CIBTech Journal of Zoology ISSN: 2319–3883
Online, International Journal, Available at http://www.cibtech.org/cjz.htm
2025 Vol.14, pp.22-29/Thenmozhi and Arul
***Research Article** (Open Access)*

ASSESSMENT OF THE DISTRIBUTION OF SINGLET, DOUBLETS AND TRIPLETS AMINO ACID RESIDUES IN THE INFLUENZA-A VIRUS'S SEQUENCE AND SECONDARY STRUCTURE ELEMENTS

***P. Thenmozhi Kanaga and S. Arul Mugilan**
Kamarajar Government Arts College, Research Scholar, PG & Research Department of Physics, Surandai-627859 (India), Affiliated to Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli-627012, Tamil Nadu (India)
**Author for Correspondence*: *thenmozhikanaga@gmail.com*

ABSTRACT
In order to effectively navigate many difficult areas of biology, such as protein engineering, structural biology, and drug discovery, it is required to comprehend and integrate the sequence and structure of proteins. Here, we explain about how the influenza-A virus's amino acid residues are distributed throughout the protein's sequence and structure. The preferential and non-preferential amino acid residues of singlet, doublet, and triplet are obtained from the influenza-A virus's sequence analysis. We are able to assess which secondary structural features are more prevalent in influenza-A virus by using these preferential and non-preferential amino acid residues. Experimental biologists employed the aforementioned approach to predict tertiary structures and establish drugs.

Keywords: *Influenza-A Virus, Sequence, Secondary Structure Elements, Frequency of Occurrence, Deviation Parameter Value*

INTRODUCTION

Influenza-A is a highly contagious virus that poses an imminent danger to global health and the economy. It constitutes a pandemic threat. Thus, while we are learning that immunization is required, we can combat this infectious disease with viruses (Taubenberger *et al.,* 2005; Lipman 2008). Therapeutic drug creation is based on scientific investigation of the influenza-A virus's sequence (Cross *et al.,* 2012; Das *et al.,* 2010). A constant goal towards this is an improved understanding of influenza. Effective techniques to address the variety in the proteome sequence for vaccine creation, as well as the immune system and mutation dynamics. Plenty of studies have been conducted about the lookup of the influenza-A virus sequence. A significant obstacle to the development of an effective vaccine is the high sequence variability of influenza-A viruses. To develop antibodies to treat influenza, conserved sequences must be used as references (Krystian Eitner *et al.,* 2010; Qianyu Lin *et al.,* 2021; Vonderviszt *et al.,* 1986; Bakalkin *et al.,* 1991; Heiny *et al.,* 2007; Andreas Kukoln *et al.,* 2014; Brendel *et al.,* 1992). The primary technique for maintained sequence searches is sequence alignment, which includes multiple sequence alignment (MSA) and pairwise alignment. Sequence alignment is vital in the field of bioinformatics. Various sequence datasets yield varying performances for distinct MSA algorithms (Edgar 2004). The most used tool for MSA is Clustal. It constructs guide trees, computes the distance matrix via pairwise alignment, and performs progressive alignment by adhering to the guide tree.Consequently, in certain cases, conventional sequence alignment techniques like pairwise and multiple sequence alignment are not appropriate for finding a viable medication for the influenza-A virus (Fabian Sievers *et al.,* 2018). So we implemented new statistical approaches for sequence analysis that include calculating deviation parameter values. This approach allows us to identify preferential and non-preferential amino acid residues. The preferred amino acid residues regulate the drug of the influenza-A virus. The influenza-A virus is not among the non-preferred amino acid residues for therapeutic development.

Subsequently we determine the preferred amino acid, which form of secondary structural elements is more relevant for the identification of drugs for the influenza-A Virus. Therefore, employing the sequence, we additionally forecast the secondary structural features. Secondary structure prediction is based on its amino acid sequence. In addition to giving information regarding protein activity, relationships, and functions, the prediction of protein secondary structure represents an important initial step towards the prediction of tertiary structure. The α-helix (H) and β-strand (E) are the two regular secondary structure states, while the random structure is the only irregular secondary structure type. In accordance with hydrogen-bonding patterns, Sander created the Dictionary of Secondary Structure of Proteins (DSSP), a mechanism for automatically assigning secondary structure into eight states (H, E, B, T, S, L, G, and I). Helix, sheet, and coil are the three states that are commonly further reduced from these eight (Kabschw *et al.,* 1983; Hooft *et al.,* 1996; Juliette Martin *et al.,* 2005).

In recent years, several methods such as statistical methods (Chou and Fasman method and GOR method), machine learning techniques (Neural Networks, Hidden Markov Model, Support Vector) were developed to predict the secondary structural elements of proteins, such as α-helix, β-strand and random structures. The Chou-Fasman method was among the first secondary structure prediction algorithms developed and relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure (Chou and Fasman 1978).

For estimating the chance of each amino acid being present in all places, three scoring matrices are created using the GOR approach. When particular residues in the designated sliding window are detected, the neural network is trained to identify the structural state of the core residue that is most likely to occur (Garnier *et al.,* 1978). Select states for the hidden Markov model to indicate transmembrane localization, local or secondary structure types, or homologous sequence positions (Zheng Yuan 2005). By applying SVM to create classifiers for distinguishing parallel and antiparallel beta sheets (Christopher Bystroff *et al.,* 2008).As psiblast profiles, sequences are encoded. JPred, PSIPRED, RaptorX and a few other programs serve to predict secondary structural aspects (Christian Cole *et al.,* 2008).

Compared all other sequence analysis and secondary structure prediction our method is most useful in drug discovery of influenza-A virus. In this research, the sequence analysis was carried out by the deviation parameter value for all possible amino acid residues for singlets, doublets and triplets of influenza-A virus sequence. In structural analysis was carried out through the percentage of amino acid residues counts of secondary structural elements.

MATERIALS AND METHODS

Sequence analysis

For the analyses, substantial numbers of protein sequence data were obtained from the Protein Data Bank (Shindyalov *et al.,* 2000; Robbie *et al.,* 2011). From the PDB, we have extracted the protein IDs and sequences of 319 influenza-A viruses. We need to count the number of singlets, doublets, and triplets' amino acid residues from the sequence. For our purpose of counting amino acid residues, we wrote a Python program. We computed the deviation parameter value using these amino acid residues.

Structural analysis

For this work, the DSSP assignment method is relied on Wolfgang Kabsch and Chris Sander created the DSSP program to enable secondary structure assignments to be standardized Protein Data Bank (PDB) entries for all proteins have secondary structure assignments (as well as much more) in the DSSP database. Seven distinct secondary structures are assigned by DSSP: H stands for alpha-helix, G for 3/10 helix, I for pi-helix, E for extended strand, B for residue in isolated beta-bridge, S for bend, and T for H-bonded turn. We require the DSSP file in order to analyse the influenza-A virus's structure. There are 319 protein IDs in the influenza-A virus. These Ids are gathered from the protein data bank. The DSSP files for the associated PDB Ids are downloaded from the DSSP database. We acquired 317 DSSP files out of 319 PDB IDs. More information is available in the DSSP file. Only the structure information of the singlet, doublets, and triplets amino acid residue counts is required. We create a Python program to count

the residues of amino acids in secondary structural elements, such as random, beta-sheet, and alpha-helix structures. Using these amino acid residues, we may calculate the percentage of amino acid residues in secondary structural elements.

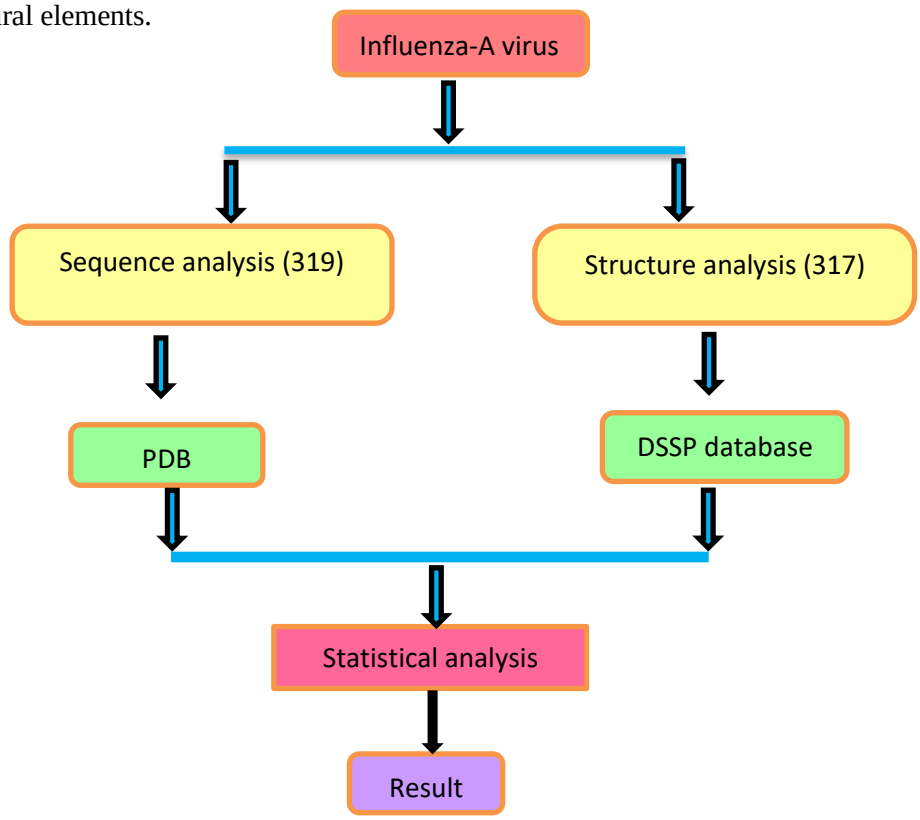

Figure 1: Flow chart

Statistical analysis

The following formula (1) can be applied to calculate the frequency of occurrence (Veluraja *et al.*, 1997; Mugilan *et al.*, 2000).

$$\text{Frequency of occurrence}=\frac{\sum N_i(S)}{\sum Y_i(S)} \quad \text{---------- (1)}$$

The same formula was used for doublets and triplets.

$N_i(S)$ - total number of individual amino acid

$N_i(D)$-total number of two consecutive amino acid residues

$N_i(T)$-total number of three consecutive amino acid residues

$Y_i(S)$, $Y_i(D)$, $Y_i(T)$- total number of amino residues in the i^{th} protein for singlets, doublets, triplets.

i-Number of protein (319)

The deviation parameter can be estimated using the formula (2)

$$\text{Deviation parameter}=\frac{\text{Observed value}-\text{Expected value}}{\text{Expected value}}\times 100 \quad \text{---------- (2)}$$

The observed value is interpreted as the frequency at which influenza-A amino acid residues occur. The expected value is defined as the frequency with which the total number of amino acid residues occurs in

viral protein. The deviation parameter values for doublets and triplets were calculated using the same formula. For Doublet the observed value is taken as frequency of occurrence of two consecutive amino acid residues, for triplets the observed value is taken as frequency of occurrence of three consecutive amino acid residues. The expected value of doublets is taken as frequency of occurrence of two consecutive amino acid residues of viral protein. The expected value of triplets is taken as frequency of occurrence of three consecutive amino acid residues of viral protein.

RESULTS

The results of the singlet, doublets, and triplets' amino acid residue analyses of the influenza-A virus were discussed in this section, along with their positive and negative significance. Additionally, we addressed about the findings concerning which amino acid residues are more preferred in alpha-helix, beta-sheet, and random structure secondary structural elements. The influenza-A virus sequence (319) has a total of 161620 singlets, 161239 doublets, and 59827 triplets' amino acid residues. The total number of amino acid residues in a singlet is 337804, a doublet is 327771, and a triplet is 316648, according to the structural information of the DSSP file (317). It includes a 20% alpha-helix, 30% beta-sheet, and 50% random structure in this total amount of residues in singlet, doublet, and triplet. We only incorporate the preferential and non-preferential amino acid residues that belong to the singlet, doublet, and triplet codes in the sequence analysis in our structure study.

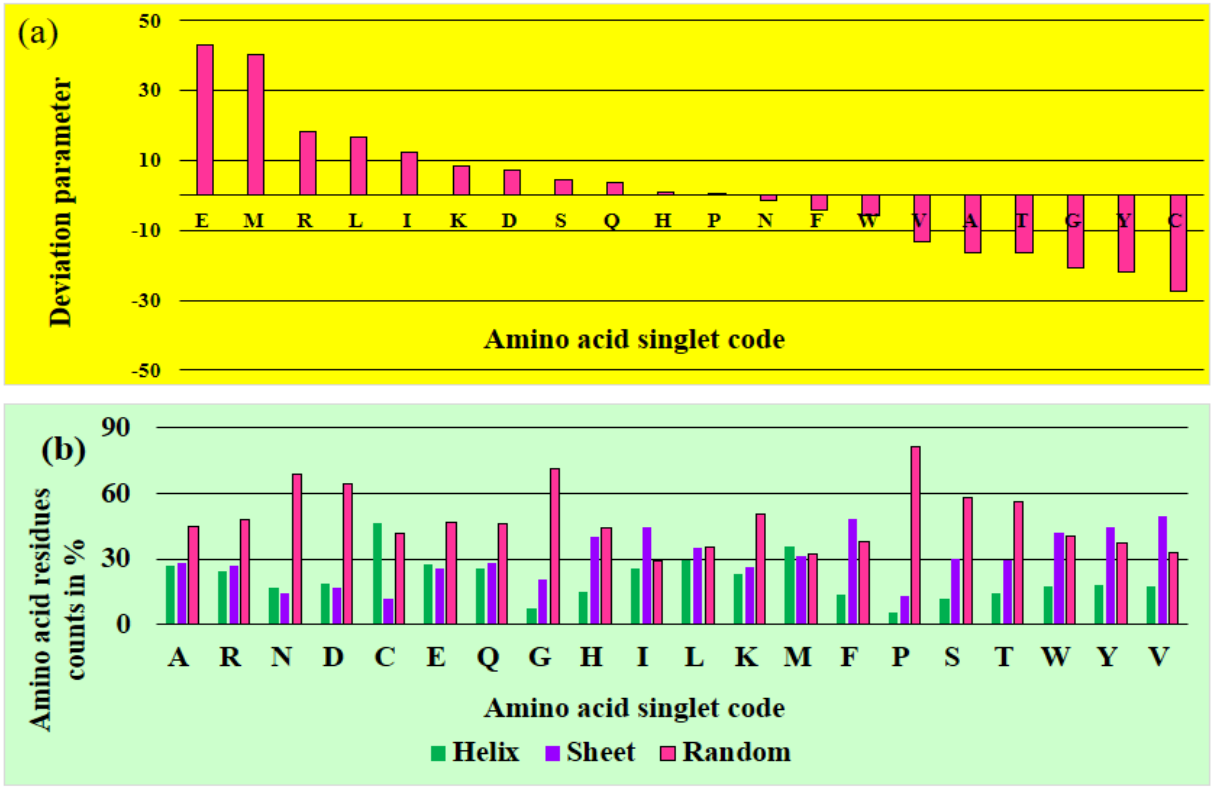

Figure 2(a): Preferential and non-preferential amino acid residues of singlet code
(b): Distribution of Secondary structural elements of singlet code

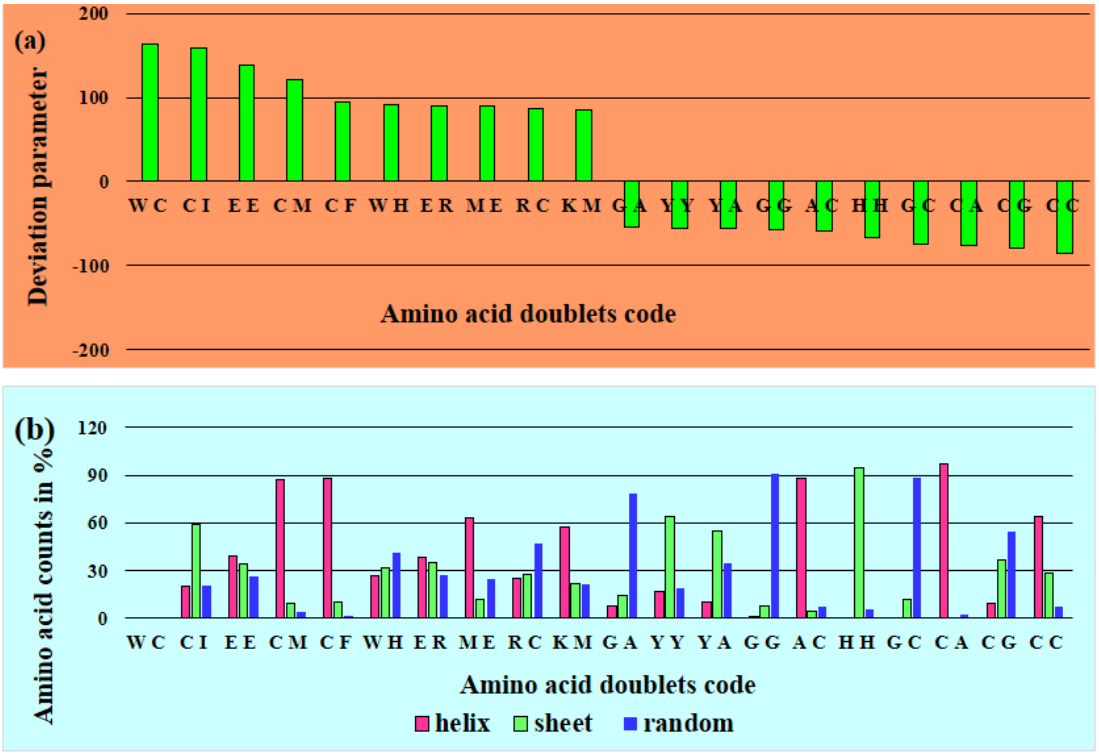

Figure 3 (a): Positive and negative significant amino acid residues doublets
(b): Distribution of secondary structural elements of doublets code

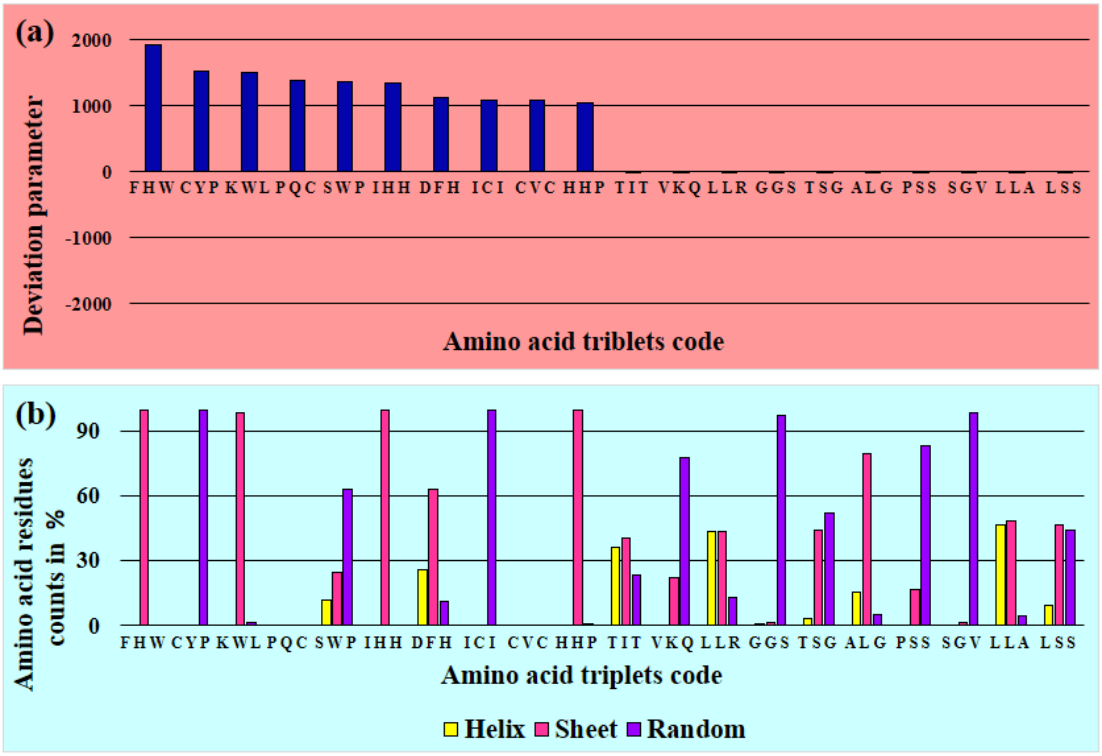

Figure 4 (a): Positive and negative significant amino acid residues of triplets
(b): Distribution of secondary structural elements of triplets code

DISCUSSION

The positive and negative sides of the aforementioned graphs are referred to as preferential and non-preferential amino acid residues, respectively. According to the graph, the first high value on the positive side represents significant (preferential) amino acid residues, and the lowest value on the negative side represents least significant (non-preferential) residues. From graph 1(a) Glutamic acid (E) is more excited in positive side. These are known as preferential amino acid residues. In the negative side region, cysteine (C) is more enthusiastic. So, it is known as non-preferential amino acid residues.

The structural analysis figure 2(b) is compared with sequence analysis figure 2(a). The preferential amino acid residues only more important role in formation of secondary structure elements of alpha-helix, beta – sheet and random structure. Glutamic acid has 27% alpha-helix, 26% beta-sheet, and 47% random structure in influenza-A Virus. Thus, non-preferential amino acid residues least role in the formation of secondary structural elements. Cysteine (C) was preferred the alpha-helix content is 46% in 12% in beta-sheet, and 42% in random structure. Only the preferred amino acid residues are taken into account when predicting the structure. Due to its higher percentage of amino acid residues than alpha-helix and beta-sheet, glutamic acid is the preferred amino acid residue in the influenza-A virus and has a preferred random structure (pi-helix, 3/10-helix, isolated-beta-bridge, and beta-turn).

Figure 3(a) suggests that, WC is more delighted in the positive side. This is the reason it's called residues of privileged amino acids. The CC is thrilled in the negative side region. Hence, non-preferential amino acid residues are the term used to describe it. The components of C and M play an important role in preferential and Y and G components are more important in non-preferential amino acid residues of influenza –A virus.

Figure 3(b) demonstrates that WC lacks any secondary structural elements. Because WC contains no amino acid residue counts of secondary structural components. However, it has amino acid counts of doublets in the sequence of influenza-A virus. As a result, we consider the subsequent highest value of preferred amino acid residues. When doublets are analysed in sequence, the next preferred amino acid residue is called CI. Among its structural components, CI is composed of 21% alpha helices, 58% beta sheets, and 21% random structures. The amino acid residues of doublets with the least significance are CC.CC is composed of 29% beta-sheets, 7% random structures, and 64% alpha-helices.

In accordance with figure 4(a), the positive range is where FHW and CYP are more activated. This is whether it's called residues of preferential amino acids. More excitement is seen in negative values for LSS and LLA. Consequently, non-preferential amino acid residues are the name given to it.

From figure 4(b), LSS prepared 9% alpha-helix, 47% beta-sheet, and 44% random structure yet FHW prepared 100% beta-sheet and CYP prepared 100% random structure.

In the previous literature assessment, individual amino acid residues of viral protein E (glutamic acid) chose the alpha-helix form, but WC pairs of amino acid residues preferred the beta-sheet structure. However, in our investigation, the random structure was selected by the influenza-A virus glutamic acid, while secondary structural elements were not preferred by WC. Because our analysis only takes the influenza-A virus. The whole viral protein was used in earlier literature surveys (Veluraja et al 1997; Mugilan 2000). We simply calculated and evaluated the first three consecutive amino acid residues of the sequence and secondary structural components using our approach other researchers discover secondary structural elements analysis of single and two consecutive amino acid residues of sequence.

CONCLUSION

To analyse the influenza-A virus's sequence preferential and non-preferential amino acid residues of singlets, doublets and triplets using the Sequence result of preferential of amino acid residues we got the result of more favoured secondary structure elements in influenza-A virus. Amino acid residues with singlet preference favoured the random structure. Preferential amino acid residues of two and three consecutive amino acid residues favoured the beta-sheet structure. From the outcome, we infer the drug of the influenza-A virus by taking into account the secondary structural components of the preferred amino

acid residues. The experimental biologist implemented the aforementioned information to aid in the creation of drugs.

ACKNOWLEDGEMENT
We would like to acknowledge my professor, Dr. S. Arul Mugilan, for his support with this work.

REFERENCES
Taubenberger JK, Reid AH, Lourens RM (2005). Characterization of the 1918 influenza virus polymerase genes. Nature **437** 889–893.
Lipman D (2008). The influenza virus resource at the National Center for Biotechnology Information. Journal of Virology **82 (2)** 596–601.
Du J, Cross TA and Zhou HX (2012). Recent progress in structure-based anti- influenza drug design. Drug Discovery Today **17(19–20)** 1111–1120.
Das K, Aramini JM, Ma LC, Krug RM, Arnold E (2010). Structures of influenza-A proteins and insights into anti-viral drug targets. Nature Structural & Molecular Biology **17(5)** 530–538.
Krystian Eitner, Uwe Koch, Tomasz Gaweda, Jedrzej Marciniak (2010). Statistical distribution of amino acid sequences: a proof of Darwinian evolution. Bioinformatics **23** 2933–2935.
Qianyu Lin, Xiang Ji, Feng Wu, Lan Ma (2021). Conserved Sequence Analysis of Influenza A Virus HA Segment and Its Application in Rapid Typing. Diagnostics **11** 1328.
Vondervisztf F, Mátrai G, Simon I (1986). Characteristic sequential residue environment of amino acids in proteins. International Journal of Peptide and Protein Research **27(5)** 483–492.
Bakalkin GY, Rakhmaninova AB, Akparov VK, Volodin AA, Ovchinnikov VV, Sarkisyan RA (1991). Amino acid sequence pattern in the regulatory peptides. International journal of peptide and protein research **38** 505-10.
Heiny AT, Miotto O, Srinivasan KN, Khan MA, Zhang GL, Brusic V, Tan TW (2007). Evolutionarily conserved protein sequences of influenza-A viruses, avian and human, as vaccine targets. PLOS ONE **2(11)** e1190.
Andreas Kukoln and David John Hughes (2014). Large-scale analysis of influenza-A virus nucleoprotein sequence conservation reveals potential drug-target sites. Virology **454** 40-47.
Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992). Methods and algorithms for statistical analysis of protein sequences. The Proceedings of the National Academy of Sciences **89(6)** 2002-6.
Edgar RC (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5** 113.
Fabian Sievers , Desmond G, Higgins (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Science **27(1)** 135–145.
Kabsch and Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22(12)** 2577-637.
Hooft RWW, Sander C, Scharf M, Vriend G (1996). The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. International Society for Computational Biology **12** 525-529.
Yuan Zhang and Celeste Saguia (2015). Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI. Journal of Molecular Graphics and Modelling **55** 72-84.
Juliette Martin, Guillaume Letellier, Antoine Marin, Jean-François Taly, Alexandre G de Brevern, Jean-Francois Gibrat (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Structural Biology 5-17.
Chou PY and Fasman GD (1978). Prediction of the secondary structure of protein from their amino acid sequence. Advances in Enzymology 45–148.

Garnier J, Osguthorpe DJ, Robson B (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology 97–120.
Zheng Yuan (2005). Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinformatics **6** 248.
Christopher Bystroff and Anders Krogh (2008). Hidden Markov Models for prediction of protein features Methods. Molecular Biology **413** 173-98.
Christian Cole, Jonathan D, Barber, and Geoffrey J, Barton (2008).The JPred 3 secondary structure prediction server. Nucleic Acids Research **36** W197–W201.
Shindyalov IN and Bourne PE (2000). The Protein Data Bank. Nucleic Acids Research **28 (1)** 235–242.
Robbie P.Joosten, Tim AH, te Beek, Elmar Krieger, Maarten L (2011).A series of PDB related databases for everyday needs. Nucleic Acids Research **39** D411–D419.
Veluraja K and Mugilan SA (1997). Amino acid doublets and triplets in protein sequences – A database analysis. Current Science **72** 572-577.
Mugilan SA and Veluraja K (2000). Generation of deviation parameters for amino acid singlets, doublets and triplets from three-dimensional structures of proteins and its implications for secondary structure prediction from amino acid sequences. Indian Academy of Sciences **25** 81–91.