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 or augmented triad based on the amino acid it codes for, individual nucleobases in introns are mapped to diminished triads, and individual nucleobases in UTRs are mapped to minor dyads. 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.

Introduction

Genes consist of sequences of DNA. The double-stranded DNA helix is shaped like a ladder, where each rung consists of two complementary 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 complementary pairings C-G and A-T. The DNA helix consists of the *coding strand* and the *template strand*. The template strand codes for mRNA during *transcription*, where each nucleotide is converted to its complement, and thymine is replaced with the base uracil.Subsequently, mRNA is converted to amino acids (the building blocks of proteins) during *translation*. Every gene codes for a single protein.

Transcription in eukaryotes includes 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*, or 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 that in the original pre-transcription DNA sequence, both the UTRs 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.

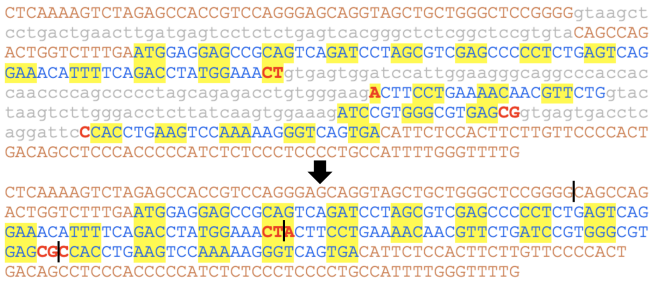

Figure 1: DNA Splicing Example

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. A cDNA sequence is complementary to its source mRNA sequence and contains thymine rather than uracil; it is identical to the template DNA strand without the introns. Biologists synthetically construct cDNA from mRNA as a more convenient way to work with the sequence.

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 ex-

ons. 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.

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 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*, TP53 Gene

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)

```
[
 ["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*, TP53 Gene

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

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 musical key in Table 1.

A	Tonic
C, T	Mediant
G	Dominant

Table 1: DNA Nucleobase Mappings

The minor dyad corresponding to each nucleobase in a UTR is constructed with its base pair using the mapping in Table 1. For instance, if the key is C minor and we encounter the nucleobase A or T, then the dyad will consist of the notes C (tonic) and E♭ (mediant). Similarly, if we encounter the nucleobase C or G, then the dyad is E♭ and G (dominant). 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 1. Initially, a minor dyad is constructed as in the UTRs. However, here we also add the diminished

seventh scale degree from the local key to each dyad to construct the diminished triads. For instance, if the key is C minor and we encounter the nucleobase C or G, then the triad will consist of the notes C (tonic), E♭ (mediant), and B𝄫 (diminished seventh). Similarly, if we encounter the nucleobase A or T, then the triad will consist of the notes E♭, G (dominant), and B𝄫. We use the term "diminished triad" loosely here, since technically only the latter triad described fits the formal definition of a diminished triad, with two minor thirds stacked on top of one another. However, as both chords outline a diminished seventh chord, their sonority is similar enough that we dub them both diminished triads as a way of distinguishing them from major or minor chords. Like UTRs, all intron triads are represented with quarter notes.

When the CDS begins, we map codons (groups of three nucleobases), rather than individual nucleobases, to chords. The first codon in the CDS is always ATG, the start codon, which is mapped to a C major triad and is given a half note duration, rather than a quarter note, to signify the start of the protein-coding region. Subsequent chords return to quarter notes. To further emphasize we are in a protein-coding region, the volume is doubled throughout the CDS. The volume is halved to its original value whenever we encounter an intron, and doubled again when we return to the CDS in the next exon.

We also consider the scenario where the final codon of an exon in the CDS is broken across a splice site (i.e. the codon is split by the intervening intron). When an incomplete codon is encountered at the end of an exon in the CDS, the subsequent exon is referenced for the remainder of the codon. The spliced codon is then represented with an eighth note triad rather than a quarter note, and the same eighth note triad is played at the beginning of the next exon to indicate that the remainder of the codon is found there. This scenario is detailed in Figure 4, where exons are represented in color and introns are in gray.

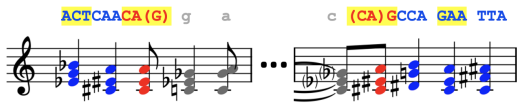

Figure 4: Spliced Codon Example

The spliced codon CAG is in red, with the nucleobases CA in the first exon and G in the second. Notice how the same A augmented eighth note triad both closes the first exon and opens the second.

Thus far, we have demonstrated a novel means to conceptualize biological splicing by differentiating between UTRs, introns, and the CDSs, as well as emphasizing spliced codons. We now want to map codons in the CDS to chords in such a way that facilitaties the perception of biological translation. To do so, we consider the amino acids the codons specify. To distinguish between amino acids, we seek an injective mapping between amino acids and musical keys. We do this by first solely considering the set of essential and nonessential amino acids (with which we can create a bijective mapping with the set of possible keys), and separately considering the set of conditionally essential amino acids (with which we can create an injective mapping with the set of possible keys).

Figures 2 and 3 explicate the chosen bijective mapping between the set of essential and nonessential amino acids and the set of possible keys. The codons that code for each amino acid are listed in the left hand column, and BioMus maps each to a major triad in the corresponding key in the right column.

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 2: Essential Amino Acids

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

Table 3: Nonessential Amino Acids

Next, we consider the set of nonessential amino acids. Tables 2 and 3 explicate the chosen injective mapping between this set and the set of possible keys. Notice that the mapping is also bijective between the nonessential amino acids and the set of natural keys.

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 4: Conditionally Essential Amino Acids

This time, BioMus maps each codon to an augmented triad, rather than a major triad, in the corresponding key in the right column.

Finally, we consider the stop codons, which do not code for amino acids and instead terminate the CDS. They are mapped to minor triads in the keys explicated in Table 5.

Each stop codon chord is given a half note duration and the volume is halved to signify the end of the protein-coding region. The subsequent minor dyads in the 3' UTR constituting the remainder of the gene return to quarter notes, and are in whatever key is dictated by the stop codon in Table 5.

UAA	C
UAG	E
UGA	G

Table 5: Stop Codons

Sample Music

Figures 5 and 6 demonstrate the sonification of key sections in the TP53 tumor suppressor gene of the zebrafish *Danio rerio*. The DNA sequence is shown above BioMus' sonification of each nucleobase or codon. Orange indicates a UTR, gray indicates an intron, and blue indicates the CDS. Red indicates CDS codons that are broken across splice sites. Alternating codons in the CDS are highlighted for visibility.

Figure 5 begins with the fragment of the 5' UTR concluding the gene's first exon. Notice how the opening key is C minor. The nucleobases A and T map to the dyad C-E♭ (tonic-mediant), and the nucleobases A and T map to the dyad C-E♭ (tonic-mediant) dyad will consist of the notes C (tonic) and E♭ (mediant). Similarly, if we encounter the nucleobase C or G, then the dyad is E♭ and G (dominant). All UTR dyads are represented with quarter notes.

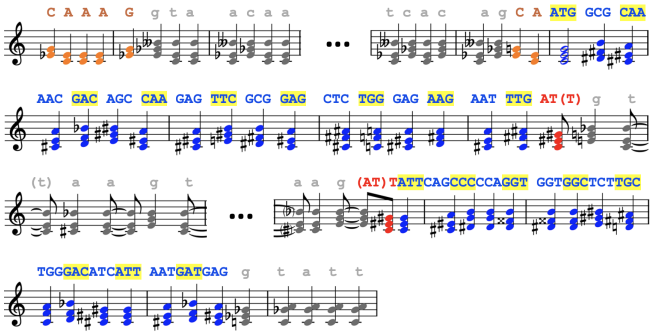

Figure 5: *Danio rerio* TP53 Gene, Beginning of CDS

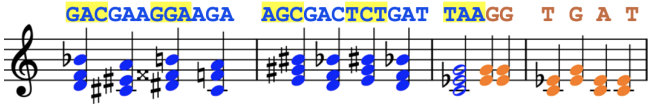

Figure 6: *Danio rerio* TP53 Gene, End of CDS

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.