Diagnostic Pathology: Open Access

Krem, Diagn Pathol Open 2021, 6:S2

Short Communication Open Access

The Application of Peptide Visualizer to Coronavirus

Samuel Krem*

Department of Science, Buckingham Browne and Nichols School, Cambridge, MA 02138, United States

Abstract

Following the publication of the DNA nucleotide footprint plotter, the peptide visualizer was developed to study peptide sequences of coronaviruses. It provides a distinctive view to peptide characteristics across different coronavirus species. The visualizer is sensitive to subtle changes of peptide sequences and has the potential to improve the diagnosis, therapy, and prevention of coronavirus infections. The application of the tool is not limited to plotting coronavirus peptides. It has general applications in visualizing changes of peptide sequences, structures, and functions in prokaryote and eukaryote organisms.

Keywords: Coronavirus; Peptide; Visualization; Computing; Mutation

Introduction

Peptides are short protein fragments consisting of various amino acids (Table 1). The amino acids contribute to different functions in proteins depending on whether they are polar or nonpolar and whether they carry positive or negative charges. By displaying the molecular weight, polarity, and changes of amino acids using color and geometric shapes, the peptide visualizer demonstrates protein characteristics in a novel and straightforward way [1].

Methods

The tool was written in Java, and the steps are shown in Figure 1. The peptide visualizer uses different colors, segment lengths, and segment angles to represent different amino acids. The length of the segment is proportional to the molecular weight of the amino acid. The color follows the amino acid color RGB tradition. For positive charged amino acids, the segment angel turns clockwise by +th1. For negative charged amino acids, the segment angel turns counter clockwise by th1. Polar amino acids turn th2 according to the original orientation, and nonpolar amino acids turn by a smaller angle, th0. The users have options to adjust the above parameters according to the need of specific effects of different plotting. In the current study, we define the parameters as: th0=50°, th1=90°, th2=60°; k=1 for positive charge, and k=-1 for negative charge. All coronavirus sequences were retrieved from NCBI virus data hub [2]. Yeast protein sequences were retrieved from NCBI genome database NCBI genome database. Plasmid protein sequences were retrieved from NCBI GenBank.

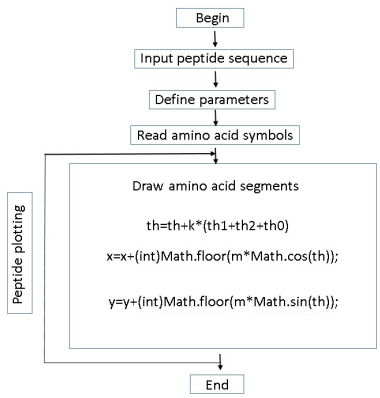

Figure 1: Flowchart of peptide visualizer.

Results

We plotted coronavirus peptides using the peptide visualizer. The different characteristics of various proteins are demonstrated in Figure 2.

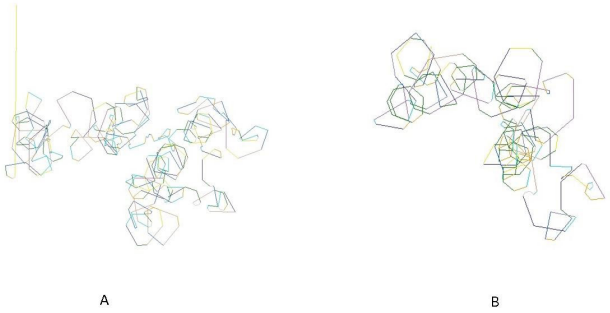

Figure 2: Peptide visualizer plots of *Scotophilus* bat coronavirus 512 nucleocapsid protein YP_1351688 BatCV_394 (A) and matrix protein YP_1351687 BatCV_227 (B).

Comparing with the DNA nucleotide footprint (Figure 3), the output of the peptide visualizer is more sensitive to demonstrating peptide sequence differences (Figure 4). While the DNA footprint plotter is useful to detect structural variations, minor differences in the sequences and protein features are more obvious in the peptide visualizer plots.

The S protein of coronaviruses, especially the receptor-binding domain (RBD), is a common target region for vaccine development [3-6]. The peptide visualizer was further applied to compare various

***Corresponding author:** Samuel Krem, Buckingham Browne and Nichols, Cambridge, MA 02138, United States, E-mail: skrem@bbns.org

Received: April 19, 2021; **Accepted:** May 03, 2021; **Published:** May 10, 2021

Citation: Krem S (2021) The Application of Peptide Visualizer to Coronavirus. Diagnos Pathol Open S2: 005.

Amino Acid	Symbol	IUPAC code	Molecular weight	Polarity and charge	θ	RGB
Alanine	Ala	A	15	nonpolar	th0	200,200,200
Arginine	Arg	R	101	positive charge	th1	20,90,255
Asparagine	Asn	N	58	negative charge	th1	0,220,220
Aspartic acid	Asp	D	58	polar	th2	230,230,10
Cysteine	Cys	C	47	polar	th2	230,230,0
Glutamic acid	Glu	E	72	polar	th2	230,230,10
Glutamine	Gln	Q	72	negative charge	th1	0,220,220
Glycine	Gly	G	1	polar	th2	235,235,235
Histidine	His	H	82	positive charge	th1	130,130,210
Isoleucine	Ile	I	57	nonpolar	th0	15,130,15
Leucine	Leu	L	57	nonpolar	th0	15,130,15
Lysine	Lys	K	93	positive charge	th1	20,90,255
Methionine	Met	M	74	nonpolar	th0	230,230,0
Methionine	Phe	F	91	nonpolar	th0	50,50,170
Proline	Pro	P	114	nonpolar	th0	220,150,130
Serine	Ser	S	31	polar	th2	250,150,0
Threonine	Thr	T	44	polar	th2	250,150,0
Tryptophan	Trp	W	130	nonpolar	th0	180,90,180
Tyrosine	Tyr	Y	107	polar	th2	50,50,170
Valine	Val	V	45	nonpolar	th0	15,130,15

Table 1: Amino acids and their features.

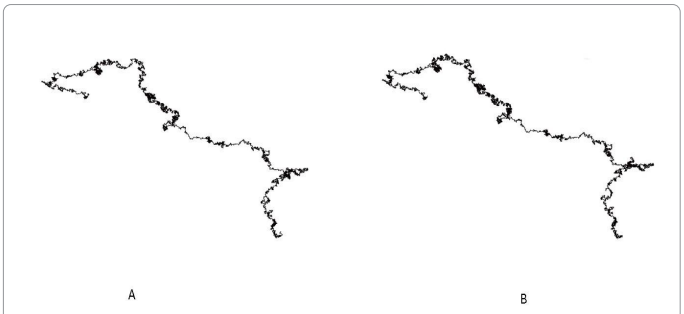

Figure 3: DNA nucleotide footprint of severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-05812/2020 MW181438 (A) and SARS-CoV-2/human/USA/UT-UPHL-200716494/2020 MW181439 (B).

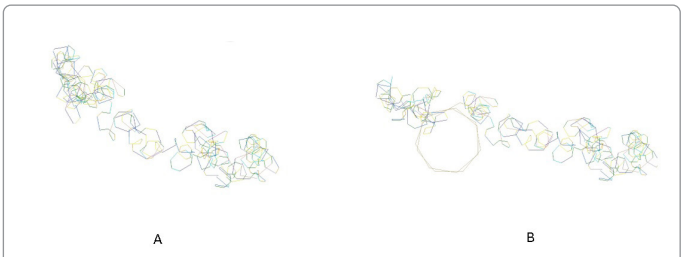

Figure 4: Peptide visualizer plots of severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-05812/2020 QOS45414 orf1a amino acid 0-500 (leader protein of polyprotein) (A) and SARS-CoV-2/human/USA/UT-UPHL-200716494/2020 QOS45427 orf1a amino acid 0-500 (B).

coronavirus strains such as OC43, HKU1, MERS-CoV (Middle East Respiratory Syndrome), and SARS-CoV. S protein amino acids 300-600 covering the RBD region were plotted in Figure 5 (DNA footprint of the corresponding species see Figure 5 of reference 1). The bigger differences between coronavirus strains and the subtle differences between coronavirus sub-strains are clearly displayed.

We also plotted plasmid and yeast proteins (Figures 6 and 7). Figure 6A is a plot of the plasmid protein IncHI1 transfer repressor of Escherichia coli strain 63743 plasmid pEQ2, GenBank accession number KF362122.2 (protein_id="AHC56037.1"). Figure 6B is the plot of the plasmid protein repHI2 of Escherichia coli strain A74 plasmid pA74T, GenBank accession number MG014720.1 (protein_id="AVR59929.1"). Figures 7A and 7B are Saccharomyces cerevisiae YJM1388 proteins Vac 17p (accession number AJQ32742.1) and Adh 7p (accession number AJQ32734.1).

Discussion

Compared with the DNA nucleotide footprint plotter [1], the peptide plotter is more sensitive to demonstrating the impact of single nucleotide mutations, insertions, and deletions. The geometric shapes change more obviously than plotting at the DNA level.

There are many essential tools to compare protein sequences such as Blast [7-27] and Clustal [28,29]. These tools can clearly display the matching and mismatching of protein sequences. But they do not demonstrate protein characteristics. Let's look at 2 mismatching situations. In the first situation, there is an amino acid difference between two protein sequences, but the mismatching amino acids have similar charge and polarity. In the second situation, the 2 mismatching amino acids have very different polarity and charges. Blast and Clustal display the mismatching of the 2 situations similarly (a letter difference). However, using peptide plotter, if the mismatching amino acids have similar polarity and charge, the plot will look similar even if the amino acids are different. If the mismatching amino acids have different polarity or charges, the plot will look very different. Peptide visualizer is an add on to the existing tools. It is very easy to run and is very fast. It helps to display peptide features in a novel and convenient way.

The peptide plots directly indicate the regions with similar or different characteristics across peptide sequences. Specific geometric shapes of coronavirus proteins distinguish the coronavirus infection from other viruses. The colorful geometric shapes can identify the mutations in corona viruses with or without functional impact. In addition, the tool can tell similar peptide regions across coronavirus

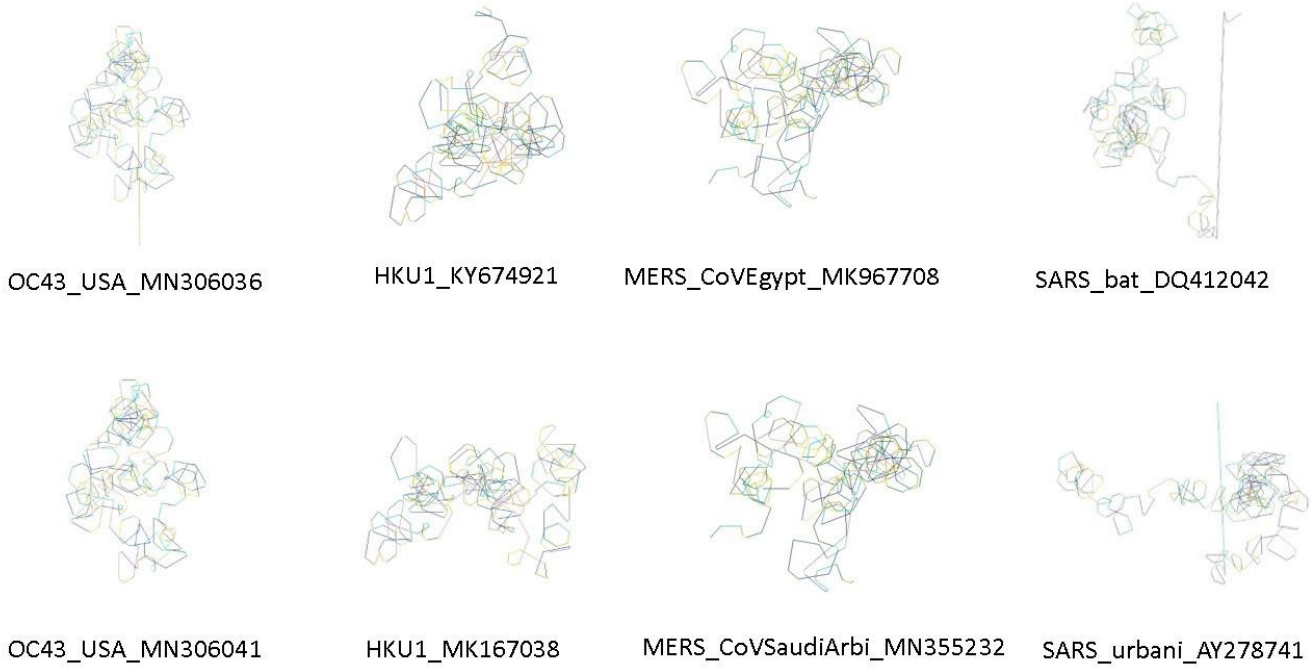

Figure 5: Peptide visualizer plots of S protein amino acids 300-600 covering the RBD region of different coronaviruses.

Figure 6: Peptide visualizer plots of plasmid proteins. A: IncHI1 transfer repressor (AHC56037.1) and B: repHI2 (AVR59929.1).

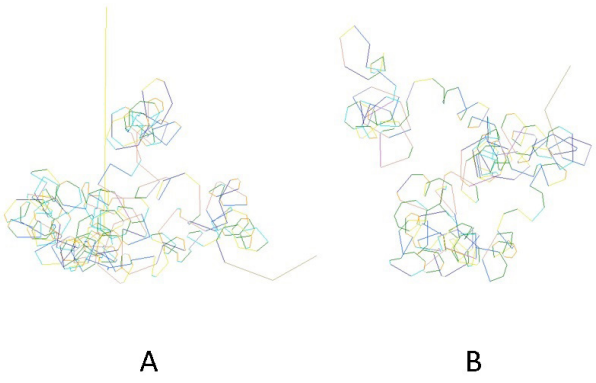

Figure 7: Peptide visualizer plots of yeast proteins. A: Vac17p (AJQ32742.1). B:Adh7p (AJQ32734.1).

species. This facilitates vaccine development and makes it possible for one vaccine to fight different coronavirus sub-species by targeting the similar regions across mutating strains.

Conclusion

Based on the large amount of peptide sequence data, we can code according to the shape of peptide sequences. This will facilitate peptide recognition and clustering using machine learning approaches such as neural networks. It will be the emphasis of my future study.

Acknowledgements

I want to thank my school Buckingham Browne and Nichols for teaching me essential knowledge in biology and Java programming skills to make this paper possible. I also want to thank my family to excuse me from house chores and allow me time to work on the project.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Krem S (2020) The application of DNA nucleotide footprint plotting in coronavirus. Inform Med Unlocked 19: 100358.
2. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, et al. (2017) Virus Variation Resource-improved response to emergent viral outbreaks. Nucleic Acids Res 45: D482-D490.
3. Samrat SK, Tharappel AM, Li Z, Li H (2020) Prospect of SARS-CoV-2 spike protein: Potential role in vaccine and therapeutic development. Virus Res 288: 198141.
4. Dai L, Gao GF (2020) Viral targets for vaccines against COVID-19. Nat Rev Immunol 18: 1-10.
5. Lan J, Ge J, Yu J, Shan S, Zhou H, et al. (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581: 215-220.
6. Tai W, He L, Zhang X, Pu J, Voronin D, et al. (2020) Characterization of the receptor-

binding domain (RBD) of 2019 novel coronavirus: Implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cell Mol Immunol 17: 613-620.

7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
8. States DJ, Gish W (1994) Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol 1: 39-50.
9. Madden TL, Tatusov RL, Zhang J (1996) Applications of network BLAST server. Methods Enzymol 266: 131-141.
10. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
11. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7: 203-214.
12. Zhang J, Madden TL (1997) PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. Genome Res 7: 649-656.
13. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, et al. (2008) Database indexing for production MegaBLAST searches. Bioinformatics 24:1757-1764.
14. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10: 421.
15. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, et al. (2012) Domain enhanced lookup time accelerated BLAST. Biol Direct 7: 12.
16. Altschul SF, Boguski MS, Gish W, Wootton JC (1994) Issues in searching molecular sequence databases. Nat Genet 6: 119-129.
17. McGinnis S, Madden TL (2004) BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20-25.
18. Ye J, McGinnis S, Madden TL (2006) BLAST: Improvements for better sequence analysis. Nucleic Acids Res 34: W6-9.
19. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, et al. (2013) BLAST: A more efficient report with usability improvements. Nucleic Acids Res 41: W29-33.
20. Shiryev SA, Papadopoulos JS, Schäffer AA, Agarwala R (2007) Improved BLAST searches using longer words for protein seeding. Bioinformatics 23: 2949-2451.
21. Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 219: 555-565.
22. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915-10919.
23. Altschul SF (1993) A protein alignment scoring system sensitive at all evolutionary distances. J Mol Evol 36: 290-300.
24. Altschul SF (1993) A protein alignment scoring system sensitive at all evolutionary distances. J Mol Evol 36: 290-300.
25. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 87: 2264-2268.
26. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994-3005.
27. Park Y, Sheetlin S, Ma N, Madden TL, Spouge JL (2012) New finite-size correction for local alignment score distributions. BMC Res Notes 5: 286.
28. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high□quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7: 539.
29. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, et al. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 38: W695-699.