International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:12, No:11, 2018

Feature Extraction Technique for Prediction the Antigenic Variants of the Influenza Virus

Majid Forghani, Michael Khachay

***Abstract*—**In genetics, the impact of neighboring amino acids on a target site is referred as the nearest-neighbor effect or simply neighbor effect. In this paper, a new method called wavelet particle decomposition representing the one-dimensional neighbor effect using wavelet packet decomposition is proposed. The main idea lies in known dependence of wavelet packet sub-bands on location and order of neighboring samples. The method decomposes the value of a signal sample into small values called particles that represent a part of the neighbor effect information. The results have shown that the information obtained from the particle decomposition can be used to create better model variables or features. As an example, the approach has been applied to improve the correlation of test and reference sequence distance with titer in the hemagglutination inhibition assay.

***Keywords*—**Antigenic variants, neighbor effect, wavelet packet, wavelet particle decomposition.

I. Introduction

OUR motivation stems from the well-known genetic prediction problem for emerging antigenic variants of the influenza virus based on the hemagglutinin protein sequence. According to the comprehensive overview given in [1], contemporary modelling techniques can be partitioned to two main clusters. While the former cluster comprises phylogenetic methods, the latter one consists of population-genetics-based techniques, including multivariate statistical learning methods. In this paper, we propose a novel statistical modelling method that belongs to the second cluster and is based on *wavelet packet decomposition*.

In order to infer relationships between the genotype and antigenicity, it is necessary to estimate the impact of specific *sites* (locations in protein sequence) on the antigenicity of the studied virus [2], [3]. According to the major conventional approach (see, e.g. [2]), only target amino acids located at the examined site are compared.

On the other hand, as widely believed in genetics, protein structure significantly depends on the mutual interaction between adjacent amino acids. The impact of neighboring amino acids on a target amino acid is referred as the nearest-neighbor effect or simply neighbor effect. The type of neighbor amino acids can be explained by their propensity to form different structures, especially secondary structure [4]. The neighbor effect has been investigated both for DNA and common proteins in application to prediction of protein structures and functions, and feasibility of amino acids substitution [4]-[8].

Although the results presented in the cited works appear to be very promising, there remains the necessity of a comprehensive method quantifying and representing the neighbor effect by means of property changes of amino acids located in an adjustable neighborhood of the site in question.

To address this drawback, we introduce a novel inter-protein distance taking into account physicochemical changes observed in neighboring amino acids before and after mutation. Considering such an effect while measuring the distance between two amino acids of the same site in protein sequences, improves the result of similarity/dissimilarity. The proposed distance is based on novel data processing technique, which we call *Wavelet Particle Decomposition*.

The proposed method is a frequency-based method that uses wavelet packet transform. The wavelet and wavelet packet (WP) transforms are powerful tools for time-frequency analysis [9]. In genetics, wavelet-based methods has been used in various studies [10] such as detection of genetic polymorphism [11], spectral properties of short genes [12], analysis of genomic sequences [13], regular patterns in DNA sequence [14] and DNA sequences classification [15]. In order to apply the transform to protein sequence, the alphabetical sequence must be converted to numerical. The choice of numerical mapping affects the reflection of protein biological properties that can be in the numerical domain [16], [17].

In our research, we choose the specific alpha-numeric conversion in accordance with the well-known AAindex database[1]. Further, for capturing the numerical representation of the neighbor effect we employ convolution procedures with experimentally chosen wavelet-packet filters. To reveal the significant information from protein structure, we propose the Wavelet Particle Decomposition (WPD) that successfully takes into account the ordering of sample representation in a wavelet packet sub-bands (from lowest to highest frequency values). Our method maps a property value of amino acid (located in a target position) into some small values, called particles. Each particle carries information related to a certain property of adjacent amino acids. We show that particles can improve the measuring of similarity/dissimilarity between sequences.

To demonstrate the performance of our method, we apply it to the well-known Hemagglutination Inhibition (HI) assay database. The HI assay is the widely known procedure aimed to measure the antigenic similarity of influenza virus strains performed by genetic experts in laboratories. It is based on the

M. Forghani is with Ural Federal University, 620002, 19 Mira street, Ekaterinburg, Russia (e-mail: majid.forqani@gmail.com).
M. Khachay is with Ural Federal University and Krasovsky Institute of Mathematics and Mechanics, Ekaterinburg, Russia (e-mail: mkhachay@imm.uran.ru).

[1]http://www.genome.jp/aaindex/

ability of antibodies produced with respect to some reference strain of the considered virus to inhibit (speed-down) the hemagglutination reaction for the test virus strain [18].

The HI assay plays the key role in vaccine virus selection. In the literature, there are known attempts to introduce mathematical models to quantify the phenotypic impact of specific amino acid substitutions on the antigenicity [3], [2], [19]. The models describe the relationship between the antigenicity and the evolution of virus populations. Although these results seem to be promising, feature extraction still remains challenging. In this paper, we show that application of the proposed features based on wavelet packet decomposition increases the overall significance of the model.

Remaining part of this paper is organized as follows: Section II gives the clear definition of the considered problem. Section III represents the necessary theoretical background of the proposed WPD method. In section IV, we illustrate the performance of the proposed method combined with known search heuristics in several numerical experiments. Finally, in Section V, we summarize the results obtained and enlist some open questions postponed to the future work.

II. Problem Statement

The HI assay data are commonly involved in the antigenicity assessment of various strains of the influenza virus. The identification and quantification of the impact of amino acid substitutions on the antigenicity of HA should help to understand the evolution of the virus and is crucial for vaccine virus selection [20], [3], [2], [19]. The most current methods consider the correlation between alphabetical changes of amino acids at fixed sites and HI titers (results of biochemical tests performed in special laboratories). The amino acid substitution leads to observable changes of some their physicochemical properties. Since the amino acid interacts with its neighbors in the protein sequence, the substitution also affects them. Therefore, it turns to be useful, in antigenic studies, to take into account the influence (neighbor effect) of amino acids located in the protein sequence nearby to the target site.

	A	B	C	D
1	virus	reference	dateOfTest	titre
2	A/BAYERN/7/95	A/BAYERN/7/95	9/5/2001	2560
3	A/JOHANNESBURG/82/96	A/BAYERN/7/95	9/5/2001	2560
4	A/BEIJING/262/95	A/BAYERN/7/95	9/5/2001	40
5	A/NEWCALEDONIA/20/99	A/BAYERN/7/95	9/5/2001	40
6	A/HONGKONG/1252/2000	A/BAYERN/7/95	9/5/2001	<
7	A/MADAGASCAR/57794/20(	A/BAYERN/7/95	9/5/2001	40
8	A/BAYERN/7/95	A/JOHANNESBURG/82/96	9/5/2001	5120

Fig. 1 Sample from the HI assay database [22]

Any record of the HI assay database refers to a couple of protein sequences of two virus strains (we call them *reference* and *test* one) combined with associated titer (see Fig. 1). Actually, an i^{th} record is a triple (T_i, R_i, y_i), where T_i and R_i, $i = 1, \dots, L$ are protein sequences of test and reference viruses, respectively, y_i is the observed value of titer, and L is a length of the database. Sequences T_i and R_i are supposed to be of equal length. Therefore, we can compare their entries $T_i(x)$ and $R_i(x)$ at any individual site x. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be some function. Following to [2], to any site x, we assign vectors $T = [T_1, \dots, T_L]$, $R = [R_1, \dots, R_L]$, $y = [y_1, \dots, y_L]$, and an absolute value $AP(f, x \,|\, T, R, y)$ of the Pearson's correlation coefficient between $f(T(x), R(x))$ and $\log_2(y)$ as follows:

$$AP(f, x \,|\, T, R, y) = \left| \frac{Cov(f(T(x), R(x)), \log_2(y))}{\sqrt{Var(f(T(x), R(x)))}\sqrt{Var(\log_2(y))}} \right|.$$

The problem is, for a given site x and given family $\mathcal{F}$ of functions, to find $f^* \in \mathcal{F}$, such that

$$AP(f^*, x \,|\, T, R, y) = \max\{AP(f, x \,|\, T, R, y) \colon f \in \mathcal{F}\}. \quad (1)$$

In our paper, the family $\mathcal{F}$ is defined by Euclidean distances between particles taken from different levels of WPD and maximization is carried out subject to WPD tree.

III. Wavelet Particle Decomposition

Suppose, we are given by a set of protein sequences of equal length N. In order to compare two protein sequences, we define a distance between amino acids located at same position in considered sequences, taking into account influence of neighboring amino acids. We refer protein sequence as discrete signal $f(x)$, where $x = 1, \dots, N$. Imagine that the signal $f(x)$ is decomposed through WP transform into $M = 2^j$ sub-bands (where $j \in \mathbb{Z}^+$ is decomposition level). To facilitate notation, for each sub-band of the last level we denote the reconstructed signal by $f_{i,x}$ for $i = 1, \dots, M$ and $1, \dots, N$. It is convenient to represent the reconstructed signals in matrix form as follows.

$$\begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,N} \\ f_{2,1} & f_{2,2} & & f_{2,N} \\ \vdots & & \ddots & \vdots \\ f_{M,1} & f_{M,2} & \cdots & f_{M,N} \end{bmatrix} \quad (2)$$

Each row of matrix (2) corresponds to the reconstructed signal for a specific sub-band. We call matrix (2) a *reconstruction matrix* or simply R-matrix.

Each column of R-matrix is equal to the result of the decomposition of a specific entry of the input signal with respect to different sub-bands. According to WP theory [21], for any entry of the signal (e.g. point x), its amplitude can be approximated by the column sum for the associated column of the R-matrix. Indeed, the decomposition and reconstruction provide a representation of each entry of the signal $f(x)$ with $M = 2^j$ points in different sub-bands. We denote this representation by Ω

$$\Omega(f(x)) = [f_{1,x}, f_{2,x}, \dots, f_{M,x}].$$

In Fig. 2 we illustrate the action of the operator Ω.

As known, the value $f(x)$ of the initial signal permits the approximate expansion as follows

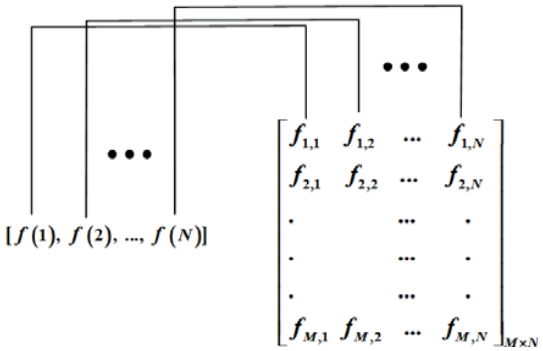

Fig. 2 Correspondence between preimage and image of the operator Ω

$$f(x) \approx \sum([f_{1,x}, f_{2,x}, \ldots, f_{M,x}]) = \sum(\Omega(f(x)) = \sum_{p=1}^{M} f_{p,x}, \quad (1 \le x \le N),$$

where the row vector $[f_{1,x}, f_{2,x}, \ldots, f_{M,x}]$ is the transpose of the x^{th} column vector of R-matrix (2).

As mentioned ago, the order of representation in Ω (from lowest frequency to highest frequency sub-bands or vice versa) keeps significant information that should be considered and extracted. Actually, we consider an image $\Omega(f(x))$ as a new signal, to which WP can be applied recursively. Applying wavelet packet decomposition and reconstruction up to the same fixed decomposition level j to $\Omega(f(x))$ we obtain a new R-matrix.

$$\begin{bmatrix} f_{1,1,x} & f_{1,2,x} & \cdots & f_{1,M,x} \\ f_{2,1,x} & f_{2,2,x} & & f_{2,M,x} \\ \vdots & & \ddots & \vdots \\ f_{M,1,x} & f_{M,2,x} & \cdots & f_{M,M,x}. \end{bmatrix} \tag{3}$$

Similarly to previous constructions we obtain

$$f_{p_1,x} = \sum(\Omega(f_{p_1,x})) = \sum_{p_1=1}^{M} f_{p_2,p_1,x},$$

where

$$\Omega(f_{p_1,x}) = [f_{1,p_1,x}, f_{2,p_1,x}, \ldots, f_{M,p_1,x}]$$

for $1 \le x \le N$ and $1 \le p_1 \le M$.

Hereinafter, we call operator Ω as *Wavelet Particle Decomposition* (WPD). Its output (image) for a certain entry we refer as particles of that value (see Fig. 3).

Using previously described procedure, the new row vector (obtained from Ω) can be decomposed to a new R-matrix, and each entry of row vector has its particles in R-matrix. The idea of point decomposition through PD is supported by the following simple statement.

Statement 1 *Suppose f is a discrete function from $L^2(R)$. For any level q of particle decomposition and any x, $1 \le x \le N$, the following equation is valid*

$$f(x) \approx \sum_{p_1=1}^{M} \sum_{p_2=1}^{M} \cdots \sum_{p_q=1}^{M} f_{p_q,\ldots,p_1,x}.$$

For the sake of brevity we skip the proof of Statement 1 that can be easily obtained by induction on q.

IV. Experiments

In order to illustrate the performance of our WPD-based technique, we carry out three numerical experiments. In the first one, we show that, even for the case of a point mutation, where two protein sequences differ between each other in a single amino acid, WPD provides a whole set of new features providing more options to find a more relevant feature for distinguishing them (in terms of equation (1)).

In the second experiment, we perform a comparison of a number of well-known search heuristics in combination with WPD to find the feature providing the most correlation improvement between protein sequence and HI titer on the database [22] (near to the optimal solution of maximization problem (1)).

To proof the significance of the relationship between the order of adjacent amino acids and obtained correlation, we perform the third experiment. In this experiment, for any database record presenting information about the certain protein, we change at random the order of amino acids neighboring to a target site. After that, we repeat the second experiment for the permuted data. The results obtained show that the violation of neighbors order leads to correlation vanishing. Thus, in the third experiment, we show that the correlation between the obtained features strongly depends on the order of adjacent amino acids.

A. Experiment 1: Neighbor Effect in a Point Mutation

We consider two close to each other strains A/NORWAY/1684/2007 and A/NORWAY /1651/2007 of the influenza virus H1N1 taken from the well-known EpiFlu database [2]. It is known that, for these strains, the first parts of their hemagglutinin (HA) protein sequences, HA1, differs at the unique site, which is a result of a single point mutation. We transform these alphabetical sequences to numeric ones using the known index from AAindex1 database. We set wavelet basis to Daubechies Db3 and WPD level to 3. Computing first two levels of WPD for these sequences, we obtain 8 plus 64 particles for each signal entry respectively.

To demonstrate the difference between the sequences in question, we subtract the second sequence from the first one. The difference is presented in Fig. 3 (a), notice the single peak that is a result of the aforementioned point mutation. Further, we put together differences of partial signals reconstructed from particles taken from the first and second level of WPD, the mutation effects are demonstrated in Figs. 3 (b) and (c), respectively.

For each level, at any point, the sum of these partial signals is approximately equal to the signal value in Fig. 3. Therefore, particle decomposition produces extra features describing additional information about the influence of the considered point mutation to adjacent amino acids.

[2]https://www.gisaid.org

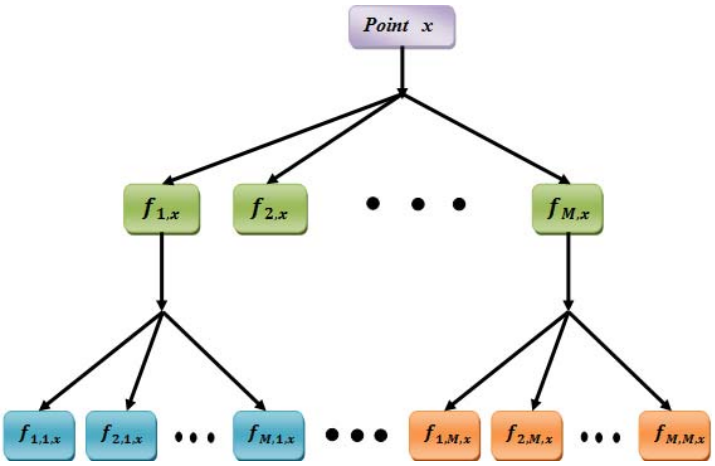

Fig. 3 Wavelet particle decomposition tree of a single point

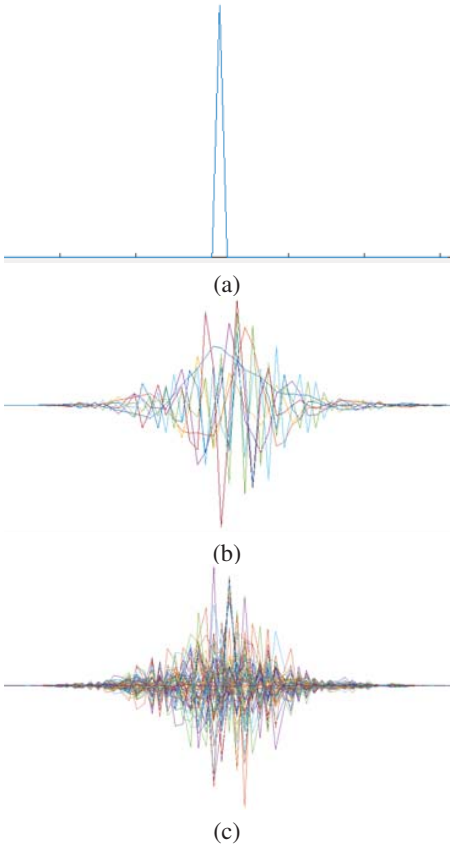

Fig. 4 Wavelet particle decomposition of a point mutation: (a) Difference between initial signals; (b) The difference between the partial signals of 1st level particles; (c) The difference between the partial signals of 2nd level particles

B. Experiment 2: Optimal Combination of Search Heuristics and WPD Technique

As shown in [3], [2], there exists a linear relationship between antigenic site substitutions and the logarithm of HI titer. Since the HA1 sequence is converted previously to numerical one, the similarity between test and reference strains in a specific site can be represented quantitatively as a difference between corresponding numerical values. This difference is also may be correlated with the logarithm of HI titer. To increase this correlation and overall significance of the feature extraction process, we use particles obtained in WPD.

Actually, we replace scalar differences between values taken in a single specific site by vector ones in a particle domain (for a certain level). Let S_1 and S_2 be two signals (sequences) of the same length k and $1 \leq x \leq k$ be some position (site) in this signals. We call the following quantity

$$D(S_1(x), S_2(x)) = \|\Omega(S_1(x)) - \Omega(S_2(x))\|$$

a vector distance between samples $S_1(x)$ and $S_2(x)$.

Generally speaking, to find the better feature, we should search through the entire tree (up to the given depth) provided by the WPD technique. Although this naive algorithm gives the best solution, it is time-consuming (since the WPD tree grows very quickly) and cannot be employed in practice. On the other hand, there are known numerous heuristic algorithms capable to provide solutions close to global optimal very efficiently. In this experiment, we consider the following heuristic algorithms to compare them with respect to the correlation improvement:

(i) Random Path (RP) from the root to a leaf (at any level, the next child node is chosen at random without dependence on the objective function)
(ii) Simple Hill Climbing (SHC) [23] (pick up the next node randomly, move to it, if it improves the objective function)
(iii) Steepest Ascent Hill Climbing (SAHC) [23] (at any level, for branching choose the child node that gives the highest positive improvement of the objective function).
(iv) Random Steepest Ascent Hill Climbing (RSAHC) (at any level, select child node for branching randomly among nodes that improve the value of the objective function)
(v) 2-top Steepest Ascent Hill Climbing (2SAHC) (at any level, try to branch at two child nodes that mostly improve the objective function)

TABLE I
THE COMPARISON OF SEARCH ALGORITHMS

	Search method						
	RP	SHC	SAHC	RSAHC	2SAHC	BFS	ES
ACI (%)	23.19	65.39	80.03	61.62	**92.38**	88.28	100
Search Time (sec)	68.93	51.40	50.72	**50.67**	95.73	104.41	266.19
ACI per sec	0.34	1.27	**1.58**	1.22	0.97	0.85	0.38

(vi) Best-first search (BFS) [24] with respect to the considered objective function
(vii) Exhaustive Search (ES) on WPD tree

Except for the RP, BFS and ES algorithm, we stop the search, if, at the current level, there is no child node that improves the objective function. The mentioned algorithms have been applied to 100 correlated sites (it is possible repeated site but with different numerical representation). To control the overall time consumption, we set the maximum depth of the WPD search tree to 5. Average Correlation Improvement (ACI) among the sites considered in percentage with respect to the result of ES algorithm together with the search time and relative ACI percentage per second are reported as follows:

The results indicate that highest ACI percentage was provided by 2SAHC heuristic, whereas the most efficient method appears to be SAHC heuristic since it gives the highest correlation with smaller time.

In order to prevent overfitting and select more relevant features for model construction, one of feature selection techniques can be applied. Earlier, the linear mixed effect model (LMEL) has been applied in mathematical modeling of virus antigenicity [2], [19]. To improve the overall modelling accuracy in terms of prediction of the antigenicity from protein sequences, we employ WPD, SHC and LMEL together, as a one combined technique.

C. Experiment 3: Significance of Neighbors Order

This experiment was organized as follows:

(i) for any site x and any alpha-numeric transformation I from AAindex database, we find well correlated feature $f[I](T(x), R(x))$ (suboptimal solution of problem (1)) using WPD (with wavelet basis Symlet sym8 and wavelet decomposition depth set to 3) in combination with SAHC searching algorithm presenting the best ACI per sec result in Experiment 2 (Tabl. I)
(ii) for a given threshold $th = 0.5$, we construct a set

$$\mathcal{XI} = \{(x, I) \colon AP(f[I], x \,|\, T, R, y) \geq th\}$$

(iii) produce a sample-set $\mathcal{XI}'$ of length 1000 by sampling randomly from $\mathcal{XI}$ without replacement
(iv) for any $(x, I) \in \mathcal{XI}'$, we perform random permutation of neighobrs amino acids and compute $AP(f[I], x \,|\, T, R, y)$ again.

The Fig. 5 demonstrates the obtained results where the blue and the orange graphs present result of correlation before and after permutation with the mean 0.57 and 0.054, respectively. The results indicate that the correlation vanishes when there is no significant order in the neighborhood.

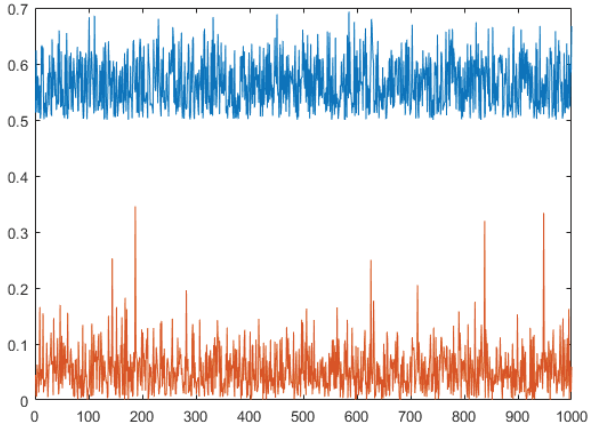

Fig. 5 Correlation vanishing after the random neighbor permutation

V. CONCLUSION

In this paper, we present the wavelet particle decomposition method, a technique for feature extraction from genetic sequences. The WPD provides the information about adjacent sites order at decomposition levels through convolution by wavelet packet filters. Despite conventional method for the study of the neighbor effect, WPD supplies adjustable neighborhood.

The results of numerical experiments reveal that WPD performance is sensitive to neighbors order. The performance of the method can be improved applying the heuristic search algorithm to decomposition tree in order to find an optimal feature. To evaluate the impact of search algorithm on WPD feature extraction, we apply them to HI assay database. According to the evaluation results the steepest ascent hill climbing algorithm is able to find an optimal feature with the highest correlation in less time. It should be better to point out that several factors can affect the obtained particles such as wavelet basis and depth of wavelet packet decomposition. It would be of interest to investigate their impact. In the future, we plan to conduct a more comprehensive comparison of heuristic search algorithms on WPD tree.

REFERENCES

[1] T. R. Klingen, S. Reimering, C. A. Guzmán, and A. C. McHardy, "In silico vaccine strain prediction for human influenza viruses," *Trends in Microbiology*, vol. 26, no. 2, 2018.
[2] W. T. Harvey, D. J. Benton, V. Gregory, J. P. Hall, R. S. Daniels, T. Bedford, D. T. Haydon, A. J. Hay, J. W. McCauley, and R. Reeve, "Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza a (h1n1) viruses," *PLoS pathogens*, vol. 12, no. 4, p. e1005526, 2016.
[3] W. T. Harvey, "Quantifying the genetic basis of antigenic variation among human influenza a viruses," Ph.D. dissertation, University of Glasgow, 2016.

[4] X. Xia and Z. Xie, "Protein structure, neighbor effect, and a new index of amino acid dissimilarities," *Molecular biology and evolution*, vol. 19, no. 1, pp. 58–67, 2002.
[5] T.-H. Kuo and K.-B. Li, "Predicting protein–protein interaction sites using sequence descriptors and site propensity of neighboring amino acids," *International journal of molecular sciences*, vol. 17, no. 11, p. 1788, 2016.
[6] W. Xue, X.-y. Hong, N. Zhao, R.-l. Yang, and L. Zhang, "Predicting protein subcellular localization by approximate nearest neighbor searching," in *Control And Decision Conference (CCDC), 2017 29th Chinese*. IEEE, 2017, pp. 2842–2846.
[7] M. Fu, Z. Huang, Y. Mao, and S. Tao, "Neighbor preferences of amino acids and context-dependent effects of amino acid substitutions in human, mouse, and dog," *International journal of molecular sciences*, vol. 15, no. 9, pp. 15 963–15 980, 2014.
[8] G.-Z. Wang, L.-L. Chen, and H.-Y. Zhang, "Neighboring-site effects of amino acid mutation," *Biochemical and biophysical research communications*, vol. 353, no. 3, pp. 531–534, 2007.
[9] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.
[10] P. Lio, "Wavelets in bioinformatics and computational biology: state of art and perspectives," *Bioinformatics*, vol. 19, no. 1, pp. 2–9, 2003.
[11] M. Cardelli, M. Nicoli, A. Bazzani, and C. Franceschi, "Application of wavelet packet transform to detect genetic polymorphisms by the analysis of inter-alu pcr patterns," *BMC bioinformatics*, vol. 11, no. 1, p. 593, 2010.
[12] R. Jiang and H. Yan, "Studies of spectral properties of short genes using the wavelet subspace hilbert–huang transform (wshht)," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 16-17, pp. 4223–4247, 2008.
[13] J. Ning, C. N. Moore, and J. C. Nelson, "Preliminary wavelet analysis of genomic sequences," in *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*. IEEE, 2003, pp. 509–510.
[14] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, "Fourier and wavelet transform analysis, a tool for visualising regular patterns in dna," *Journal of Theoretical Biology*, vol. 206, no. EPFL-ARTICLE-86700, pp. 323–326, 2000.
[15] J. Zhao, X. W. Yang, J. P. Li, and Y. Y. Tang, "Dna sequences classification based on wavelet packet analysis," in *Wavelet Analysis and Its Applications*. Springer, 2001, pp. 424–429.
[16] E. R. Dougherty, X. Cai, Y. Huang, S. Kim, and R. Yamaguchi, "Editorial [hot topic: Genomic signal processing: Part 1 (guest editors: Er dougherty, x. cai, y. huang, s. kim and r. yamaguchi)]," *Current Genomics*, vol. 10, no. 6, pp. 364–364, 2009.
[17] H. K. Kwan and S. B. Arniker, "Numerical representation of dna sequences," in *Electro/Information Technology, 2009. eit'09. IEEE International Conference on*. IEEE, 2009, pp. 307–310.
[18] G. K. Hirst, "The quantitative determination of influenza virus and antibodies by means of red cell agglutination," *Journal of Experimental Medicine*, vol. 75, no. 1, pp. 49–64, 1942.
[19] R. Reeve, B. Blignaut, J. J. Esterhuysen, P. Opperman, L. Matthews, E. E. Fry, T. A. De Beer, J. Theron, E. Rieder, W. Vosloo *et al.*, "Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus," *PLoS computational biology*, vol. 6, no. 12, p. e1001027, 2010.
[20] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *science*, vol. 305, no. 5682, pp. 371–376, 2004.
[21] D. K. Ruch and P. J. Van Fleet, *Wavelet theory: An elementary approach with applications*. John Wiley & Sons, 2011.
[22] V. Gregory, W. T. Harvey, R. S. Daniels, R. Reeve, L. Whittaker, C. Halai, A. Douglas, R. Gonsalves, J. J. Skehel, A. J. Hay, and J. W. McCauley, "Human former seasonal influenza A(H1N1) haemagglutination inhibition data 1977-2009 from the who collaborating centre for reference and research on influenza – London, UK," University of Glasgow, Tech. Rep., 2016.
[23] M. Harman, "The current state and future of search based software engineering," in *2007 Future of Software Engineering*. IEEE Computer Society, 2007, pp. 342–357.
[24] N. R. Vempaty, V. Kumar, and R. E. Korf, "Depth-first versus best-first search." in *AAAI*, 1991, pp. 434–440.