

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 both scientists and the general public to perceptualize genomics in a novel and illuminating manner. This paper presents BioMus, an original model converting DNA to musical data as MIDI piano chords. Gene sequences are sourced from Ensembl, a genome database of the European Bioinformatics Institute, and are parsed into their constituent exons and introns. Exons are further parsed into their 5' and 3' untranslated regions (UTRs) and 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, and individual nucleobases in introns and UTRs are respectively mapped to diminished triads and dyads outlining a minor triad. Rhythmic alterations indicate when CDS codons are spliced. To further emphasize protein-coding regions, all CDS chords are at a higher volume. By mapping nucleobases and codons to musical chords, BioMus introduces a novel and straightforward means of conceptualizing the process of biological splicing and translation that is accessible to users of all scientific backgrounds.

Introduction

Bioinformaticists constantly seek new ways to computationally represent and interpret genomic data. Representation of genetic sequences is primarily visual, with tools such as FluEntDNA that visualize DNA sequences with the four nucleobases (adenine (A), cytosine (C), guanine (G) and thymine (T)) in colors. Such methods generally involve a high degree of technical biological background to comprehend and are not accessible to a non-scientific audience. BioMus presents a novel means of conceptualizing genomics aurally, rather than visually, in a straightforward manner that emphasizes the processes of biological splicing and translation. BioMus' sonification model allows for a general audience, ranging from nontechnical to bioinformaticist, to integrate the nontraditional sense of hearing in understanding essential biological processes. It also offers visually impaired users who cannot access visual methods a newly illuminating perspective on genomics.

BioMus takes in a DNA sequence of a single gene and converts it to a series of MIDI piano chords. Genes consist

of sequences of double-stranded DNA, which are built with pairs of complementary nucleotides (C-G and A-T). DNA is converted to mRNA during *transcription*, and mRNA is converted to amino acids (the building blocks of proteins) during *translation*. Every gene codes for a single protein.

Eukaryotic transcription includes a stage called *RNA splicing*, where regions called *introns* are removed, or "spliced out," from the DNA sequence. The final "mature" mRNA solely consists of the remaining regions, called *exons*, strung together. The exons comprising mature mRNA consist of two *untranslated regions*, or UTRs, flanking the *CoDing Sequence*, or CDS, which is the protein-coding region. In the CDS, nucleotides are 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*.

BioMus obtains gene sequences from Ensembl, a genome database of the European Bioinformatics Institute, using the Ensembl REST API. These sequences are subsequently parsed into their constituent exons and introns, and exons are further parsed into their 5' and 3' UTRs and CDS. To convey the structure of the gene, individual nucleobases in introns and UTRs are respectively mapped to diminished triads and dyads outlining a minor triad. To convey translation, in the CDS, codons rather than individual nucleobases are mapped to major or augmented triads given the amino acids they code for.

Note that in the original DNA sequence, both the UTRs and the CDS may be spliced over multiple exons (i.e. they may be broken up 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.

Figure 1 shows the splicing process for a sample gene. The top sequence in Figure 1 presents the pre-transcription DNA sequence, and the bottom sequence presents the result of splicing using *complementary DNA*, or cDNA. A cDNA sequence is complementary to its source mRNA sequence; it is identical to the noncoding DNA strand without introns. Biologists synthesize cDNA from mRNA to work with the sequence more conveniently.

In Figure 1, exons are in color, while introns are gray. The UTRs are orange, and the CDS is blue or red. Red nucleobases in the CDS indicate codons that are broken across

CTCAAAAGTCTAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGgtaagct
 cctgactgaacttgatgagtcctctctgagtcacgggctctcggctccgtgtaCAGCCAG
 ACTGGTCTTTGAATGAGGAGCCGCGAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAG
 GAAACATTTCTAGACCTATGGAAGCTgtgagtggaatccattggaaggcagggccaccac
 caacccagccccctagcagagacctgtgggaagACTTCCTGAAACAACTTCTGgtac
 taagtcttgggacctcttatcaagtggaaagATCCGTGGGCGTGAGCGgtgagtgacctc
 agggattcCACCTGAAGTCCAAAAGGGTCAGTGAATTCTCCACTTCTGTTCCCCACT
 GACAGCCTCCACCCCATCTCTCCCTCCCTGCCATTTTGGGTTTTG

↓

CTCAAAAGTCTAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGCAGCCAG
 ACTGGTCTTTGAATGAGGAGCCGCGAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAG
 GAAACATTTCTAGACCTATGGAAGCTCTTCCTGAAACAACTTCTGATCCGTGGGCGT
 GAGCGCACCTGAAGTCCAAAAGGGTCAGTGAATTCTCCACTTCTGTTCCCCACT
 GACAGCCTCCACCCCATCTCTCCCTCCCTGCCATTTTGGGTTTTG

Figure 1: DNA Splicing Example

splice sites. Notice how the 5' UTR is broken across the first two exons, and the CDS spans the final four exons. Alternating highlights in the CDS are used for visibility of codons. After excising the introns from the gene, the result is the contiguous sequence of exons at the bottom of Figure 1. The black vertical bars indicate splice sites. To perceptualize splicing, BioMus splits the chord for a spliced codon into half its original duration and places it at each end of the splice site: the chord will both close the exon that begins the codon and open the next exon that completes the codon.

In this paper, we begin by examining relevant work in the field, and subsequently discuss the process of sourcing genetic data from the Ensembl database and its conversion to musical chords. Finally, examples of BioMus' sonification model are presented.

Related Work

Sonification of genetic material has been approached from a variety of angles. To specifically address accessibility of science to a wider audience, Takahashi and Miller converted genome-encoded protein sequences to piano notes in order to produce musical patterns that still adhere to the structure of the sequences. Like BioMus, Takahashi and Miller mapped codons in the CDS to chords. However, their scheme mapped pairs of amino acids to triads, with differing inversions distinguished individual amino acids within the pairs. The duration of each triad correlated to the frequency of its corresponding codon in the CDS. Unlike BioMus, Takahashi and Miller only consider the CDS, and do not consider how to represent spliced codons.

In another sonification model, Ingalls et al. consider the sequence alignment problem, in which sequences from the same gene but from different species are compared to identify regions of similarity. Their tool COMPOSALIGN translates genome-wide aligned data into a musical composition. Sequence alignment . They present a unique mapping of alignment information onto musical features, sonifying the presence and absence of nucleobases in the alignment

Large-scale genenomic data, for instance, that sonifies large-scale genomic data

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 chromosomal 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.....GACAGCCTTCCCTGGATTGG"},
{"gtaagctcctgactgaactt.....ccccacttttccctcttgcag"},
{"CAGCCAGACT.....GGTCACTGCCATGGAGGAGC.....TATGGAAGCT"},
{"gtgagtggaatccattggaag.....ttctgtctctgtcttcttgcag"},
{"ACTTCCTGAAACAACTTCTG"},
{"gtaaggacaagggttgggct.....ctcttttaccacatctacag"},
...
{"gtaagcaagcaggacaagaa.....tttcccttgcctcttcttctag"},
{"CACTGCCCAACACACCAGC CACTGGATGGAGAATATTTACCCTTCAG"},
{"gtactaagctcttgggacctc.....ccctcctctgttctgtcag"},
{"ATCCGTGGGCGTGAGCGCTT.....GGGAGCAGGGCTCACTCCAG"},
{"gtgagtgacctcagccctt.....cctgtctctgtctctacag"},
{"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 Ensembl, 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.

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. All chords are built within the octave C4-B4. This means that all chords built on the notes C through E will be in root position, and the remainder will be in inversion.

```
[
  ["CTCAAAGTCTAGAGCCACC.....GACACGCTTCCTGGATTGG"],
  ["CAGCCAGACTGCCTTCGGGTCACTGCC"]
]
[
  [""],
  ["gtaagctcctgactgaactt.....ccccacttttctcttgcag"],
  ["ATGGAGGAGC.....TATGGAACCT"],
  ["gtgagtgatccattggaag.....ttctgtctctgtcttctag"],
  ["ACTTCCTGAAACAACGTTCTG"],
  ["gtaaggacaagggtgggct.....ctcttttcacccatctacag"],
  ...
  ["gtaagcaagcaggacaagaa.....tttccttgctcttctctag"],
  ["CACTGCCCAACAACACCAGC CACTGGATGGAGAATATTCACCTTCAG"],
  ["gtactaagctcttgggacctc.....ccctcctctgttgcag"],
  ["ATCCGTGGGCGTGAGCGCTT.....GGGAGCAGGGCTCACTCCAG"],
  ["gtgagtgacctcagccctt.....cctgctctgtctcctacag"],
  ["CCACCTGAAG.....CTCAGACTGACATTCTCCAC.....TTTGTGCCA"]
]
```

Figure 3: *Homo sapiens*, TP53 Gene

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 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 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 \flat (mediant). Together, the dyads outline a minor triad in the current key. Similarly, if we encounter the nucleobase C or G, then the dyad is E \flat 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 dyad is constructed as in the UTRs (except with the dominant scale degree lowered a half step to form a diminished rather than a perfect fifth with the tonic). Then, 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 \flat (mediant), and B \flat (diminished seventh). Similarly, if we encounter the nucleobase A or T, then the triad will consist of the notes E \flat , G \flat (dominant), and B \flat . 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, as a signpost of the beginning of translation. 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 facilitates 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 on the corresponding key in the right column. Next, we consider the set of nonessential amino acids. Figures 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.

This time, BioMus maps each codon to an augmented

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

Table 2: Essential Amino Acids

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

Table 3: 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 4: Conditionally Essential Amino Acids

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.

UAA	C
UAG	E
UGA	G

Table 5: Stop Codons

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.

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.

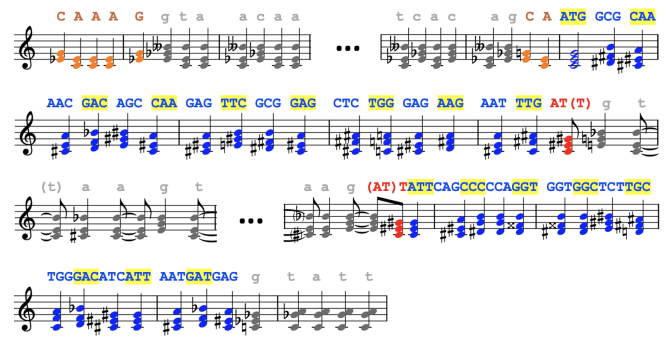


Figure 5: *Danio rerio* TP53 Gene, Beginning of CDS

Notice how the opening key is C minor. The nucleobases A and T map to the dyad C-Eb (tonic-mediant), and the nucleobases A and T map to the dyad C-Eb (tonic-mediant). We then enter the gray region of the first intron, where diminished triads outline the C diminished seventh chord (notice how the perfect fifth G above the tonic becomes the diminished fifth Gb, and each triad includes the addition of the diminished seventh Bbb).

The second exon begins with the last two nucleobases of the 5'UTR, which again outlines the C minor triad, before beginning the CDS with the start codon ATG, which maps to a C major triad in Table 2. Notice how ATG maps to a half, rather than a quarter, note to signify the start of the protein-coding region.

From here, each CDS codon maps to a major or augmented triad based on tables 2 - 4. For instance, GCG codes for the nonessential amino acid alanine, which maps to a B major triad in Table 3, while GAG codes for the conditionally essential amino acid glutamine, which maps to an A augmented triad in Table 4.

Then, notice how the codon ATT, which codes for the essential amino acid isoleucine and maps to the C# major triad in red, is broken across the splice site between the second and third exons. The intervening intron assumes the key of the final codon of the previous exon and thus outlines a C# diminished seventh chord.

Finally, notice how the third exon begins with the other "half" (i.e. remaining eighth note) of the C# major triad of the spliced ATT codon. This exon concludes with the complete codon GAG, which again codes for glutamine and maps to an A augmented triad in Table 4. The following intron assumes the same key of A, and outlines an A diminished seventh chord. Notably, if played, the volume in the CDS would be doubled relative to the rest of the sequence.

Figure 6 shows a fragment of the exon containing the end of the CDS.



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

The CDS terminates with the stop codon TAA, which maps to a C minor triad in Table 5. Notice how TAA maps to a half, rather than a quarter, note to signify the end of the

protein-coding region. After this, we immediately enter the 3' UTR, which assumes the key of the stop codon and thus produces dyads outlining a C minor triad.

Conclusion

BioMus serves as a bridge between bioinformatics, computer science, and music by giving both scientists and the general public a novel, creative means to aurally conceptualize the biological processes of DNA splicing and translation. The structure of the gene and its splice regions are illuminated by respectively mapping individual nucleotides in UTRs and introns to diminished chords and minor dyads, and breaking chords across splice sites just as their corresponding codons are. Translation is conveyed by mapping codons to major and augmented chords in the CDS based on the amino acids they code for. The chosen injective mapping between codons and musical keys also allows the listener to distinguish between essential, nonessential, and conditionally essential amino acids. BioMus' straightforward sonification model allows users to conceptualize the structure of DNA and the processes of splicing and translation whether or not they are career scientists, and opens to the door for visually impaired users to access these concepts without being constricted by traditional visual methods of genomic representation.

In the future, BioMus would ideally sonify a greater variety and granularity of genomic structures. For instance, known mutations could be marked with dissonance. Furthermore, given a sample sequence and a target sequence, the similar regions of the sample sequence resulting from alignment with the target could be marked aurally by both new harmonies and rhythms. Finally, BioMus' current sonification scheme is intentionally linear to align with its vision of straightforward representation. Going forward, BioMus could benefit from modifying this scheme to integrate more complex rhythms and a larger sonic range, while still maintaining a clear musical distinction between genetic structures, in order to increase the musicality of its results.

Acknowledgments

I am very grateful to Professor Zachary Dodds of Harvey Mudd College for his 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.