 + 2\sum_{u=1}^{s} h_u \frac{\sin u\delta/2}{u\delta/2} \cos u\omega. \tag{9}$$

The multiplier $\frac{\sin 2\pi u/(2s+1)}{2\pi u/(2s+1)}$ is an example of convergence factor. The smoothed transfer function, which are initially 1 and decay to smoothly 0, are a smooth approximation to the ideal filter $H(\omega)$. Fig. 2.2 shows the smoothed transfer functions for $s = 5$ (an 11-term filter) and $s = 20$ (a 41-term filter), and the ideal transfer function.

It is important to note that the amplitudes obtained with the use of complex demodulation are relative rather than absolute measures. This is due to several factors, including the following: (i) the signal is not exactly sinusoidal. And (ii) the absolute measures of the amplitude represents the sum of high and low frequency periodic components, however, complex demodulation separates out the amplitudes at each frequency.

2.2. Computational procedures

First, we transform the biological sequences into numeric values of entropy and Factor I [5]. The variability of protein multiple alignments is measured as a numeric array of Boltzmann–

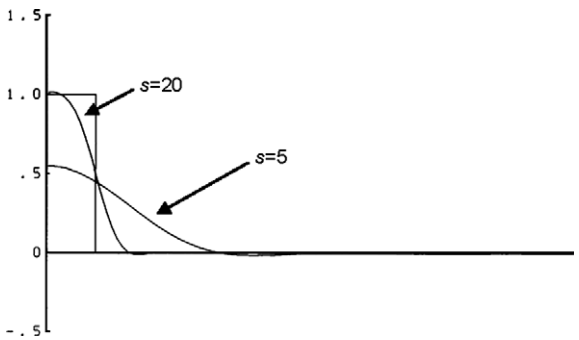

Fig. 2.2. Transfer functions of least squares low-pass filters with convergence factors applied, $s = 5$ and $s = 20$ (figure from [8]).

Shannon entropy values. The Boltzmann–Shannon entropy E is used to quantify sequence variability of amino acid residues at each aligned amino acid site [4]. It is calculated as $E(p) = -\sum_{j=1}^{21} pj\log_2(pj)$, where p_j is the probability of a residue being a specific amino acid or a gap, and $0 \leqslant E(p) \leqslant 4.39$. Five major patterns of amino acid attribute covariation that summarize the most important physiochemical aspects of amino acid covariability were interpreted as follows: Factor I = a complex index reflecting highly intercorrelated attributes for polarity, hydrophobicity, and solvent accessibility. Factor II = propensity to form various secondary structures, e.g. coil, turn or bend versus alpha helix frequency. Factor III = molecular size or volume, including bulkiness, residue volume, average volume of a buried residue, side chain volume, and molecular weight. Factor IV = relative amino acid composition in various proteins, number of codon coding for an amino acid, and amino acid composition. Factor V = electrostatic charge including isoelectric point and net charge. There may be association among different factors, therefore we focus on Factor I to see if we could find secondary structural information from it. Factor II–V is not examined in this paper.

Second, spectral analysis is conducted to produce power spectral density (PSD) plot (as shown in Fig. 2.1). Through examining the PSD plot, certain periodic component of interests can be selected for the following CDM procedure.

Third, given a periodic component of interests, CDM is applied to produce a plot of instantaneous amplitude and phase of the periodic component of interest as a function of particular amino acid sites.

2.3. Data types

The data utilized in this study are the basic region-leucine zipper (bZIP) protein domain and the basic region-helix-loop-helix-PAS (bHLH-PAS) protein domain. The bZIP proteins and bHLH-PAS are both very important transcription factors. bZIP proteins contain a basic region mediating sequence-specific DNA-binding, followed by a leucine zipper region, which is required for dimerization [22]. Both the basic region and the leucine zipper region have a helix form. Binding to DNA induces a coil-to-helix transition of the basic DNA-binding region. The leucine zipper region exhibits a stable helix form. The bZIP domain is one of the simplest types of DNA-binding domains. However, the bZIP transcription factors are capable of recognizing a diverse range of DNA sequences and regulate the gene transcription. The collection of 321 bZIP protein sequences (clad bzip_2) was retrieved from the database Pfam (Dec, 2004).

The second group of proteins to be analyzed are the bHLH-PAS proteins. They are a family of sensor proteins involved in signal transduction in a wide range of organisms. The bHLH-PAS domains contain a structurally conserved α/β-fold. There are basic region-helix-loop-helix motif, PAS-1 and PAS-2 motifs in the domain. Both PAS-1 and PAS-2 motifs contain a five-stranded antiparallel β-sheet with one face flanked by several α-helices. The PAS-1 and PAS-2 motifs are connected by a short linker.

The choice of bZIP and bHLH-PAS proteins is based on their structural and functional attributes. Since there are subtle differences among different regions of the sequences, it is intriguing to distinguish the differences between these various regions and to explore a novel approach to identifying the boundary of each region. It is hypothesized that CDM can distinguish the subtle differences among the structural and functional regions of sequences with a similar helix

conformation. If CDM is able to distinguish the subtle differences, it is expected to work better for regions of sequences have more structural and functional differences. bHLH-PAS contains complex α/β-fold which provides us with a complicated data set to examine the CDM procedure.

3. Results

3.1. bZIP protein domain

An entropy profile was calculated (Fig. 2.3) based on the method described by [4]. Such a profile is a numeric representation of the residue diversity at each amino acid site in a set of aligned proteins. Large entropy values represent high variability for that site while small values represent low variability. In our aligned sequence database, the basic region in bZIP proteins ranges from residue 1 to 27 while the leucine zipper region extends from residue 28 to 55. There are interesting oscillations of the entropy values and the periodic component was identified by spectral analysis [8].

A spectral density plot (Fig. 2.4) for this entropy profile was produced by the spectral analysis method-Fast Fourier Transformation [8] using SAS software (PROC SPECTRA). The peak at around 3.6 aa indicates that there is a major significant periodic component at that point in the entropy profile. Increases of spectral density in the period range from 13.75 to 56 aa indicate that there is also low-frequency periodic component, whose period estimate is much larger.

We focus on the high-frequency periodic component at around 3.6 aa that conforms to the average 3.6 aa per turn for an ideal α-helix. The CDM procedure was then used to analyze the amplitude of the 3.6-period component as a function of amino acid sites. We are particularly interested in locating the boundary between the basic region and the leucine zipper region.

The amplitude of the periodic component at 3.6 aa vs. residue is plotted in the dotted line in Fig. 2.3. We found there is amplitude decrease in the boundary region between the basic and leucine

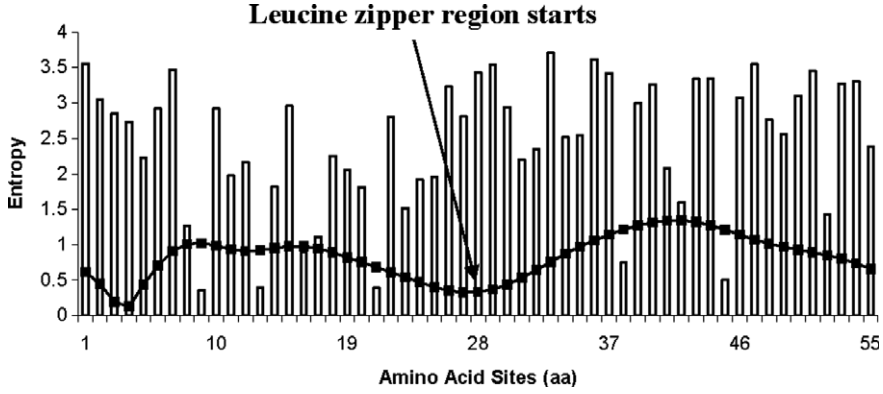

Fig. 2.3. Entropy profile of bZIP protein domains and the amplitude of the 3.6-aa periodic component. The entropy profile is represented by the histogram. Large entropy value represents high variation at that residue while small one represents low variation. Basic region: residue 1–27 Leucine zipper region: residue 28–55. The amplitude of the 3.6-aa periodic component vs. amino acid site is in dotted curve produced by CMD method.

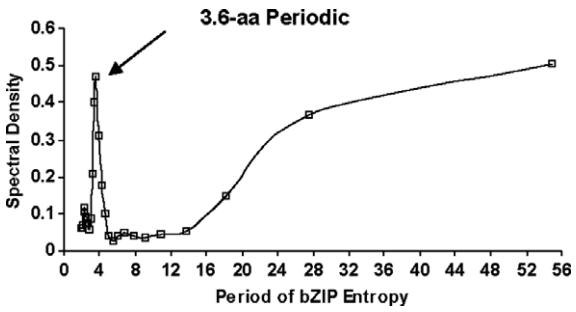

Fig. 2.4. Spectral density plot of the entropy profile of bZIP protein domains in the range from 2 to 10 aa (the periodic component of around 3.6 aa period is labeled).

zipper regions. The entropy amplitude at residue 27 achieved a local minimum. Structural studies indicate that residue 27 is the last residue of the basic region and the ZIP region starts at residue 28 (Pfam, 2005).

These results indicate that the entropy amplitudes of residues located in the end of the basic region and near the start of the ZIP region are significantly smaller than the average values (t-test: p-value < 0.05). The maximal entropy amplitude occurs at residue 42. These latter findings indicate that some residues in leucine zipper region are highly conserved while others are highly variable. The latter results in a large amplitude which is reflected by large fluctuations in the entropy values.

These findings suggest that the existence of a local minimum entropy amplitude identifies the boundary of specific structural or functional regions. It is interesting that the entropy amplitude at residue 4 has the global minimal amplitude and the amplitude increases beyond residue 4. This observation suggests a functional and structural difference of these residues that warrants further investigation.

Next, we investigate the Factor I profile [5] of the bZIP domain of the well-studied transcription factor C-fos shown in Fig. 2.5. The sequence of the domain (primary accession number: P01100; secondary accession number: P18849, [10]) is:

139-KRRIRRERNKMAAAKCRNRRREL|TDTLQAETDQLEDEKSALQTE IANLLKEKEKLEFI LAAH-200
Basic Region | Leucine Zipper

The spectral plot of c-fos factor I profile does not reveal a significant periodic component at around 3.6 aa (Fig. 2.6). However, implementing the CMD method and assuming that there is a periodic component of 3.6 aa, we obtain the amplitude for this relevant frequency (dotted curve in Fig. 2.5).

Based on the known structure of the c-fos protein [10], the leucine zipper region starts at residue 162 (labeled in Fig. 2.5). However, the amplitude of the 3.6-aa periodic component of Factor I at residue 162 is not significantly different from the average (t-test, $p > 0.05$) and indeed the local minimum is not at residue 162. However, the minimal amplitude of factor I occurs at residue 164, which is close to the leucine zipper starting residue of 162. This result suggests that the CDM method may be useful to predict the start point of a new structural or functional region, even if there is no significant 3.6-aa periodic component of Factor I. The deviation of the leucine zipper start residue from the residue with a minimal amplitude is possibly related to the absence of

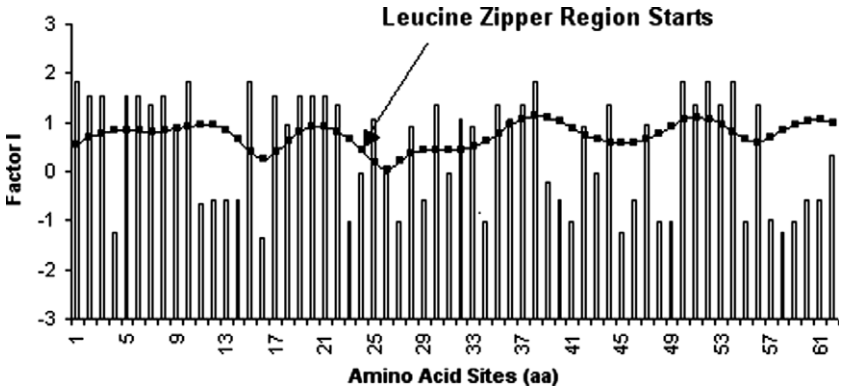

Fig. 2.5. Factor I profile of a bZIP protein domain of transcription factor c-fos protein. The amplitude of the periodic component at 3.6 aa vs. amino acid site is in dotted curve produced by CMD method.

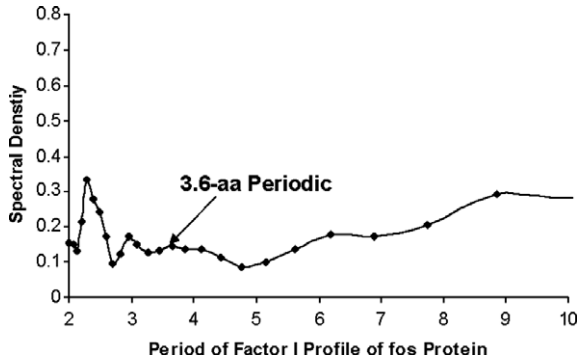

Fig. 2.6. Spectral density plot of the factor I profile of a bZIP protein domain of transcription factor c-fos protein in the range from 2 to 10 aa (the periodic component of around 3.6 aa period is labeled).

a statistical significant 3.6-aa periodic component of Factor I. This observation provides us with an interesting topic that may trigger further investigation.

3.2. PAS protein domain

Within the bHLH/PAS proteins the PAS region is involved in protein dimerization with another protein of the same family [13,23,28]. It has also been associated with light reception, light regulation and circadian rhythm regulators (clock). In bacteria, the PAS repeat is usually associated with the input domain of a histidine kinase, or a sensor protein that regulates a histidine kinase. 77 bHLH-PAS protein domains were obtained from PFAM database (version 17.0 May, 2005). The entropy profile of 77 bHLH-PAS domains is shown in Fig. 2.7.

The spectral density plot of the entropy profile of bHLH-PAS protein domains is given in Fig. 2.8. Only short-range periodicity (i.e. high-frequency components) is shown in Fig. 2.8. There

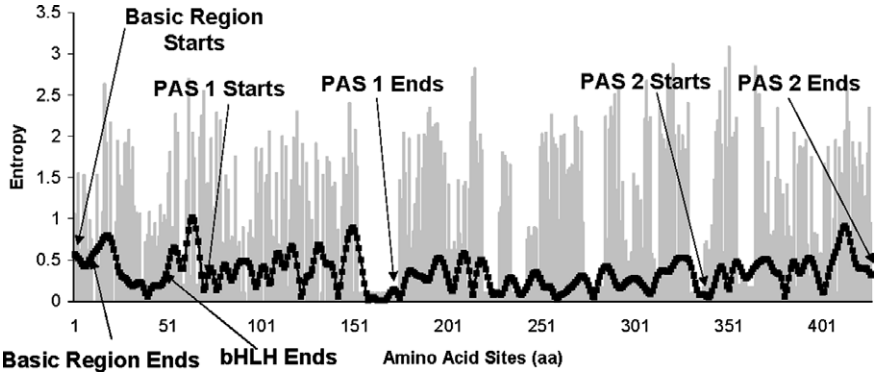

Fig. 2.7. Entropy profile of bHLH-PAS protein domains and the amplitude of the 3.6-aa periodic component. The entropy profile is represented by the histogram. The amplitude of the 3.6-aa periodic component vs. amino acid site is in dotted curve produced by CMD method.

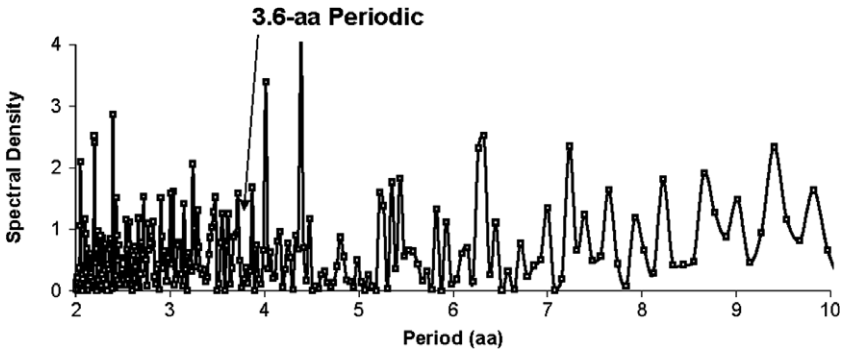

Fig. 2.8. Spectral density plot of the entropy profile of bHLH-PAS protein domains in the range from 2 to 10 aa (the periodic component of around 3.6 aa period is labeled).

are peaks located in the 3.40–3.91 aa range that signal the existence of an α-helix [18]. Therefore, CMD method is conducted to produce the amplitude of the 3.6-aa periodic component as a function of amino acid site (in dotted curve in Fig. 2.7).

Further, we investigate the Factor I profile of a well-known bHLH-PAS protein Arnt_human protein (p27540/ gi:114163), whose secondary structure has been determined [15]. It is a 789 aa-length protein containing bHLH/PAS1/PAS2 domains (Basic region: 90...102; Helix-loop-helix region: 103...143; PAS 1 domain: 161...235; PAS 2 domain 349...419). Regions 1–50 and 468–789 are removed because of they are included in the bHLH/PAS domain. The estimated secondary structure of the Arnt protein is obtained via Prediction protein web server.

The spectral density plot of the Factor I profile of Arnt protein has been produced in Fig. 2.9. There are some peaks located in the 3.40–3.91 aa range which signals the existence of α-helix, especially there is a large peak at around 3.6 aa [18].

The Factor I profile of Arnt protein domain is show as a histogram in Fig. 2.10. The CMD method produces the amplitude as a function of amino acid sites for the 3.6-aa periodic component (curve in Fig. 2.10).

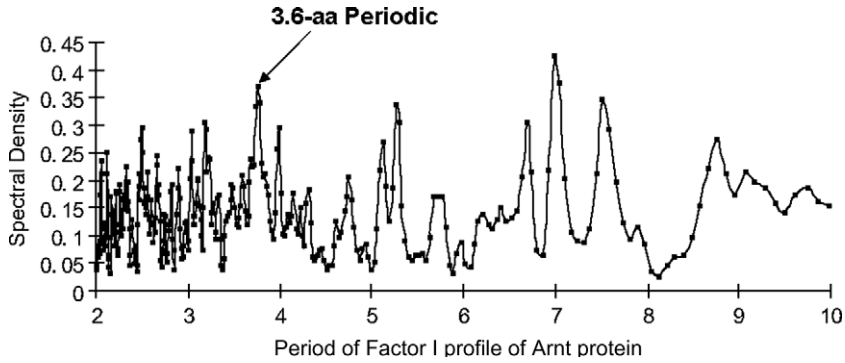

Fig. 2.9. Spectral density plot of the Factor I profile of Arnt protein in the range from 2 to 10 aa (the periodic component of around 3.6 aa period is labeled).

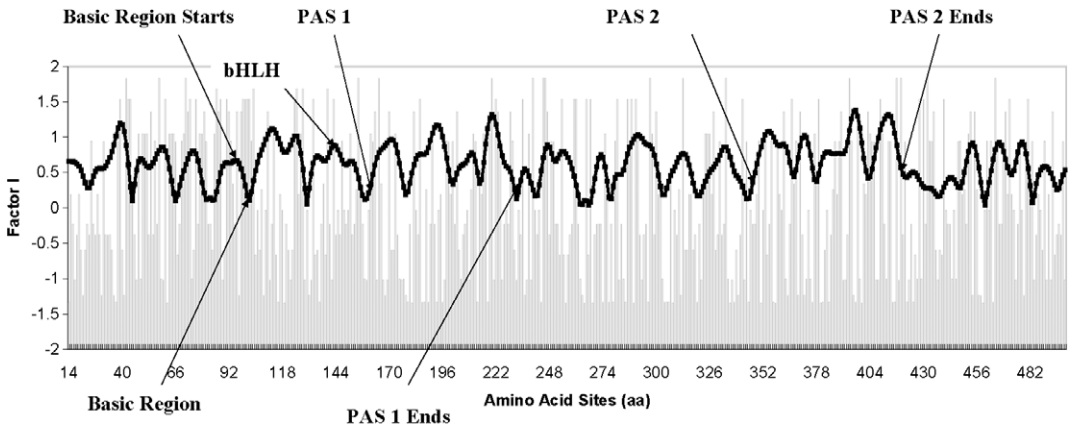

Fig. 2.10. Factor I profile of Arnt_human protein and the amplitude of the 3.6-aa periodic component (residue 14–499, other residues are trimmed in the process of CDM). The changing amplitude as a function of amino acid sites is produced by CDM and plotted as the curve.

Many local amplitude minimums have implications for the boundary of the α-helix regions or β-sheets (Table 2.1). Local minimum residue 102 is the ending residue of the basic region of the bHLH conserved domain. Local minimum 158 is close to the beginning residue 161 of PAS domain 1 ranging from residue 161 to 235. Local minimum 344 is close to the beginning residue 349 of PAS domain 2 ranging from residue 349 to 419.

Most local amplitude maximums have implications for the boundary of the α-helix regions or β-sheets (Table 2.1). The results are: Local maximum residue 143 is the ending residue of the 2nd helix of the HLH region. Local maximum residue 171 is the ending residue of an α-helix; local maximum residue 291 is close to the ending residue of a predicted α-helix; local maximum residue 273 is located between two β-sheets; local maximum residue 313 is close to the ending residue of a predicted β-sheet; local maximum residue 334 is the beginning residue of a predicted β-sheet; local maximum residue 355 is the beginning residue of a predicted β-sheet; local maximum residue 397 is the 2nd beginning residue of a predicted β-sheet. The only exceptions are: local maximum

Table 2.1
Summary of the locations of local amplitude minimums and maximums of the 3.6 aa periodic component of the Factor I profile

Residue	Local Min	Local Max	Location in the secondary structure
102	Y		End of the basic region
158	Y		Close to the start residue 161 of PAS 1 region
344	Y		Close to the start residue 349 of PAS 2 region
143		Y	End of the HLH region
171		Y	End of a helix
273		Y	Between two β-sheets
291		Y	Close to the end of a helix
313		Y	Close to the end of a β-sheet
334		Y	Start of a β-sheet
355		Y	Start of a β-sheet
397		Y	2nd start residue of of a β-sheet

It is found that the local amplitude minimums and maximums often occur in the boundary area of the α-helices and β-sheets.

residue 220 is located within a predicted α-helix region; local maximum residue 413 is located in a predicted β–sheet region.

4. Discussion

The finding of motifs and protein secondary structure prediction are important problems in bioinformatics. Generally α-helices are easier to predict than β-sheets with 9.5% more accuracy [1]. Recently the protein structure prediction has been dramatically improved by better remote homology detection (e.g., using PSI-BLAST [2] or hidden Markov models [17]), and larger sequence databases [9]. For instance, [16] obtained over 75% secondary structure prediction accuracy using a similar neural network architecture, where homologs are first detected via PSI-BLAST. A coiled coils prediction program, Paircoil2, achieves 98% sensitivity and 97% specificity on known coiled coils [3]. However, while secondary structure prediction methods continue to improve, there is still room for further improvement in structure prediction. Hidden Markov models and neural network methods may identify patterns without an understanding of what causes are generating these detectable differences. The complex demodulation method described in this paper attempts to fill this gap and explore the relationship between the underlying physiochemical traits and the detectable pattern. We have shown in this paper that the amplitude of the physiochemical frequency component is associated with the secondary structure or pattern differences. It is interesting to examine the amplitude of frequency component and check its role in protein pattern recognition. The relatively fixed period (3.6 aa) of α-helices in protein structures gives us the opportunity to investigate a specific frequency component with a 3.6 aa period in the CDM method. The amplitude of physiochemical frequency component can give us a picture of the resultant variation.

From the case study of the well-known protein sequences of bZIP and bHLH-PAS proteins, the amplitude of certain periodic component is shown to contain meaningful biological information.

The complex demodulation procedure is able to quantify the amplitude and phase of periodic components of protein sequences. It is the first time for the introduction and illustration of the applications of complex demodulation on protein sequences. These analyses reveal that the minimums or maximums of amplitudes of the 3.6-aa periodic component of protein profiles (i.e. entropy and factor I profiles) are predictors of the boundaries of helices secondary structures. The results in the paper should promote scientific interest in investigating the application of the CDM method in computational biology and bioinformatics. Possibly the analyses of the amplitudes or phases of periodic components through the CDM method could reveal important functional and structural information. Also it may be useful for improving the accuracy of the prediction for N-termini of α-helices because the current prediction accuracy is just 38% [27].

Hayano [12] has addressed three concerns of the CDM performance. (1) Is the resolution sufficient enough to distinguish between the low-frequency (LF) and high-frequency (HF) components? (2) Is the estimation of amplitude robust against alterations in the frequencies of the components? (3) What is the upper limit of rapid changes in amplitude that can be detected by the analysis?

Hayano reported that the CDM can not only distinguish the low-frequency and high-frequency amplitudes but also exclude the influence of the DC trends (those frequency <0.022 Hz) on the low-frequency amplitude. To examine the alterations in the frequencies of the component, Hayano simulated a signal by adding two sine waves whose frequencies were fluctuating between 0.06 and 0.12 Hz and between 0.18 and 0.44 Hz. The low-frequency and high-frequency amplitudes are then calculated by CDM. The results showed slight fluctuations of only 1.6 and 4.5%, whereas the power spectral plot showed wide-based multiple peaks reflecting the fluctuating frequencies of the components. These results indicate that the amplitude estimated by CDM is sufficiently robust against the alterations in the frequency of the signal. Further, the simulation of Hayano suggests that CDM provides a reliable estimate of amplitude when the frequency of amplitude fluctuations was below 0.034 Hz for the LF component and below at least 0.040 Hz for the HF component.

During the filtering procedure, the input signal is truncated at both ends because the use of the low-pass filter is analogous to the use of a data window. In the practice of analyzing series, the standard way is to extend the input signal with arbitrary numeric sequences like 0000000 or 1111111111 so that the original data is not truncated. We realize that these arbitrary numeric sequences may affect the accuracy of amplitude estimates. Therefore, a moving average window can be considered as an alternative of these arbitrary numeric sequences. Also in this research, the phase plot produced by the CDM method is not included because [12] has stated that the phase alternation (i.e. the frequency alternation) has little influence on the amplitude estimation. The results in this research are based on case studies on the bZIP and bHLH-PAS protein domains. More representative protein sequences should be examined with this CDM method in the future. The CDM method alone may not be able to accurately predict the exact boundary of the secondary structure blocks. The residues with local minimal or maximal amplitude may be not boundary residues (Table 1). More work remains to be done to improve the prediction accuracy of CDM methods. The CDM method assumes that there is a mean to the oscillations. However, [20] has argued that the fractal analysis of sequences has been more appropriate than standard descriptive statistics of mean and variance because with so many phylogenetically diverse sequences both mean and variance can go to zero or infinity and it first has to be shown that standard descriptive statistical assumptions apply to each case rather than be simply assumed.

Although the CDM method has not been extensively investigated, it appears to be a promising computational procedure to quantify the amplitude of the numeric profiles of protein sequences, which contains a lot of unknown biological information and signals. Many follow-up applications of the CDM methods are expected to promote our understanding of the complex protein sequences and structures.

Acknowledgments

Thanks for helpful suggestions from Andrew Fernandes. This research was supported by a grant from the National Institutes of Health (GM45344) to W.R.A.

References

[1] P. Aloy, A. Stark, C. Hadley, R. Russell, Prediction without templates:New folds, secondary structure, and contacts in CASP5, Proteins: Struct. Funct. Bioinformatics 53 (2003) 436.
[2] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389.
[3] Andrew V. McDonnell, Taijiao Jiang, Amy E. Keating, Bonnie Berger, Paircoil2: improved prediction of coiled coils from sequence, Bioinformatics 23 (2006) 356.
[4] W.R. Atchley et al., Correlations among amino acid residues in bHLH protein domains: an information theoretic analysis, Mol. Biol. Evol. 17 (2000) 164.
[5] W.R. Atchley et al., Solving the protein sequence "metric" problem, Proc. Natl. Acad. Sci. USA 102 (2005) 6395.
[6] W.R. Atchley, A.D. Fernandes, Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network, Proc. Natl. Acad. Sci. 102 (2005) 6401.
[7] H. Babkoff, T. Caspy, M. Mikulincer, H.C. Sing, Monotonic and rhythmic influences: a challenge for sleep deprivation research, Psychol. Bull. 109 (1991) 411.
[8] P. Bloomfield, Fourier analysis of time series: an introduction, Wiley, New York, 1976, pp. 1–150.
[9] J. Cuff, G. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, Proteins: Struct. Funct. Genet. 40 (2000) 502.
[10] J.N. Glover, S.C. Harrison, Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA, Nature 373 (1995) 257.
[11] C.W.J. Granger, M. Hatanaka, Spectral Analysis of Economic Time Series, Princeton University Press, 1964.
[12] J. Hayano, J.A. Taylor, A. Yamada, S. Mukai, R. Hori, T. Asakawa, K. Yokoyama, Y. Watanabe, K. Takata, T. Fujinami, Continuous assessment of hemodynamic control by complex demodulation of cardiovascular variability, Am. J. Physiol. 264 (1993) 1229.
[13] M.H. Hefti, K.J. Francoijs, S.C. de Vries, R. Dixon, J. Vervoort, The PAS fold: a redefination of the PAS domain based upon structural prediction, Eur. J. Biochem. 271 (2004) 1198.
[14] H. Herzel, O. Weiss, E.N. Trifonov, 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding, Bioinformatics 15 (1999) 187.
[15] E.C. Hoffman, H. Reyes, F.F. Chu, F. Sander, L.H. Conley, B.A. Brooks, O. Hankinson, Cloning of a factor required for activity of the Ah (dioxin) receptor, Science 252 (5008) (1991) 954.
[16] D. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195.
[17] K. Karplus, C. Barret, R. Hughey, Hidden Markov models for detecting remote protein homologies, Bioinformatics 14 (1998) 846.
[18] J. Kyte, Structure in Protein Chemistry, Garland Publishing, Inc., New York and London, 1995, pp. 201.
[19] L.A. Lipsitz, J. Hayano, S. Sakata, A. Okada, R.J. Morin, Complex demodulation of cardiorespiratory dynamics preceding vasovagal syncope, Circulation 98 (10) (1998) 977.

[20] L.S. Liebovitch, Fractals and Chaos Simplified for the Life Sciences, Oxford University Press, 1998.
[21] C.M. Pasquier et al., A web server to locate periodicities in a sequence, Bioinformatics 14 (1998) 749.
[22] L.M. Podust, A.M. Krezel, Y. Kim, Crystal structure of the CCAAT box/enhancer-binding protein beta activating transcription factor-4 basic leucine zipper heterodimer in the absence of DNA, J. Biol. Chem. 276 (1) (2001) 505.
[23] C.P. Ponting, L. Aravind, PAS: a multifunctional domain family comes to light, Curr. Biol. 7 (1997) R674.
[24] S. Rutherford, S. D'Hondt, Early onset and tropical forcing of 100,000-year Pleistocene glacial cycles, Nature 408 (2000) 72.
[25] P. Schieg, H. Herzel, Periodicities of 10–11 bp as indicators of the supercoiled state of genomic DNA, J. Mol. Biol. 343 (4) (2004) 891.
[26] K. Shin, E. David, Baroreflex sensitivity assessed by complex demodulation of cardiovascular variability, Hypertension 1997 (29) (1997) 1119.
[27] C.L. Wilson, P.E. Boardman, A.J. Doig, S.J. Hubbard, Improved prediction for N-termini of alpha-helices using empirical information, Proteins 57 (2004) 322.
[28] I.B. Zhulin, B.L. Taylor, R. Dixon, PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox, Trends Biochem. Sci. 22 (1997) 331.