BioMusic: Making Proteins Sing

Haylee Ham **Chelsea Ernhofer**

Abstract

For this project, we explore protein sonification- that 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

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 have benefits. Rather than learning about individual elements of sequences, sonification is useful when determining patterns and potential mutations across the entire DNA or protein strand. Sonification can also be used to aid visually impaired individuals, allowing them to hear the sequence when they are not able to see it.

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; these algorithms generate reading frames and some even produce music with interlaced melodies to represent the three possible reading frames present in a DNA strand. 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 10 proteins in the sequence. The proteins are divided into two gropus 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. Tempos are 120, 250, 400 and 600. These two features

ensure that pieces of music generated from proteins are variable and musically interesting.

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 number of notes in final piece of music is equal to the total number of acids in the sequence. 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.

Our group 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 and are not dependent on the specific amino acid or surrounding amino acids.

The web app builds off of this algorithm. Users are initially prompted to input an accession number. Using the e-utilities package, we retrieve the DNA sequence from NCBI and pass it to the

4 Results

5 Discussion

6 References

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