Let The DNA Speak
Project Proposal

Stefan Appelhoff, Kim Philipp Jablonski,
Nina Krüger, Sourabh Lal,
Tom Wiesing & Mengyuan Zhang

March 17, 2014

Contents

1 Background & Motivation

When we were discussing about what we wanted to do as a group for our final sonification project, we all on the one hand wanted to deal with complex data that would not be easily analyzed in any other way than sonification and on the other hand we wanted the whole project to also have a relevance to a certain academic field. Because our group involves diverse majors such as computer sciences, psychology, mathematics, and computational biology, it was a bit difficult at first to find a common ground with a topic that could spark an interest for everyone. However, after some brainstorming we agreed to the topic of analyzing DNA.

During the lectures we have had so far, an intriguing purpose of sonifying data for us was the increase of performance during the workflow. Being able to attend to different tasks, while simultaneously checking a rather monotone, different task sounded to us like something worth working for. With this notion and our more specific interest in the topic of DNA, we quickly came up with the idea of building a tool that allows researchers to analyze different representations of DNA.

2 Specific Focus: Comparing DNA Representations

One of the most obvious reasons to use different representations of DNA is to compare it. Ranging from graphical devices like codon atlas or chromatograms, over more data centered possibilities like simple tables, they offer a wide variety of learning and perceiving the structure of DNA through many different senses. However, we think that one promising opportunity has been wantonly neglected, namely the way of sonification. In our opinion this offers a whole new spectrum of interaction and understanding of the DNA you probably only know from images or graphs. In order to demonstrate the applicability of our approach, we are going to compare several DNA strands. The main difficulty then was to find DNA sequences which are similar enough to be comparable, but different enough to have a noticeable auditory difference. In order to achieve that goal we chose genes which encode for the same function but exist in different organisms. Another approach would be to compare mutated and healthy genes and thus deduce their possible severity. Hereby we hope to observe a general similarity of the compared strands, but with significant differences in distinct positions, which will then lead to a new auditory result.

3 Research Question

Can comparing DNA strands of genes which are encoding for similar functions in different organisms produce an audible result when the method of sonification is applied?

4 Methods

The required features will be using JavaScript HTML and CSS. We are going to create a website where can interact with the sonified DNA. We will provide a lot of sample DNA as well as allow the end user to input their own DNA sequences. The website will be available at [12]. We are using two different libraries for implmenting the sonfication, timbre.js from [13] and MIDI.js from [14]. The libraries allow us to both play sine waves and thereby construct all kinds of sound manually as well as to play notes on different instruments directly.

On the website there will be different modes to sonify DNA. One mode is to group the DNA into triplets and then assign a note to each triplet. Another mode is to play DNA on a per-base basis, i. e. assign each "letter" in the input string a note and then play that. Apart from having different modes, it will also be possible to "play" DNA using different instruments such as piano, xylophone, violin and trumpet. When playing DNA it is also possible to play different DNA at the same time. When using different instruments for different DNA, it will be possible to easily compare DNA by hearing if the instruments play in sync.

We will get our data from the European Nucleotide Archive. The European Nucleotide Archive is a publicly available database of Nucleotide sequences. Individual sequences have a length of around 300000 characters. Sonifying these long datasets will be very difficult because we will have to select only small fractions of them to use. We expect to find large parts of the sequences to be random because not all information contained in them is actually required by the organism the sequences come from.

5 Why Sonification?

Our program mainly allows for fast comparison of very large DNA sequences. For instance, imagine you are in the laboratory replicating a huge amount of genes by cloning. Of course you always want to clone the exact same gene, meaning they should all have the exact same DNA sequence. But how do you want to make sure, that this is the case? DNA sequencing is relatively easy. However, you still have to compare every single nucleotide of your original sequence to every single nucleotide of your cloned gene. Naturally, this would be a huge waste of time. Now further imagine you could listen to music while you work. Then, suddenly the music sounds a little off and exactly that tells you that there is a wrong triplet in your cloned DNA sequence. In exactly the same way you could compare several sequences at the same time and as soon as you hear something weird, you can just check the corresponding nucleotide triplets, which are depicted on the screen and easily identify the faulty one.

6 Problems with our approach

In several psychological studies, listening to music while working has been shown to disrupt the work flow (Thompson, Schellenberg & Letnic, 2012) and impairing the performance. The "Cognitive Capacity Hypothesis" is often put forward to explain for these detrimental effects of multitasking (Baddeley, 2003): In summary, the hypothesis describes our attentional system as a limited channel, that only has a certain capacity. Information or any sensory input from the environment to the human is regarded as a flow of information that has to go through the limited channel of our attention. In the case of music we are dealing with a specifically problematic case, because music is using the same modular brain regions as speech (Peretz & Coltheart, 2003). This hypothesis is put forward by comparing musical entities such as chords to entities of speech such as words. Reading alone places a high demand on attention (Carretti, Borella, Cornoldi, & De Beni, 2009), not to speak of the combined attentional costs of reading, writing, listening to music and perhaps even talking or checking a computer screen. With this prospect, it seems very likely that the combined attentional costs due to this multitasking exceed our "limited capacity" of attention. The result would not be the wanted overall gain for performance through combining tasks with simultaneously checking DNA through music, but rather an overall loss in efficiency.

7 Possible Solutions

However daunting the problem of the "Cognitive Capacity Hypothesis" might sound, there are also positive sides that need to be considered. Firstly, it has recently been put forward that due to the drastic increase in media use and the closely connected increase in daily activities of multitasking behavior, people might increase their general cognitive performance in multitasking (Alzahabi & Becker, 2013). Generally Alzahabi & Becker argue that the behavior of being active on so many diverse devices and tasks such as writing text messages, watching TV, checking emails, etc. is so ubiquitous, that a training effect can result in a cortical reorganization (Draganski & May, 2008). Next to the prospect that people, due to the incresingly demanding environment, get better at multitasking, there is another argument that might solve the problems suggested by the "Cognitive Capactiy Hypothesis": The ability to perceive incoming information and manipulate it in a "mental workspace" is called working memory. Working memory is one of the most important executive cognitive functions and heavily involved in task-switching and multitasking behavior. The problem of the "Cognitive Capacity Hypothesis" as outlined above describes a disruption of this system due to an exceeding inflow of information. However, research on the Irrelevant Sound Effect has suggested that to disrupt the working memory system with auditory stimuli (e.g., music), these auditory stimuli need to have a considerable amount of acoustical variation over time (Perham & Vizard, 2010). Taking into account that the music produced by our DNA sequences will be repetitive and inconclusive, until a mistake or misalignment is there, it could well be that the working memory performance stays intact despite the musical input until the music suggests a unnormality.

8 Who does what

TBD Mengyuan

References

[1] MIDI.js Authors. *MIDI.js*. 2010-2013. URL: https://github.com/mudcube/MIDI.js.

[2] BBC. *Sonification: Representing data through music*. Mar. 2014. URL: http://www.bbc.com/news/magazine-25975712.

[3] EMBL-EBI. *European Nucleotide Archive*. Mar. 2014. URL: http://www.ebi.ac.uk/ena/.

[4] *European Nucleotide Archive*. 2014. URL: http://www.ebi.ac.uk/ena/.

[5] Catherine Faas. *17 interesting facts about DNA*. Mar. 2014. URL: http://holykaw.alltop.com/17-interesting-facts-about-dna.

[6] Anne Marie Helmenstine. *10 Interesting DNA Facts*. Mar. 2014. URL: http://chemistry.about.com/od/lecturenoteslab1/a/10-Interesting-Dna-Facts.htm.

[7] *Let The DNA Speak*. 2014. URL: http://letthednaspeak.tk/.

[8] John G. Neuhoff Thomas Hermann Andy Hunt. *The Sonification Handbook*. Berlin: Logos Publishing House, 2011. ISBN: 978-3-8325-2819-5.

[9] nao yonamine. *timbre.js*. 2012. URL: http://mohayonao.github.io/timbre.js/.