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 the CDS is mapped to a major chord based on the amino acid it codes for, individual nucleobases in introns are mapped to diminished chords, and individual nucleobases in UTRs are mapped to minor chords. 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 nucleobases and codons to chords and analyzing genetic material as music, BioMus thus gives scientists a novel and straightforward means to conceptualize the process of biological splicing and translation.

ONE PARAGRAPH, max 250words

Introduction

Related Work

Ingalls et al. present

Background

Genes consist of sequences of DNA. The double-stranded DNA helix is shaped like a ladder, where each rung consists of two nucleotides paired together. Every nucleotide is built on a *nucleobase*, which in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T), and have the pairings C-G and A-T. The process of converting DNA to mRNA is called *transcription*, and the process of converting mRNA to amino acids (the building blocks of proteins) is called *translation*. Every gene codes for a single protein.

Transcription contains a stage called *RNA splicing*, where certain regions called *introns* are removed, or "spliced out," from the original DNA sequence. The final "mature" mRNA solely consists of the remaining regions, called exons, that are connected to each other. The exons comprising mature mRNA consist of two *untranslated regions* UTRs and the *CoDing Sequence*, or CDS.

The CDS is a sequence of nucleotides that corresponds with the sequence of amino acids in a protein during translation. It begins with the start codon (ATG) and terminates with a stop codon (TAA, TAG, or TGA). In the CDS, nucleotides grouped into three to create *codons*. Each codon codes for a single amino acid. The UTR preceding the CDS is called the *5' UTR*, and the UTR following the CDS is called the *3' UTR*.

Note in the original DNA sequence before transcription, both the UTR and the CDS may be spread across multiple exons (i.e. they may be broken into multiple sections by intervening introns). In the CDS, this means that individual codons may be broken across splice sites: the codon will begin at the end of one exon, and conclude at the beginning of the next exon.

Figure 1 shows the splicing process for a sample gene.

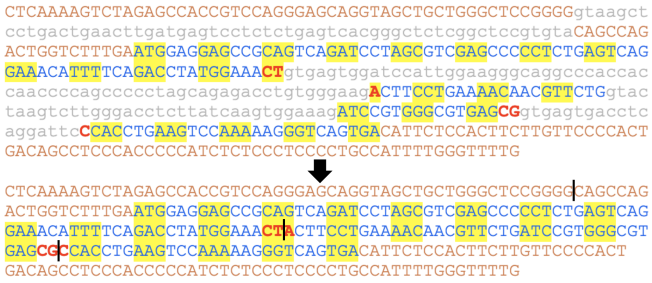

Figure 1: Homo sapiens

The top sequence in Figure 1 presents the gene before transcription, and the bottom sequence presents the result of splicing represented with *complementary DNA*, or cDNA. cDNA is analogous to mRNA; it is synthetically constructed from mRNA by biologists for sequencing as a more convenient way to work with the sequence. A cDNA and mRNA sequence are identical except with the substitution of uracil for thymine in the mRNA.

In Figure 1, exons are in color with capitalized nucleobase, while introns are lowercase and gray. The UTRs are in orange, and the CDS is in blue and red. Red nucleobases in the CDS indicate codons that are broken across splice sites. Notice how the 5' UTR is broken across the first two

two exons, and the CDS is broken across the final four exons. In the CDS, codons are explicated by the alternating highlights. After we have spliced the introns from the gene, we end up with the contiguous sequence of exons in the bottom sequence in Figure 1. Splice sites are indicated with the black vertical bars.

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 nucleobase sequences for each region, and the result is a list of alternating exons and introns. For instance, consider the abbreviated sequence obtained from Ensembl for the *Homo sapiens* (human) TP53 tumor suppressor gene in Figure 2. Ellipses indicate omitted nucleobases for the sake of example.

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

Figure 2: 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 sequence of the canonical transcript. 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 2. As seen in Figure 3, 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.

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

Figure 3: Homo sapiens

Conversion to MIDI

MIDI conversion begins by using the MIDIUtil Python library to create a MIDI track with the tempo set to ♩ = 200 BPM. Subsequently, individual nucleobases in UTRs are mapped to minor dyads (two-note chords), codons in the CDS are mapped to major triads, and individual nucleobases in introns are mapped diminished triads.

Sonification of the gene begins with the 5' UTR, and the initial key is set to C minor. No key changes occur in UTRs or introns; the key is exclusively changed in the CDS and is determined by the codon. Thus, before the CDS begins, all introns and the 5' UTR remain in the initial key of C minor. After the CDS begins, the key of an intron is determined by the final codon in the preceding exon (whether or not that codon is broken across the splice site). The stop codon terminating the CDS then determines the key of all subsequent introns as well as the 3' UTR.

To construct the minor dyads in the 5' and 3' UTRs, BioMus defines the following mapping of individual nucleobases to musical notes relative to the local key in Table 2.

C	Tonic
G,A	Mediant
T	Dominant

Table 1: DNA Base Mappings

The minor dyad for each nucleobase in a UTR is constructed with its base pair using the mapping in Table 2. For instance, if the key is C minor and we encounter the nucleobase C, then the dyad will consist of the notes C (tonic) and E♭(mediant). Thus, all dyads consists of tonic-mediant or mediant-dominant pairs. All UTR dyads are represented with quarter notes.

To construct the diminished triads in introns, BioMus again uses the mapping of individual nucleobases to musical notes in Table 2. Initially, a minor dyad is constructed

C	Tonic
G,A	Mediant
T	Dominant

Table 2: DNA Base Mappings

The minor dyad for each nucleobase in a UTR is constructed with its base pair using the mapping in Table 2. For

instance, if the key is C minor and we encounter the nucleobase C, then the dyad will consist of the notes C (tonic) and E♭(mediant). All UTR dyads are represented with quarter notes.

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.

ATG	Methionine/ Start Codon	C
ATT, ATC, ATA	Isoleucine	C♯
AAA, AAG	Lysine	D
ACT, ACC, ACA, ACG	Threonine	E♭
TTT, TTC	Phenylalanine	E
TGG	Tryptophan	F
TTA, TTG, CTT, CTC, CTA, CTG	Leucine	F♯
CAT, CAC	Histidine	G
GTT, GTC, GTA, GTG	Valine	A♭

Table 3: Essential Amino Acids

AAT, AAC	Asparagine	A
GAT, GAC	Aspartate	B♭
GCT, GCC, GCA, GCG	Alanine	B

Table 4: Nonessential Amino Acids

TAT, TAC	Tyrosine	C
TGT, TGC	Cysteine	D
TCC, TCT, TCA, TCG, AGT, AGC	Serine	E
AGA, AGG, CGT, CGC, CGA, CGG	Arginine	F
CCT, CCC, CCA, CCG	Proline	G
CAA, CAG, GAA, GAG	Glutamine/ Glutamic acid	A
GGT, GGC, GGA, GGG	Glycine	B

Table 5: Conditionally Essential Amino Acids

UAA	C
UAG	E
UGA	G

Table 6: 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.