

Demo: Sonic Catalog of Rare Diseases

Stephen Taylor
University of Illinois
Urbana, IL, USA
staylor7@illinois.edu

Aditi Kantipuly
McGill University
Montreal, Quebec, Canada
aditi.kantipuly@mail.mcgill.ca

Abstract

This 15-20 minute demo presents our work in progress, a sonic catalog of rare diseases, along with prior work of data-driven music based on genetic sequences from SARS CoV-2. These data-driven compositions are created from spreadsheets, imported into Max and Kyma, and mapped to musical sound.

CCS Concepts: • Applied computing → Sound and music computing.

Keywords: datasets, Max, Kyma, sonification, data-driven music

ACM Reference Format:

Stephen Taylor and Aditi Kantipuly. 2023. Demo: Sonic Catalog of Rare Diseases. In *Proceedings of the 11th ACM SIGPLAN International Workshop on Functional Art, Music, Modelling, and Design (FARM '23)*, September 8, 2023, Seattle, WA, USA. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3609023.3609807>

1 Introduction

We present our work in progress, a sonic catalog of rare diseases featuring genetic promoter sequences from 230 rare diseases, based on the Malacards human disease database [3] and the Monarch Initiative [4]. We built a spreadsheet with 10 columns for each disease, including the disease category, its associated promoter sequence, number of phenotypes, number of genes, number of variants, etc. Then, we imported the spreadsheet into Max (<https://cycling74.com>); with Max, Kyma (<https://kyma.symbolicsound.com>) and Apple's Logic Pro we created a data-driven composition lasting a little over seven minutes. The current version (fixed-media electronics) can be found here:

<https://www.youtube.com/watch?v=wz1fEO-jyr8>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. *FARM '23*, September 8, 2023, Seattle, WA, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0295-2/23/09...\$15.00

<https://doi.org/10.1145/3609023.3609807>

The video shows the Max patch, moving sequentially through each disease. We hear high plucked or struck notes, representing the DNA of the promoter sequence for each disease, accompanied by a lower, sustained harmony derived from the category of disease (metabolic, endocrine, etc.). After each DNA sequence we hear a series of percussive hits; these represent the number of phenotypes for each disease. The music begins in the high register, and descends as we move through the different categories; then it ascends once more at the end.

2 Mapping Rare Diseases to Indian Ragams

One of our goals in this sonic catalog is to use ideas from Indian classical music. Using the established disease groupings from "The National Economic Burden of Rare Disease Study" (2021), we systematically integrated these into the established twelve-raga system, relying on classical vedic texts, to inform our classification approach. Twelve different categories of disease, linked to different bodily systems, are each associated with a different chakra, derived from the Melakarta ragams. This is a collection of 72 scales, each with eight notes; each scale is divided into a lower and upper tetrachord (a group of four notes). The 72 scales are divided into twelve categories, or chakras, based on the lower tetrachord. For each of these twelve lower tetrachords, there are six possible upper tetrachords, to make a total of 72 scales. The scales are based on the pitches Saa Re Ga Ma Pa Dha Ni Saa; most of the syllables (except for Saa and Pa – Do and Sol in western music) have multiple versions, similar to sharps and flats in western music.

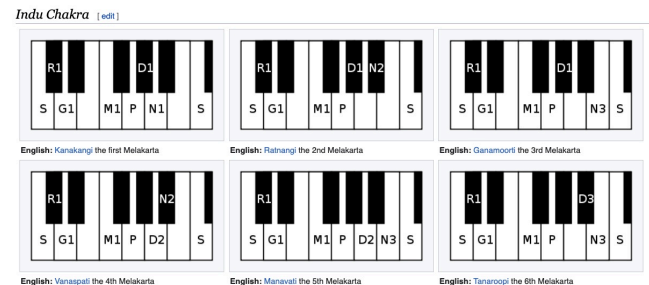


Figure 1. The first six Melakarta ragams; the lower tetrachord is the same for each scale. [Public domain], via Wikimedia Commons. ([https://commons.wikimedia.org/wiki/Melakarta_ragams_\(svg\)](https://commons.wikimedia.org/wiki/Melakarta_ragams_(svg))).

For our sonification scheme, the lower tetrachord for each chakra provides the harmonic background for each of the twelve disease categories. The smallest category, skin (represented by the Aditya chakra) consists of three diseases; the largest, chromosomal disorders and congenital malformations (the Brahma chakra), has 65. Meanwhile, the six upper tetrachords (versions of Pa Dha Ni Saa) remain the same for each chakra. These upper tetrachords represent the four DNA bases, which use a plucked or struck timbre to play part of the DNA promoter sequence for each disease. The amount of DNA letters played depends on the number of genes associated with the disease, ranging from 1 to 54. Table 1 shows a summary of the mapping.

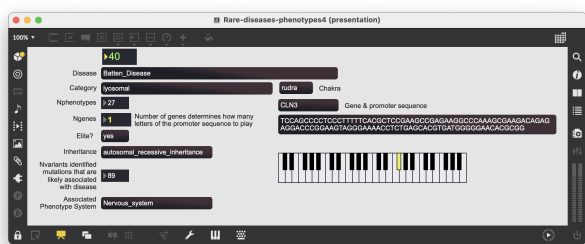


Figure 2. Screenshot of the Max patch for the catalog of rare diseases.

The Max patch is triggered by user input. Every time they hit the space bar (or type a number into the top number box), we hear a new disease in the catalog. In the video linked above, we are hitting the space bar to control how fast we move through each disease. One of the problems with a catalog of 230 diseases (or any large dataset) is how to get through the data in a reasonable amount of time. One way to solve this problem is by not playing the whole promoter sequence; instead, the amount of the promoter sequence depends on the number of genes associated with each disease (doing this also adds variety to the musical sound).

Another problem for a listener is how to know "where they are" in the dataset. We address this problem by using musical register: the music starts high, gradually descends to a low harmony on C2 (the Brahma chakra, the longest part of the piece), then climbs again to the end. This movement from high to low to high is inspired by a clock's hour hand in its circular journey; when the music hits its lowest point, it's about halfway done. Taylor has used a similar pitch trajectory in an earlier data-driven composition for carillon bells, based on the genome of the archaea *M. Jannaschii* [6]. For our rare disease catalog, this descent and ascent could also possibly recall a visitor's experience visiting Maya Lin's Vietnam Memorial; the visitor begins at ground level and walks underground at the memorial's midpoint, then back to the surface.

Also, for musical reasons we are adding a solo viola part, derived from the data but not as strictly as the electronic accompaniment; this will provide a sense of continuity, musical pacing, and a thread of human connection. We also plan to make an interactive version (without viola), where the user can explore the catalog on their own, without having to step through the dataset sequentially. Letting the user control the pace is another way to address the problem of large datasets.

3 Sonifying Proteins from the Coronavirus

Besides this sonic catalog, we also show previous related work by one of the authors (Taylor), using spreadsheets processed by Kyma to create short music videos of protein structures from the coronavirus (see Figure 2). Here, Kyma reads through a spreadsheet to create fixed-media video and audio. The collection of 18 short videos can be found here:

<http://www.stephenandrewtaylor.net/corona-vimeo-index.html>

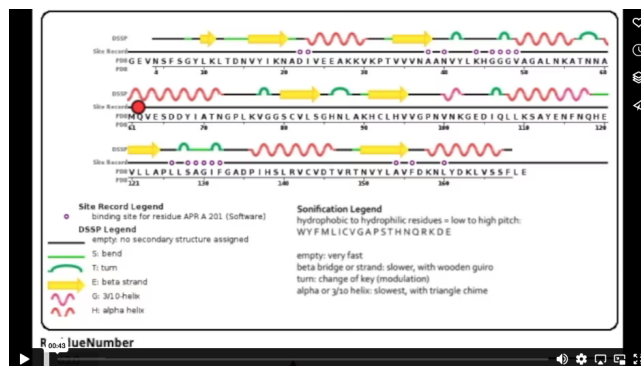


Figure 3. Screenshot of the Kyma video window for data-driven music from the coronavirus.

These pieces are derived from protein structures in SARS CoV-2, using data from the Coronavirus3D project [2] and the Protein Data Bank [5], with sounds made by the Kyma sound design environment. The spreadsheets here are much less complex; they consist only of the amino acid sequence for each protein, with an additional column for secondary structure (now with DeepMind's AlphaFold [1], which can predict secondary structure from the amino acid sequence, it might be possible to use only the single column).

Amino acids are represented by pitches (lower = hydrophobic, higher = hydrophilic), with different timbres: hydrophobic acids sound more "oily", while the electrically charged hydrophilic acids ring out. For secondary structure, a triangle chime marks each alpha helix (red curvy lines), while a wooden guiro marks beta sheets (yellow arrows). Turns (shown in green) are represented by modulations, so the music changes to a new key. Our hope is that these sounds help provide an understanding of how chains of amino acids are built — not just for the coronavirus, but for any protein. Another goal is for these to work as music: a collection of short

Table 1. Mapping Scheme for the Rare Diseases Spreadsheet

Category of disease	Chakra (lower tetrachord, sustained harmony)
DNA promoter sequence	Plucked timbres (upper tetrachord)
Number of phenotypes (1 to 15,000)	Percussion (the more hits, the more phenotypes)
Number of genes (1 to 54)	Amount of DNA promoter sequence (plucked sounds) to play
Associated phenotype system	Timbre of plucked instrument (harp, guitar, etc.)
Inheritance, Elite genes	Percussion timbre

pieces, sort of like an updated Bach cello suite, exploring different tunings (including microtonality) and timbres.

One common problem with data-driven music or sonification is that the tempo never changes: it advances one data point at a time at a constant speed. For some musical styles, of course, this is fine; but the rhythmic consistency could entrain or lull the listener, so they listen to the rhythm, not the data, as it were. Here, even though these protein-derived pieces progress through the chain of amino acids one at a time, the speed changes depending on secondary structure. Beta sheets (with the guiro) are a little slower; alpha helices (with the triangle) are slower still; and turns slow down a little as well. These swerves and abrupt changes of speed may help the listener experience the protein chain as a miniature, 3-dimensional roller coaster.

4 Sonification and Data-Driven Music

For both the catalog of rare diseases, and the music from SARS CoV-2, the mapping schemes are strict. If you know the mapping, you can infer quite a bit about the dataset just by listening. In this sense, these works are sonifications, “the data-dependent generation of sound, if the transformation is systematic, objective, and reproducible” [9]. But if you don’t know the mapping, our hope is that the result still sounds musical. In this sense, sonification is like visualization: if you don’t know the rubric (what the charts, colors or animations represent), the result is an abstract design which can be more or less aesthetically pleasing.

Following Carla Scaletti [7], we prefer to think of these as data-driven music. The main goal is not to arrive at a new scientific understanding of these phenomena, but to invite reflection. For the catalog of rare diseases, our fondest wish would be to bring hope to the rare disease community. In a way, one could say these pieces are an updated kind of program music [8]. Instead of Beethoven composing music based on a thunderstorm or babbling brook, we present music

that is not only inspired by natural phenomena, but derived from its underlying data.

Acknowledgments

Thanks to Carla Scaletti and Kurt Hebel for their invaluable technical advice with data-driven music from the coronavirus, and deep insights on sonification and music; and thanks to bioinformatician Teofil Nakov, who pulled out the promoter sequences for the corresponding genes and the respective diseases. Thanks also to the anonymous referees for their helpful comments and suggestions, both on the music and the presentation.

References

- [1] [n. d.]. *AlphaFold*. Retrieved July 16, 2023 from <https://www.alphafold.ebi.ac.uk>
- [2] [n. d.]. *Coronavirus3D*. Retrieved June 8, 2023 from <https://coronavirus3d.org/index.html>
- [3] [n. d.]. *MalaCards human disease database*. Retrieved June 8, 2023 from <https://www.malacards.org>
- [4] [n. d.]. *The Monarch Initiative*. Retrieved June 8, 2023 from <http://legacy.monarchinitiative.org>
- [5] [n. d.]. *Protein Data Bank*. Retrieved June 8, 2023 from <https://www.rcsb.org>
- [6] University of Illinois at Urbana-Champaign. 2015. *Archaea* by Stephen Taylor. Retrieved July 16, 2023 from <http://150.illinois.edu/archaea/>
- [7] Carla Scaletti. 2018. Sonification ≠ Music. In *The Oxford Handbook of Algorithmic Music*, Roger T. Dean and Alex McLean (Eds.). Oxford University Press, 363–386. <https://doi.org/10.1093/oxfordhb/9780190226992.013.9>
- [8] Stephen Taylor. 2017. From Program Music to Sonification: Representation and the Evolution of Music and Language. In *Proceedings of the 23rd International Conference on Auditory Display (ICAD 2017)* (Pennsylvania State University). International Community for Auditory Display. <https://doi.org/10.21785/icad2017.060>
- [9] Bruce N. Walker and Michael A. Nees. 2011. Theory of Sonification. In *The Sonification Handbook*, Thomas Hermann, Andy Hunt, and John G. Neuhoff (Eds.). Logos Verlag, 8–40. <https://sonification.de/handbook/>

Received 2023-06-01; accepted 2023-07-01