BioMusic: Making Proteins Sing

Haylee Ham **Chelsea Ernhofer**

Abstract

For this project, we explore protein sonification, which is the translation of protein sequences to music. To achieve this goal, we investigate sonification algorithms and well as music theory in order to produce pieces of music that are both effective representations of sequences and melodiously pleasing. We then create a web application to demonstrate the process and share the generated music with users. We employ Javascript libraries to produce auditory and written music in browser. The app can be found at https://biomusic.herokuapp.com/.

1 Introduction

Bioinformatics involves analyzing complex biological data through computational methods. Biological structures such as DNA strands and protein sequences are historically presented and analyzed visually. Through visual models, one can notice specific forms and constructs of different biological elements. This area has been heavily studied and multiple tools exist to visually model biological data. Considerably less investment has been placed in discovering novel means whereby to study biological data. One such new method is sonification.

2 Background

In the context of bioinformatics, sonification refers to the attempt to capture the information content of DNA and protein sequences auditorily. Although sequences are traditionally analyzed visually, using auditory methods to inspect DNA and proteins has benefits. Sonification is useful when determining patterns and potential mutations across the entire DNA or protein strand and can be used to aid visually impaired individuals.

The majority of work in sonification has been done with DNA sequences. There are six main algorithms currently used in the DNA sonification process. These algorithms are based on biological features and rules, such as the distinction between introns and exons in a DNA sequence. This ensures that music produced through sonification contains applicable information about the structure of the sequence in question. Most DNA sonification algorithms analyze nucleotides in groups of three to simulate the structure of codons. Simpler algorithms use hash tables to map each of the four nucleotide bases to one of four musical notes or pairs of nucleotides to 16 possible musical notes.

Our group takes advantage of these two main principles by analyzing protein sequences (translated nucleotides that form a coding region) and mapping individual amino acids to specific notes. This captures the biological context of the protein and provides unique information for each protein.

3 Methods

The specific algorithm we implemented uses features of the major DNA sonification algorithms. First, our algorithm determines a key signature and tempo based off of the first few amino acids in the sequence. These acids are divided into two groups of 5 and the unicode point of each protein character is found. Sums of both groups are taken and hashed to a specific key and tempo. Keys included in the final version of our project are C major, D major, G major, D minor, and G minor. Once the key is selected, potential notes are restricted to only those notes included within the specific scale of the chosen key. Tempos are 60, 125, 200, and 300. These two features ensure variability in the music generated.

The most obvious connection between our algorithm and those previously mentioned is the use of hash tables to map specific proteins to individual notes. Our algorithm iterates through the protein sequence and assigns a note to each individual amino acid. The duration of the note is determined based off of the overall distribution of amino acids in a given protein sequence. First, counts for each amino acid are calculated. This distribution is then divided into quartiles and note durations are assigned uniformly to each quartile with the most common amino acids receiving the longest durations. The durations used are whole, half, quarter, and eighth note.

We also decided to differentiate between conserved domains and unconserved domains. Sections of the protein which are conserved are represented with chords, while other sections of the protein are comprised of single notes in the final piece of music. Chords are created using harmonious intervals based off of music theory. These conserved domains are also coloured teal in the textual representation of the protein sequence.

The web app builds off of this algorithm. Users are initially prompted to input an accession number. Using the NCBI e-utilities package, we retrieve the XML document for the protein sequence from NCBI and parse it to retrieve the protein sequence and conserved domains. This information is passed to the sonification algorithm. After hashing the sequence and mapping it to the key, tempo, and musical notes, the audio is produced using the webaudio Javascript API and the sheet music is produced using the abcjs Javascript package. The user is shown a page that contains the sequence, with portions of the sequence colored to indicate a conserved domain, as well as two buttons that will play the song and reveal the sheet music.

4 Results

The results of this work can be found at https://biomusic.herokuapp.com/. After inputting an accession number, a user is able to listen to the protein sequence and see the accompanying sheet music. After running multiple examples, we find that we are able to hear patterns in protein sequences as past research would suggest. The patterns that are, perhaps, most evident are the most common amino acids which are represented as whole notes. There is often repitition of the same pitch and duration present in the music, indicating a very common amino acid. Additionally, we notice that protein sequences do not appear to start with conserved domains; there is at least amino acid before the conserved domain begins. We are happy with our results since they make evident the form and structure of the protein as well as the distribution of amino acids within the protein.

5 Discussion

DNA and protein sonification are promising methods whereby researchers can analyze sequences in novel ways. Through this project, our group was able to produce pieces of music that are both biologically accurate and melodious. We hope that this small web app can encourage interest in protein sequencing and help researchers to identify patterns in proteins sequences through auditory measures.

6 References

Temple, Mark D. ”An auditory display tool for DNA sequence analysis.” BMC bioinformatics 18.1 (2017): 221.