Sonification of Proteins for Analysis

Ananya Gupta and Suhanee Zaroo

San Jose State University

CS 22B: Python Programming for Data Analysis

Aarohi Chopra

Table of Contents

List of Figures

Background

Data Sonification refers to converting data to sound and analyzing that to discover patterns or anomalies. As most computational scientists rely on visual imaging techniques to study large datasets, using sound is a relatively newer technique for data analysis (Kaper et al., 1999). However, many benefits are being discovered for sonification. As the volume of data to be analyzed grows, sonification can be a more efficient method of analyzing large amounts of data. It can also be a valuable tool to make the data more comprehensible for a wider audience by utilizing human psychoacoustic intuitions (Martin et al., 2021).

At the moment, protein structure is an emerging and intensively researched topic in biology. Proteins are large, complex, species-specific molecules that perform various critical roles in an organism, including structure, function, and regulation of tissue development. It is translated from an RNA gene sequence into a sequence of amino acids joined together. Twenty different amino acids constitute a protein. The prediction of protein three-dimensional structure from amino acid sequence has been a challenge in computational biology for decades, owing to its intrinsic scientific interest and the many potential applications for robust protein structure prediction algorithms, from genome interpretation to protein function prediction. More recently, there has also been a growing interest in designing an amino acid sequence that will fold into a specified three-dimensional structure as a potential route to engineering proteins with functions useful in biotechnology and medicine (Kuhlman & Bradley, 2019).

The idea of conversion of data from a protein sequence to sound was first documented in a 1980 book, *Godel, Escher, Bach: An Eternal Golden Braid,* owing to various similarities between a musical composition and protein sequences. The first paper on utilizing this technique was published in 1996, and ever since, computational biologists have come up with more sophisticated models to map the structure of proteins to music. Markus Buehler's team at MIT has pioneered an artificial intelligence system to study this music catalog generated from different proteins (Yu & Buehler, 2020).

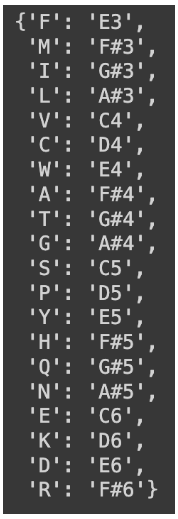
Hypothesis

For this study, we present a program that maps a protein sequence to a sequence of musical notes based on the Goldman-Engelman-Steitz (GES) hydrophobicity scale of the amino acids. By doing so, we hypothesize a clear audio difference in the protein sequences of β-actin and GAPDH in humans, which assists in understanding the positioning of hydrophobic amino acids in a sequence of two standard human proteins with known structures. We also aim to compare sonification and visualization methods of understanding the underlying chemical properties of amino acids with hydrophilicity plots using the Hopp-Woods scale (Hopp et al. (1989). Hydrophobicity is a property of amino acids responsible for stability during protein folding and has a critical role in determining a protein's architecture and potentially predicting it for unknown sequences (Moelbert et al., 2004).

Methods

The amino acid sequences of two human proteins, β-actin and GAPDH, were obtained using the Biopython Library. The Biopython Project is an international association of developers of freely available Python tools for computational molecular biology (Cock et al., 2009). The goal of this library is to make day-to-day bioinformatics tasks easier with the use of biology-based, reusable objects and classes. The protein sequences of interest were obtained in Python by programmatically accessing the NCBI protein database using Bio's Entrez.efetch package using their unique GI numbers. This generated SeqRecord objects, which were then parsed using the SeqIO.read() method. Finally, the amino acid sequences were saved as strings for further analysis.

The library used for musical mapping was music21, an open-source toolkit in Python for analyzing, searching, and transforming musical data (Cuthbert et al., 2011). Amino acids were mapped to a musical note based on their hydrophobicity, as determined by the Goldman-Engelman-Steitz (GES) hydrophobicity scale. (Martin et al., 2021). Each amino acid was mapped to a corresponding note and stored in a dictionary using the Musical Instrument Digital Interface (MIDI) standard. This standard transmits and stores information about a musical note, timing, and pitch and can be used to play notes in Google Colab or opened in a separate application to modify the audio file further (Wright, 2023). The MIDI standards for each amino acid were determined from a previous study by Martin et al. in 2021, where the hydrophobic amino acids in the list were assigned lower pitches. For this study, we assigned the amino acids to MIDI numbers 30-70 and converted them to their corresponding notes, as shown in Figure 1.



```
{'F': 'E3',
 'M': 'F#3',
 'I': 'G#3',
 'L': 'A#3',
 'V': 'C4',
 'C': 'D4',
 'W': 'E4',
 'A': 'F#4',
 'T': 'G#4',
 'G': 'A#4',
 'S': 'C5',
 'P': 'D5',
 'Y': 'E5',
 'H': 'F#5',
 'Q': 'G#5',
 'N': 'A#5',
 'E': 'C6',
 'K': 'D6',
 'D': 'E6',
 'R': 'F#6'}
```

Fig 1: Mapping of Amino Acids to MIDI notes

Then, a function was defined to convert a sequence of amino acids into a sequence of music notes, which was applied to the two proteins imported previously. Two separate lists were made with the same musical parameters (tempo and duration) determined from music21, to which the notes of the sequences were appended. These lists were written and downloaded as MIDI files, opened in GarageBand in macOS, and modified with the addition of percussion instruments. The files were also streamed on Google Colab.

Hydrophilicity plots of both the protein sequences were constructed to compare the effectiveness of sonification methods to visualization methods of studying amino acid properties. The Hopp and Woods hydrophilicity scale was used for the same. The necessary Python libraries and subsequent modules were imported. Using the ProtParam module of Biopython, the ProteinAnalysis function was used to create sequence objects from the amino acid strings. The plotting points of the amino acid residues were generated using the protein_scale(Scale, WindowSize, Edge) method on these sequence objects. To make the visualization process easy to compare to the music generated by protein sequences, the Hopp and Woods scale of hydrophilicity was used. The plots were generated in Python using the Matplotlib Library. A real-time graph of these hydrophilicity plots was also made using the FuncAnimation module from matplotlib.animation. The hydrophilicity measurements and residues of the two sequences were saved in Pandas DataFrames with column names “Residues” and “Hydrophilicity.” The real-time plots were displayed using the HTML display method in the Google Colab notebook (YouTube, 2022).

Results

The twenty amino acids in the protein sequences imported were mapped between MIDI numbers 30-70. The sequence of notes obtained from these MIDI numbers was played through music21 in Colab, downloaded as a MIDI file, and opened in GarageBand using

MacOS. The piano instrument was accompanied by in-built percussion options provided by GarageBand to sound pleasant.

The modulation in pitch could be heard in the output music file obtained. This pitch modulation, representative of the hydrophobic and hydrophilic nature of amino acids, was visualized in hydrophilicity plots. The audio's pitch corresponded to the peaks on the graphs plotted. The positioning of the hydrophobic residues in the sequence, corresponding to the lower pitch notes, gives the user a rough idea about the position of the residue within the folded structure of the protein. Through sonified sequences, the differences in their hydropathic nature were understood a lot more effectively than using just visual methods such as plotting.

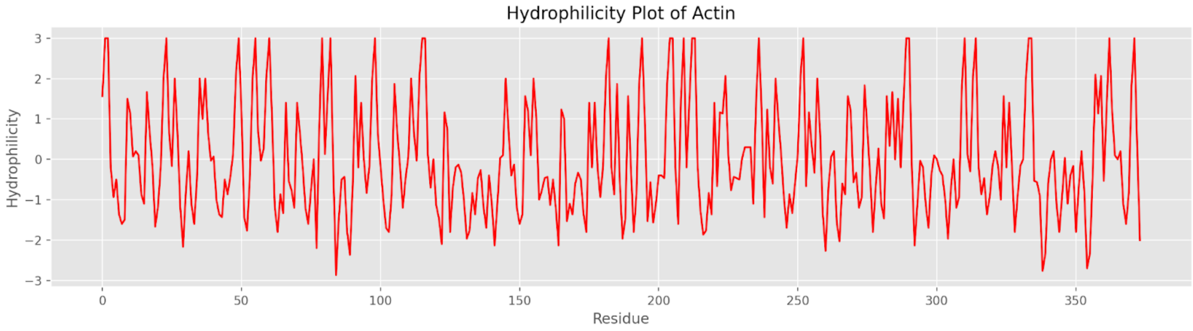


Fig 2: Hydrophilicity Plot of Actin

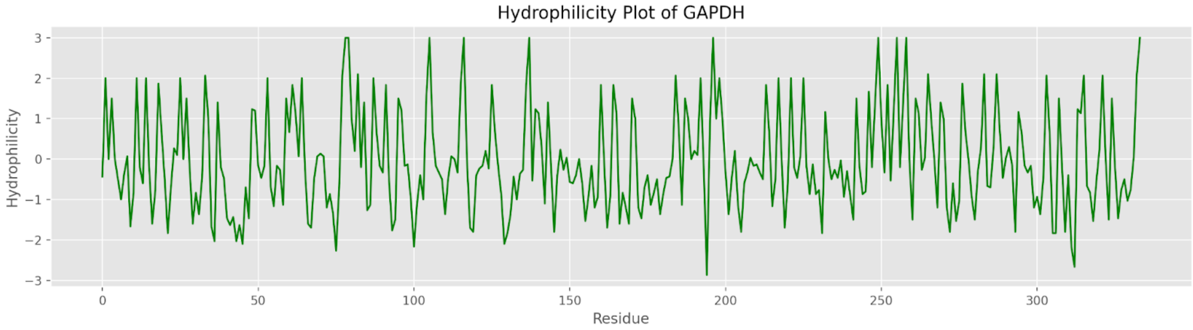


Fig 3: Hydrophilicity Plot of GAPDH

Conclusions

The sonification of data is an essential tool in understanding complex data sequences and making the data understandable for a broader range of audiences. Proteins are a class of molecules with crucial functions in all organisms, and understanding their structure and corresponding functions is vital in bioengineering proteins of our choice to treat diseases and mutations across all species. Converting data from proteins to music is a unique and new concept in computational biology. It has great potential in understanding the arrangement of amino acids in proteins and how they correspond to particular structures and functions. For this project, we directly imported human β-actin and GAPDH protein sequences using theirs by programmatically accessing NCBI Entrez using the Biopython library. The amino acids of these protein sequences were then converted to a corresponding musical note using music21 in Python based on the GES hydrophobicity of the residues. The resulting musical sequence was played in Colab and also modified in GarageBand. The audio obtained provided a sonic determination of the positioning of the hydrophobic amino acid residues in the protein sequence, which was also reaffirmed by comparing the results to hydrophilicity plots. The sound of low-pitch notes corresponds to more hydrophobic amino acid residue in the sequence and assists the user in visualizing the residue within the protein structure. Dissimilarities in the chemical nature of amino acids were identified more effectively in the plots when the sequence MIDI files were played, which indicates the assets of using auditory information. In the future, the amino acids can be mapped based on different criteria, like the size of residues or the known structure of the protein (primary, secondary, tertiary, or quaternary), to provide a more in-depth understanding of protein functionality.

References

Dunh, J., & Clark, M. A. (1999). *The sonification of proteins*. JSTOR. https://www.jstor.org/stable/1576622

Kaper, H., Weibel, E., & Tipei, S. (1999, July). *Data Sonification and Sound Visualization*. IEEE Xplore. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10077454

Rimland, J., Ballora, M., & Shumaker, W. (2013, May 28). *Beyond visualization of Big Data: A multi-stage data exploration approach using visualization, sonification, and Storification*. SPIE Digital Library. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8758/87580K/Beyond-visualization-of-big-data---a-multi-stage/10.1117/12.2016019.short?SSO=1

Zhang, P., & Chen, Y. (2021, September 29). *The music of proteins is made audible through a computer program that learns from Chopin*. The Conversation. https://theconversation.com/the-music-of-proteins-is-made-audible-through-a-computer-program-that-learns-from-chopin-168718#:~:text=Protein%2Dto%2Dmusic%20mapping%20can,pitch%2C%20note%20lengths%20and%20chords.

The hydrophobic effect is a principal force stabilizing tertiary and quaternary structures. LabXchange. (2023). https://www.labxchange.org/library/items/lb:LabXchange:a9290de2:html:1

Martin, E. J., Meagher, T. R., & Barker, D. (2021, September 23). *Using sound to understand protein sequence data: New sonification algorithms for protein sequences and multiple sequence alignments*. BMC bioinformatics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8459479/

Cuthbert, M., Ariza, C., & Friedland, L. (2011). *Feature extraction and machine learning on symbolic music using music21 toolkit*. MIT. http://web.mit.edu/music21/papers/Cuthbert_Ariza_Friedland_Feature-Extraction_ISMIR_2011.pdf

Wright, G. (2023, January 24). *What is MIDI (Musical et al.)?*. WhatIs.com. https://www.techtarget.com/whatis/definition/MIDI-Musical-Instrument-Digital-Interface#:~:text=Musical%20Instrument%20Digital%20Interface%20(MIDI,from%20its%20own%20sound%20library.

Kuhlman, B., & Bradley, P. (2019, August 15). *Advances in protein structure prediction and Design*. Nature News. https://www.nature.com/articles/s41580-019-0163-x

Yu, C.-H., & Buehler, M. J. (2020, March 17). *Sonification based de novo protein design using artificial intelligence, structure prediction, and analysis using molecular modeling*. AIP Publishing. https://pubs.aip.org/aip/apb/article/4/1/016108/23083/Sonification-based-de-novo-protein-design-using

Moelbert, S., Emberly, E., & Tang, C. (2004, March). *Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins*. Protein science : a publication of the Protein Society. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2286732/#:~:text=Hydrophobicity%20is%20thought%20to%20be,surface%20of%20a%20folded%20protein.

YouTube. (2022). *Pull up for precise seeking 0:16 0:01 / 6:23 How to make Animated plot with Matplotlib and Python - Very Easy . YouTube*. Retrieved December 3, 2023, from https://www.youtube.com/watch?v=QAqi77tA_1s.

TP, H. (1989). *Use of hydrophilicity plotting procedures to identify protein antigenic segments and other interaction sites*. Methods in enzymology. https://pubmed.ncbi.nlm.nih.gov/2481215/

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … others. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423.