Sonification of Multiple Sequence Alignments for Data Exploration in Bioinformatics

Edward J. Martin

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2019

Abstract

Making multiple sequence alignments (MSAs) is a key method for comparative analysis of protein sequences. However, MSA data exploration is difficult due their complex nature and the diversity of research questions.

Parameter Mapping Sonification (PMSon) is a form of auditory display in which data features are mapped onto sound synthesis parameters, using non-verbal sound to convey information. This allows a novel mode of interaction with data for domain experts.

Five original PMSons for data exploration in bioinformatics are proposed and implemented as software in the *Perl* and *Sonic Pi* programming languages, three for protein sequences and two for multiple sequence alignments.

Much bioinformatics data exploration software and PMSons lack end-user input in evaluation, which leads to uncertainties in their development and causes issues in uptake and efficiency. This project implements qualitative research methods, including survey research, the NASA TLX measure of subjective workload, and a focus group, to make end-users the centre of the evaluation process of this software.

Acknowledgements

I'd like to thank Daniel Barker, Heleen Plaisier, and Thomas R. Meagher for their prior work on this project.

I'd also like to thank Stevie Bain for her support with the evaluation section of this work.

Table of Contents

Chapter 1

Introduction

1.1 Sonification

Sonification is the use of non-verbal sound to convey information[26]. More generally, it refers to a range of different approaches which work analogously to data visualisation in the sonic domain. In doing this sonification uses key features of psychoacoustics to facilitate knowledge transfer[47]. Sonification is divided into five sub-domains: Audification, Auditory Icons, Earcons, Parameter Mapping Sonification, and Model-Based Sonification[23].

Parameter-Mapping Sonification (PMSon) conveys information by mapping data features into sound synthesis parameters[16]. Figure 1.1 demonstrates the process of PMSon design, illustrating how effective PMSon design involves the interplay of data processing, sound synthesis, and human perception, in both the data and sound domains. In this paper, I will use the term sonification to refer to parameter mapping sonification unless otherwise indicated, as both this work and most relevant background work concerns this domain.

1.2 Multiple Sequence Alignment

Proteins are essential to life, and perform a vast array of functions within living creatures. They are constructed of sequences of amino acids, and a protein is defined by the sequence of amino acids of which it is composed. In biology it is standard to represent each of these amino acids by a single, capitalised letter. There are 20 main amino acids coded by genes. Different amino acids have different physico-chemical properties. Figure 1.2 demonstrates some of the complex properties of different amino acids, and

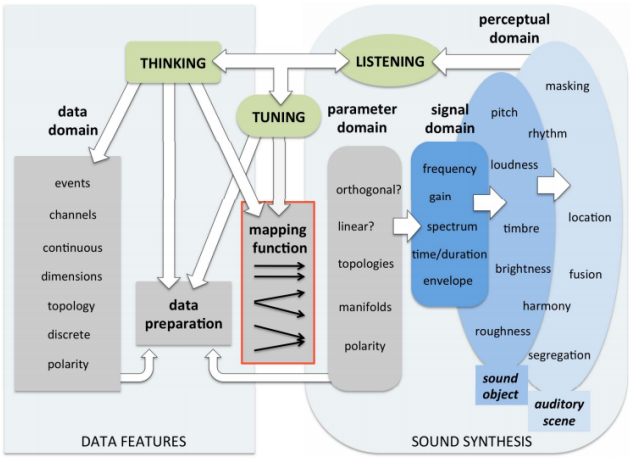


Figure 1.1: Visualisation of general design process of PMSon, reproduced from *Parameter-Mapping Sonification* in *The Sonification Handbook*[16]. PMSon involves mapping data features (left) to sound synthesis parameters (right). Data and numerical controls are in grey, sound and auditory factors are in blue. The green boxes indicate the contribution of human perception.

also some of the groupings which biologists use to help understand similarity between different amino acids. The relationships between the different amino acids in a protein determine the 3-dimensional structure and the function of the protein.

By assuming that protein sequences that share similarity also share function, biologists can use comparative analyses of proteins to find sites with significant evolutionary conservation. A major tool for this type of comparative analysis is the creation of *multiple sequence alignments (MSAs)*. An MSA is an alignment of three or more protein sequences, which uses the following algorithmic approach: given n protein sequences $S_i =; i = 1,\ldots,n$ of corresponding length $m_1, m_2, \ldots, m_n$, of the form:

$$S := \begin{cases} S_1 = (S_{11}, S_{12}, \ldots, S_{1m_1}) \\ S_2 = (S_{21}, S_{22}, \ldots, S_{2m_2}) \\ \vdots \\ S_n = (S_{n1}, S_{n2}, \ldots, S_{nm_n}) \end{cases}$$

An MSA is created by inserting gaps into each of the S_i sequences until the modified sequences, $S_i^{'}$ conform to length $L \geq max\{m_j | j = 1,\ldots,n\}$ and no columns of $S_i^{'}$ consist of only gaps. These gaps are inserted to optimise a cost function, such as a score

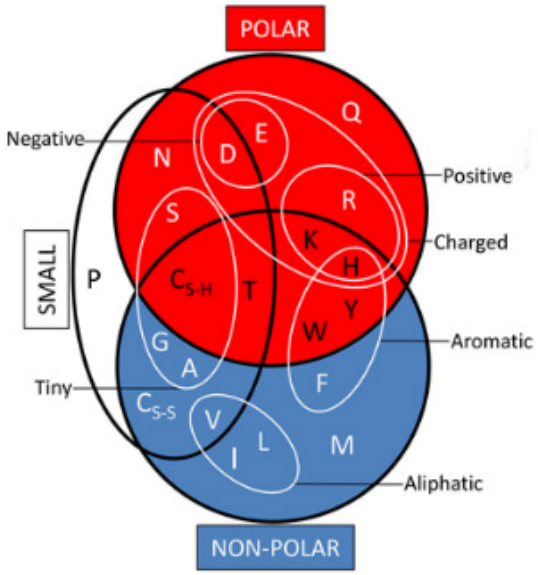


Figure 1.2: Venn diagram of physico-chemical properties of the 20 amino acids. Hydrophillicity(red) and hydrophobicity(blue) are correlated with polar and non-polar amino acids respectively, reproduced from *Unearthing the Root of Amino Acid Similarity* J. Stephenson et. al.[42].

derived from substitution matrices[21]. The output of the process is a two dimensional matrix. Approaches to multiple sequence alignment are generally heuristic, due to the complexity of the optimisation problem[48]. Researchers use MSAs to infer sequence homology; infer evolutionary relationships; predict aspects of protein structure; discern protein disorder, function, and localisation; understand genomic rearrangements; and estimate evolutionary rates[51]. Visualising MSAs is key to connecting domain experts with the data, and there are many different software approaches to visualising MSAs. Most use colour to represent different aspects of the information presented. Figure 1.3 demonstrates one approach to the visualisation of MSAs. Each row corresponds to a protein in the alignment, with the capital letters representing the amino acids, and a '-' character representing gaps inserted by the algorithm for the purposes of alignment. Biologically, gaps represent insertion mutations in the longer sequences, and/or deletion mutations in the shorter sequences. Due to the complex properties of amino acid interactions visualisations often end up complex or overloaded[39].

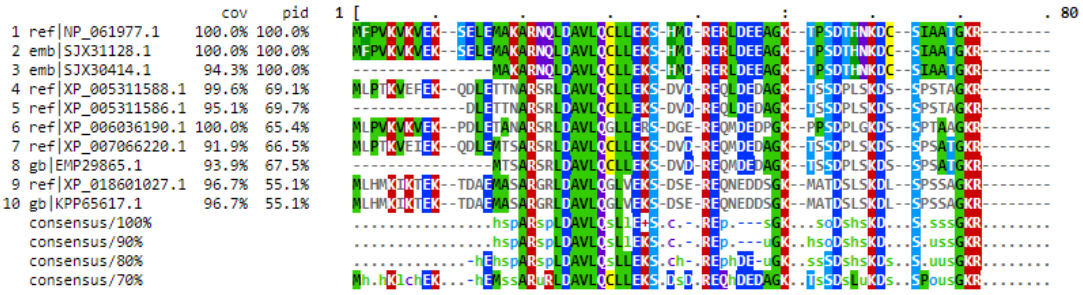


Figure 1.3: Small section of a 10 protein example multiple sequence alignment visualised with MView[30]. Residues in rows 2 to 10 are coloured if they match with the reference sequence in row 1, the colours are determined by the amino rsidue.

1.3 Hypotheses

The project revolves around three hypotheses.

1. The informative content of protein sequences can be conveyed through mapping amino acids to sounds using PMSon methods.
2. Sonifying small multiple sequence alignments can provide researchers with useful complementary information to current visualisation methods for data exploration purposes.
3. The development of a qualitative evaluation framework centred on end-users will improve the uptake and quality of both sonification approaches and software for data exploration in bioinformatics.

1.4 Objectives

Each of the hypotheses corresponds to a measurable objective which will be tested in the evaluation stage of this research. The first objective is that the sonifications produced by this work will be able to transfer protein sequence information to researchers in the field of bioinformatics. The second objective is that researchers in the field will be able to glean useful information about a multiple sequence alignment using the sonifications produced by this work in conjunction to visualisation methods. The third objective, despite extending beyond the scope of this work, is that the outcomes of the evaluation process improve the usability and uptake of the software being developed, and that the evaluation framework is used by the bioinformatics and sonification communities.

Novel contributions of the work are:

- Three novel analysis-driven mappings of the amino acids of a single input protein sequence into sound and their implementation as software:
  - A novel hydrophobicity scale mapping from the amino acids to pitch which improves on previous methods by being less musically focused, data-driven, and motivated by a experimental hydrophobicity values.
  - The use of a biochemically significant and significantly smaller reduced alphabet than previously used in encoding the amino acids to sound.
  - A combination approach incorporating both of these methods using different sound characteristics to incorporate both approaches.
- Two novel sonifications of protein multiple sequence alignments and their implementation as software, which apparently no one has ever done before:
  - A column-based entropy sonification of the conservation of an MSA. This is the first entirely automatic protein mapping using an automated mathematical analysis process.
  - A sonification using the hydrophobicity scale developed for the protein sonification.
- The implementation of end-user based qualitative research to conduct phenomenological evaluation of bioinformatics software for data exploration.
  - Survey research of end-users.
  - Implementing the NASA Task Flow Index for evaluating subjective workload of data exploration tasks in bioinformatics.
  - Focus group research for diverse and in-depth qualitative research in a group setting.

1.5 Motivation

Although rapid advances in sequencing technologies are causing a flood of sequence data, it is the complexity of these highly-interconnected data sets that provide the greatest challenge for data exploration of genomic data[35]. Some of this complexity is demonstrated in Figure 1.2, where some the complicated interrelations of the amino acids are shown. Other complexities of these data sets involve the varying effects of substituting amino acids for one another, where the effects of substitution can range from major to minor depending on the amino acid. Some of this substitution information is captured by *substitution matrices* such as PAM and BLOSUM, which contain

210 numbers to represent the probabilities of transformation using log-odds scores[21].

Another example of this complexity is in the task of identifying amino acid repeats (AARs). AARs are repeated sequences of amino acids found within proteins that have specific roles in protein function and evolution, however their evolutionary and functional scenarios are poorly understood, especially in eukaryotic proteins. Finding repeated patterns is an NP-hard problem in principle, and the difficulty of identification, despite different algorithmic strategies, has contributed to an insufficiency of understanding[28]. Furthermore, biologists are notorious for asking their data complex questions[39]. The use of a multi-modal approach, whereby sonification is used alongside visualisation methods could help meet these needs for innovation in data exploration. Previous work has demonstrated that sonification techniques can convey a range of genomic information[45][46]. PMSons are well suited to dealing with complex and interconnected data sets [16]. The use of sonification as part of a multimodal approach to data exploration is a way to bring the required innovation to this enterprise, and to many other problems within the field of data exploration of protein sequence data in bioinformatics.

Further to this, sonification research is often lacking in evaluation methods[33]. Although there are some examples of sonifications of protein sequence data with evaluation processes involving experiments, such as the PROMUSE system detailed in Section 2, often these do not involve end-users and usability, but are limited to questions of the form 'can you recognise this characteristic from this sound?'. Evaluation processes are also seemingly absent from data exploration software research for bioinformatics: a 2010 review titled *Visualization of multiple alignments, phylogenies and gene family evolution* made no reference to feedback or evaluation of any of the ($>$15) visualisation approaches featured[39]. Qualitative research methods can meet this need by using the phenomenological input of end-users to evaluate the success of methods in the development of these software approaches[43].

1.6 Results Achieved

All five software implementation successfully produce sound from a range of Fasta sequence data files. The success of the first objective is supported by both the focus group and questionnaire results, which provides evidence that the informative content of protein sequences can be conveyed through mapping amino acids to sounds using PMSon methods. The evaluation process has provided some evidence that the sec-

ond objective can be met, however there is more work to be done to justifiably say that sonifying small multiple sequence alignments can provide researchers with useful complementary information to current visualisation methods for data exploration purposes. However, clear feedback from the evaluation process has identified the future work necessary to meet this objective. The success of the third hypothesis is harder to evaluate, however the evaluation process of this work has given clear and useful feedback. I believe that this process will improve the uptake and quality of this sonification approach. Judgements on the effect on the wider sonification and bioinformatics communities are beyond the scope of this work.

1.7 Structure of Paper

This paper will begin with a critical discussion of the background work to the project. This will be followed by a description of the implementation process of the sonification software. This will include a discussion of hardware, programming languages, improvements to previous work on the project, and the sound synthesis approach in *Sonic Pi*. A description of the implementation of each of the five sonification algorithms will follow, including conceptual work and pseudocode. The section on evaluation will describe the qualitative methods used in the questionnaire, NASA Task Load Index, and focus group.Results and a discussion will follow. The conclusion will include unsolved problems and future work.

Chapter 2

Background

The metaphor of biological sequence data as musical score is a popular one. Douglas Hofstader imagined the RNA transcription process as to a tape recorder where amino acids synthesise the music of proteins[24]. This wow-factor has also led much aesthetic work in the sonification of biological sequence data for the purposes of teaching and scientific outreach[12]. Geneticist Susumu Ohno suggested that the common presence of repetition in both biological sequence data and music implied that repetition was more fundamental to world than randomness[36]. The pioneering work of mapping amino acids to pitch was performed by Kenshi Hayashi and Nobuo Munakata in 1995[34]. They injectively mapped 20 amino acids to 20 musical notes using a hydrophobicity scale adjusted to preserve the groupings of similar amino acids. Their mapping extended earlier work, where nucleotides were mapped to notes of a major scale[20]. The tonal range of the mapping was capped at a fifth because of an analogy with the tonal range of speech. However, when expanding on this work to apply it to amino acids they decided to use the same range five times in succession. This resulted in a large 44 semitone pitch range. The mapping subsequently had irregular gaps between notes for which there was no physico-chemical metaphor. Musical complexity for aesthetic purposes was included at the cost of analytic considerations. Although the authors suggested that different instruments could play different sequences concurrently for comparison using MIDI sequencers and synthesizers, this was purely theoretical work with no software produced.

King and Angus introduced the use of a reduced alphabet to the sonification of proteins[25]. The 20 amino acids were mapped surjectively into seven different groups representing physico-chemical properties: polar, hydrophobic, charged, positive, aliphatic, aromatic, and tiny. The mapping imposed that each of the first six groups were consid-

ered to be mutually exclusive and corresponded to the first six notes of a major scale, with tiny proteins played an octave higher. This approach simplifies complex memberships of these groups shown in Figure 1.2. This makes the metaphor of the mapping less true to the physical nature of the amino acids. This approach was implemented as software.

Although concerning structural protein alignments, a different analytical method than MSAs, the PROMUSE software included some innovation[17]. A higher level view of the data is given by sonifying an analysis of the data, rather than the data itself. The paper also conducted an experiment to see if the participants were successful in identifying data features from the sonification. However, the participants in this evaluation were not end-users but computer science students. The sonification approach itself was musical, mapping data aspects to the four components of a jazz quartet: bass, drums, chord, and a lead instrument.

Takahshi and Miller created software that mapped similar amino acids to the same root note, creating a musically inspired 12 note basis for their sonification. This lacked a clear metaphor for analysis. They used metrical variation and different chord inversions to distinguish between different amino acids[45]. More recent work by Robert P. Bywater uses sonification of higher level structural information for melodic pattern identification to identify 3d structural motifs in sequence data[11]. This work was evaluated by a perceptual study focused on whether participants could associate sound with 3d models of proteins, although 55% of participants having less than a year's experience with proteins. A web tool for the sonification of DNA and protein sequences was developed that argued for the inclusion of sonification within sequence browsers[46]. The amino acids were mapped injectively onto musical notes. The tool provided a complementary visual display and also allowed users options in the sound parameters used in each application. It also included several examples of protein features that could be heard via the sonification like a tutorial. This seems an important method for engaging new users with the approach. Unpublished work from Daniel Barkers research group and colleagues, by Heleen Plaisier of the Royal Botanic Garden Edinburgh (formerly of the University of Edinburgh), Thomas R. Meagher of the University of St Andrews, and Daniel Barker of the University of Edinburgh is also key to this project. The unpublished software is comprised of *Perl* code which takes a DNA sequence in Fasta format as input and maps it injectively into sound. The output is a looping sonification in *Sonic Pi* software. This is the approach on which I will be basing my implementation, however with more ambitious objectives.

Chapter 3

Implementation

#	Name	Input	Output	Mappings
3.5	Hydrophobicity Scale	Protein	2	Hydrophobicity to Pitch
3.6	Reduced Alphabet	Protein	5	Reduced Alphabet to Pitch
3.7	Hydrophobicity and Reduced Alphabet	Protein	5	Hydrophobicity to Pitch Reduced Alphabet to Synth.
3.8	Entropy	MSA	1	Positional Entropy to Pitch
3.9	MSA Hydrophobicity Scale	MSA	$n+1$	Hydrophobicity to Pitch Consensus to Loudness

Table 3.1: Summary of Sonification Implementations. *Output* refers to number of *list* objects created in *Sonic Pi* code to be played simultaneously, n refers to number of proteins in the MSA).

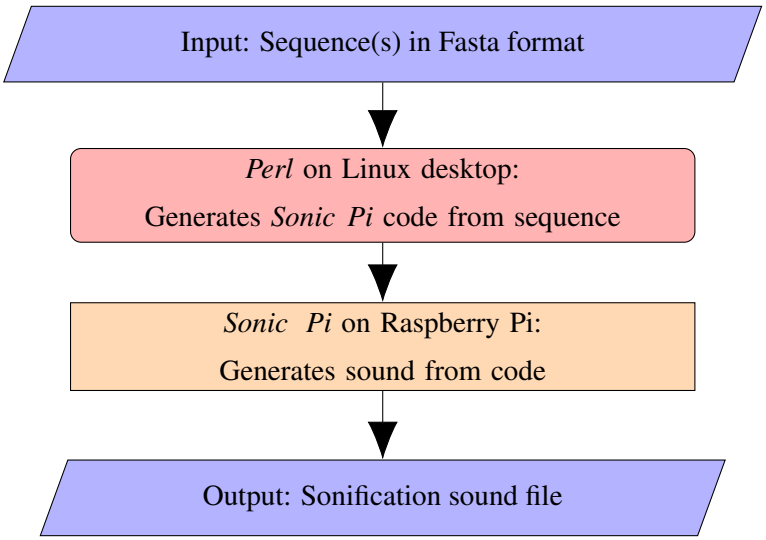
In this section I will first discuss my choice of programming languages. I will first detail the changes I've made to previous work on the project and the *Sonic Pi* methods that are common to all five of my sonifications. I will then detail the hardware, programming languages, and *Sonic Pi* sound synthesis approach. This will be followed by the descriptions of my five algorithms for sonification, including conceptual design work and pseudocode, including problems and difficulties encountered, my suggested solutions, and alternative solutions and their evaluation.

3.1 Hardware and Operating Systems

I used two separate pieces of hardware, a Raspberry Pi computer and a desktop. The desktop was running *Scientific Linux 7.6 (Nitrogen)* distribution[2]. The Linux kernel version was Linux 3.10.0-957.12.1.el7.x86_64 x86_64. The Raspberry Pi used version

1.5 of the 4273Π variant of the Raspbian operating system[9]. The reason for using the Raspberry Pi was that I was unable to get *Sonic Pi* to work on the *Scientific Linux* desktop, as it is not supported by the developer. I tried to use WINE to facilitate installation but was unsuccessful[5]. As the host research group has access to Raspberry Pi machines and they are supported by the *Sonic Pi* developer, I decided that this was the best solution for the current project.

3.2 Programming Languages



The implementation of this work makes use of two programming languages. I am using the *Perl* programming language to create my scripts, specifically *Perl 5*, version 16, subversion 3 (v5.16.3)[4]. This is predominantly because previous unpublished work on the project has used *Perl*, and not because of any inherent advantage in the *Perl* language. Also *Perl* handles regular expressions well and is widely used in bioinformatics, especially the *BioPerl* packages which I used to improve the succinctness and human readability of the code for future collaborative working purposes[41]. Any interpreted language would have essentially similar performance and speed is not a concern currently. I am using *Sonic Pi* (v3.1.0) software for sound synthesis[7]. This is again, in part, due to the use of *Sonic Pi* in previous work. The synthesis aspects of *Sonic Pi* are sufficient for the current work. I had also anticipated that the buffer-and-thread design of *Sonic Pi* would act as a metaphor for the different proteins in the structure of multiple sequence alignments. However, I quickly abandoned this approach during the early stages of development. *Sonic Pi* did provide some limitations to the project. As mentioned previously, *Sonic Pi* is not supported for all distributions

of *Linux* and therefore my implementation was complicated by the use of separate hardware. *Sonic Pi* also provided a limitation of data size, however this was overcome by using smaller multiple sequence alignments. This was appropriate for the project, as small MSAs were the main subject of investigation, although future work may have to find a way to deal with large alignments.

3.3 Initial improvements from previous work

The initial work from Daniel Barker's laboratory, mentioned in Section 2, was a script which translated a DNA sequence to a sequential four sound output using an injective mapping. The script was successful but there were some improvements to be made. The *Sonic Pi* function *play_pattern_timed* that was being used to initiate sound synthesis was redundant, with *Sonic Pi* Developer Sam Aaron describing it as bad on many levels on the *Sonic Pi* Issues thread on *Github*[6].

My first solution was to store the protein pitch synthesis parameters in an immutable *ring* data object. These could then be played using the function *tick*. Although this solution seemed more elegant from programming perspective, this approach caused difficulty when playing two or more data structures simultaneously. This was an aim for later work with MSA sonification. I also considered using nested lists to create two-dimensional arrays within *Sonic Pi* as a potential solution. However, this compromised user readability of pitch synthesis parameters meaning for harder future debugging of code and another layer of explanation for anyone wishing to understand how the code worked. I decided to use the simpler *list* data object, creating a separate list for each set of pitch synthesis parameters. This allowed for simultaneously playing of many separate data structures and also easy human readability for debugging during development.

The previous code had used the default synthesizer in *Sonic Pi*. I altered the code such that these synthesizer parameters were included. This means that synthesizers should be more consistent for future *Sonic Pi* versions if the default parameters change. I tried many of the synthesizers and parameters that the software offered, first settling on the *pluck* synthesizer. This was partly due to the quick onset of the note, and a percussive quality to the sound which I felt marked each note onset with the feeling of a beat, thus making them noticeably distinct from one another which serves as an audio metaphor for the discrete nature of protein sequence data. I later found that the *pluck* sound was quite annoying when played repeatedly, which made it harder to maintain

attention during sonifications. I decided that my sonifications would use *sine* and *saw* synthesizers. When the attack parameter of these synthesizers was set to a value of 0, it kept the sense of a beating pulse which maintained the audio metaphor. The selection of two separate synthesizers was partly for myself during development and partly for participants in my evaluations, so that each sonification had its own unique characteristic to disambiguate the sounds. In future implementation these could be encompassed by different options for end-users.

The previous work made use of *Sonic Pi*'s *live_loop* functionality. However, as my evaluation was going to use sound files recorded from *Sonic Pi* and played outwith the software package, I saw no benefit to the looping attributes of this functionality. This also allowed me to overcome the tricky early problem of marking the start and end of a looped protein sequence. My early solution attempts had not been satisfactory: using rests seemed to overload the mapping, as rests already represented gaps in the sequence; using low notes or high notes similarly seemed to break the metaphor of pitch corresponding to some attribute of the amino acid; and using a sample to mark the start and end did not flow well with the rest of the work. By removing the looping function, the start and stop were marked by the beginning and end of the sound file. This seemed to work well with the audio metaphor of *playing* the protein sequence: one sequence of amino acids makes one sequence of sound. However, in future implementations this means that navigation of the proteins would need to be easily controlled by the user. Also, the automatic looping of shorter sequences facilitated multiple listens, encouraging better engagement from listeners. This could be implemented as a function in future software development. The previous code had superfluous input for DNA inputs, which I removed. The previous code also used letters for pitch notation in *Sonic Pi*. I altered this to MIDI numbers to allow mathematical mapping procedures, allowing for innovation from the exclusively manual mapping procedures detailed in Section 2.

3.4 *Sonic Pi* code

All the 5 implementations in Sections 3.5 to 3.9 output similar code in *Sonic Pi*. Pitch synthesis parameters are created as *list* data structures. Following this play instructions are given, with instructions to play all the lists once, simultaneously, and with appropriate sound synthesis parameters. The play instructions are enclosed in a for-loop. All the following algorithms use this approach to writing play instructions for *Sonic*

Pi. Algorithm 1 contains example pseudocode.

Algorithm 1 *Sonic Pi* play instructions

```
1: Print "length.time do ||i||"
2: for j from 1 to number of output strands do
3:     Print "use_synth :sine newline play strandj[i], release: 0.5, attack: 0, cutoff: 80, amp: 85"
4: end for
5: Print "end"
```

3.5 Protein: Hydrophobicity Scale

This is an injective mapping of the 20 amino acids to 20 MIDI pitches. It is inspired by the hydrophobicity scale approach of Hayashi and Manakata[34]. However, I wanted a pitch range smaller than their 44 semitones to aid in pattern recognition for listeners. I also wanted to create a data-driven generative approach to mapping, to remove some of the false implied-relationships caused by the musicality-first approach they had taken.

n	*Difference* Range	Increase	Count	MIDI span
1	$0 \leq x < 0.5$	1	11	11
2	$0.5 \leq x < 1$	2	6	12
3	$1 \leq x < 2$	3	0	0
4	$2 \leq x < 4$	4	3	12
			total	35

Table 3.2: First Mapping approach: the upper limit for the n^{th} *difference* range was mapped to 2^{n-2} and the *increase* value was n.*n* is an index. x refers to the *difference* column in Table 3.3.

First, I needed hydrophobicity data. I used experimental work performed by Engelmann et. al. detailed in Table 3.3[14]. This data includes a measure of the sum of experimental hydrophobicity and hydrophillicity values, called the *Water-oil* scale. I sued this as the basis for my mapping. I attempted a few naive implementations. Mapping to western tonal musical scales such as a minor pentatonic scale seemed to give rise to musical metaphors which represented the data falsely. Using equal temperament semi-tones was a better approach, however it was hard to distinguish the degree of hydrophobicity from the sound. I needed a process that separated amino acids that had very similar *Water-oil* values, kept a reasonable pitch range overall, and captured larger jumps in the *Water-oil* scale.

Code	Amino Acid	Hydrophobic	Hydrophillic	Water-Oil	Difference	MIDI
F	Phenylalanine	-3.7		-3.7		50
M	Methionine	-3.4		-3.4	0.3	51
I	Isoleucine	-3.1		-3.1	0.3	52
L	Leucine	-2.8		-2.8	0.3	53
V	Valine	-2.6		-2.6	0.2	54
C	Cysteine	-2.0		-2.0	0.6	55
W	Tryptophan	-4.9	3.0	-1.9	0.1	56
A	Alanine	-1.6		-1.6	0.3	57
T	Threonine	-2.2	1.0	-1.2	0.4	58
G	Glycine	-1.0		-1.0	0.2	59
S	Serine	-1.6	1,0	-0.6	0.4	60
P	Proline	-1.8	2.0	+0.2	0.8	61
Y	Tyrosine	-3.7	4.0	+0.7	0.5	62
H	Histidine	-3.0	6.0	+3.0	2.3	65
Q	Glutamine	-2.9	7.0	+4.1	1.1	66
N	Asparagine	-2.2	7.0	+4.8	0.7	67
E	Glutamate	-2.6	10.8	+8.2	3.4	71
K	Lysine	-3.7	12.5	+8.8	0.6	72
D	Aspartate	-2.1	11.3	+9.2	0.4	73
R	Arginine	-4.4	16.7	+12.3	3.1	77

Table 3.3: Amino acid hydrophobicity values adapted from previous experimental work by Engelmann et. al[14]. *Water-oil* is calculated by a sum of the hydrophobic and hydrophillic values. The *difference* column has been added to indicate the increment from the previous amino acid when ordered by the *water-oil* value, and the MIDI column has been added to demonstrate the note attributed by my method.

n	*Difference* Range	Increase	Count	MIDI span
1	$x < 1$	1	17	17
2	$x \geq 1$	2	3	6
			total	23

Table 3.4: Second Mapping approach: the *differences* were divided into two groups, where x is less than 1 and where x is greater than 1. The *increase* MIDI step size in pitch equal to n. The *difference range* corresponds to the *Difference* in Table 3.3.

I developed a generative process that created MIDI values from the *Water-oil* values. I wanted to capture the magnitude of hydrophobicity in the sonification. I calculated the difference between the *Water-oil* value for each residue and the one preceding it in the ranking, this can be seen in the *Difference* column of Table 3.3. I used these values to create three possible mappings of the 20 amino acids to MIDI pitches, detailed in Tables 3.2 to 3.5.

n	***Difference* Range**	**Increase**	**Count**	**MIDI span**
1	$0 \leq x < 1$	1	17	17
2	$1 \leq x < 2$	2	0	0
3	$2 \leq x < 3$	3	1	3
4	$3 \leq x < 4$	4	2	8
			total	28

Table 3.5: Third Mapping approach: the upper limit for the n^{th} *difference* range was mapped to n and the *increase* value was n. x refers to the *difference* column in Table 3.3.

I evaluated all three mapping based on my own judgement, considering pitch range and how large and small differences were represented. The third approach in Table 3.5 worked best. The resultant mapping from amino acids to MIDI notes is detailed in the *MIDI* column of Table 3.3. It is worth noting here that this approach allocates arginine as the most hydrophillic, despite it also having the greatest hydrophobic effect of any amino acid. Arginine has been suggested to be an anomaly in the genetic code, which somewhat explains these peculiar properties[8]. The effect of this peculiarity is something I wish to investigate in my evaluation.

I also needed to deal with gaps in the protein. To do this, I created a second *list* data structure which would handle gaps by playing a different sound. Any gap characters were parsed into a second list which would play simultaneously, using the *noise* synthesizer with an attack of 0.5, release of 0.2, cutoff of 95 and an amplitude of 0.5 This created a sound that clearly different to the sine synthesizer of the amino acids. Therefore, gaps sound distinct from amino acids which serves to avoid breaking the metaphor of mapping hydrophobicity to pitch. In the *Perl* script I used a hash to map the amino acids to notes using the mapping detailed in Table 3.3. I used regular expressions to parse the Fasta format file and used the *chomp* function to remove spaces. I then indexed through each residue in the protein, using the hash to assign a MIDI pitch value to the acids list, or to the gap list. The two lists were printed before the code finished by printing the *Sonic Pi* play instructions detailed in Algorithm 1. Pseu-

Algorithm 2 Protein: Hydrophobicity Scale algorithm

Require: Fasta format file input, beat duration and line length

Create sound mapping as a hash
Read file
$note\_count \leftarrow 0$, $line\_count \leftarrow 0$, $Gap\_output \leftarrow blank$, $Acid\_output \leftarrow blank$
while letters in file **do**
    $line\_count + 1$
    remove all spaces from file
    if "$\wedge >$" is in line **then**
        Print *Sonic Pi* open list statement
    else
        create array by splitting residues
        for each residue in array **do**
            capitalise residue
            append Gap output and Acid output with "," **if** $note\_count > 0$
            append Gap and Acid outputs with newline to match line length
            print warning and replace with gap **if** incorrect character present
            if residue is a gap **then**
                append Gap output with sound mapping of residue
                append Acid output with rest
            else
                append Acid output with sound mapping of residue
                append Gap output with rest
            end if
            $note\_count + 1$
        end for
    end if
end while
print $Acid\_output$, Gap_output, and *Sonic Pi* play instructions

docode for the algorithm is given in Algorithm 2. Sound examples of this approach can be found in the supplemetary materials (hydroPro_ESX1.wav, hydroPro_ANKR1.wav, and Task_1_Protein.wav).

3.6 Protein: Reduced Alphabet

A major complexity of protein data sets concerns the physico-chemical similarities and differences of the amino acids, and the implications that this has for substitution rates and the forming of protein structures. Measuring this similarity can be done by considering their molecular attributes, their role in proteins, or some combination of both. A popular approach is to represent these relationships by dividing the 20 amino acids into groups, thus creating a simplified amino acid alphabet or *reduced alphabet*[50][38][42]. This approach was taken previously by King and Angus[25]. The first reduced alphabet that I attempted to use was a 5 letter alphabet derived from a genetic algorithm[27]. However, after some initial testing, I found that this alphabet lacked both a molecular and a protein-role metaphor, which made interpretation difficult. I changed to using the reduced alphabet seen in Table 3.6, which has a molecular metaphor of hydrophobicity ranking[49].

I had originally attempted to use different instruments for the reduced alphabets to aid in discernibility. I conducted ongoing and informal feedback, by reviewing examples myself and sharing them with other students in order to carry out lightweight evaluation of new ideas. From this process, I found that people found it harder to follow the *flow* of the sound with the quick changing of instruments. I also tried using audio panning to make the sounds from each group appear from a different *direction* to the listener, however this made it harder to follow the *flow* of the sonification. I therefore decided to only use pitch variety to represent the reduced alphabet. In implementation, I created arrays for each of the reduced letters, and then queried membership of each group as I indexed through the protein sequence. I mapped each sound to a note within a C pentatonic scale, to give a gap between notes of more than one semitone. The output was of five separate lists, each to be played simultaneously. One list represented one reduced 'letter' in the alphabet, and the fifth was for gaps or unknown symbols and used the same gap noise as in the previous implementation. The pseudocode for this is given in Algorithm 3. An example of the sound produced by this algorithm is given in the supplementary materials (onlyRedu_SCTR.wav).

Amino Acids	MIDI Pitch in Section 3.6	Synth. in Section 3.7
FILVWY	67	piano
ACGMP	64	sine
KQST	62	pluck
DEHNR	60	tb303

Table 3.6: Reduced Alphabet reproduced from [49]. Ordered from most hydrophobic to most hydrophillic.

Algorithm 3 Protein: Reduced Alphabet algorithm

Require: Fasta format file input, beat duration and line length

Create 5 arrays containing acids for reduced alphabet (including gap array)
Set pitch value for each array
Read file
$note\_count \leftarrow 0$, $line\_count \leftarrow 0$
create 5 strings, one for each letter in reduced alphabet
while letters in file **do**
    $line\_count + 1$
    remove all spaces from file
    if "$\wedge >$" is in line **then**
        Print *Sonic Pi* open list statement
    else
        create array by splitting residues
        for each residue in array **do**
            capitalise residue
            append all strings with "," **if** $note\_count > 0$
            append strings with newline to match line length
            print warning and replace with gap **if** incorrect character present
            for i from 1 to 5 **do**
                if residue is in $array_i$ **then**
                    append $string_i$ with pitch value for $array_i$
                else
                    append all other strings with a rest
                end if
            end for
            $note\_count + 1$
        end for
    end if
end while
print strings and *Sonic Pi* play instructions

3.7 Protein: Hydrophobicity and Reduced Alphabet

I decided to combine the approaches for the sonification of a single protein sequence in Sections 3.5 and 3.6. As both involved mapping to pitch, I had to change one of the mappings. I decided to keep the hydrophobicity to pitch mapping from Section 3.5, and alter the mapping of the reduced alphabet by using different instruments to represent the reduced alphabet. Table 3.6 demonstrates the mapping that I used. I selected the instrument such that they had distinct *timbres* and there would be no ambiguity between them. I had to alter the *amp* (amplitude) parameter in *Sonic Pi* to ensure that the sounds were perceived as of similar importance. I had to do this based on my own subjective psychoacoustic judgement, although I am well aware that this may not be universal.

This algorithm is similar to Algorithm 3 detailed in Section 3.6. The key changes are: a hash must be created for the sound mapping as in line 1 of Algorithm 2; when each string is appended in line 20, instead of the reduced alphabet pitch value, that must be the mapping value for the residue derived from the hash; and finally the play instructions will include the different instruments detailed in Table 3.6. The pseudocode is given in the appendix due to its similarity to Algorithms 2 and 3, it is under Algorithm 6 in Appendix A.3. An audio example of a sonification produced with this method is given in the supplementary materials (redu_SCTR.wav).

3.8 Multiple Sequence Alignment: Entropy Approach

I was interested in creating an approach that would take a multiple sequence alignment and output a single thread of sound from which researchers could tell conversation of the proteins. This sonification approach follows the work of Bywater and the PROMUSE software in sonifying the higher level information of protein alignments[11][17]. To do this I needed to use a measure of variety at each column within the multiple sequence alignment. I pursued the idea of using substitution matrices such as PAM and BLOSUM to score the differences between the positions[21]. However, I thought that the idea of entropy was more immediately available to researchers rather than introducing a new idea of *variability score* for each column. To meet this issue I decided to use Shannon entropy as a measure of variety in each column[40]. Shannon entropy for the i-th column, H_i, is defined as:

$$H_i = -\sum_j p_{ij} log_2(p_{ij})$$

where $j \in \{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V,-\}$ and $p_{ij} = \frac{\tilde{j}_i}{n}$, where $\tilde{j}_i$ is the count of j in the i-th column and n is the number of proteins in the MSA.

My new script would calculate H_i for each column i in the multiple sequence alignment. To be able to make them audible, my set of entropy values H needed to be transformed to audible MIDI notes. I did this by using z-score standardisation on the set of column entropy values, which I then scaled to a listenable range and took the floor value to give an integer MIDI number:

$$H^* = \left\lfloor \frac{H_i - \bar{H}}{\sigma_H} \cdot 10 + 60 \right\rfloor$$

where H^* is the transformed data set, H_i is the datapoint being transformed, $\bar{H}$ is the mean of H, and σ_H is the standard deviation of H The 10 corresponds to the spread which I wanted, to maintain my notes within a diverse but listenable range. The 60 represents the mean MIDI pitch I wanted my values to be centred around, which is middle C. These choices are based on my own preferences when listening to the sonifications. They could theoretically be any values which map the set of data H to integers between 0 and 127.

There was a problem of how to appropriately deal with gaps within my entropy sonification. My current approach deals with all gaps in a column as a single gap state. This means that a column containing a single arginine residue and 9 gaps would have the same entropy as a column with a single arginine residue and 9 leucine residues. This seems counter-intuitive, as a large number of gaps seems to have much more variety than more instances of the same residue. My idea to solve this was that each gap could be considered a separate state, thus changing the number of possible states and making for a more interpretable measurement of entropy. However, the problem with this is that it could cause trouble for the standardisation process when using the script for large inputs as the parameters are automated and MIDI notes are limited to values 1 to 127. It could cause very large variations in sound caused by wide ranging values for entropy due to variety of gaps, but leave the important information of alignment lost in the maelstrom. I decided to maintain my original and naive approach to dealing with gaps, where they all represent the same 'gap' state within the entropy calculation, however this is an area for further research. Another area for future research is how the approach deals with extreme examples with very diverse sequences or outlying

examples. The issue of dealing with gaps is a significant one for both of my MSA sonifications. Gaps are necessary for understanding MSAs but constantly 'sounding' them could distract from the information that is present. It is an issue of sonifying absent data, one which is very tricky. Maybe with the development of a tool, whereby users can turn on/turn off sounds associated with gaps could be a solution.

To perform my z-test, I needed to calculate the mean and standard deviation of my entropy values. I did this by using the Statistics::Descriptive packages *mean()* and *standard_deviation()* functions[3]. To quantize my notes, I needed to truncate the values given by my z-test standardisation process. I did thing using the *floor()* function from the POSIX package[1]. To parse the MSA in Fasta format, I used *BioPerl SeqIO* functions *new* to read the file and *next_seq* to index through the sequences. I used *BioPerl Seq* function *length* to identify the length of the MSA[41]. The pseudocode for this is detailed in Algorithm 4. An example of this approach can be found in the supplementary materials (Entropy_GCPR.wav).

3.9 MSA: Hydrophobicity Scale

I wanted to create a sonification of an MSA that would preserve the largest amount of information into the sound file. The mappings here are the same used in Section 3.5. Each protein in the MSA is sonified simultaneously using the hydrophobicity mapping in Table 3.3. However, as all proteins or strands of the MSA are being sonified simultaneously, the result is much more cacophonous.

I attempted to use different instruments for this purpose, but it was difficult to distinguish different instruments playing the same note. Instead I chose to use a single instrument, but use volume to communicate consensus between proteins in the MSA. I have chosen to use the superposition feature *Sonic Pi* to modulate the volume. Where two sounds of the same pitch and the same synthesizer are played simultaneously the sound is amplified, and this conveys consensus within the MSA. So a loud single note sounding represents a consensus, as opposed to a quiet single note which conveys gaps in the other strands. The choice of loudness to represent consensus instead of quietness seemed a natural audio metaphor to myself, however some psychoacoustic research has suggested that the polarity of sound features has little bearing on magnitude estimation[15].

In using this approach I discovered a problem with the *Sonic Pi* software. It seemed that, depending on the length of the alignment, I would sometimes hit upon some size

Algorithm 4 MSA: Entropy Sonification algorithm

Require: Fasta format file input, beat duration and line length

Create array 'acids' of all acid letters
Load file
$note\_count \leftarrow 0$, $strand\_count \leftarrow 0$ ▷ create matrix of MSA inputs
create 2-dimensional array MSAarray
while sequences in MSA **do**
    $note\_count \leftarrow 0$
    $len \leftarrow lengthofsequence$
    create array by splitting residues
    for each residue in array **do**
        insert residue into MSAarray[$strand\_count$][$note\_count$]
        $note\_count + 1$
    end for
    $strand\_count + 1$
end while
create $output\_array$ ▷ calculate entropy
for i from 0 to $len - 1$ **do**
    $Entropy \leftarrow 0$
    $string \leftarrow$ ""
    for j from 0 to $strand\_count - 1$ **do**
        $string \leftarrow string.MSAarray[j][i]$
    end for
    for each k in acids **do**
        $count_k \leftarrow$ count k in string
        if $count\_k > 0$ **then**
            $Entropy \leftarrow Entropy - \frac{count_k}{strand\_count} log_2(\frac{count_k}{strand\_count})$
        end if
    end for
    Push $output\_array$, $Entropy$
end for
for each l in $output\_array$ **do** ▷ Standardise and map to MIDI values
    subtract mean, divide by st. dev, multiple by 10, add 60, take the floor value
end for
print $output\_array$ and *Sonic Pi* play instructions

limit for *Sonic Pi* lists when multiple were played simultaneously. Obvious this is a problem for the scalability of the technology, and suggests that further research should be done with different software. However, as the focus of the current research is small MSAs, and due to time constraints, this research shall continue using the *Sonic Pi* software. The implementation used *BioPerl Seq* and *SeqIO* in the same way as in Section 3.8 to parse the MSA. Then each protein was mapped to the hydrophobicity scale using a hash, as in Section 3.5. Gap characters did not sound, although that is an issue that requires further research. The pseudocode for this algorithm is detailed in Algorithm 5. Sound examples generated using this process can be found in the supplementary materials (Multi_WD40.wav and Task_2_MultipleSequenceAlignment.wav).

Algorithm 5 MSA: Hydrophobicity algorithm

Require: Fasta format file input, beat duration and line length

```
Create sound mapping
Read file
strand_count ← 0, output ← ""
while sequences in MSA do
    strand_count + 1
    note_count ← 0
    len ← length of sequence
    create array by splitting residues
    for each residue in array do
        append output with "," if note_count > 0
        append output with newline to match line length
        append output sound mapping of residue
        note_count + 1
    end for
    Print output with Sonic Pi play instructions
    output ← ""
end while
for i from 1 to strand_count do
    print Sonic Pi play instructions for strand_i
end for
```

Chapter 4

Evaluation

4.1 Introduction

The evaluation of these methods will consist of qualitative research using bioinformatics researchers. It was be comprised of a questionnaire focusing on two of my sonification methods, one protein based and one MSA based, incorporating the NASA Task Flow Index workload evaluation, and a focus group discussing all five sonification methods. This project has ethics approval from the School of Informatics Research Ethics Process at the University of Edinburgh, with application reference number is 2019/29456. This gave us permission to hold and audio record a focus group and also distribute a questionnaire. The Consent Form is reproduced in Appendix B.2 and the Participant Information Sheet is reproduced in Appendix B.3.

4.2 Methods

4.2.1 Questionnaire Design

I designed two tasks for my participants to complete, one concerning the sonification of a single protein sequence using the hydrophobicity scale approach detailed in Section 3.5, and another concerning multiple sequence alignment sonification using the hydrophobicity scale approach detailed in Section 3.9. I chose to evaluate only these two approaches to keep the cognitive workload for participants low in order to ensure quality responses.

In the first task the participants were given a hydrophobicity based sonification of a major Human Prion Protein (https://www.uniprot.org/uniprot/P04156), created

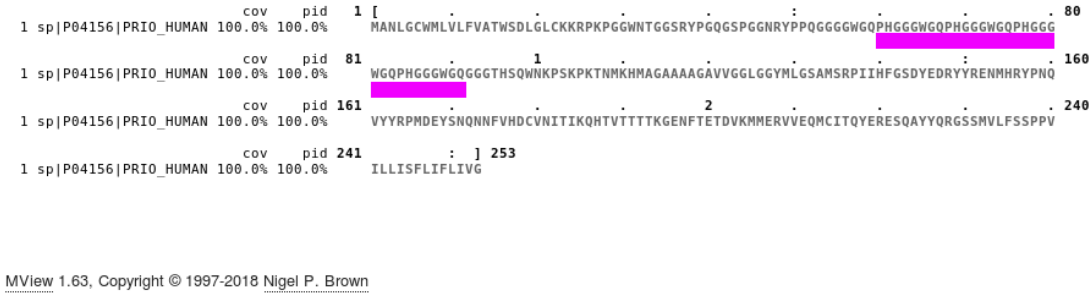


Figure 4.1: PRIO_HUMAN protein from UniProt, adapted from MView with the addition of highlighted magenta region[30].

using the algorithm described in Section 3.5. This is titled Task_1_Protein.wav and is included in the supplementary material. This protein features an 8 letter word (PHGGGWGQ) repeated 4 times in tandem from position 60 to position 92 which can be seen in Figure 4.1[28]. The (HGGGW) segment is a copper binding site, where copper ions can bind to the protein[13]. Participants were informed that the protein contained "a short (<20 letters) amino acid motif, or word, repeated four times". They were also supplied with a visualisation of the protein from MView, similar to Figure 4.1 although without the magenta highlighted region. They were then asked to enter their guess at the repeated motif. Participants were then shown Figure 4.1, with the repeated motif highlighted on the MView visualisation. They were then asked three questions:

- "Did the sonification sound file help you identify the repeated motif?"-Yes or No response.
- "What was the best thing about the protein sonification?"-Free text response.
- "What was the worst thing about the protein sonification?"-Free text response.

This first task had several purposes. Firstly, I wanted a task that demonstrated whether the sonification of a protein sequence conveyed information successfully, as per my first hypothesis. I also wanted this task to be easy for participants, but of a format that is computationally difficult. As mentioned in Section 1.5, identifying amino acid repeats is computationally difficult. Secondly, I wanted a task that was achievable by my participants. This was to engage them, give them a chance to understand the viability of the methods, and to encourage them to put more effort into the more challenging second task. I also wanted direct feedback on the approach to improve it during further work on the project.

The second task followed a similar structure, but concerned a multiple sequence

alignment. In creating the task for this, I created a MSA using 5 proteins which all contained two examples the SH3_1 conserved domain and one example of the SH2 conserved domain. These can be seen in Figure B.1 in Appendix B.5. The SH3_1 domain indicates a "beta-barrel fold that consists of five or six -strands arranged as two tightly packed anti-parallel sheets" and it is "ancient fold found in eukaryotes as well as prokaryotes"(http://pfam.xfam.org/family/PF00018). The SH2 domain "contains 2 alpha helices and 7 beta strands"(http://pfam.xfam.org/family/PF00017). These 3-dimensional structures can be seen in Appendix B.6. The sound file for the sonification is titled Task_2_MultipleSequenceAlignment.wav and is included in the supplementary material.

Participants were told that "this MSA contains 3 conserved domain (<50 letters)", and were asked to try and identify them. They were supplied with the visualisation in Figure B.1 to assist them, though the highlighted conserved domains were omitted. Participants were asked the same three questions:

- "Did the sonification sound file help you identify the repeated motif?"-Yes or No response.
- "What was the best thing about the protein sonification?"-Free text response.
- "What was the worst thing about the protein sonification?"-Free text response.

In creating my second task I was interested in creating a difficult task. I wanted participants to engage more than the relatively simple first task of identifying a repeated pattern. The difficulty of this task meant that I was not expecting success of my participants. However, I felt that engagement with the task would be higher with a difficult task, and this would give more honest and useful responses about the sonification methods. This choice in task corresponds to a difficulty in the qualitative evaluation of data exploration techniques. Much qualitative research is rooted in phenomenology as a philosophy[32]. "By phenomenology, Husserl (1913) meant the study of how people describe things and experience them through their senses. His most basic philosophical assumption was that *we can only know what we experience* by attending to perceptions and meanings that awaken our conscious awareness"[37]. There is therefore a sense that in evaluating a simple task, the response we get from participants is only based on valid experience of this simple task, and is therefore limited. There is a difficulty in creating artificial examples of real-world situations for phenomenological evaluation in a controlled method. This is why my choice of task for participants was difficult. I wished to move some way to simulating a more complex environment for phenomenological evaluation of the method. Future evaluations would involve deployment of the

technology in real-world scenarios, however that is beyond the scope of this project.

My call for participation in the online questionnaire was distributed via three mailing lists for bioinformatics researchers in Scotland. *Ashworth Bioinformatics Club* consists of staff and students of the University of Edinburgh based in the Ashworth buildings who work with biological sequence data. *Edinburgh Bioinformatics* is a mailing list of predominantly Edinburgh University staff who work with biological sequence data. *NextGenBUG*, which stands for the 'next generation sequencing bioinformatics users group', consists of professionals across Scotland whose work pertains to next generation sequencing of biological data sets. The call for participants can be found in Appendix B.1. My questionnaire drew 5 participants, 2 of which have experience with biological sequence data equating to PhD study (6-8 years) and 3 of which had experience beyond PhD level (>8 years). Two possessed musical experience equivalent to undergraduate study (2-4 years), two possessed musical experience equivalent to High School Study (0-2 years) and one had little or no experience (0 years).

4.2.2 NASA Task Load Index

To evaluate the subjective mental workload of the task, I made use of the NASA Task Load Index (TLX)[19]. It is a subjective, multi-dimensional assessment tool that assesses the perceived workload of a task. It is constructed of two parts. Firstly, the workload is divided into six sub-scales: frustration, effort, own performance, temporal demand, physical demand, and mental demand. The participant is asked to rate each of the six sub-scales on a twenty one point scale, which represents 0-100 divided into intervals of size 5. The second part of TLX creates individual weightings for the subscales then derived from pairwise comparisons of each of the factors by participants. The weighted score for each subscale is the number of times it is chosen as more relevant to workload of the task. The weighted score is then multiplied by the scale score as created in the first part, and then divided by 15 to give a workload score from 1 to 100. The choice of this phenomenological evaluation method came after surveying the literature of methods of evaluation. Evaluating novel methods for data exploration is difficult as insight is hard to quantify. The TLX is simple to implement, cost-effective, and has 30 years of research and applications supporting it.

4.2.3 Focus Group

The main difference between focus groups and other qualitative research setting is that the data collection is facilitated by a group setting[43]. A focus group is an interview with a group who have knowledge of a topic, and the interaction between participants creates data not accessible through individual interviews. Participants share their views, hear the views of others, and refine their view in light of what they hear[22]. This will provide a different approach to the purely phenomenological approach taken in the questionnaire. "Focus groups work best for topics people could talk about to each other in their everyday lives-but do not"[29]. I feel that a novel mode of data exploration fits this criteria well, as is is commonly interesting to practitioners of bioinformatics, but also diversionary from their everyday work.

My focus group had 5 attendees, plus myself as moderator. I used purposeful sampling to select attendees with an appropriate knowledge of protein sequences and for whom attendance would be possible, alongside a more general call for participation via email reproduced in Appendix B.1. The focus group had four participants, all of who are engaged in active bioinformatics research.In order to minimise the task of moderator preparation by avoiding educating someone to be an 'expert' in the methodology, I took the role myself. I familiarised myself with the group processes and moderator roles from appropriate literature prior to the event[43][22][32]. I showed my focus group six examples of sound files during the session: the first two corresponded to Section 3.5, the third to Section 3.6, the fourth to Section 3.8, the fifth to Section 3.9, and the sixth to Section 3.7. Each of my sonification algorithms was represented. Participants were each given a printout of MView visualisations of the proteins/MSAs under consideration[30]. The printout is included in Appendix B.4. The sound files played during the session will accompany the submission of this work.

The focus group was sound recorded completely, and I transcribed the audio using techniques from my research. I decided against using transcription software as I had participants from a range of linguistic background and wanted to develop familiarity with my data. I am using the *scissor-and-sort* technique of *content analysis* to analyse my focus group data as it is quick, cost-effective, and efficient[43]. There are risks of subjectivity and potential bias in this approach as it relies heavily on my own judgement, although these are present in many more time-consuming and sophisticated approaches and I will be mindful of them during the analysis process[43].

4.3 Results

4.3.1 Questionnaire Results

For the first task, all of my respondents were able to find the repeated motif. All of my respondents answered 'yes' to the question "Did the sonification help you identify the repeated motif?" This provides evidence that the first objective for my project has been met.

"What was the best thing about the protein sonification?" Responses:

- "Tuneful"
- "Makes repeated patterns obvious"
- "Repeated pattern easy to hear"
- "The ability to hear 'riffs' created by motifs. That is, it is easier to notice repetitive sequence from repetitive sound than it is from eyeballing letters."
- "Made it easier to identify the repeat motif (though combining the sound file with corresponding position in the sequence in visual form would have made this even more straightforward)."

Although the first response was primarily aesthetic, the other 4 responses are all analytic and agree with my hypothesis that "The informative content of protein sequences can be conveyed through mapping amino acids to sounds". The fourth response states that this approach was easier than visual analysis. The fifth response make a recommendation for the improvement of the technology, that is to make it more complementary to visual analysis by integrating positional information into the software.

"What was the worst thing about the protein sonification?" Responses:

- "Short lived"
- "No way to navigate the sound file easily (e.g. by selecting what residue to start with). Would be nice if you could navigate to different parts of the audio file by selecting residue instead of time."
- "Hard to map the location of the repeated sound to the actual sequence"
- "It is difficult to know where you are within the sequence. I had to go back to the beginning of the track and count the number of bars until the motif was reached and then work out approximately what number amino acid this was equivalent to, which was quite awkward."
- "Took a while to listen to, compared with just looking at the sequence."

Again, the first response seems primarily aesthetic, which is beyond the scope of this research. It is worthwhile noting that this respondent gave aesthetic judgements

to all four questions. This feedback is not what I am seeking to receive as it has no bearing on my research hypotheses. However, it may have value for future directions of the research. Three of the responses concern knowing location within the sequence. Two of the responses refer to improving navigation within the sound file. The final response concerns the length of the sonification in a negative sense. For the second task, when asked the questions "Did the sonification help you identify the conserved domains in the multiple sequence alignment?", all respondents answered "No". That is, for this particular task, this sonification was not useful. This is not surprising, as the task was intentionally difficult. However, the feedback from the task is insightful in understanding the issues in the completion of this difficult task.

"What was the best thing about the multiple sequence alignment sonification?" Responses:

- "The single sequence parts"
- "Made gaps in the alignment obvious"
- "Complementary way of representing the data"
- "The relative harmonization did help me identify the third domain, and was noticeable."
- "It was broadly possible to identify more conserved regions by paying attention to the volume."

The third response is evidence of positive feedback for the project. Other than the aesthetic first response, the other three are positive comments on aspects of the use of the sonification for analysis and completion of the task. This is promising, as it shows that the "No" responses are not rejections of the method as worthless, but more likely insufficient for the task set. These three responses, the second, fourth, and fifth, go someway to answering the second hypothesis of the project: Sonifying small multiple sequence alignments can provide researchers with useful complementary information to current visualisation methods for data exploration purposes.

"What was the worst thing about the multiple sequence alignment sonification?" Responses:

- "Very unpleasant to listen to"
- "Hard to hear conserved areas"
- "It is hard translating the sounds to the sequence"
- "It was difficult to keep track of my place in the sequence. Clashes between notes were very noticeable and off-putting and perhaps incorrectly made me discard two of the three true domains."

- "It was not readily possible (at least for me) to identify clear patterns in the sound beyond just hearing that certain areas were louder and quieter, equivalent to information I could easily have got just by looking at the sequence alignment (i.e. looking for conserved, non-gapped regions)..."

The difficulty of the task is prevalent in three of the responses here. The second response is a terse statement of the difficulty of identifying conserved ares. This is evidence that although I have moved someway towards my second hypothesis, it is not the case that this has been achieved. The third response also mirrors this sentiment. The fourth response contains recommendations from the first task feedback about location within the sequence. The fourth also contains some subjective psychoacoustic feedback, which is the only feedback from the survey concerning perception. Although I did ask participants for any other comments, the responses to this are not of relevance to this work. I will not reproduce responses to my questionnaire in full in the appendices as this violates the ethics agreement which I made with participants.

4.3.2 NASA Task Load Index Results

Factor	Weights	Ratings	Weighted Workload Score
Physical Demand	0.4	7	2.8
Mental Demand	3.8	75	285
Temporal Demand	2.4	32	76.8
Performance	2	74	144
Effort	3	60	180
Frustration	3.4	56	190.4
Mean	2.5	50.67	146.5
Standard Deviation	1.22	26.49	97.72
Tukey's Fences	-0.1 and 5.5	-31 and 137	-93.6 and 360.8

Table 4.1: NASA TLX weights and Ratings from evaluation. The weights must sum to 15, the ratings are out of 100.

The results of the TLX are detailed in Table 4.1. The weights from this evaluation identify the subscales the participants considered most important scale for this task: with Mental Demand, Frustration, and Effort all weighted higher than the mean. The ratings demonstrate which factors were deemed the most difficult: with Mental Demand, Performance, Effort, and Frustration all having ratings above the mean. The weighted workload score captures the workload contribution of the subscales. Mental Demand is clearly the largest contributor to workload, it is almost 1.5 times as large as

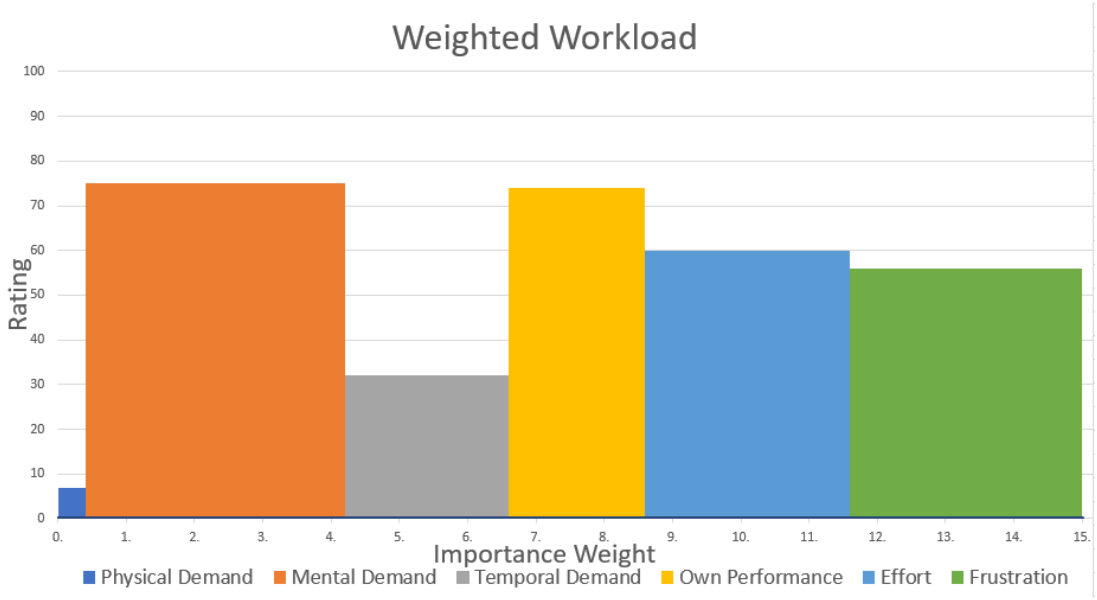


Figure 4.2: Graphic composition of weighted workload score. Values are given in Table 4.1. The width of the subscale bars represents the importance (weight) of each factor, the height represents its magnitude (rating), and the area represents the weighted workload score.

the second largest contributor, Frustration. Expanded results are reported in Table B.1 in Appendix B.7.

The mean rating was 51 out of 100. However, by taking the mean weighted workload score of 146.5 and dividing by the mean weight of 2.5, we get a score that represents the mean rating. In this method, the subscales with larger weightings have more of an impact on the score. In this case the value would be 58.6. This seems a more realistic overall workload score, one which plays down the contribution of the Physical demand and takes the weights into consideration. These scores of 51 and 58.6 are out of 100 and measure the overall difficulty of the task, which means we can have an understanding of how difficult the participants found the use of the sonification for this task. However, a score of over 50 is interpretable as a difficult task, especially as our participants are research scientists in the field. And though there has not been success in developing a *redline* for the overall workload in the NASA TLX (a point whereby any score above this means a workload is too high), a score of 51 or 58.6 out of 100 would seem to be too low to be considered a workload that is too high[18].

Many researchers do not use the second part of the TLX and report unweighted responses, this is referred to as 'Raw TLX'[18]. There is evidence that 'Raw TLX' may increase experimental validity, and allow for easier comparison between studies[10].

My 'Raw TLX' scores are reported in Table 4.1. A popular way to understand the results of the NASA TLX is to consider the component subscales separately[18]. From the raw scores we can see the relative difficulty of each subscale on the task. What will be interesting is using these results in comparison with future work on the project, or compared with similar sonification tasks. Certain subscales of the TLX may be discarded for certain tasks[18]. This may be appropriate for future evaluations of these methods, especially as Physical Demand seems inappropriate for a PC-based analysis task. I conducted analysis on the data in Table 4.1 using Tukey's Fences, a common method of outlier detection, to see whether the data also suggested that the Physical Demand subscale was an outlier:

Lower Tukey's Fence$= Q1 - 1.5 \cdot IQR$

Upper Tukey's Fence$= Q2 + 1.5 \cdot IQR$

where IQR is the inter-quartile range, $Q1$ is the first quartile, and $Q3$ is the third quartile of the data. Using this analysis, the values for Physical Demand are not outliers, so we would be unjustified in removing them. The Physical Demand results also fall within 2 standard deviations of the mean, which is not good evidence for them being outliers. There is also a rationalisation for the inclusion of Physical Demand. Although the Physical Demand of the task may be small, it can refer to the physical skill of listening and controlling the sonification and also of interactions with speakers or headphones, and should therefore be maintained in analysis.

4.3.3 Focus Group Results

There are four stages to performing *scissor-and-sort content analysis*: First you must determine which parts of the transcript are important. You must then develop a categorisation system for the topics discussed. Next you select representative statements regarding the topics, and finally you develop an interpretation of what it means. My categorisation system for *content analysis* consists of 5 categories: *aesthetic judgements*, feedback concerning the beauty and feeling of the sonifications; *project judgements*, feedback on the project as a whole, and on the new mode of data representation for protein sequence data; *analytic judgements*, these represent feedback on the information carrying aspects of the sonification; *psychoacoustic judgements*, these represent the subjective psychoacoustic responses of participants as to how they experienced the sound; and *suggestions for improvements or future work*. The representative statements regarding each of these topics are reproduced in Appendix B.8.

4.3.3.1 Analytic Judgements

The analytic judgements of my focus group participants were my main priority in moderating the discussion, the representative statements from my participants are reproduced in Table B.4 in Appendix B.8. These responses were quite specific to the different implementations that were played to them, so I will write a paragraphs specific to each one.

I played two sonifications of single proteins using the hydrophobicity scale approach detailed in Section 3.5. Participants reaction to the first sonification (hydroPro_ANKR1.wav) were concerned with the speed of the sonification, with opposing opinions on whether they would rather the sonification be faster or slower. The argument for faster concerned the limits of attention and the time spent listening. The argument for slower was to give more time for analysis and understanding the sound. This gives more evidence that customisable and interactive sound parameters will be an important future direction for the work, as the different perceptions and purposes of the technology are important to different users.

After playing the second sonification (hydroPro_ESX1.wav) I was expecting reactions more specific to the clear AAR present in the protein. The motif (PPxxPxPPx), where the x can refer to any amino acid, is repeated nine times in tandem in the protein. All participants agreed that they could "really hear it this time", and they engaged in some of what sociologist of sonification Alexandra Supper calls *sonification karaoke* by singing what they heard to be relevant[44]. This conveys the enthusiasm of the participants for the method, and excitement at hearing the motif. This provides evidence for my first hypothesis. The other analytic feedback I received in response to this sonification was about a participant struggling to pay attention and focus in detail, they then said "if I had more of a purpose then I would have focused more or gone back to re-listen". Although this speaks about the artificial nature of the focus group environment, it also provides more evidence for the want for control of the playing of the sonifications.

The third sonification (onlyRedu_SCTR.wav) is then played for the participants. This sonification is detailed in Section 3.6, and uses a reduced alphabet to map amino acids to four pitches. This is a sonification of a seven trans-membrane protein, which has seven alternating regions of hydrophobicity and hydrophillicity as the protein weaves in and out of a cell. I had hoped that this would be easy to hear with the reduced alphabet with lots of repeated HIGH notes. However participants agreed that

"it was not easy to hear the motif change". Participants also noticed a regular high note sounding throughout the sonification. This is the arginine residue, which is both the most hydrophobic and hydrophillic residue. This caused disruption when trying to identify regions of hydrophobicity. I had anticipated that this may cause issues and this need further work. Participants also wanted a greater variety in pitch and felt that this would make it easier to understand. In the first sonification, the pitch range is 27 semitones. In this sonification, that range is 7 semitones.

The fourth sonification (Entropy_GCPR.wav) was detailed in Section 3.8. Participants were in agreement that it was "easy to discern between the highly conserved and not highly conserved regions". This is evidence towards my second hypothesis. However, participants also agreed that it was not easy to understand the sound in between the higher and lower ends, which represented the most and least conserved areas.

The fifth sonification (Multi_WD40.wav) was detailed in Section 3.9. Participants found it less clear than the others. They agreed that using the sonification as an only source of information was difficult, and that they could not tell what they were listening to without a location indicator on a visual accompaniment. This provides evidence for more sophisticated complementary approach to visual representation than the static printouts used in this process. Participants stated that they could use this to give "a flavour of what the [MSA] is like" or as "an approximation just to get an initial idea" of the alignment, which corresponds to my main application of the software for initial data exploration and provides some evidence for my second hypothesis. One participant said that this was easier to pay attention to this sonification as it was more challenging.

The sixth sonification (redu_SCTR.wav) is detailed in Section 3.7 and uses instruments to represent the reduced alphabet mapping, while pitches still map to the hydrophobicity scale. Participants all agreed that the different instruments communicated the different reduced alphabet groups very clearly, however hearing the difference in pitch within the instruments was much harder. Generally the insight that arose from discussion was that using the reduced alphabet was meant to simplify the sound for the listener, but by including all the separate hydrophobicity pitches, that complication is still maintained. This seems more evidence to me that the different instruments for reduced alphabet is a worthwhile approach, however doing it simultaneously is a lot for listeners to take in. The ability for users to control the mappings seems a logical solution to this.

4.3.3.2 Project Judgements

Participants naturally gravitated to giving feedback on the project and idea of sonification in bioinformatics as a whole rather than the individual implementations. This was probably due to using myself as a moderator for the discussion rather than someone less involved with the project. Despite this not being the primary aim of the focus group, these responses were informative and the main representative statements are reproduced in Table B.3 in Appendix B.8. The first piece of project feedback was very positive, stating "I think that with your project you need to think not about whether it is possible, because you've proved that it is, but can you compete against what is used nowadays". This seems firm evidence towards my first hypothesis.

Participants discussed the utility of the approach for visually impaired scientists and were very positive about the prospect. This would be an interesting avenue for future research with evaluation processes conducted with visually impaired scientists. Participants often remarked that a particular approach, such as the reduced alphabet was a good idea, but might not be useful in every circumstance. This backs up that customisability is a good direction of future research. I received feedback on how the software could be used, with participants identifying those with long MSAs or lots of MSAs who "just want to know if there is a conserved domain" or as an initial "way of filtering" their data. They also suggested that development should start by concentrating on a "broad" approach to find conserved regions and then get more specific as more people are using the software with "different ideas or different needs".

4.3.3.3 Psychoacoustic Judgements

The subjective psychoacoustic responses of the participants was something that the focus group brought up independently of questioning, and the importance of this to the project became clear through the responses, which can be read in full in Table B.6 in Appendix B.8. There was much disagreement between the participants, who often would characterise the sounds in polar opposite terms. Broadly, the metaphor of high pitch corresponding to high hydrophillicity was easy for the participants to grasp, the same with the entropy to pitch metaphor, however, the metaphor of different instrument corresponding to a reduced alphabet letter (and implicitly hydrophobicity) they considered difficult. Participants revealed more general insights into how they consider their audio perception, saying that they can "recognise what is present" in the moment through sound, but cannot remember it after a short time. This is more

evidence for a complementary system of visual and audio parts.

4.3.3.4 Improvements Judgements

The participants were keen to contextualise the methods into their standard work flows and gave clear ideas for future work on the project. The representative statements can be found in Table B.7 in Appendix B.8. To improve the entropy sonification, participants suggested using categories representing the proportion of proteins that were conserved across the MSA, which would limit the variety of notes and remove a lot of the variety in the middle pitches which participants found difficult to interpret. They then raised concerns about the loss of information in this approach, but to me this is more evidence for a customisable interface for end-users where they could choose whether they wanted the categories or not. It seems that researchers want to ask very diverse questions of their data, and they want to be able to manipulate the tools they use to ask these questions in a variety of ways. During the focus group, participants said they would want to be able to control: the polarity of the sound, the speed of the sonification, the navigation of the sound, the categorisation of the sounds into step sizes, the instrumentation, and the use of reduced alphabets. Another repeated and key improvements from participants was a visual representation of where in the alignment or protein the sound was being generated from. One participant even suggested using "one of those balls like at karaoke". Despite the light tone, loaction information was something all participants wanted.

4.3.3.5 Aesthetic Judgements

The complete representative statements of the participants' aesthetic judgements can be found in Table B.5 in Appendix B.8. Generally, participants were positive in their aesthetic judgements of the sonification, describing it as "chaotic, but not completely chaotic" and "more diverse than expected". Occasionally participants drew direct comparison to music, specifically "scary movie" soundtracks, such as those of filmmaker John Carpenter, and a participant wisecracked about "easy-listening proteins". These responses are indicative of the enthusiasm the focus group showed for the sonifications. In response to the last two sonifications played, those which I had identified as the least musical, responses such as "the weirdest sound", "doesn't conform to the normal structure of music", and "like someone bashing at notes" captured participants initial responses, although all of these were followed by a more in depth and positive

response to the sound.

4.4 Discussion

There is a consensus of evidence between the different forms of evaluation that supports the claim that I have met my first hypothesis, agreeing that protein information can be conveyed sonically. There is contrasting evidence towards my second hypothesis. The evidence supports the claim somewhat, however questionnaire participants were unable to complete the MSA task using the sonification. Further work is needed to satisfactorily establish that MSA information can be conveyed reliably through sonifications and complementary visualisations. There are key improvements needed on the project to improve the conveyance of MSA information via the sonifications. Firstly, the complementary visualisations must become more sophisticated and must communicate location accurately. Secondly, there must be better customisability of the software for users. This includes control over the navigation of the sound, the speed of the sound, and the sonification parameters in use. A software package giving users control over these will meet many of the issues that came up in the evaluation.

The main utility of these methods seems to be in the initial and broad scale exploration of data, especially in seeking amino acid repeats (AARs) and conserved domains. Aiding visually impaired scientists seems a promising area of future research for this project. The results of the NASA TLX identify the most important scales for evaluation of this task as: Mental Demand, Performance, Effort, and Frustration. However, more work must be done to contextualise these results as these are difficult to contextualise without results to compare to for similar data exploration tasks. In response to my third hypothesis, I claim that I have received good feedback from my qualitative evaluation process. This has identified key successes and failures of the implementations, general feedback on the project, and key ideas on how to move forward with the research.

Chapter 5

Conclusions

The *Sonic Pi* software had several problems: scripts could become redundant if it is updated, and the program goes through quite regular update processes. There was a capacity issue when working with larger alignments. Also the incompatibility issue of *Sonic Pi* with *Scientific Linux*. These could be solved by using another sound synthesising language, such as SuperCollider[31]. This would also facilitate direct piping, which would remove the problems associated with extra software and improve usability. There remains a problem with the potential redundancy of *Perl 5*, however it is still a popular language in bioinformatics and its development is far less volatile than *Sonic Pi*.

Gaps in MSAs are hard to deal with. Sonifications can end up being very long and there is no way for users to navigate the sound easily or change the speed. Users have very different questions they wish to ask of data, and change their data exploration needs regularly. Listeners 'get lost' when listening to alignments, and struggle to maintain attention. The individual psychoacoustic experience of every individual is unique, and they do not experience sound in the same way. Future work should work towards the solution of these problems by developing a customisable and interactive software package for sonifying proteins and MSAs. This would allow the control of speed, easy navigation of the sonification, and user-controlling of all sound synthesis parameters, allowing them to create the right sonification to answer their current question. This must also include the development of dynamic complementary visualisation.

Future work would also involve evaluation of these methods with visually impaired scientists. Also the creation of tutorials to demonstrate different examples of how the software can be used. Surveys must also be conducted to identify need for this software.

Bibliography

[1] POSIX - metacpan.org.

[2] Scientific Linux.

[3] Statistics::Descriptive - Module of basic descriptive statistical functions. - metacpan.org.

[4] The Perl Programming Language - www.perl.org.

[5] WineHQ - Run Windows applications on Linux, BSD, Solaris and macOS.

[6] Sam Aaron. Sonic Pi Github Issues Thread - https://github.com/samaaron/sonic-pi/issues/735#issuecomment-147209392.

[7] Samuel Aaron and Alan F. Blackwell. From sonic Pi to overtone. In *Proceedings of the first ACM SIGPLAN workshop on Functional art, music, modeling & design - FARM '13*, page 35, New York, New York, USA, 2013. ACM Press.

[8] P Baldi, S Brunak, and F Bach. *Bioinformatics: the machine learning approach*. 2001.

[9] Daniel Barker, David EK Ferrier, Peter WH Holland, John BO Mitchell, Heleen Plaisier, Michael G Ritchie, and Steven D Smart. 4273π: Bioinformatics education on low cost ARM hardware. *BMC Bioinformatics*, 14(1):243, 12 2013.

[10] Ernesto A. Bustamante and Randall D. Spain. Measurement Invariance of the Nasa TLX. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52(19):1522–1526, 9 2008.

[11] Robert P Bywater and Jonathan N Middleton. Melody discrimination and protein fold classification. *Heliyon*, 2(10):e00175, 10 2016.

[12] Mary Anne Clarke and John Dunn. Life Music: The Sonification of Proteins. *Art and Biology*, 1999.

[13] Colin S. Burns, , § Eliah Aronoff-Spencer, , Christine M. Dunham, , Paula Lario, Nikolai I. Avdievich, Ł William E. Antholine, Marilyn M. Olmstead, Alice Vrielink, Gary J. Gerfen, Jack Peisach, William G. Scott, , and Glenn L. Millhauser*. Molecular Features of the Copper Binding Sites in the Octarepeat Domain of the Prion Protein. 2002.

[14] D M Engelman, T A Steitz, and A Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual review of biophysics and biophysical chemistry*, 15(1):321–53, 6 1986.

[15] Jamie Ferguson and Stephen Brewster. EVALUATING THE MAGNITUDE ESTIMATION APPROACH FOR DESIGNING SONIFICATION MAPPING TOPOLOGIES. pages 23–27, 2019.

[16] Florian Grond and Jonathan Berger. Parameter mapping sonification. In *The Sonification Handbook*, chapter 15, pages 363–397. Berlin: Logos Publishing House, 2011.

[17] Marc D. Hansen, Erik Charp, Suresh Lodha, Doanna Meads, and Alex Pang. PROMUSE: A System For Multi-Media Data Presentation Of Protein Structural Aalignments. *Biocomputing '99*, pages 368–379, 12 1998.

[18] Sandra G. Hart. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 10 2006.

[19] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52:139–183, 1 1988.

[20] Kenshi Hayashi and Nobuo Munakata. Basically musical. *Nature*, 310(5973):96–96, 7 1984.

[21] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–9, 11 1992.

[22] Monique M. Hennink. *Focus group discussions*. 2014.

[23] Thomas Hermann, Andy Hunt, and John G. Neuhoff. *The Sonification Handbook*. Logos Verlag, Berlin, 2011.

[24] Douglas R. Hofstadter. An Eternal Golden Braid. In *Gödel, Escher, Bach*. Penguin Books, 1980.

[25] Ross D King and Colin G Angus. PM-Protein music. *Cabios Applications Note*, 12(3):251–252, 1996.

[26] Gregory Kramer, Bruce Walker, Terri Bonebright, Perry Cook, John Flowers, Nadine Miner, and John Neuhoff. Sonification Report: Status of the Field and Research Agenda. *Faculty Publications, Department of Psychology*, 3 1999.

[27] Jacek Lenckowski and Krzysztof Walczak. Simplifying Amino Acid Alphabets Using a Genetic Algorithm and Sequence Alignment. In *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics*, pages 122–131. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[28] H. Luo and H. Nijveen. Understanding and identifying amino acid repeats. *Briefings in Bioinformatics*, 15(4):582–591, 7 2014.

[29] P Macnaghten, G Myers Qualitative Research Practice: Concise, and Undefined 2006. Focus groups. *books.google.com*.

[30] Fabio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, and Rodrigo Lopez. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research*, 4 2019.

[31] James McCartney. Rethinking the Computer Music Language: SuperCollider. *Computer Music Journal*, 26(4):61–68, 12 2002.

[32] SB Merriam and EJ Tisdell. *Qualitative research: A guide to design and implementation*. 2015.

[33] G I Mihalas, Minodora Andor, Anca Tudor, and S Paralescu. Potential Use Of Sonification For Scientific Data Representation. 28:45–57, 2018.

[34] Nobuo Munakata and Kenshi Hayashi. Gene Music: Tonal Assignments of Bases and Amino Acids. In *Visualizing Biological Information*, pages 72–83. World Scientific, 12 1995.

[35] Sen I. O'Donoghue, Benedetta Frida Baldi, Susan J. Clark, Aaron E. Darling, James M. Hogan, Sandeep Kaur, Lena Maier-Hein, Davis J. McCarthy, William J. Moore, Esther Stenau, Jason R. Swedlow, Jenny Vuong, and James B. Procter. Visualization of Biomedical Data. *Annual Review of Biomedical Data Science*, 1(1):275–304, 7 2018.

[36] Susumu Ohno and Midori Ohno. The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition. *Immunogenetics*, 24(2):71–78, 8 1986.

[37] Michael Quinn Patton. *Qualitative research & evaluation methods : integrating theory and practice*. 2015.

[38] Eric L. Peterson, Jan Kondev, Julie A. Theriot, and Rob Phillips. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics*, 25(11):1356–1362, 6 2009.

[39] James B Procter, Julie Thompson, Ivica Letunic, Chris Creevey, Fabrice Jossinet, and Geoffrey J Barton. Visualization of multiple alignments, phylogenies and gene family evolution. *Nature Methods*, 7(3):S16–S25, 3 2010.

[40] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 7 1948.

[41] J. E. Stajich, David Block, Kris Boulez, Steven E Brenner, Stephen A Chervitz, Chris Dagdigian, Georg Fuellen, James G R Gilbert, Ian Korf, Hilmar Lapp, Heikki Lehväslaiho, Chad Matsalla, Chris J Mungall, Brian I Osborne, Matthew R Pocock, Peter Schattner, Martin Senger, Lincoln D Stein, Elia Stupka, Mark D Wilkinson, and Ewan Birney. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10):1611–1618, 10 2002.

[42] James D. Stephenson and Stephen J. Freeland. Unearthing the Root of Amino Acid Similarity. *Journal of Molecular Evolution*, 77(4):159–169, 10 2013.

[43] David W. Stewart and Prem N. Shamdasani. *Focus groups : theory and practice*. 2015.

[44] A Supper. *Lobbying for the ear: The public fascination with and academic legitimacy of the sonification of scientific data*. 2012.

[45] Rie Takahashi and Jeffrey H Miller. Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns. *Genome Biology*, 8(5):405, 5 2007.

[46] Mark D. Temple. An auditory display tool for DNA sequence analysis. *BMC Bioinformatics*, 18(1):221, 12 2017.

[47] Bruce N Walker and Michael A Nees. Theory of Sonification. In *The Sonification Handbook*, chapter 2. Logos Verlag, 2011.

[48] LUSHENG WANG and TAO JIANG. On the Complexity of Multiple Sequence Alignment. *Journal of Computational Biology*, 1(4):337–348, 1 1994.

[49] Edward A. Weathers, Michael E. Paulaitis, Thomas B. Woolf, and Jan H. Hoh. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Letters*, 576(3):348–352, 10 2004.

[50] William C. Wimley and Stephen H. White. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Structural & Molecular Biology*, 3(10):842–848, 10 1996.

[51] Guy Yachdav, Sebastian Wilzbach, Benedikt Rauscher, Robert Sheridan, Ian Sillitoe, James Procter, Suzanna E. Lewis, Burkhard Rost, and Tatyana Goldberg. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, 32(22):btw474, 7 2016.

Appendix A

First appendix

A.1 Abbreviations and Acronyms

	Meaning
AAR	Amino Acid Repeats
BLOSUM	Blocks Substitution Matrix
DNA	Deoxyribonucleic acid
Fasta	Fast-All
MIDI	Musical Instrument Digital Interface
MSA	Multiple Sequence Alignment
NASA	National Aeronautics and Space Administration
PAM	Point Accepted Mutation
PMSon	Parameter Mapping Sonification
RNA	Ribonucleic acid
Synth	Synthesizer
TLX	Task Load Index
WINE	Wine Is Not an Emulator

Table A.1

A.2 List of Sound Files in Supplementary Materials

Filename	**Description**
Entropy_GCPR.wav	Entropy based sonification of 5 protein MSA of 7 trans-membrane proteins using method detailed in Section 3.8
hydroPro_ESX1.wav	Hydrophobicity based sonification of HUMAN_ESX1 protein using method detailed in Section 3.5
hydroPro_ANKR1.wav	Hydrophobicity based sonification of HUMAN_ANKR1 protein using method detailed in Section 3.5
Multi_WD40.wav	Hydrophobicity based sonification of 5 protein MSA containing WD40 conserved domain using method detailed in Section 3.9
onlyRedu_SCTR.wav	Reduced Alphabet based sonification of HUMAN_SCTR protein using method detailed in Section 3.6.
redu_SCTR.wav	Reduced Alphabet and Hydrophobicity based sonification of HUMAN_SCTR protein using method detailed in Section 3.7.
Task_1_Protein.wav	Hydrophobicity based sonification of HUMAN_PRIO protein using method detailed in Section 3.5
Task_2_MultipleSequenceAlignment.wav	Hydrophobicity based sonification of 5 protein MSA containing SH3 and SH2 conserved domains using method detailed in Section 3.9

Table A.2

A.3 Protein: Hyrdrophbocity and Reduced Alphabet Algorithm

Algorithm 6 Protein: Hydrophobicity and Reduced Alphabet algorithm

Require: Fasta format file input, beat duration and line length

Create 5 arrays containing acids for reduced alphabet (including gap array)
Create sound mapping as a hash
Read file
$note\_count \leftarrow 0$, $line\_count \leftarrow 0$
create 5 strings, one for each letter in reduced alphabet
while letters in file **do**
    $line\_count + 1$
    remove all spaces from file
    if "$\wedge >$" is in line **then**
        Print *Sonic Pi* open list statement
    else
        create array by splitting residues
        for each residue in array **do**
            capitalise residue
            append all strings with "," **if** $note\_count > 0$
            append strings with newline to match line length
            print warning and replace with gap **if** incorrect character present
            for i from 1 to 5 **do**
                if residue is in $array_i$ **then**
                    append $string_i$ with sound mapping from hash for residue
                else
                    append all other strings with a rest
                end if
            end for
            $note\_count + 1$
        end for
    end if
end while
print strings and *Sonic Pi* play instructions with different instruments

Appendix B

Evaluation Materials

B.1 Questionnaire Call for Participation

Subject: **Sonifying Proteins: Can you hear Bio-information?**

Sonification is the use of non-speech audio to convey information or perceptualize data.

I am conducting research as part of my MSc Informatics project into the sonification of biological sequence data, specifically multiple sequence alignments. I have developed several sonification approaches and now I am at an evaluation stage. I'm looking for people experienced in the use of biological sequence data to complete two sonification tasks and an accompanying questionnaire to evaluate some of my approaches.

The entire process can be done on your own PC and should take less than 15 minutes, and you will get to listen to a novel sonification of a protein sequence and multiple sequence alignment. Speakers or headphones are required, and participants must be based in the UK.

To participate, please follow the link below:

https://edinburgh.onlinesurveys.ac.uk/sonifying-proteins

The questionnaire will be open for a week from now, and will close on Wednesday 24th July at 4pm BST. Feel free to share the link with appropriate colleagues.

I am also seeking participants for a focus group session tomorrow on Thursday 18th July from 10.30-11.30 in Ashworth 3 room 250. This will be a more in depth look at a wider range of my sonification approaches and the discussion will be recorded for feedback and analysis purposes. Please email me at E.j.martin@sms.ed.ac.uk if you are interested in participating.

Thank you for your support,

Best wishes,

Edward Martin School of Informatics Edinburgh University

B.2 Consent Form

Participant number:_____

Participant Consent Form

Project title:	The use of parameter-mapping sonification to facilitate knowledge discovery from protein multiple sequence alignments in bioinformatics
Principal investigator (PI):	Daniel Barker
Researcher:	Edward Martin
PI contact details:	Daniel.Barker@ed.ac.uk

Please tick yes or no for each of these statements.

		Yes	No
1.	I confirm that I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.		
2.	I understand that my participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.		
3.	I agree to being audio recorded. (Only applicable for focus group participants)		
4.	I consent to my anonymised data being used in academic publications and presentations.		
5.	I understand that my anonymised data can be stored for a minimum of two years		
6.	I allow my data to be used in future ethically approved research.		
7.	I agree to take part in this study.		

Name of person giving consent	Date	Signature

Name of person taking consent	Date	Signature
Edward Martin		

THE UNIVERSITY of EDINBURGH informatics

B.3 Participant Information Sheet

Participant Information Sheet

Project title:	The use of parameter-mapping sonification to facilitate knowledge discovery from protein multiple sequence alignments in bioinformatics
Principal investigator:	Daniel Barker
Researcher collecting data:	Edward Martin
Funder (if applicable):	N/A

This study was certified according to the Informatics Research Ethics Process, RT number **2019/29456.** Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The sole researcher is Edward Martin, an MSc student in the School of Informatics at University of Edinburgh. The supervisor for the project is Daniel Barker, a Reader in Bioinformatics in the School of Biological Sciences. The data may be accessed by other members of Daniel Barker's research group: Maria Mantas and Joseph Guscott, PhD students, and Stevie Bain, a postdoctoral researcher.

What is the purpose of the study?

Multiple sequence alignments are an essential tool for comparative analyses of protein sequences. However, their interpretation can be difficult for researchers, due to their size, complexity, and the limitations of visualisation software. Parameter-mapping sonification is a technique from the field of auditory display which uses sound for data exploration.

This project aims to develop novel parameter-mapping sonification software to help researchers better understand multiple sequence alignments by encoding the amino acids into sound. The project will also provide a model for the evaluation of bioinformatics data exploration software and exploratory data sonification.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or

withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

Participants will listen to sonification audio files presented with accompanying text. They will then complete a questionnaire, which will collect data on questions regarding the sonification. The questionnaire will also ask qualitative and quantitative questions about participants' opinions on the sonifications. This will take no more than 20 minutes.

A subset of participants will also be part of a focus group, which they will know in advance. The focus group discussion will follow the survey and will consist of further examples and questions from the researcher, Edward Martin. These questions will be qualitative and will regard the participants' opinions and level of understanding of the sonifications. This session will involve audio recording for the purposes of transcription and analysis. The questionnaire and focus group will together take no longer than 80 minutes.

Are there any risks associated with taking part?

There are no significant risks associated with participation.

Are there any benefits associated with taking part?

There will be no tangible benefits associated with participation.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will always be anonymous. With your consent, information can also be used for future research. Your data may be archived for a minimum of two years.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a

unique participant number rather than by name. Your data will only be viewed by the researcher/research team: Edward Martin, Daniel Barker, Maria Mantas, Joseph Guscott, and Stevie Bain.

All electronic data will be stored on a password-protected computer, on the School of Informatics' secure afs file servers, the University's DataStore Storage Area Network, or on the University's secure encrypted cloud storage services (ownCloud or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Edward Martin, E.j.martin@ed.ac.uk.

If you wish to make a complaint about the study, please contact:

Daniel Barker, Daniel.Barker@ed.ac.uk, and inf-ethics@inf.ed.ac.uk.

When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheets will be made available on https://4273pi.org

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Edward Martin, E.j.martin@sms.ed.ac.uk.

General information.

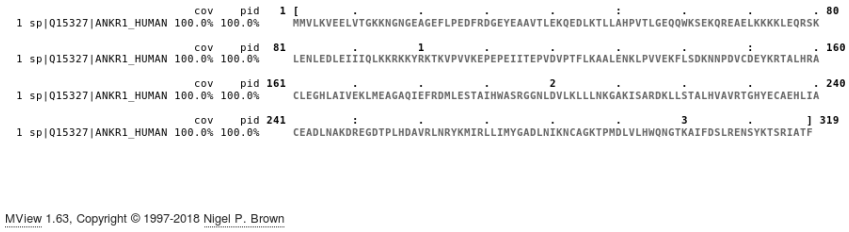
THE UNIVERSITY of EDINBURGH informatics

For general information about how we use your data, go to: edin.ac/privacy-research

THE UNIVERSITY *of* EDINBURGH
informatics

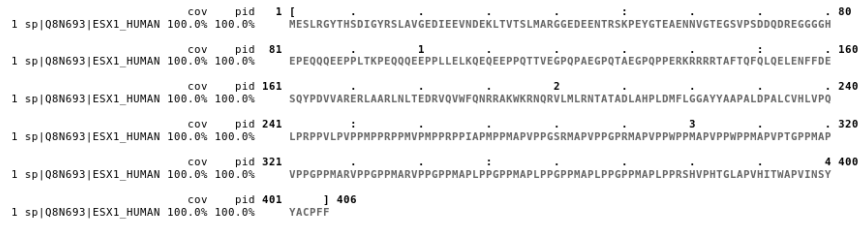
B.4 Focus Group Handout

ANKR1_HUMAN



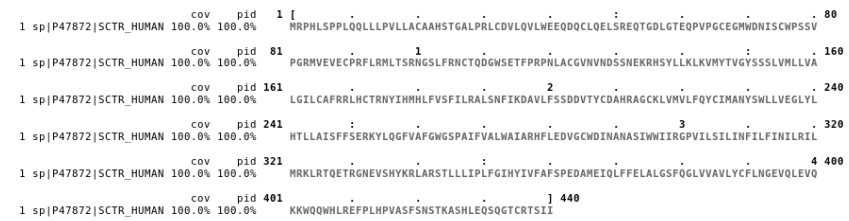
MView 1.63, Copyright © 1997-2018 Nigel P. Brown

ESX1_HUMAN

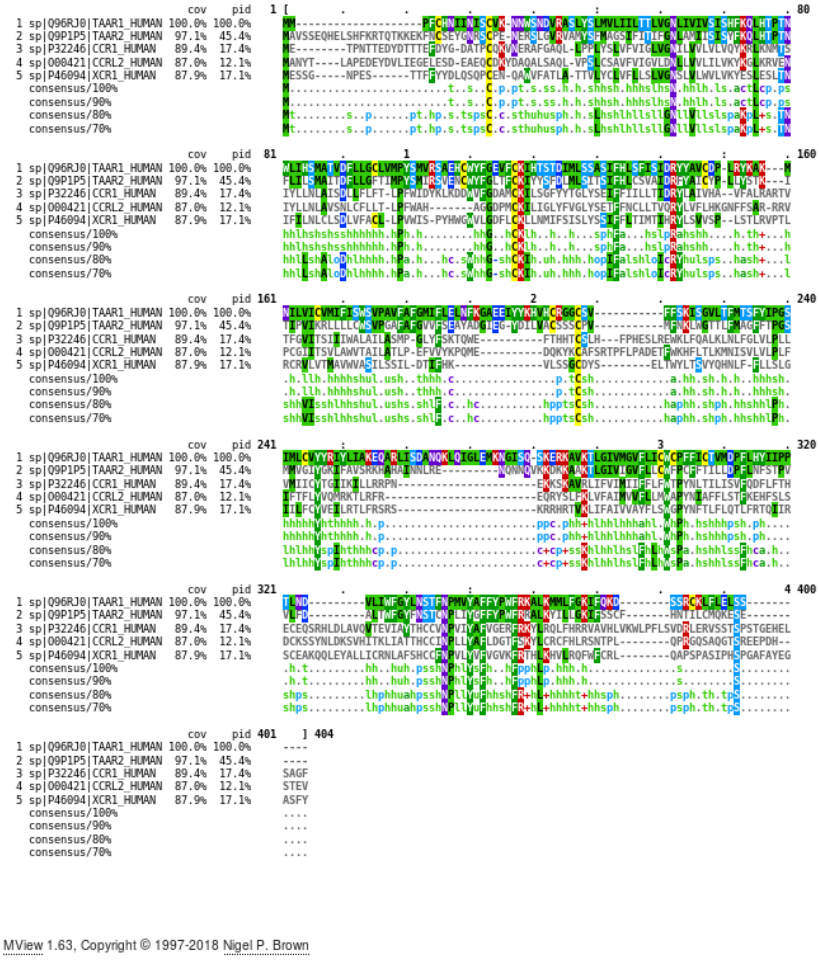


MView 1.63, Copyright © 1997-2018 Nigel P. Brown

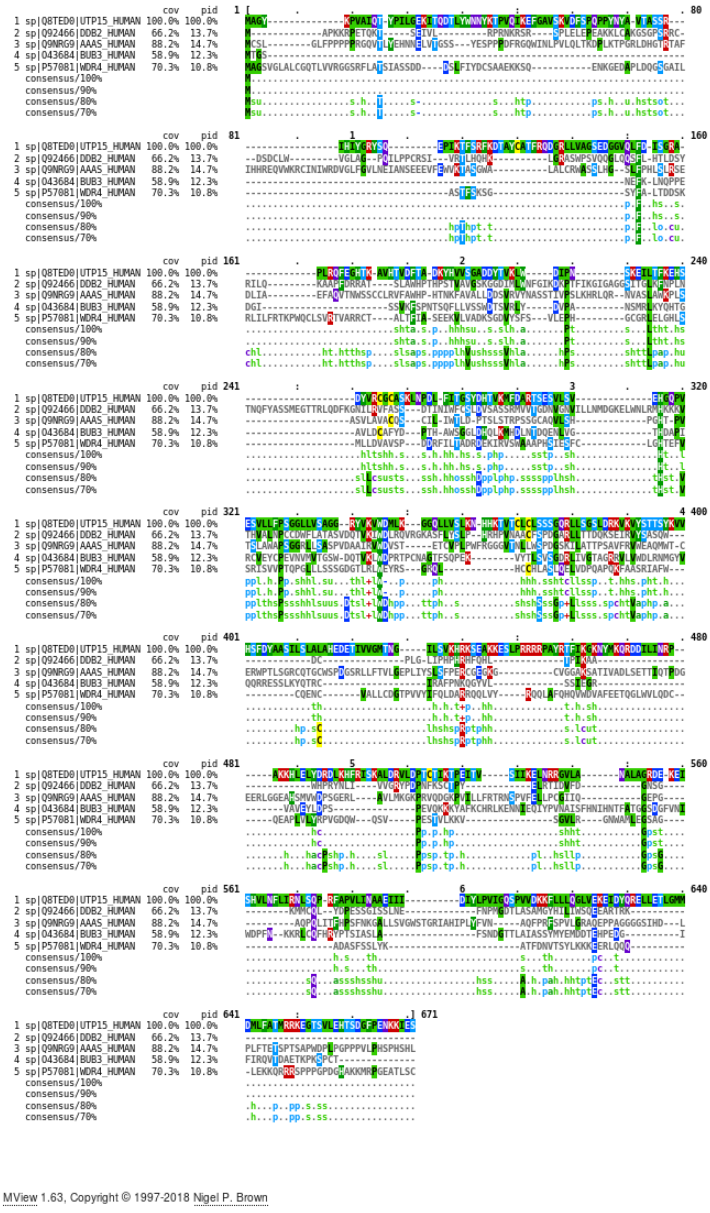
SCTR_HUMAN



MSA_GPCR



MSA_WD40



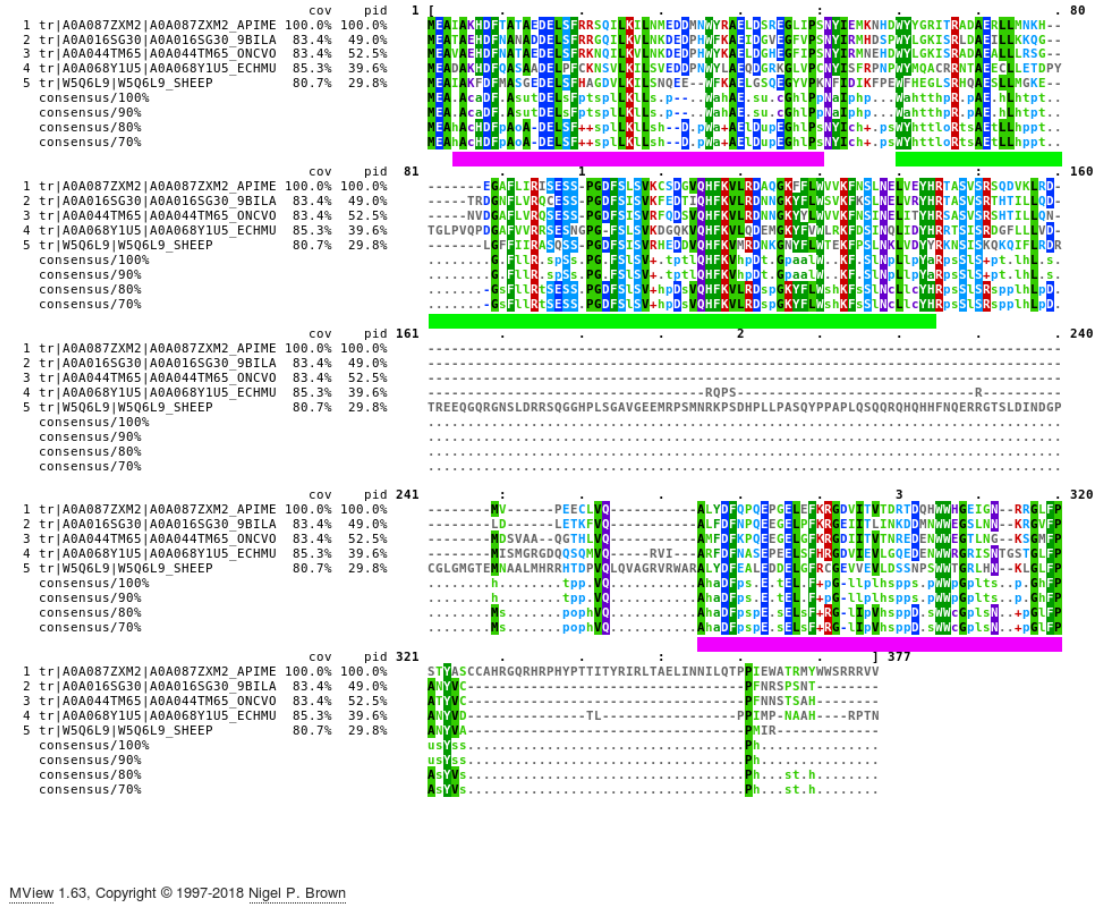


Figure B.1: MSA of 5 proteins all containing 3 conserved domains. The two SH3_1 domains are highlighted in magenta and the SH2 domain is highlighted in green. All the data for the MSA is from pfam. This alignment is adapted from MView with the addition of highlighted magenta and green region[30].

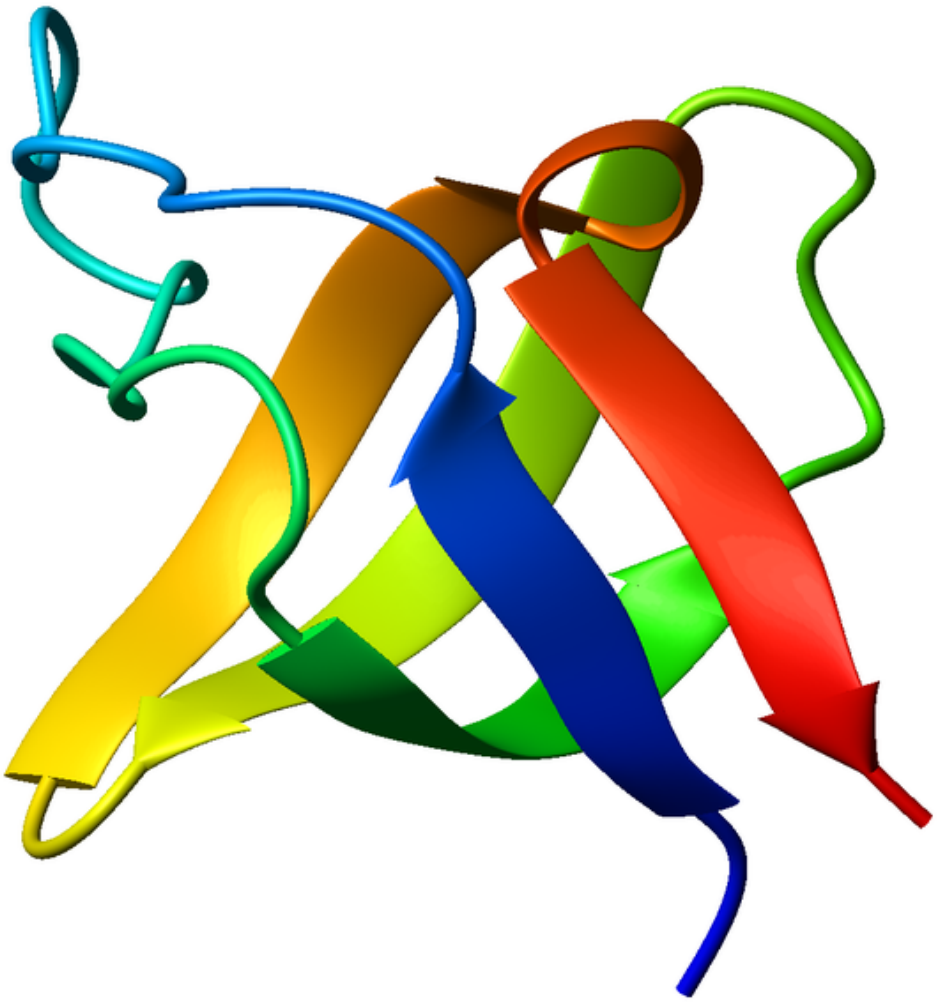


Figure B.2: “Ribbon diagram of the SH3 domain, alpha spectrin, from chicken (PDB accession code 1SHG), colored from blue (N-terminus) to red (C-terminus)”https://pfam.xfam.org/family/SH3_1

B.5 Questionnaire MSA handout

B.6 Questionnaire 3d Structures

B.7 NASA TLX Results

B.8 Extracts from Transcript of Focus Group

These extracts are representative statements regarding the five topics of my categorisation system. They are divided into five subsections, correspondingly. For reference,

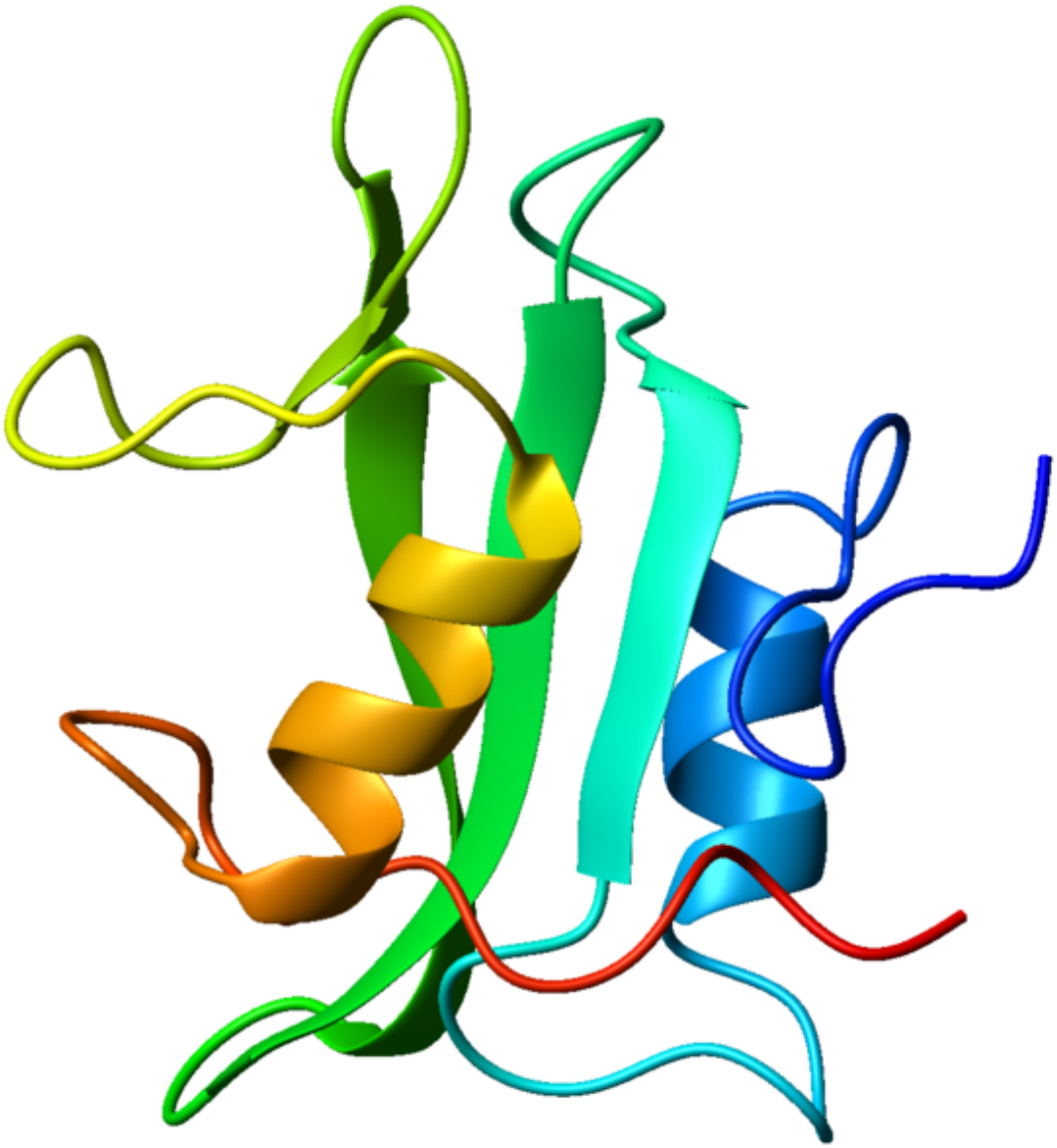


Figure B.3: “Crystallographic structure of the SH2 domain. The structure consists of a large beta sheet (green) flanked by two alpha-helices (orange and blue)”http://pfam.xfam.org/family/PF00017

Factor	Weights	Tally	Rating	Rating S.D	W. W. Score
Physical Demand	0.4	2	7	8.37	2.8
Mental Demand	3.8	19	75	15.41	285
Temporal Demand	2.4	12	32	16.81	76.8
Performance	2	10	74	25.35	144
Effort	3	15	60	30.21	180
Frustration	3.4	17	56	34.89	190.4

Table B.1: These are expanded results from the NASA TLX detailed in Table 4.1. *Tally* refers to how many times each subscale was selected in the comparison process. There were 5 participants. The *Rating S.D.* is the standard deviation of the ratings. *W.W.Score* refers to the weighted workload score created by multiplying the rating by the weight for each subscale.

here are the line numbers at which the six sonifications were played. These are marked with horizontal lines in Tables B.3 - B.7.

Line	Filename	Sonification Description
16	hydroPro_ANKR1.wav	protein ANKR_1_HUMAN with approach in Section 3.5.
63	hydroPro_ESX1.wav	protein ESX1_HUMAN with approach in Section 3.5.
93	onlyRedu_SCTR.wav	protein SCTR_HUMAN with approach in Section 3.6.
144	Entropy_GCPR.wav	MSA: GPCR proteins with approach in Section 3.8.
215	Multi_WD40.wav	MSA: WD40 domain with approach in Section 3.9.
314	redu_SCTR.wav	protein SCTR_HUMAN with approach in Section 3.7.

Table B.2: Transcript line numbers detailing when each sonification was played during focus group. The filenames correspond to files accompanying this submission.

Line	Statement
37	I think that with your project you need to think not about whether it is possible, because youve proved that it is, but can you compete against what is used nowadays.
38	You need to make it so that the biologist or bioinformatician is keener on using this approach than on the approach that is being used already.
51	Is [the speed of the sonification] something that youre looking to have control of?
54	You might not be interested in the whole protein anyway. Most people are not.
114	I thought that reducing the alphabet was a good idea, as it would be less confusing than hearing a lot of notes.
120	I appreciate that I couldnt get all seven of them, but I think there is still something good abut grouping the notes into groups. Unless you are specifically looking for that, you arent bothered about the different hydrophobicity or hydrophilicity of all the 20 amino acids.
125	I think that the grouping would still be a good idea, depending on the analysis that you were trying to get from it.
172	...but I dont see how that would be better than just looking at the alignment as that is very easy to see white and not white. It is much faster. I dont see what would be the added value of listening to the alignment.
178	That is if you can see. If you were a visually impaired scientist, I think this would be a really good thing to have.
184	Its a trade-off between including everything and having more information, or making it more understandable but losing information.
218	[in response to learning that they'd only heard half the alignment] God!
233	I think its useful but only to give a rough approximation to what the alignment is like compared to other alignments.
266	You might not be able to reproduce things between different people as easily if you get that difference in the way people hear different pitches. [Other participant:]Maybe you could customise it. [First Participant:] Yeah, I was going to say!
274	Theres nothing about the sound which is biological, it would just be a preference.
278	People are interested in different aspects of the same data.
279	I could see people choosing to using this technique to if they have a really long multiple sequence alignment or a lot of multiple sequence alignments and they just want to know if theres a conserved domain: just as a way of filtering through the data that theyve already got and they dont want to look up every single one. Theyre just going to use this as an initial filtering approach to look for conserved domains, so the flexibility to customise it would be good.
288	So, if you start really broad it helps you get to conserved regions, then as you get more people using it who have different ideas or different needs then you can start to get more specific, if possible.
330	But I think implementing different instruments might be a good shout, but you might not want to overcomplicate it.
334	It would take a bit of getting used to: knowing what instruments corresponds to which level of hydrophobicity. That would be quite difficult to get a handle on. If its just pitch then its easier to get a handle on at some level, but with different instruments corresponding you would have to practice and train yourself. Im not sure how easy that would be, and whether people would bother.
352	I think it comes down to trying to attune yourself to an instrument representing a level of hydrophobicity.

Table B.3: *Project Judgements*: representative statements from focus group transcript concerning feedback on the project as a whole, and on the new mode of data representation for protein sequence data. Horizontal lines indicate when each sonification was played during focus group, details of these are in Table B.2.

Line	Statement
39	I think it is kind of fast, if you want to make things out of it. Id rather hear it with more spacing between each sound, if I was to analyse it. But then again, there is a difference between each thing. So, I know that this is one thing, and this is the next thing. So maybe more space...
43	Not sure I agree, it seemed to go on for a long time
49	Im not sure youd always want to be going slower, I think you wouldnt have time if you had several proteins to look at.
64	You could really hear it this time. The hydrophobicity at least, the part that was low pitched
66	Yeah, you could hear low-low-low then high. [second participant:]...Then little high one.
68	I would not say a conserved domain, but a repeated motif.
72	The only thing I noticed that was different was that the high-pitched ones were a bit more frequent than the first one. Probably I wasnt focusing enough in detail and that could be a flaw of the method. It depends on your purpose for listening to it...
95	Im struggling to hear repeated patterns, to be honest. There could have been something there, but it was hard to discern. You had that high note going all through it...
105	It was not easy to hear the motif change.
112	It just made it sound like there was a hydrophillic thing coming in all the time every 2 amino acids, which stopped you picking up on the longer structures.
115	I could hear subtle changes, but I think it was that high note throughout that was throwing me off. You could tell there was a pattern, but in terms of the number of times that pattern came up, or even what that pattern was, Im not sure you could hear that. [Second Participant:] Yeah.
120	I appreciate that I couldnt get all seven of them
122	If you could make the difference in pitch more pronounced. In the previous example the lower pitch felt lower.
124	I think the same, there didnt seem to be a huge difference between the highs and the lows.
125	Here, even the lowest pitch felt quite high.
146	I think you can really hear where the conserved regions are. It does sound very different, going along high and the you get the *dum-dum-dum* for a bit which gives you an idea where to look, or if there is a conserved region. If that was just your initial question. I felt that I could hear that more clearly than the first ones...
152	I think the extremes are very easy to tell, but everything that isnt really low or high is a blur. It was easier to understand that one than the previous.
171	I think as [other participant] said, its easy to discern between the highly conserved and not highly conserved...
185	I think its very easy to understand the very high pitch and the very low pitch, but everything in between is not at all.
221	But you still can pick out conserved regions. Every now and again you can hear that one note coming out strong. Especially when it happened a lot, you got it over a few amino acids. I dont think it was as clear as the previous, as all the notes playing at the same time is quite a lot to take in. I dont know if I could tell, with everything else going on, the difference between where all the amino acids were conserved, or whether there was one and gaps.
226	I think all of this gives a flavour of what the protein is like, but Id struggle to get detailed information out of it. Maybe Im just too stupid, but I think Id struggle to compare this alignment to another alignment that youd hear just afterwards. You might get a general idea of which one was more conserved...
236	This was easier to discern than the previous one because it was challenging and the other was boring. As in why am I putting attention?. As well I was trying to find a pattern.
247	And while I do agree that it is more challenging, there are positives with it being boring, as Pablo said, with regards to analysis.
251	Whereas when I hear the same sound, I can more easily find a pattern across what Im listening to, and usually patterns are what youre looking for in this kind of stuff.
258	I think that one is better as long as the sound within the spectrum are more different.
260	I do think there are positives to listening to less sounds than to more difference.
284	I was going to say about the difficulty to getting detailed information out of it, so you might want to use it as an approximation just to get an initial idea about something.
296	As an only source of information this is difficult
302	As I cant tell what Im listening to.
325	I did like the fact that there were different instruments, as I think you can tell the difference more clearly between different instruments than just different notes. The whole difference in sound? I dont know.
328	I dont know if I could take on board both things.
329	I can definitely tell the different instruments, but the different sound within the same instruments? No, not really.
333	Yeah, its a lot to take in, I think.
338	I think it was the most informative, because of the different instruments. I think it is more easily distinguishable...
348	Trying to do the reduced alphabet is to simplify things, but then using different instruments with different pitches within the reduced alphabet undoes that work of simplification. Then you get more complex again. Maybe having the full alphabet with full pitches is easier than reduced with different pitches.

Table B.4: *Analytic Judgements*: representative statements from focus group transcript concerning feedback on the information carrying aspects of the sonification. Horizontal lines indicate when sonifications are played, details of these are in Table B.2.

Line	Statement
24	It is much more diverse that I expected. I was expecting something very boring, but it has a lot of diversity.
30	Its slightly chaotic, but not completely chaotic. Its like a scary movie tune, like a soundtrack
31	Its very Carpenter!
44	It goes on for a long time, but I think that that is nice. When you work with any genetic data, you can forget how big it is. When you look at the 4 lines on the page, it looks like a small piece of protein. But listening to every single one of these amino acids as a note, even going quite fast, you see that it is a lot of information. I think it is good for giving you an appreciation of how big that dataset is.
97	You had that high note going all through it, a beep, higher than the others. And there were bits where it sounded like a guitar coming in like acoustic guitar. That was a bit gentler
102	[in response to being asked if it was more musical than the previous example] Yeah, probably was to me. I found it easier to relate to: a repetitive musical structure. It was a bit easier to listen to easy listening proteins!
115	It sounded very nice
149	I felt that I could hear that more clearly than the first ones, but maybe thats because it sounded less like music and more discrete, it was just high-high-high then low-low-low.
220	I thought it was going to be chaos, especially when it started-it was just like someone bashing at notes.
236	and the [entropy sonification] was boring.
339	I think it is more easily distinguishable, but its also the least musical. Its the weirdest sound.
340	Its the most variable sound, so it doesnt conform to the normal structures of music.

Table B.5: *Aesthetic Judgements*: representative statements from focus group transcript concerning feedback concerning the beauty and feeling of the sonifications. Horizontal lines indicate when sonifications are played, details of these are in Table B.2.

Line	Statement
150	Maybe that was easier for my brain to associate with things going on. Though that could be very subjective.
155	The thing I pay attention to is the high notes, and I zone out for the low notes. So, if youre really keen on hearing the conserved bits, then you could flip it round and make those the high notes, so they stand out. If what youre interested in is breaks in your conservation, then it makes sense to make those as the high notes.
165	I dont know, I guess youd get used to it I suppose. I still think theres a tendency for my hyperactive mind to zone out unless theres something to hang onto, so I might still struggle to pay attention to the low bits. That could be good if you dont want to pay attention to the low bits, but it might be better to be the other way around. Even if I listened a lot, I might well still zone out on the low notes.
223	. I dont think it was as clear as the previous, as all the notes playing at the same time is quite a lot to take in.
240	that in relation to the strong sound and lower sounds; for me, not sure about other people, but for me a high lower sound is louder than a strong high sound.
244	A loud low pitch sound is louder than a loud high pitch sound. Hen it is a louder sound, I can tell more easily the lower pitched one than the higher pitched one.
247	For me it wasnt easy to tell a louder high pitch, compared to a louder low pitch.
249	My visual memory is much better than my listening memory, I can tell what Im listening to at the moment and I can recognise what is present in all of them or is hydrophobic, but I wont remember this two seconds later. Whereas when I hear the same sound, I can more easily find a pattern across what Im listening to, and usually patterns are what youre looking for in this kind of stuff. I do think there might be a point for it being more of the same sound than more different things, as I do think there might be a threshold for how much you can tell by listening to this sort of thing.
262	I find the opposite pattern with hearing loudness. I think higher pitches are easier for me to find as louder.
264	Which could be a flaw, as some people are more biased towards the hydrophobic notes, and some people towards the hydrophilic ones, or some people towards the more conserved areas, and the less conserved.
274	At the end of the day it doesnt really matter what the sound is. Theres nothing about the sound which is biological
291	Maybe people who are blind would hear it more accurately. Maybe their senses are more attuned.
335	It would take a bit of getting used to: knowing what instruments corresponds to which level of hydrophobicity. That would be quite difficult to get a handle on. If its just pitch then its easier to get a handle on at some level, but with different instruments corresponding you would have to practice and train yourself. Im not sure how easy that would be, and whether people would bother.
354	It might be possible but might take a lot of training. It might be something a computer would be better at than a human like machine learning which might defeat the purpose slightly!

Table B.6: *Psychoacoustic Judgements*: representative statements from focus group transcript concerning feedback concerning the subjective psychoacoustic responses of participants as to how they experienced the sound. Horizontal lines indicate when sonifications are played, details of these are in Table B.2.

Line	Statement
40	Id rather hear it with more spacing between each sound, if I was to analyse it.
123	If you could make the difference in pitch more pronounced. In the previous example the lower pitch felt lower.
156	So, if youre really keen on hearing the conserved bits, then you could flip it round and make those the high notes, so they stand out. If what youre interested in is breaks in your conservation, then it makes sense to make those as the high notes.
186	If you took a similar approach to the last method, and instead of calculating mathematically the pitches, you grouped them with discrete sounds, you could then say whether it is all conserved or 25% conserved etc. You would lose information as you are grouping things but... [Moderator:] But you think it might aid in understanding? [Participant:] Yeah.
192	Yeah, like a percentage cut-off for conservation, like 0% to 100%
194	And then it would be easier to find things that are very conserved, things that are maybe more or less But then it depends on what you are interested in, as you are losing information. But just listening to very different sounds might make it more difficult to make associations between things within the main alignment.
230	I think I would need to look at something and complement it with something else to get the detail to an adequate level and have something that is measurable rather than just a getting a feeling that something this is different.
253	I do think there might be a point for it being more of the same sound than more different things, as I do think there might be a threshold for how much you can tell by listening to this sort of thing.
258	I think that one is better as long as the sound within the spectrum are more different.
260	I do think there are positives to listening to less sounds than to more difference. But thats me.
269	If I want to listen to the hydrophobic as the high pitched one then I could pick that and if you wanted the opposite you could pick that instead.
276	And you could customise it so its just four sounds, or to listen to everything thats conserved, or nothing thats conserved, or quartiles in between. As its very dependent on what your listening for and for the data. People are interested in different aspects of the same data.
293	Id love to be hearing this at the same time as knowing exactly where the note was coming from
297	if you were to implement this along with a visual analysis... As [participant] said, I cant see where I am: I try and then nope. How would you make it so that Im looking at something and I know? Would you have an arrow pointing at it?
301	You could have one of those balls like at karaoke.
309	[Moderator:]You could have a bar that goes across, like in so many music applications. That could be on top one of these [MView] and go across. Also, sonically we could put in clicks. On this viewer, do you see the little 1 and 2 on top? These correspond to the 100th and 200th residue. The idea is that a little click would allow people to realign, though it might be superfluous once youve got the viewer. Also, it might be good for sectioning off so you can identify that it occurs between 1 click and 2 clicks. Then you can just go back to click-click. This might help perceive size. [participant:] Thats something to think about if you have a really long alignment.
350	Trying to do the reduced alphabet is to simplify things, but then using different instruments with different pitches within the reduced alphabet undoes that work of simplification. Then you get more complex again. Maybe having the full alphabet with full pitches is easier than reduced with different pitches.

Table B.7: *Further Work*: representative statements from focus group transcript concerning suggestions for improvements or further work. Horizontal lines indicate when sonifications are played, details of these are in Table B.2.