Converting DNA to Music: Sonifying Splicing and Translation

Ilana Shapiro
Computer Science Department
Pomona College
Claremont, CA 91711 USA
issa2018@mymail.pomona.edu

Abstract

The sonification of genetic material is a little-explored mode of unconventional computation that bridges the divide between bioinformatics, computer science, and music, allowing bioinformaticians to perceptualize their data in a novel and illuminating manner. This paper presents BioMus, an original model for converting DNA to musical data in the form of MIDI piano chords. Gene sequences are sourced from Ensembl, a genome database of the European Bioinformatics Institute, and are parsed into exons and and introns. Exons are further parsed into their 5' and 3' untranslated regions (UTRs) and their CDS (CoDing Sequence, i.e. the spliced exons constituting the amino acid-coding sequence after UTRs are removed). Then, each codon in a CDS region is mapped to a major chord, individual nucleotides in introns are mapped to diminished chords, and individual nucleotides in UTRs are mapped to minor chords. CDS codons are mapped to chords based on the amino acids they code for, and rhythmic alterations indicate when CDS codons are broken across splice sites. To further emphasize protein-coding regions, all CDS chords are also at a higher volume. By mapping nucleotides and codons to chords and analyzing genetic material as music, BioMus thus gives scientists a novel means to conceptualize the process of biological splicing and translation.

ONE PARAGRAPH, max 250words

Introduction

Related Work

Ingalls et al. present

Converting DNA to Music

Obtaining Genetic Data

BioMus's process of DNA sonification begins with the user specifying a desired species and gene. This information is passed to Ensembl's REST API to obtain Ensembl's choromomal coordinates of the gene's exons. These coordinates are sourced from the gene's *canonical transcript*, the gene's transcript in Ensembl that is overall the most conserved and highly expressed, has the longest CDS, and is also represented in other major databases such as the NCBI (Ensembl 2023b). The exon coordinates also define the chromosomal coordinates of the intervening introns. Each pair of exon and intron coordinates is passed back to the Ensembl REST API to obtain the nucleotide sequences for each region, and the result is a list of sequence regions alternating exons and introns. For instance, consider the abbreviated sequence obtained from Ensembl for the *Homo sapiens* (human) TP53 tumor suppressor gene in Figure 1. Ellipses indicate omitted nucleotides for the sake of example.

```
[
 ["CTCAAAAGTCTAGAGCCACC..................GACACGCTTCCCTGGATTGG"],
 ["gtaagctcctgactgaactt..................ccccacttttcctcttgcag"],
 ["CAGCCAGACT..........GGTCACTGCCATGGAGGAGC..........TATGGAAACT"],
 ["gtgagtggatccattggaag..................tttctgctcttgtctttcag"],
 ["ACTTCCTGAAAACAACGTTCTG"],
 ["gtaaggacaagggttgggct..................ctcttttcacccatctacag"],
 ...
 ["gtaagcaagcaggacaagaa..................tttccttgcctctttcctag"],
 ["CACTGCCCAACAACACCAGC CACTGGATGGAGAATATTTCACCCTTCAG"],
 ["gtactaagtcttgggacctc..................cccctcctctgttgctgcag"],
 ["ATCCGTGGGCGTGAGCGCTT..................GGGAGCAGGGCTCACTCCAG"],
 ["gtgagtgacctcagcccctt..................cctgcttctgtctcctacag"],
 ["CCACCTGAAG..........CTCAGACTGACATTCTCCAC..........TTTGCTGCCA"]
]
```

Figure 1: Homo sapiens

Exons are in color, while introns are in grey. Within the exons, the orange regions are the UTRs, and the blue regions are the CDS. Notice how the CDS begins with the bolded start codon ATG and ends with the bolded stop codon TGA.

Now, the goal is to extract the 5' UTR from the list. In order to ensure the correct CDS, is necessary to verify the 5' UTR with Ensemble, since it may include multiple ATG start codons that do not signal the start of the CDS. This is not necessary for the 3' UTR, since once we know we are in the CDS, the first stop codon will always signal the end of the CDS and the beginning of the 3' UTR. We obtain the 5' UTR from Ensembl by querying the Ensembl REST API for the cDNA (complementary DNA) sequence of the canonical transcript. cDNA is identical to the CDS, but includes the UTRs (Ensembl 2023a). By requesting Ensembl to "mask" the UTRs of the cDNA sequence by representing them in lowercase, we can successfully isolate the 5' UTR and extract it from the list in Figure 1. As seen in Figure 2, we end up with a list of the 5' UTR regions (top), and a list of the CDS regions and introns (bottom).(Miranda 2020)

We now have all the tools necessary to translate the genetic data into music.

Conversion to MIDI

Next, a MIDI track is created using the MIDIUtil library with the tempo set to ♩ = 200 BPM, and the DNA sequence is transcribed to RNA with Biopython. Recall the bases used

Figure 2: Homo sapiens

in RNA are adenine (A), cytosine (C), guanine (G) and uracil (U), and have the pairings C-G and A-U. BioMus defines the following mapping of individual bases to musical notes relative to the local key in Table 1.

C	Tonic
G,A	Mediant
U	Dominant

Table 1: RNA Base Mappings

The musical key is minor in non-coding regions and major in protein-coding regions, with the initial key set to C minor. Before the start codon AUG is encountered, individual nucleotides outline the 3 notes of minor triad of the local key (base pairs form dyads, or 2-note chords). Then, during translation, the key is major and the volume doubles. The key is determined by the codon (see the AMINOACIDS dictionary above). Individual nucleotides no longer determine the notes; this is now dictated by the codons. Each time the key changes based on the codon, the major triad of that key is played, until a stop codon is reached. Then we remain in the same minor key as the previous stop codon, but the volume is halved again and individual nucleotides once again outline the notes of the new minor triad. Then this process can repeat if another start codon is then encountered'

The tonic, mediant, and dom

Throughout this paper, the chosen gene for MIDI generation and analysis is TP53, which codes for the tumor suppressor protein p53.

AUG	Methionine/ Start Codon	C
AUU, AUC, AUA	Isoleucine	C♯
AAA, AAG	Lysine	D
ACU, ACC, ACA, ACG	Threonine	E♭
UUU, UUC	Phenylalanine	E
UGG	Tryptophan	F
UUA, UUG, CUU, CUC, CUA, CUG	Leucine	F♯
CAU, CAC	Histidine	G
GUU, GUC, GUA, GUG	Valine	A♭

Table 2: Essential Amino Acids

AAU, AAC	Asparagine	A
GAU, GAC	Aspartate	B♭
GCU, GCC, GCA, GCG	Alanine	B

Table 3: Nonessential Amino Acids

UAU, UAC	Tyrosine	C
UGU, UGC	Cysteine	D
UCC, UCU, UCA, UCG, AGU, AGC	Serine	E
AGA, AGG, CGU, CGC, CGA, CGG	Arginine	F
CCU, CCC, CCA, CCG	Proline	G
CAA, CAG, GAA, GAG	Glutamine/ Glutamic acid	A
GGU, GGC, GGA, GGG	Glycine	B

Table 4: Conditionally Essential Amino Acids

UAA	C
UAG	E
UGA	G

Table 5: Stop Codons

Sample Music

Conclusion

BioMus serves as a bridge between bioinformatics, computer science, and music by giving scientists the creative means to

Acknowledgments

I am very grateful to Professor Zachary Dodds of Harvey Mudd College for his invaluable mentorship throughout this project.

References

Ensembl. 2023a. Ensembl Glossary. *European Molecular Biology Laboratory - European Bioinformatics Institute*.

Ensembl. 2023b. Transcript flags. *European Molecular Biology Laboratory - European Bioinformatics Institute*.

Miranda, E. R. 2020. Genetic Music System with Synthetic Biology. *Artificial Life* 26(3):366–390.