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 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 both the structure of a gene and the processes 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)) n colors. Comprehending such visual methods generally requires a high degree of technical background, thus limiting the 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. By bridging the divide between bioinformatics, computer science, and music, BioMus' sonification model makes the structure of a gene and the essential biological processes of splicing and translation accessible to a wide-ranging audience. It particularly offers visually impaired users 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. Eukaryiotic 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, that codes for proteins. In the CDS, nucleotides are grouped into three to create codons, each of which 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. These sequences are parsed into their constituent exons and 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 mapped to diminished triads and dyads (two-note chords) outlining a minor triad, respectively. To convey translation, codons rather than individual nucleobases in the CDS are mapped to major or augmented triads based on the amino acids they code for.

In the original DNA sequence, both the UTRs and the CDS may be broken over multiple exons (i.e. they may be split 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.

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

Sonofication of genetic material has been approached from a variety of angles. To address accessibility of genomics 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. Their scheme mapped pairs of amino acids to triads, with differing inversions distinguishing 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 when sonifying the gene, and they do not consider spliced codons.

Plaisier et al. also seek to increase accessibility to genomics by proposing a sonification model that is both entertaining and informative. They map nucleotides to specific notes to transcend the monotonous appearance of traditional DNA sequence visualizations and create the excitement expected by a public audience. By using the Sonic Pi program for sonification, they crucially support real-time customization of the program, providing a link between DNA and live programming to further enhance public engagement. BioMus' sonification model perceptualizes a wider variety of bioligical structures than Plaisier et al.'s model.

In another sonification model, Ingalls et al. consider the sequence alignment problem, where sequences from the same gene but different species are compared to identify overlapping regions. Their tool COMPOSALIGN translates genome-wide aligned data into a musical composition by mapping alignment information onto musical features. Their approach sonifies the presence and absence of characters (nucleotides or amino acids) in the alignment such that their assignment to the corresponding sequence (i.e. species) is clear. By mapping each character to a measure-long motif, rather than a single note or chord, COMPOSALIGN achieves a mapping that is modular and flexible. Sequence alignment is outside the scope of BioMus' current scheme.

Converting DNA to Music

Obtaining Genetic Data

BioMus's sonification of DNA begins with the user specifying a desired species and gene. This information is passed to Ensembl's REST API to obtain the chromosomal coordinates of the gene's exons from its canonical transcript, the gene's most conserved and highly expressed transcript in Ensembl that has the longest CDS and is also represented in other major databases such as the NCBI (Ensembl 2023b). The exon coordinates also define the coordinates of the intervening introns. Each pair of exon and intron coordinates is passed back to the Ensembl REST API to query for the nucleobase sequences of 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* TP53 tumor suppressor gene in Figure 1. Ellipses indicate omitted nucleobases for the sake of example.

Exons are in color, while introns are in grey. Within the exons, orange indicates UTRs and blue indicates 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 start of the CDS, it 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

```
[
 ["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*, TP53 Gene

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 verify the 5' UTR by querying the Ensembl REST API for the cDNA sequence of the canonical transcript. A cDNA sequence is complementary to its source mRNA sequence; it is identical to the original DNA without introns. Biologists synthesize cDNA from mRNA to work with the sequence more conveniently. 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. 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).

```
[
 ["CTCAAAAGTCTAGAGCCACC..................GACACGCTTCCCTGGATTGG"],
 ["CAGCCAGACTGCCTTCCGGGTCACTGCC"]
]
[
 [""],
 ["gtaagctcctgactgaactt..................ccccacttttcctcttgcag"],
 ["ATGGAGGAGC.........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

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. Then, we begin mapping nucleobases and codons to chords 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.

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.

Individual nucleobases in the 5' and 3' UTRs are mapped

to dyads outlining a minor triad. BioMus maps nucleobases to notes relative to the local musical key as in Table 1.

A	Tonic
C, T	Mediant
G	Dominant

Table 1: DNA Nucleobase Mappings

The dyad for to each nucleobase 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). Together, they outline a C minor triad. All UTR dyads have quarter note durations.

Individual nucleobases in introns are mapped to diminished triads. Initially, a dyad is constructed using the mapping from Table 1. The dominant scale degree is then lowered a half step to form a diminished rather than a perfect fifth from the tonic, and finally the diminished seventh scale degree is added to each dyad. 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, since technically only the latter chord has the two minor thirds stacked on top of one another that formally define it as a diminished triad. However, both chords outline a diminished seventh chord and have similar enough sonority that we dub them both diminished triads as a way of distinguishing them from major, minor, or augmented triads. Like UTRs, all intron triads have quarter note duration.

When the CDS begins, we map codons, 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, rather than a quarter note, duration 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.

To accurately convey splicing, we also consider the scenario where codons in the CDS are broken across a splice site (i.e. the codon is split between two exons by an 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 repeated at the beginning of the next exon to indicate that the remainder of the codon is found there. This is detailed in Figure 3, where exons are in color and introns are gray.

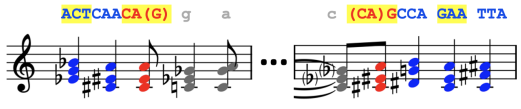

Figure 3: Spliced Codon Example

The spliced codon CAG is red, with CA in the first exon and G in the second. Note 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 the structure of a gene and the splicing process by differentiating between UTRs, introns, the CDS, and spliced codons. We now want to map codons in the CDS to chords to convey 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 then 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 BioMus' bijective mapping between the set of essential and nonessential amino acids and the set of possible keys. The codons for each amino acid are listed in the left 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 BioMus' injective mapping between this set and the set of possible keys. Notice that the mapping is also bijective between the set of 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

Now, BioMus maps codons to augmented triads, rather than major triads, in the key in the right column to avoid overlap with the essential and nonessential amino acids.

Finally, we consider the stop codons, which solely serve to terminate the CDS. They are mapped to minor triads in the

keys 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 dyads in the 3' UTR constituting the remainder of the gene return to quarter notes, and are in whatever minor key is dictated by the stop codon in Table 5.

UAA	C
UAG	E
UGA	G

Table 5: Stop Codons

Sample Music

Figures 4 and 5 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 UTRs, gray indicates introns, 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 4 begins with the fragment of the 5' UTR concluding the gene's first exon.

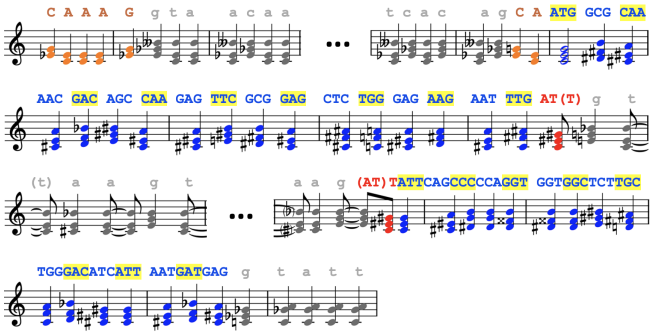

Figure 4: *Danio rerio* TP53 Gene, Beginning of CDS

Notice how the opening key is C minor. The nucleobases A and T map to the dyad C-E♭ (tonic-mediant), and the nucleobases G and G map to the dyad E♭-G (mediant-dominant). 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 G♭, and each triad includes the addition of the diminished seventh B𝄫).

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 as in Table 2. Notice how ATG maps to a half note, rather than a quarter note,s 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 as in Table 4.

Then, notice how the codon ATT, which codes for the essential amino acid isoleucine and maps to the red C♯ major triad, is broken across the splice site between the second and third exons. The intervening intron assumes the key dictated by this final (albeit incomplete) codon, and thus outlines a C♯ diminished seventh chord. The third exon then 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 as 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.

Figure 5 shows part of the exon with the end of the CDS.

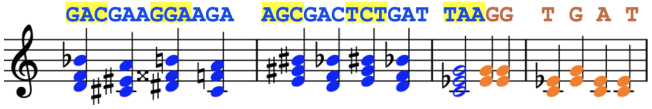

Figure 5: *Danio rerio* TP53 Gene, End of CDS

The CDS terminates with the stop codon TAA, which maps to a C minor triad as in Table 5. Notice how TAA maps to a half note, 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 structure of a gene and the biological processes of splicing and translation. The structure of the gene and its splice sites are illuminated by mapping individual nucleotides in UTRs and introns to diminished chords and minor dyads, respectively, 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.