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

Effects of windowing and zero-padding on Complex Resonant Recognition Model for protein sequence analysis

Charalambos Chrysostomou[1], Huseyin Seker[1]* and Nizamettin Aydin[2]

Abstract— **Signal processing techniques such as Fourier Transform have widely been studied and successfully applied in many different areas. Techniques such as zero-padding and windowing have been developed and found very useful to improve the outcome of the signal processing methods. Resonant Recognition Model (RRM) and Complex Resonant Recognition Model (CRRM) that are based on the discrete Fourier Transform and widely used for the analysis of protein sequences do not consider such methods, which can however improve or alter the features extracted from the protein sequences. Therefore, in this paper, an extensive analysis was carried out to investigate into the influence of the zero-padding and windowing on the features extracted from the Complex Resonant Recognition Model. In order to present such effects, five different classes of influenza A virus Neuraminidase genes, which include H1N1, H1N2, H2N2, H3N2 and H5N1 genes, were used as a case study. For each of the Influenza A subtypes, two sets of Common Frequency Peaks (CFP) were extracted, one where windowing is applied and the other one where windowing is suppressed, for each signal length set for the analysis. In order to make all the signals (protein sequence) the same length, zero-padding was used. The signal lengths used in this study are set to 470, which is the maximum protein length, and also 512, 1024, 2048, 4096, 8192 and 16384 for further analysis. The results suggest that the windowing and zero-padding have key impact on CFP extracted from the Influenza A subtypes as the best match with CFP extracted from influenza A subtypes using CRRM is when the signal length of 4096 and windowing were both applied. Therefore, the outcome of this study should be taken into consideration for more accurate and reliable analysis of the protein sequences.**

Index Terms— **Complex Resonant Recognition Model, Discrete Fourier Transform, Influenza A Virus, Neuraminidase, Resonant Recognition Model, Windowing, Zero-Padding**

I. INTRODUCTION

In bioinformatics several approaches have been developed in order to be able to identify rules that guide proteins to interact with their environment or other proteins directly from protein primary structure. One widely used tool is Basic Local Alignment Search Tool (BLAST) [1] which is based on the search of similarities in the arrangements between a group of proteins. In addition, another tool used to extract structural and physicochemical features from primary protein structure is Protein Feature Server (ProFeat) [2]. In recent years, signal processing techniques in application to bioinformatics are also used to extract features that are able to match proteins biological functions. Such a method is Resonant Recognition Model (RRM) [3]. By using signal processing techniques significant features can be identified on the frequency spectrum which can be related to a specific protein function. In case of RRM that is based on Discrete Fourier Transform (DFT), each of the protein classes analysed can correspond to a unique peak or a set of unique peaks in the frequency spectrum and can be related to the biological function that the class analysed represents. RRM uses only absolute spectrum to be able to extract features from the protein sequences and not real or imaginary frequency spectrum that DFT can generate. A new method has been developed, Complex Resonant Recognition Model (CRRM) [4], to take additional advantage of real and imaginary spectra as they can provide additional information to the analysis.

Signal processing techniques such as Fourier Transform have widely been studied and successfully applied in many different areas. Techniques such as zero-padding and windowing have been developed and found very useful to improve the outcome of the signal processing methods. Resonant Recognition Model (RRM) and Complex Resonant Recognition Model (CRRM) that are based on the discrete Fourier Transform and widely used for the analysis of protein sequences do not consider such methods, which can however improve or alter the features extracted from the protein sequences. Therefore, in this paper, an extensive analysis was carried out to investigate into the influence of the zero-padding and windowing on the features extracted from the Complex Resonant Recognition Model. In order to present such effects, five different classes of influenza A virus Neuraminidase genes, which include H1N1, H1N2, H2N2, H3N2 and H5N1 genes, were used as a case study.

II. PROTEIN DATA

The Influenza A virus genome consists of eight genes; hemagglutinin (HA) gene, neuraminidase (NA) gene, nucleoprotein (NP) gene, matrix proteins (M) gene, non-structural proteins (NS) gene and three RNA polymerase (PA, PB1, PB2) genes. For this analysis NA gene is selected as it is the target for current antiviral drugs, called neuraminidase inhibitors [5]. Influenza A H1N1, H2N2 and H3N2 subtypes caused three major pandemics during the twentieth century where H1N2 and H5N1 are considered as current pandemic threads.

Protein sequences used in this paper were collected from the Influenza Virus Resource data set [6]. Details of the influenza A subtypes and their related proteins in respect to the NA gene are as follows:

[1]Charalambos Chrysostomou and Huseyin Seker are with the Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK cchrysostomou@dmu.ac.uk, hseker@dmu.ac.uk

[2]Nizamettin Aydin is with Department of Computer Engineering, Yildiz Technical University, Turkey naydin@yildiz.edu.tr

*Correspondence to Huseyin Seker (hseker@dmu.ac.uk)

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

For H1N1 subtype, 200 NA proteins that appeared in 2009 were retrieved. H1N1 influenza A subtype from 2009 was the result of reassortment between Eurasian H1N1 swine virus and H1N2 swine virus [7]. The NA gene from the H1N1 swine virus was retained.

For H1N2 subtype 27 NA proteins were retrieved from the period 2001-2004. H1N2 influenza A subtype from the period 2001 - 2004 was the results of reassortants of classical swine H1N1 virus and triple reassortant H3N2 viruses [8]. The NA gene from the H3N2 virus was retained.

For H2N2 subtype, 76 NA proteins that appeared in the period of 1957-1968 were retrieved. H2N2 influenza A subtype from the period 1957-1968 was the results of reassortment between human H1N1 virus and avian H2N2 viruses [7]. The NA gene from H2N2 virus was retained.

For H3N2 subtype, 200 NA proteins that appeared in the period of 1968-2000 were retrieved. H3N2 influenza A subtype from the period 1968-2000 was the result of reassortment between human H2N2 and avian H3 viruses [7]. The NA gene from H2N2 virus was retained.

For H5N1 subtype, 70 NA proteins that appeared in the period of 2005-2009 were retrieved. H5N1 influenza A subtype was the results of the combination of various influenza A subtype viruses [9] where the NA gene was retained from the avian H1N1 virus.

By using CLUSTALW, an online tool [7], pairwise percent identity of all protein sequences was calculated. Percentage identity is a measurement used to determine the similarity between protein sequences. Table I shows the average percentage identity between all the influenza A neuraminidase subtypes

TABLE I
AVERAGE PERCENT IDENTITY

	H1N1	H1N2	H2N2	H3N2	H5N1
H1N1	93%	-	-	-	-
H1N2	40%	98%	-	-	-
H2N2	42%	86%	96%	-	-
H3N2	40%	88%	86%	94%	-
H5N1	83%	41%	43%	41%	96%

As Table I shows, the percent identity within each individual influenza subtype class is very high with 93%, 98%, 96%, 94% and 96% for H1N1, H1N2, H2N2, H3N2 and H5N1 NA influenza A subtypes, respectively. In contrast to the average percent identity of individual subtypes, average percent identity between different classes may vary significantly. High average percent identity can be observed between H1N1 and H5N1 with 83%, H1N1 and H2N2 with 86%, H1N2 and H2N2 with 88% and H2N2 and H3N2 with 86%. In addition, low average percent identity can be observed between H1N1 and H1N2 with 40%, H1N1 and H2N2 with 42%, H1N1 and H3N2 with 40%, H5N1 and H1N2 with 41%, H5N1 and H2N2 with 43%, and finally H5N1 and H3N2 with 41%.

III. PRE-PROCESSING OF PROTEIN SEQUENCES

Before applying the informational spectrum analysis to protein sequences, techniques such as zero-padding and windowing used in signal processing needs to be considered.

A. Zero-padding

The first technique used is zero-padding where a specified number of zero elements is added to the end of each sequence in order to increase signal length before DFT is applied to the signals. This technique is essential for CRRM as the given protein sequences may not be of the same length that makes CRRM unrealisable. Seven different resolutions were used for the analysis of influenza A neuraminidase proteins. The first signal length used are 470, which is the maximum protein length of the influenza A protein subtypes. The remaining six signal lengths used are 512 (2^9), 1024 (2^{10}), 2048 (2^{11}), 4096 (2^{12}), 8192 (2^{13}) and 16384 (2^{14}).

B. Windowing

The second technique considered is windowing, which tries to reduce spectral leakage [10] by multiplying a pre-calculated window with the encoded numerical sequences. Spectral leakage is caused when processing finite-length signals using frequency analysis of infinite signals where it seems like some energy leaked out of the primary signal spectrum into neighbour frequencies. As the literature shows in other applications [11], [12] where frequency analysis and DFT are used windowing can reduce or even eliminate spectral leakage. In this case, Hamming window [13] is used and can be defined in equation 1

$$w = 0.54 - 0.46cos(\frac{2\pi(N)}{N}) \quad (1)$$

where N is the total amino acids in a protein sequence.

IV. COMPLEX RESONANT RECOGNITION MODEL

For complex resonant recognition model (CRRM) [4] to be able to be applied alphabetical protein sequences need to be covered to numerical sequences. In the literature although more than 500 scales [14] can be used to encode protein sequences the electron-ion interaction potential (EIIP) amino acid scale [15], as given in Table II, is used as RRM is based on the DFT and EIIP.

TABLE II
EIIP VALUES

Amino acid	EIIP	Amino acid	EIIP	Amino acid	EIIP
Leu	0.0000	His	0.0242	Ser	0.0829
Ile	0.0000	Lys	0.0371	Cys	0.0829
Asn	0.0036	Ala	0.0373	Thr	0.0941
Gly	0.0050	Tyr	0.0516	Phe	0.0946
Glu	0.0057	Trp	0.0548	Arg	0.0959
Val	0.0058	Gln	0.0761	Asp	0.1263
Pro	0.0198	Met	0.0823		

By using Equation 2 each of the corresponding numerical values for the amino acids was normalized

$$E' = \frac{E - \mu(E)}{\sigma(E)} \quad (2)$$

where σ correspond to standard deviation and μ to mean value of EIIP values.

The Discrete Fourier Transform (DFT) is defined as follows

$$X(n) = \sum_{m=0}^{N-1} x(m)e^{-j(2\pi/N)nm} \quad n = 1, 2, ..., N/2 \quad (3)$$

where $x(m)$ is the mth member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. The following formula determines the maximal frequency in the spectrum $F = \frac{1}{2d}$ where F is the maximal frequency of all signals and d is the distance between points of the sequence. Assuming that all points of the sequence are equidistant with distance $d = 1$ then the maximum frequency in the spectrum can be found as $F = \frac{1}{2(1)} = 0.5$.

The output of DFT is a complex sequence and can be represented as follows

$$X(n) = (R(n) + I(n)j), \quad n = 1, 2, ..., N/2 \tag{4}$$

where $R(n)$ is the Real part of the sequence and $I(n)j$ the Imaginary part.

The aim of CRRM is to determine a Characteristic Frequency Peak (CFP) for each spectrum, absolute, real and imaginary, that correlates with a biological function expressed by a set of protein sequences using informational spectrum analysis. The absolute, real and imaginary informational spectrum can be formulated as follows

Absolute Spectrum:

$$S_a(n) = X(n)X^*(n) = |X(n)|^2, \quad n = 1, 2, ..., N/2 \tag{5}$$

where S_a is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X*(n)$ are the complex conjugate.

Real Spectrum:

$$S_r(n) = |R(n)|^2, \quad n = 1, 2, ..., N/2 \tag{6}$$

where S_r is the Real spectrum for a specific protein, $R(n)$ are the real parts of DFT coefficients $X(n)$.

Imaginary Spectrum:

$$S_i(n) = |I(n)j|^2, \quad n = 1, 2, ..., N/2 \tag{7}$$

where S_i is the imaginary spectrum for a specific protein, $I(n)$ are the Imaginary parts of DFT coefficients $X(n)$.

Informational Spectrum:

$$C_a = \Pi S_{(a)}(m), \quad m = 1, 2, ..., M \tag{8}$$

$$C_r = \Pi S_{(r)}(m), \quad m = 1, 2, ..., M \tag{9}$$

$$C_i = \Pi S_{(i)}(m), \quad m = 1, 2, ..., M \tag{10}$$

where C_a is the absolute informational spectrum (AIS), C_r is the real informational spectrum (RIS), C_i is the imaginary informational spectrum (IIS) and M is the number of protein sequences used for a specific class. Equation 11 is used to scale AIS, RIS and IIS.

$$V = \frac{\sqrt{\sum_{n=0}^{L} C_{a,r,i}(n)}}{L} \tag{11}$$

Where L is the number of points in the Absolute (C_a), Real (C_r) and Imaginary Informational Spectrum (C_i).

CFP pursuant to the AIS, RIS and IIS analysis can be used to characterise and distinguish the proteins. However, the following conditions should be fulfilled for the CFP to be related to a biological function. For a group of protein sequences that share the same biological function only one CFP should exist. No CFP should exist for biologically unrelated protein sequences. CFP is expected to be different for dissimilar biological functions.

V. RESULTS

Tables III, IV and V show the results for CRRM AIS, RIS and IIS respectively. For each of the Influenza A subtypes two CFP are extracted for each signal length used; one where windowing is applied (w) and one where windowing is suppressed (ψ). As the results demonstrate when CRRM is applied windowing and zero-padding can have key impact on CFP extracted from Influenza A neuraminidase subtypes

Significant changes in the AIS, RIS, and IIS can be observed for all Influenza A NA subtypes by increasing signal length from 470 to 16384. For CFP in AIS major changes are observed in subtype H1N1 (0.1681 to 0.0730), H5N1 (0.1681 to 0.4839), H1N2 (0.3170 to 0.3970) and H3N2 (0.3170 to 0.3970). For RIS the CFP shifted in subtype H1N1 from 0.4106 to 0.1687, H5N1 from 0.4830 to 0.3181, H1N2 from 0.3170 to 0.3965 and H2N2 from 0.3170 to 0.4091. For RIS the CFP shifted in subtype H5N1 from 0.3170 to 0.4844 and H3N2 from 0.3447 to 0.3972.

Additionally, significant changes in the AIS, RIS, and IIS can be observed for all Influenza A NA subtypes by applying the hamming window to the encoded protein sequences before CISA. For CFP in AIS where windowing is applied significant changes are observed in subtype H1N1 (0.1687 to 0.0732), H5N1 (0.4839 to 0.0739), H1N2 (0.3970 to 0.4584), H2N2 (0.3972 to 0.4859), and H3N2 (0.3970 to 0.4586). For RIS the CFP shifted in subtype H5N1 from 0.3181 to 0.0731, H1N2 from 0.3965 to 0.4587, H2N2 from 0.4091 to 0.4863 and H3N2 from 0.3171 to 0.4586. For RIS the CFP shifted in subtype H5N1 (0.4844 to 0.0742), H1N2 (0.3973 to 0.4577), H2N2 (0.3973 to 0.4852) and H3N2 (0.3972 to 0.1889).

By considering the average percentage identity shown in Table I and the information retrieved from the literature regarding the Influenza A NA proteins the best match with CFP extracted from influenza A subtypes using CRRM as shown in Tables III-V is obtained when the signal length of 4096 and windowing were both applied.

VI. CONCLUSIONS

In this paper, an extensive analysis was carried out to investigate into the influence of zero-padding and windowing on the features extracted from CRRM in AIS, RIS and IIS. For this analysis five different classes of influenza A virus were used; H1N1, H1N2, H2N2, H3N2 and H5N1 subtypes. For each of the Influenza A subtypes two sets of CFP's were extracted, one where windowing was applied (w) and one where windowing was suppressed (ψ), for each signal length used. The signal length used in this study was set at 470 (maximum protein length), 512 (2^9), 1024 (2^{10}), 2048 (2^{11}),

TABLE III

ABSOLUTE SPECTRA RESULTS

N	H1N1		H5N1		H1N2		H2N2		H3N2	
	ψ	w	ψ	w	ψ	w	ψ	w	ψ	w
470	0.1681	0.0745	0.1681	0.0745	0.3170	0.4574	0.4085	0.1894	0.3170	0.1894
512	0.0742	0.0742	0.4844	0.0742	0.3965	0.4590	0.4082	0.1504	0.3965	0.4590
1024	0.1689	0.0732	0.4844	0.0742	0.3975	0.458	0.3975	0.1504	0.3975	0.4590
2048	0.0737	0.0737	0.4839	0.0737	0.3970	0.4585	0.3970	0.4858	0.3970	0.4585
4096	0.0737	0.0735	0.4839	0.0740	0.3970	0.4585	0.3972	0.4858	0.3970	0.4585
8192	0.0737	0.0735	0.4839	0.0739	0.3970	0.4584	0.3972	0.4860	0.3971	0.4585
16384	0.0737	0.0735	0.4839	0.0739	0.3970	0.4584	0.3972	0.4859	0.3970	0.4586

TABLE IV

REAL SPECTRA RESULTS

N	H1N1		H5N1		H1N2		H2N2		H3N2	
	ψ	w	ψ	w	ψ	w	ψ	w	ψ	w
470	0.4106	0.4915	0.4830	0.4872	0.3170	0.5000	0.3170	0.5000	0.3170	0.3170
512	0.2188	0.2188	0.3184	0.3184	0.3965	0.4590	0.3926	0.4863	0.3965	0.4590
1024	0.0732	0.0732	0.4834	0.0732	0.3965	0.4590	0.4092	0.4863	0.3174	0.4590
2048	0.0732	0.0732	0.4834	0.0732	0.3965	0.4585	0.4092	0.4863	0.3169	0.4585
4096	0.1687	0.0732	0.3181	0.0732	0.3965	0.4587	0.4092	0.4863	0.3171	0.4585
8192	0.1687	0.0731	0.3181	0.0731	0.3965	0.4587	0.4091	0.4863	0.3171	0.4586
16384	0.1687	0.0732	0.3181	0.0731	0.3965	0.4587	0.4091	0.4863	0.3171	0.4586

TABLE V

IMAGINARY SPECTRA RESULTS

N	H1N1		H5N1		H1N2		H2N2		H3N2	
	ψ	w	ψ	w	ψ	w	ψ	w	ψ	w
470	0.0745	0.0745	0.3170	0.0745	0.4085	0.4574	0.4085	0.4851	0.3447	0.4596
512	0.0742	0.0742	0.4844	0.0742	0.4082	0.2227	0.4082	0.1777	0.4082	0.2109
1024	0.0742	0.0742	0.4844	0.0742	0.3975	0.4580	0.3975	0.4854	0.3975	0.3975
2048	0.0742	0.0742	0.4844	0.0742	0.3975	0.4575	0.3975	0.4854	0.3975	0.1890
4096	0.0742	0.0742	0.4844	0.0742	0.3972	0.4578	0.3972	0.4854	0.3972	0.1890
8192	0.0741	0.0741	0.4844	0.0742	0.3973	0.4578	0.3972	0.4852	0.3972	0.1890
16384	0.0741	0.0741	0.4844	0.0742	0.3973	0.4577	0.3973	0.4852	0.3972	0.1889

4096 (2^{12}), 8192 (2^{13}) and 16384 (2^{14}). The results suggest that windowing and zero-padding can have key impact on CFP extracted from Influenza A subtypes. The best match with CFP extracted from influenza A subtypes using CRRM as shown in Tables III-V is obtained when the signal length of 4096 and windowing were both applied. Therefore, the outcome of this analysis should be taken into consideration for a more reliable and accurate analysis of different sets of protein sequences.

REFERENCES

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, p. 403–410, 1990.

[2] Z. Li, "Profeat: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence." *Nucleic Acids Res.*, vol. 34, p. 32–37, 2006.

[3] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules? Theory and applications." *IEEE transactions on bio-medical engineering.*, vol. 41, p. 1101–1114, 1994.

[4] C. Chrysostomou, H. Seker, N. Aydin, and P. Haris, "Complex resonant recognition model in analysing influenza a virus subtype protein sequences," in *IEEE International Conference on Information Technology and Applications in Biomedicine*, Greece 2010, p. 1–4.

[5] A. Moscona, "Neuraminidase inhibitors for influenza," *New England Journal of Medicine*, vol. 353, no. 13, p. 1363–1373, 2005.

[6] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *Journal of virology*, vol. 82, no. 2, p. 596–601, 2008.

[7] Morens, D.M. and Taubenberger, J.K. and Fauci, A.S., "The persistent legacy of the 1918 influenza virus," *The New England journal of medicine*, vol. 361, no. 3, p. 225–229, 2009.

[8] Y. Choi, S. Goyal, M. Farnham, and H. Joo, "Phylogenetic analysis of H1N2 isolates of influenza A virus from pigs in the United States," *Virus research*, vol. 87, no. 2, p. 173–179, 2002.

[9] M. M. Mukhtar, S. T. Rasool, D. Song, C. Zhu, Q. Hao, Y. Zhu, and J. Wu, "Origin of highly pathogenic H5N1 avian influenza virus in China and genetic characterization of donor and recipient viruses," *JOURNAL OF GENERAL VIROLOGY*, vol. 88, no. Part 11, p. 3094–3099, NOV 2007.

[10] A. Girgis and F. Ham, "A quantitative study of pitfalls in the FFT," *Aerospace and Electronic Systems, IEEE Transactions on*, no. 4, p. 434–439, 1980.

[11] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, p. 51–83, 1978.

[12] D. Agrez, "Improving phase estimation with leakage minimization," *Instrumentation and Measurement, IEEE Transactions on*, vol. 54, no. 4, p. 1347–1353, 2005.

[13] R. Blackman and J. Tukey, "The Measurement of Power Spectra, p. 190," *New York*, 1958.

[14] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "Aaindex: amino acid index database, progress report 2008," *Nucleic acids research*, vol. 36, no. suppl 1, p. D202, 2008.

[15] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. LalovicC, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?" *IEEE Transaction on Biomedical Engineering*, vol. 32, no. 5, p. 337–341, 1985.