Investigation into the Effects of an Individual Amino Acid on Protein Function by Means of a Resonant Recognition Model

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

[1] Bio-Health Informatics Research Group, Centre for Computational Intelligence, Department of Informatics, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK
[2] Department of Computer Engineering, Yildiz Technical University, Turkey
{cchrysostomou,hseker}@dmu.ac.uk, naydin@yildiz.edu.tr

Abstract. Upon identification of a new protein it is important to single out these amino acids responsible for the structural classification of the protein as well as the amino acids contributing to the protein's specific biological characterisation. A novel approach is presented to identify and quantify this cause and effect relationship between amino acid and protein. This exploits the Common Frequency Peak (CFP) that it extracts from the Resonant Recognition Model (RRM). Applicability and robustness of the method are shown on a case study where five different protein families of the influenza A virus Neuraminidase (NA) genes are studied. They include H1N1, H1N2, H2N2, H3N2 and H5N1. The analyses identified five segments, namely three between H1N1 and H5N1 and two between H1N2, H2N2 and H3N2 and suggested that they play a key role in Influenza A NA gene functionality and can potentially be considered as target areas for future antiviral drugs and vaccines such as neuraminidase inhibitors.

Keywords: Resonant Recognition Model, Effects of individual amino acids, Discrete Fourier Transform, Influenza A Virus, Neuraminidase.

1 Introduction

A number of tools capable of identifying directly the mechanism by which proteins interact with their environment have been developed. These directly identify rules from primary protein structure. Two popular tools are Basic Local Alignment Search Tool (BLAST) [1] and Protein Feature Server (ProFeat) [5]. The former searches for similarities in the arrangements of amino acids in proteins while the latter extracts structural and physicochemical features from protein sequences. Matching biological function to features that are extracted by signal processing is another technique that is available. An example of this is the Resonant Recognition Model (RRM) [8,3] that uses Discrete Fourier Transform (DFT). Significant features can be identified by signal processing the frequency spectrum which is related to a specific protein function. With RRM each of the protein classes under study can be related to the biological function that they represent, with the unique peak or set of peaks extracted from the frequency spectrum. Each unique peak extracted using RRM is called a Common Frequency Peak (CFP).

⋆ Corresponding Author

G. Lee, D. Howard, and D. Ślęzak (Eds.): ICHIT 2011, LNCS 6935, pp. 229–236, 2011.

A question that arises when signal processing techniques are used is which amino acid of the proteins sequences contributes most to specific feature extracted? One approach [11,10] in the literature uses RRM to determine which section of protein is a dominant contributor to CFP. This section can be identified by measuring the magnitude of DFT at the CFP position and selecting the window with the highest value. In this study, a new approach that measures the effect of individual amino acids of protein sequences upon CFP extracted from RRM is developed and presented. For this analysis five different protein classes of influenza A virus Neuraminidase (NA) gene, which includes H1N1, H1N2, H2N2, H3N2 and H5N1 subtypes, are used to show applicability and robustness of the method developed.

2 Methodology

2.1 Resonant Recognition Model

In this paper electron-ion interaction potential (EIIP) amino acid scale [9], as shown in Table 1, encodes alphabetical protein sequences to numerical sequences for RRM. In the literature more than 500 unique amino acid scales [4] are available to encode protein sequences.

Table 1. EIIP Values

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

Each of the corresponding numerical values for the amino acids was normalized by using z-score,

$$E' = \frac{E - \mu(E)}{\sigma(E)} \tag{1}$$

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

The Discrete Fourier Transform (DFT) is defined as follows

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

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

$$F = \frac{1}{2d} \tag{3}$$

where F is the maximal frequency of all signals and d is the distance between points of the sequence. If it is assumed that distance $d = 1$ then the maximum frequency in the spectrum can be found as $F = 1/2(1) = 0.5$. The output of DFT is a complex sequence and can be characterized 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 RRM is to determine a CFP in the absolute spectrum that correlates with a biological function expressed by a set of protein sequences using informational spectrum analysis. The absolute informational spectrum can be formulated as follows

Absolute Spectrum

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

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.

Informational Spectrum

$$C_{(n)} = \Pi S_{(n)}(m), \quad m = 1,2,...,M \tag{6}$$

where $C_{(n)}$ is the informational spectrum and M is the number of protein sequences used for a specific class. Equation 7 is used to scale Informational Spectrum

$$V = \frac{\sqrt{\sum_{n=0}^{N/2} C(n)}}{N/2} \tag{7}$$

where L is the number of points in the Informational Spectrum (C).

CFP pursuant to the absolute informational spectrum analysis can be used for characterising and distinguishing the proteins. However, the following conditions should be fulfilled for the CFP to be related to a biological function:

1. For diverse biological functions CFP is expected to be dissimilar.
2. For biologically dissimilar protein sequences no CFP should exist.
3. For a collection of protein sequences that allocate the same biological function single CFP should exist.

2.2 Influence of Individual Amino Acid to Common Frequency Peak

This algorithm must be followed to determine how individual amino acid can affect the magnitude of DFT at the CFP position.

STEP 1: Calculate CFP position for given protein sequences.

STEP 2: Calculate the magnitude of DFT at CFP position for all original protein sequences.

STEP 3: Remove single amino acid at position X from original protein sequences with length N.
STEP 4: Recalculate the magnitude of DFT at CFP position for all modified protein sequences.
STEP 5: Compare DFT at CFP position between ORIGINAL and modified protein sequences.
STEP 6: Repeat STEP 3 to 5 for all N amino acids in protein sequences.

After all six steps are completed the outcome is a measurement for all individual amino acids of all given protein sequences.

3 Protein Data

For the analysis five Influenza A Neuraminidase (NA) genes were retrieved from the Influenza Virus Resource data set [2]. The protein sequences are the following

- A/Aarhus/INS242/2009(H1N1) [ADK33724]
- A/Anhui/1/2005(H5N1) [ADG59213]
- A/Egypt/84/2001(H1N2) [CAD29972]
- A/Albany/3/1958(H2N2) [ABO52305]
- A/Hong Kong/1-2-MA21-2/1968(H3N2) [ACF22356]

A pairwise percent identity for all influenza NA genes could be calculared with CLUSTALW [6]. Table 2 shows the percent identity for all influenza A NA protein sequences.

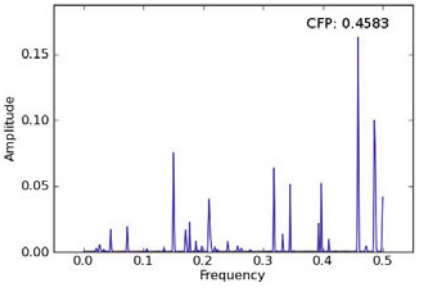
Table 2. Percent Identity

	H1N1	H1N2	H2N2	H3N2
H1N2	40%	-	-	-
H2N2	42%	85%	-	-
H3N2	39%	87%	94%	-
H5N1	87%	41%	42%	41%

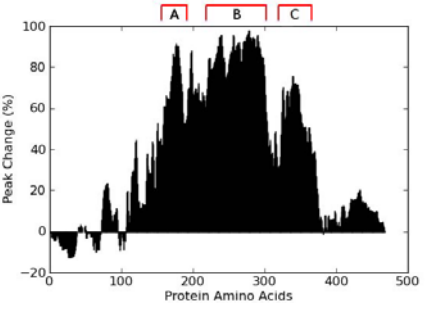
As Table 2 shows a high percent identity is observed between H1N1 and H5N1 (87%), H1N2 and H2N2 (85%) and H1N2 and H3N2 (87%) NA genes. Low percent identity is observed between H1N1 and H1N2 (40%), H1N1 and H2N2 (42%), H1N1 and H3N2 (39%), H5N1 and H1N2 (41%) and H5N1 and H2N2 (42%), H5N1 and H3N2 (41%).

4 Results

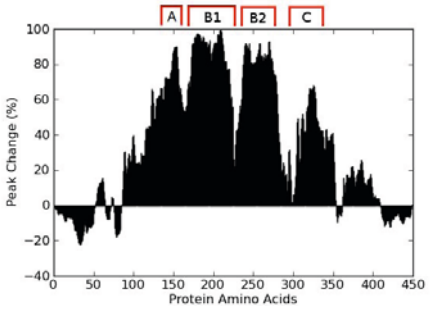
By following the method described in the previous sections with steps 1 to 6 for the five influenza A NA genes the following results are obtained. Step 1 requires that the



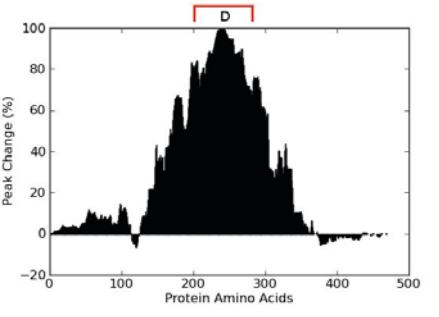
(a) Informational Spectrum Analysis Results



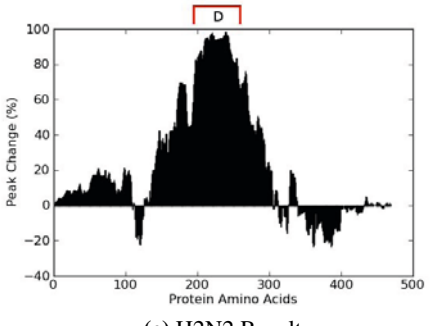
(b) H1N1 Results



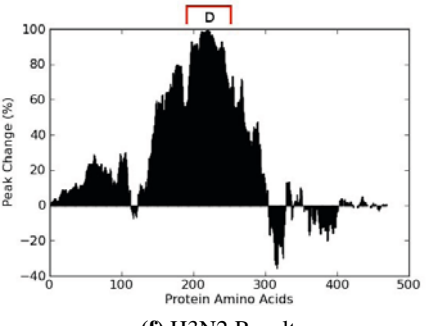
(c) H5N1 Results



(d) H1N2 Results



(e) H2N2 Results



(f) H3N2 Results

Fig. 1. Results of the RRM-based analyses of all the protein classes

CFP position for the protein sequences is calculated. For the five influenza A NA genes, Figure 1a shows the CFP calculated to 0.4583. By following the remaining Steps 2 to 6 the influence of all amino acids of Influenza NA genes are calculated. Figures 1b, 1c, 1d, 1e and 1f show the results for H1N1, H5N1, H1N2, H2N2 and H3N2 NA genes where each point in figures gives the impact of a particular amino acid on DFT at the CFP position. All the results are presented in percentage to the original value of DFT at the CFP position.

Five key areas are identified by a 60% thresholding of results. These present the highest impact on DFT at CFP position. The key areas: A, B and C as displayed in figures 1b and 1c between H1N1 and H5N1 NA genes. Area D in figures 1d, 1e and 1f is also identified between H1N2, H2N2 and H3N2 NA genes. All amino acids that correspond to areas A, B, C and D can be found in Tables 3, 4, 5, 6 and 7 for H1N1, H5N1, H1N2, H2N2 and H3N2, respectively .

Table 3. Hight impact areas for H1N1 NA gene

Area	Residues	Sequence
A	159-185	MSCPIGEVPSPYNSRFESVAWSASACH
B	193-301	IGISGPDNGAVAVLKYNGIITDTIKSWRN
		NILRTQESECACVNGSCFTVMTDGPSD
		GQASYKIFRIEKGKIVKSVEMNAPNYH
		YEECSCYPDSSEITCVCRDNWHGSNR
C	325-348	NPRPNDKTGSCGPVSSNGANGVKG

Table 4. Hight impact areas for H5N1 NA gene

Area	Residues	Sequence
A	138-160	LMSCPVGEAPSPYNSRFESVAWS
B1	167-223	GTSWLTIGISGPDNGAVAVLKYNGIITDT
		IKSWRNNILRTQESECACVNGSCFTVMT
B2	235-280	IFKMEKGKVVKSVELNAPNYHYEECSC
		YPDAGEITCVCRDNWHGSN
C	319-328	SPNGAYGIKG

Table 5. Hight impact areas for H1N2 NA gene

Area	Residues	Sequence
D	193-299	CVTGDDKNATASFIYNGRLVDSIGSWS
		KKILRTQESECVCINGTCAVVMTDGSA
		SGKADTKILFIEEGKIGHTSLLSGSAQ
		HVEECSCYPRYPGVRCVCRDNWKGSN

Table 6. Hight impact areas for H2N2 NA gene

Area	Residues	Sequence
D	195-270	TGDDRNATASFIYDGRLVDSIGSWSQN ILRTQESECVCINGTCTVVMTDGSASG RADTRILFIKEGKIVHISPLSG

Table 7. Hight impact areas for H3N2 NA gene

Area	Residues	Sequence
D	192-256	VCITGDDKNATASFIYDGRLVDSIGSW SQNILRTQESECVCINGTCTVVMTDGS ASGRADTRILF

By using the identified areas as shown in Tables 3-7 three segments that exist unchanged in influenza A genes between H1N1 and H5N1 and two segments between H1N2, H2N2 and H3N2 NA genes are identified. For H1N1 and H5N1 NA genes the following segments are identified

- **PSPYNSRFESVAWS** from A area,
- **IGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTVMTS** from B1 area and
- **EITCVCRDNWHGSN** from B2 area.

For H1N2, H2N2 and H3N2 NA genes the following segments are identified

- **GRLVDSIGSWS** and
- **ILRTQESECVCINGTC** from area D.

5 Conclusions

Upon identification of a new protein it is important to single out these amino acids responsible for the structural classification of the protein as well as the amino acids contributing to the protein's specific biological characterisation. In this paper, a novel approach is presented to identify and quantify this cause and effect relationship between amino acid and protein. This exploits the CFP that it extracts from the RRM. Applicability and robustness of the method are shown on a case study where five different protein families of the influenza A virus NA genes, which includes H1N1, H1N2, H2N2, H3N2 and H5N1 NA gene, are studied. For each of the Influenza A NA genes studied, areas A, B, C and D that have a high impact on DFT at CFP position as shown in figures 1b to 1f are selected. The analyses identified five segments, namely three between H1N1

and H5N1 and two between H1N2, H2N2 and H3N2 and suggested that they play a key role in Influenza A NA gene functionality and can potentially be considered as target areas for future antiviral drugs and vaccines such as neuraminidase inhibitors [7]. The biological functionality which the amino acid scale used to encode protein sequences to numerical sequences represents, can directly be linked to the identified segments of the protein sequences. In this study EIIP was used but in the literature more than 500 unique amino acid scales [4] exist. These scales can be used to encode protein sequences and utilised for further analysis that may help reveal additional important segments of the protein sequences studied.

References

1. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
2. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman, D.: The influenza virus resource at the National Center for Biotechnology Information. Journal of Virology 82(2), 596 (2008)
3. Cosic, I.: Macromolecular bioactivity: is it resonant interaction between macromolecules?Theory and applications. IEEE Transactions on Bio-Medical Engineering 41, 1101 (1994)
4. Kawashima, S., Ogata, H., Kanehisa, M.: AAindex: amino acid index database. Nucleic Acids Research 27(1), 368 (1999)
5. Li, Z.: Profeat: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 34, 32–37 (2006)
6. Morens, D., Taubenberger, J., Fauci, A.: The persistent legacy of the 1918 influenza virus. The New England Journal of Medicine 361(3), 225 (2009)
7. Moscona, A.: Neuraminidase inhibitors for influenza. New England Journal of Medicine 353(13), 1363 (2005)
8. Pirogova, E.: Analysis of amino acid parameters in the resonant recognition model. In: Proceedings of the International Conference on Bioelectromagnetism, p. 71 (1998)
9. Veljkovic, V., Cosic, I., Dimitrijevic, B., LalovicC, D.: Is it possible to analyze DNA and protein sequences by the methods of digital signal processing? IEEE Transaction on Biomedical Engineering 32(5), 337–341 (1985)
10. Veljkovic, V., Veljkovic, N., Este, J., Huther, A., Dietrich, U.: Application of the EIIP/ISM bioinformatics concept in development of new drugs. Current Medicinal Chemistry 14(4), 441–453 (2007)
11. Veljkovic, V., Veljkovic, N., Muller, C., Muller, S., Glisic, S., Perovic, V., Kohler, H.: Characterization of conserved properties of hemagglutinin of h5n1 and human influenza viruses: possible consequences for therapy and infection control. BMC Structural Biology 9(1), 21 (2009)