2003-01-01

Rhythmic Parsing of Sonified DNA and RNA Sequences

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

Eugene Coyle
Technological University Dublin, Eugene.Coyle@tudublin.ie

Rhythmic Parsing of Sonified DNA and RNA Sequences

Charlie Cullen[ϕ] and Eugene Coyle*

[ϕ]Department of Music,
Dublin Institute of Technology, Dublin
IRELAND
E-mail: [ϕ]charlie.cullen@dit.ie

** Department of Control Engineering,*
Dublin Institute of Technology, Dublin
IRELAND
*E-mail: *eugene.coyle@dit.ie*

Abstract **-- Sonification allows existing mathematical data to be used as the model for audio output, notably that the audio produced is related to or representative of that data in some way. Existing work in the field has been largely focused on the aesthetic tailoring of the output audio for compositional benefit rather than as a framework for audio representation and analysis.**

It is the goal of this research to apply existing techniques for pitch substitution to an analytical method that seeks to define and represent patterns within existing data sets (primarily DNA and RNA sequences). It is often the case that sonified audio has little or no rhythmic component, and it is felt that as rhythm is such an important part of the musical analysis process it should be given far more serious consideration when representing mathematical data as audio.

In order to adequately analyse the different rhythms and time signatures that can be used to parse sonified data, a piece of software has been developed called DNASon that allows for basic time interval and signature definitions for application to a sonified DNA/RNA sequence.

I SONIFICATION

Sonification can be defined as the use of nonspeech audio to convey information [1]. The modern means of sonification usually involve some form of computer processing to substitute audio signals for existing mathematical data, but existing methods such as sonar and the Geiger counter define the use of audio to represent data for analysis and understanding without the specifics of software programming. The rapid increase in the amounts of data that many fields of modern research and endeavour have to deal with has led to the need for better tools and means of representation, and to this end audio representation is a viable and often powerful means of analysing data.

It is suggested that sound can convey significant amounts of information [2] and thus could be of great benefit in the analysis and understanding of data, perhaps in some instances in tandem with other visual techniques as a means of further enriching the perception and understanding of data structures.

Pattern matching is an area of data analysis that relies heavily on computational methods to distinguish possible trends or behaviours within a sequence or data set. It is suggested that as human intelligence and perception are designed specifically for the purpose of pattern matching it may be of use to convey the data in a format that lends itself towards human pattern recognition. Visual techniques for conveying data are well known (in anything from a line graph to fractal modelling) but sonification for analysis is still relatively undervalued as a means of representing data for human recognition and understanding.

II SONIFCATION OF DNA AND RNA SEQUENCES

The analysis and understanding of deoxyribonucleic acid (DNA) is one of the most pertinent topics in modern biology, with the recent decoding of the human genome paving the way for many advances within the field of genetics. At ground level DNA is a four base 3-digit code for all life that exists on this planet, with the bases Adenine, Guanine, Cytosine and Thymine arranged in groups of three known as codons that define the relevant amino acids which combine to make the proteins found in all living organisms.

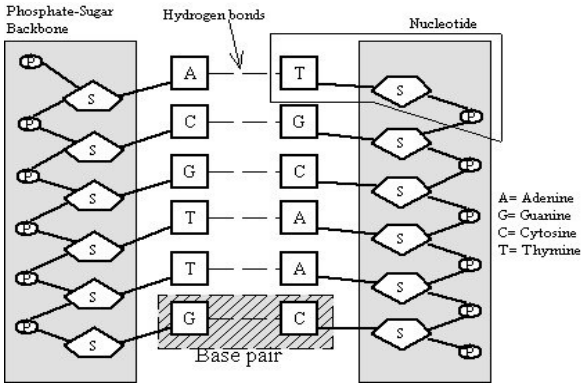

Figure 1: Example of a DNA base sequence.

This compactness of data transfer is further illustrated by the three to one redundancy of codons to amino acids, whereby the 19 amino acids found in proteins are coded from 64 possible codon combinations (4 bases, 3-digits 4x4x4 = 64). Ribonucleic acid (RNA) is the truncated product of DNA, wherein the section of the DNA sequence relevant to amino acid scripting is transcribed into RNA sequences that substitute Uracil for Thymine, before being translated into proteins.

Sonification of DNA and RNA was perhaps first suggested by Douglas R. Hofstadter [3] in 1979 by way of regarding the RNA transcription mechanism as similar to that of a tape recorder. The notion of actually representing DNA sequences by pitch substitution was defined by Nobuo Munakata and Kenshi Hayashi [4] in 1984 with the suggestion that the four bases of DNA could be allocated pitches within the interval of a fifth.

Further, it was also suggested by Munakata and Hayashi that the corresponding amino acids coded from RNA sequences could also be represented by pitches within the same output sonification[5]. With these basic tenets as a guideline for the sonification of DNA/RNA sequences many other composers have produced works using DNA/RNA as a template.

Much of this work has been primarily for compositional purposes, with the sonified DNA/RNA often being edited or adapted for aesthetic reasons. To this end, it is proposed that this research uses the sonified output of DNA/RNA sequences for pattern matching and analysis purposes rather than as frameworks for composition.

III DNA/RNA SONIFICATION FOR ANALYSIS

The goal of this research is to provide a means of sonifying data for analysis within as straightforward a framework as possible, so that it may be used as an analysis tool regardless of musical training or ability. In many fields where sonification could be of great benefit the very notion of using audio for analysis purposes is difficult enough to accept, and to this end any tools or principles developed must strive to narrow this gap in perception as quickly as possible.

For the sonification of DNA/RNA in rhythmic intervals it was felt that specific software needed to be developed, rather than using existing tools which although often very accomplished were not primarily focused on rhythmic *and* pitch-based analysis. The main aim of any application would be to make the sonification of DNA/RNA sequences as user-friendly as possible so that the software could be used regardless of musical and indeed technical knowledge as a means of analysis.

The DNASon software developed as part of this research endeavours to provide a quick and efficient means of setting DNA/RNA sequences to music within a Graphical User Interface (GUI) based application. The program was developed as a Macromedia Flash front-end that dialogs with back-end Visual Basic code via the FS command structure of Macromedia Flash.

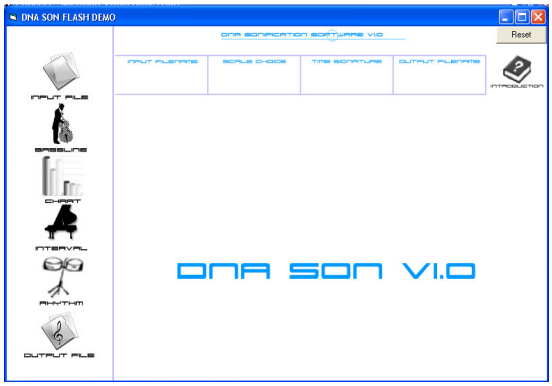

Figure 2: The DNASon main dialog screen.

The development of a separate front-end provided several advantages in that the graphical content of the GUI could be more advanced than that which could be catered for within the confines of Visual Basic. Also, due to the nature of the FS command structure elements of Visual Basic dialogs could be incorporated into the GUI as required, allowing for standard Windows File Input/Output dialogs to be

called at runtime to specify the files used by the code.

The use of Windows drag and drop functionality is also utilised by calling Visual Basic Forms instead of Macromedia Flash movies as a means of allowing the user to allocate pitches to amino acids.

The initial dialog screen of the GUI allows the user to view several short Macromedia Flash movies that cover aspects of DNA, RNA and Sonification that are relevant to the software. This is also the main dialog for the software, and the first option available is the specification of the DNA/RNA input sequence file to be sonified (of type .fasta).

The user is next presented with selections defining the notes representing the DNA/RNA bases and also chord intervals for the amino acids they code for. The frequency of each amino acid in the sequence is also displayed as a graph to allow the user to make an informed choice of chord for each amino acid based on its relative occurrence.

The user can now assign a different general midi instrument to the bassline and chord intervals of the sonified output sequence. The allocation of different sounds to different aspects of the sonified data is of great benefit in helping users to distinguish between the data in various terms. The bases in a sequence are now a conveyed by a different instrument from that which represents the amino acids, and this seeks to further illustrate the different information contained within the DNA/RNA code.

The sequence can then be parsed into rhythmic intervals to allow for easier understanding at output. Specifically, the bass notes and chords representing nucleotide bases and their corresponding amino acids can be played in 3/4 (3 note groups) or 4/4 (3 notes followed by a rest) time so that each amino acid in the sequence can be better analysed in isolation or indeed as part of a sequence as required.

This use of rhythm is intended to break up the output information into smaller chunks that can better be understood by the human brain. A relevant analogy would be the use of punctuation in a letter as means of indicating words and sentences rather than a continuous stream of letters that would have little or no meaning to the reader.

The 3/4 time signature seemed appropriate due to the 3-base grouping of codons within an RNA sequence, and it was felt a good way of accenting each codon within the output sonified sequence. The 4/4 time signature was also included as three beats followed by a rest, and this proved to be an even better way of defining codon base groupings and their corresponding amino acids as three crotchets and a dotted minim respectively.

The sequence is then ready to be written to output as a Standard Format 1 Midi File. Midi files of Format 1 allow for up to 16 distinct tracks to be assigned within a sequence file, and this proves to be of use in allocating the bassline of the base sequence to one track and the chord intervals of the amino acid sequence to another. With the output file now written to disk, the user can choose to exit the software or reset all previously chosen parameters and begin sonification of a sequence again.

IV CONCLUSIONS AND FUTURE WORK

The development of the DNASon software application has been the main focus of the research thus far, and it is hoped that with the basic tool for rhythmic sonification up and running analysis can now begin to take place.

The next aim of this research is to examine how well the sonified data is conveyed when parsed rhythmically into different time signatures and note groupings. It can be suggested that there are as many different note groupings as there are musical compositions, and it may prove after further work to be of benefit to allow the user to define the output sequence on a note by note basis using some form of pattern definition within the software. It could also prove useful to allow the user to define their own bass notes within a key rather than working from the arbitrary assignments of 1^{st}, 3^{rd}, 5^{th} and 7^{th} that are available at present.

It is also seen of potential benefit to incorporate elements related to the final proteins created from the chains of amino acids by means of further analysis. The structure of proteins is defined in many ways by the folding patterns that take place along the chain of amino acids, and these structures can often fold in up to four different levels of complexity. The effects of protein folding could be of interest relative to the original base sequence that coded the amino acids that formed the protein itself. Some form of melodic or rhythmic element could be incorporated into the output sonified sequence (perhaps as a modulation of key or change in rhythm or tempo) as an attempt to represent the entire process of translating RNA to protein in some way.

The noted potential drawback is that this could conceivably shift the focus of the software into composition rather than analysis, and as such some form of documentation may be required as an annotation of the sonified output sequence in order that anyone seeking to use a particular output

sequence for analysis is first made aware of the tonal and rhythmic definitions within it. To this end it hoped to include within the code some provision for appending the input DNA/RNA sequence alongside the parameters chosen to the output file, so that any further work with the file will take these factors into account.

It is also hoped to define some form of provision for playback of sequences either within the code or as a separate application within a suite of sonification tools. This playback tool could also conceivably contain some form of graphical display element that would convey the annotated parameters mentioned above alongside the sonified data sequence for an even richer understanding of the data being analysed.

REFERENCES

[1] Sonification Report: Status of the Field and Research Agenda, International Conference on Auditory Display (ICAD) 1997

[2] H. G. Kaper, S. Tipei and E. Wiebel, "Data Sonification and Sound Visualization" Computing in Science and Engineering, Vol. 1, No. 4 (July/August '99), 48-58

[3] Douglas R. Hofstadter, GEB: an Eternal Golden Braid (pg 518-519), Penguin Books ISBN 0-14-028920-8

[4] Nobuo Munakata and Kenshi Hayashi, "Basically Musical", Nature 310, 96 (1984)

[5] Nobuo Munakata and Kenshi Hayashi, Gene Music: Tonal assignments of Bases and Amino Acids (pg 72-83) In *Visualizing Biological Information* (Ed. C. Pickover),World Scientific, Singapore(1995). (ISBN 9810214278)