Auditory Display of Genome Data: Human Chromosome 21

Sook Young Won

Center for Computer Research in Music and Acoustics,
Department of Music,
Stanford University
Stanford, California 94305, USA
sywon@ccrma.stanford.edu

ABSTRACT

The motivation for this paper is to systematically explore the efficacy of mapping data to sound parameters with the specific aim of sonifying statistical trends and hearing the 'gist' [1] of the data. In this paper, we consider the task of searching through a gene sequence of the human chromosome 21 for CpG islands and type of gene evidence. Musical intervals and rhythm is proposed for detecting CpG islands and musical timber is used for representing the gene evidence. We extract human genome data from the NCBI (National Center for Biotechnology Information) [2] database and use list processing and synthesis capabilities of CLM [3].

1. INTRODUCTION

This project was developed as part of the *Auditory remapping of bioinformatics* course [4] at CCRMA, Stanford University. An objective of the course was to explore effective means of auditory display of human genome. For that reason, we describe using parameter mapping to express auditory display of statistical searching through large databases. As an example of statistically detected features from data, we consider the task of searching human genome data with the aim of detecting the presence of CpG islands using auditory cues.

CpG is a site where cytosine (C) is neighbor to guanine (G) and connected by a phosphodiester bond (p). Detection of regions of genomic sequences that are rich in the CpG pattern (known as CpG islands) is important because they indicate areas of interest along the genome. Experimental results suggest that all housekeeping genes[1] and 40% of the tissue specific genes[2] in humans have as associated CpG islands [5] [6]. Adding that, we display type of evidence used to construct gene model.

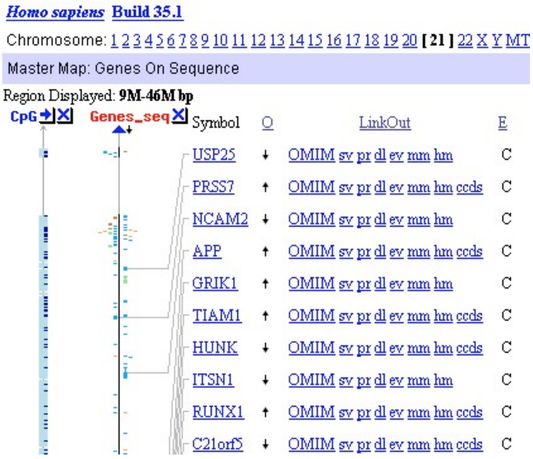
2. SONIFICATION PROCESS

2.1. Sound synthesis in CLM

Common Lisp Music (CLM), a music specific flavor of LISP, is appropriate for handling large data sets such as the human DNA sequence. CLM's extensive library of synthesis and signal processing functions allows for arbitrarily complex and refined control of music synthesis. CLM includes a set of example synthesis instruments useful in efficient mapping of sound dimensions and rapid prototyping of parameter mapping using duration, frequency, amplitude and timbre in musical constructs of rhythm and tempo, pitch, dynamic and instrument respectively.

2.2. Data description

In the next stage, we search multiple dimensions of a human gene and select features for sonification. We explored a variety of databases for the genetic information currently being generated through the Human Genome Project, provided by NCBI. We search databases for genomic features a sequence with various map options [7]. Among the 36 types of maps, we focus upon the gene-sequence map and the CpG island map. The gene-sequence map shows the physical map of the gene distribution, the verbose description about each gene, and the link for downloading the sequence data. And the CpG island map shows regions of high G+C content and CpG content on the genome sequence.



Homo sapiens Build 35.1

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 [21] 22 X Y MT

Master Map: Genes On Sequence

Region Displayed: **9M-46M bp**

Symbol	O	LinkOut	E
USP25	↓	OMIM sv pr dl ev mm hm	C
PRSS7	↑	OMIM sv pr dl ev mm hm ccds	C
NCAM2	↓	OMIM sv pr dl ev mm hm	C
APP	↑	OMIM sv pr dl ev mm hm ccds	C
GRIK1	↑	OMIM sv pr dl ev mm hm	C
TIAM1	↑	OMIM sv pr dl ev mm hm ccds	C
HUNK	↓	OMIM sv pr dl ev mm hm ccds	C
ITSN1	↓	OMIM sv pr dl ev mm hm	C
RUNX1	↑	OMIM sv pr dl ev mm hm ccds	C
C21orf5	↓	OMIM sv pr dl ev mm hm ccds	C

Figure1. *The gene-sequence map and the CpG island map of human chromosome 21 in NCBI*

Researchers in genetics and bioinformatics have developed various methods to analyze and organize human genome resource for identifying features of DNA sequence.

The human genome is made up of 3 billion bases of DNA and split into 46 chromosomes, pairs of chromosomes 1-22 and a pair of sex chromosomes – XX or XY [8]. We limit our sonification experiment to the shortest chromosome, chromosome 21 consisting of 46Mbase including 337 genes and 4999 CpG islands.

[1] Housekeeping genes are genes that are trascriptionally active (i.e. produce proteins) in cells throughout the body.

[2] Tissue specific genes are transcriptionally active only in certain cells.

2.2.1. CpG Islands

CpG islands are designated according to a number of criteria. Gardiner-Garden and Frommer [9] defined a CpG island as being at lest 200 bases long, having at least 50% G+C content and an observed CpG/expected CpG ratio of at least 0.6. Takai and Jones criteria [10] are a length of at least 500 bp, G+C content of > 55% and observed CpG/expected CpG ratio of >= .65. There are about 29,000 CpG islands in the human genome. We consider the search for CpG islands of illustrative value since there is an (perhaps naïve) analogy to be drawn between a CpG island and a musical motive. This analogy lies in the fact that humans are adept at identifying similarities between musical motives that may have a large degree of variance between recurrences. Schematic expectations and auditory stream segregation, two extensively studied aspects of music perception and cognition, provide useful bases for auditory representation of statistically driven identification of particular conditions within a data set.

2.3. Mapping Algorithm

2.3.1. Using auditory stream segregation to detect clusters of G + C content

The first criteria of CpG islands is that the G+C content is over 50% thus we only consider the number of Cytosine and Guanine in the genomic sequence. We divided four types of nucleotide into two groups, {A, T} and {C, G}, and then assigned C3 and G5 to each group respectively. The relatively wide musical interval of a twelfth makes the statistical clustering of G+C immediately apparent even to untrained listeners.

2.3.2. Adding duration to articulate likelihood of CpG islands

The definition of CpG islands also considers an observed CpG/expected CpG. We accordingly mark dinucleotides having one of CpG pattern - CC, CG, GC, or GG - by longer notes. In detail, as the pointer reads the sequence linearly, it decide whether a previous nucleotide and a current nucleotide compose CpG pattern or not. If the dinucleotide is CpG, the current nucleotide converts a combination of a 100ms long-note and a 100ms long-rest while other dinucleotides generate 50ms long-notes. Therefore, when we hear groups of high-pitched notes with long durations, we can easily recognize CpG regions.

2.3.3. Using timbre to distinguish details of the gene

The other interesting organizational datum in the NCBI genome database is the degree of confirmed knowledge about the data. Genes are represented in six CLM instruments as done in six colors on the gene-sequence map. And junk DNA region which do not have any evidnce is mapped to Plucked String model.

Instrument	**Gene Color**	**Type of Evidence**
Bell	Blue	Confirmed gene model
Violin	Light Green	EST only
Oboe	Dark Brown	Predicted + EST
Flute	Light Brown	Predicted Only
Drum	Orange	Conflict
Piano	White	Interim LocusID

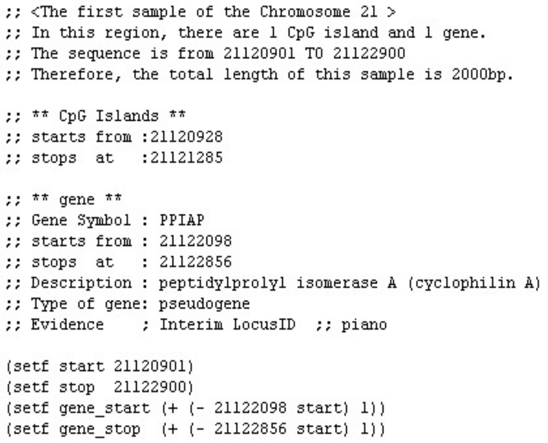
Table 1. *Mapping CLM instruments onto the types of evidence used to construct the gene model*

Additional Notes:

- The blue gene model is based on the clean alignment between a RefSeq or GenBank mRNA sequence and the genomic sequence.
- The light green gene model is based on EST evidence only.
- The dark brown gene model is predicted by Gnomon [6] and EST.
- The light brown gene model is predicted by Gnomon only.
- The orange gene model means there is some discrepancy between the mRNA sequence and the gene model.
- Models with Interim LocusIDs may be paralogs, genes not yet curated, duplications because of assembly errors, or pseudogenes.

2.4. Implementation

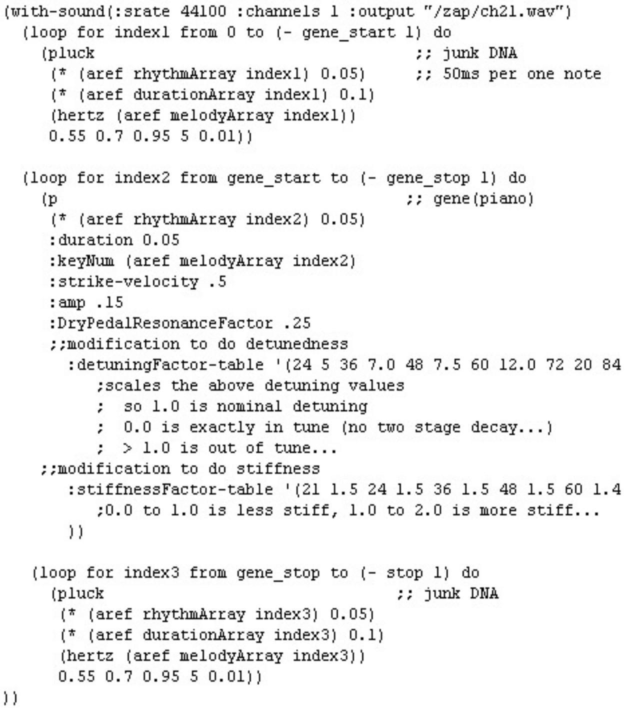
The sample sound file (~~http://ccrma.stanford.edu/~sywon/sonification/ch21.wav~~) is about the region from 21120901bp to 21122900bp.



```
;; <The first sample of the Chromosome 21 >
;; In this region, there are 1 CpG island and 1 gene.
;; The sequence is from 21120901 TO 21122900
;; Therefore, the total length of this sample is 2000bp.

;; ** CpG Islands **
;; starts from :21120928
;; stops  at   :21121285

;; ** gene **
;; Gene Symbol : PPIAP
;; starts from : 21122098
;; stops  at   : 21122856
;; Description : peptidylprolyl isomerase A (cyclophilin A)
;; Type of gene: pseudogene
;; Evidence    ; Interim LocusID  ;; piano

(setf start 21120901)
(setf stop  21122900)
(setf gene_start (+ (- 21122098 start) 1))
(setf gene_stop  (+ (- 21122856 start) 1))
```

Figure 2. *Description about sonifying the sample region of a DNA sequence in CLM*

As the mapping algorithm describes below, we assign the values including gene information to musical variables. Then we generate the sample sound file using piano and plucked string instruments in CLM.



```
(with-sound(:srate 44100 :channels 1 :output "/zap/ch21.wav")
  (loop for index1 from 0 to (- gene_start 1) do
    (pluck                                ;; junk DNA
     (* (aref rhythmArray index1) 0.05)   ;; 50ms per one note
     (* (aref durationArray index1) 0.1)
     (hertz (aref melodyArray index1))
     0.55 0.7 0.95 5 0.01))

  (loop for index2 from gene_start to (- gene_stop 1) do
    (p                                   ;; gene(piano)
     (* (aref rhythmArray index2) 0.05)
     :duration 0.05
     :keyNum (aref melodyArray index2)
     :strike-velocity .5
     :amp .15
     :DryPedalResonanceFactor .25
     ;;modification to do detunedness
       :detuningFactor-table '(24 5 36 7.0 48 7.5 60 12.0 72 20 84
          ;scales the above detuning values
          ;  so 1.0 is nominal detuning
          ;  0.0 is exactly in tune (no two stage decay...)
          ;  > 1.0 is out of tune...
    ;;modification to do stiffness
       :stiffnessFactor-table '(21 1.5 24 1.5 36 1.5 48 1.5 60 1.4
          ;0.0 to 1.0 is less stiff, 1.0 to 2.0 is more stiff...
       ))

   (loop for index3 from gene_stop to (- stop 1) do
     (pluck                              ;; junk DNA
      (* (aref rhythmArray index3) 0.05)
      (* (aref durationArray index3) 0.1)
      (hertz (aref melodyArray index3))
      0.55 0.7 0.95 5 0.01))
))
```

Figure 4. Sound generating part of the CLM code

3. CONCLUSION

In this project, we attempt to integrate principles of music perception, specifically auditory stream segregation with traditional constructs of music composition to create a meaningful sonification of the statistical presence of trends in data. We use the task of identifying CpG islands in searching through human genome data as an example application. Auditory scene analysis and the experiential ability to distinguish between varied repetition of melodic fragments are useful methods in sonifying data trends.

4. ACKNOWLEDGMENTS

The author would like to thank professor Jonathan Berger for his help and contribution to this project. The author also would like to thank Byung-Jun Yoon for providing the knowledge about human genome.

5. REFERENCES

[1] Z. Haixia, C. Plaisant, B. Shneiderman, and R. Duraiswami, “Sonification Of Geo-Referenced Data For Auditory Information Seeking: Design Principle And Pilot Study”, *Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display*, Sydney, Australia, July 2004.

[2] NCBI Human Genome Resources, ~~http://www.ncbi.nlm.nih.gov/genome/guide/human~~

[3] B. Schottstaedt, *Common Lisp Music Documentation*, ~~http://www-ccrma-stanford.edu/software/clm~~ CCRMA-Stanford University, 2000.

[4] J. Berger, *Auditory Remapping of Bioinformatics*, ~~http://ccrma.stanford.edu/courses/120~~, CCRMA-Stanford University, 2004.

[5] A.P. Bird, “Functions for DNA Methylation in Vertebrates”, *Cold Spring Harbor Symposia on Quantitative Biology*, 1993.

[6] E. C. Rouchka, R. Mazzarella, and D. J. States, “Computational Detection of CpG islands in DNA”, *Technical Report*, Washington University, Department of Computer Science, WUCS – 97 – 39, 1997.

[7] S. M. Dombrowski and D. Maglott, *Using the Map Viewer to Explore Genomes*, NCBI, 2003.

[8] The Wellcome Trust, *Key facts about the human genome*, ~~http://www.wellcome.ac.uk/en/genome~~, 2001.

[9] M. Gardiner-Garden and M. Frommer, “CpG islands in vertebrate genomes”, *Journal of Molecular Biology* 196, pp. 261-282, 1987.

[10] D. Takai and P.A. Jones, “Comprehensive analysis CpG islands in human chromosomes 21 and 22”, *Proceedings of the National Academy Sciences*, 99(6), pp. 3740-3745, 2002.

[11] John Dunn and Mary Anne Clark, *Life Music: The Sonification of Proteins*. ~~http://mitpress2.mit.edu/e-journals /Leonardo/isast/articles/lifemusic.html~~

[12] B. Yoon and P.P. Vaidynathan, “Identification of CpG islands using a bank of IIR lowpass filters”, *Proceedings of 11th Digital Signal Processing Workshop*, New Mexico, 2004.