Musical Synthesis of DNA Sequences

Peter Gena, Ph.D., Charles Strom, M.D., Ph.D.

School of the Art Institute of Chicago, Illinois Masonic Medical Center
Chicago, Illinois USA

Abstract

As a consequence of the Human Genome Project, there has been an explosion of primary DNA sequencing data available on CD ROM. This includes complete genomes of viruses, partial genomes of bacterias, and complete sequences for hundreds of human proteins. Consequently, we began to envision a type of computer-generated music that would take cues for its musical parameters directly from the physiological ones present in DNA. A DNA sequence consists of a specified order for the production of amino acids. The physical properties of amino acids (dissociation constant, molecular weight, and chemical class) combined with the properties of the individual bases (melting temperatures) provide the basis for inheritance and evolution and our musical compositions. The converted results, one for each codon, represent distinct musical actions in MIDI note events. Thus far, we have generated musical compositions from several human, viral, and bacterial sequences. This paper outlines our research.

1 Introduction: DNA

The genetic code is an alphabet made up of four chemical compounds which form the nucleotide bases—adenine (A), cytosine (C), guanine (G), and thymine (T). These bases are linked in a specific order to form the double helical structure known as deoxyribonucleic acid, or DNA. Each individual living organism has a unique order of bases that completely determines its physical structure. The four nucleotides are arranged in three-letter units known as codons. Each codon specifies one of nineteen amino acids. When they are grouped by chemical type, there are eight such categories. The DNA template, located in the nucleus of each cell, acts as a blueprint that directs the production of proteins. DNA is translated into messenger ribonucleic acid, or mRNA that is in turn serially scanned by ribosomes, organelles located in the cell's cytoplasm. Ribosomes use the mRNA as a template to direct the synthesis of proteins.

2 Physio-musical conversion

The initial programming task was to write an algorithm that converts the list of sixty-four codons into distinct musical events according to physical properties. A look-up table of the codons and their corresponding amino acid types, followed by the dissociation constant or pK(a) and molecular weight, was constructed as a data-base. There are eight basic musical timbres; one for each of the eight classes of amino acids. Each of the nineteen amino acids has a distinct pK(a) that helps define pitch. Additional modifications involve physical properties of the molecular bonding occurring in the codon itself, independent of what amino acid it codes for. Using 7.0 as the neutral point in acid/base equilibrium point, pK(a)'s below 7.0 are acidic while those above are basic. Hence, there are two equations for each codon: one correlates higher pitch with acidity, the other with base. The algorithm makes a binary choice with each selection.

2.1 Pitch

$$f = ((p (4G + 2T) +12) + k$$

$$f1= (([p - 7.0]) (4G +2T) +12) + k$$

where:
f & $f1$ = MIDI pitches
p = pK(a)
G = $\sum$ G + C per codon
T = $\sum$ A + T per codon
k = Constant (Hydrogen Bonds): [AA = -2, TT = -1,
CC = +1, CG = +2, GC = +3, GG = +4].

Additional pitch-bend commands for each note place the music in just intonation.

2.2 Intensity

Intensities (velocity) are also adjusted according to the hydrogen bonding occurring in each codon. As with pitch, there are two corresponding equations for each codon and a binary choice is made with each selection.

$$I = 6H$$

$$I1 = 109 - I$$

where:
I and I1 are MIDI velocity levels
H is proportional to codon melting temperature and Hydrogen-bond strength per codon (each G = +8, C = +6, A = +4, T = +1).

2.3 Duration

The pK(a) and atomic weights of the amino acids determine durations.

$$D = 0.01pM + 0.1Sk$$

where:
D = duration in clock ticks
p = pK(a)
M = molecular weight of amino acid
$S = f$ (sum of hydrogen bonds per codon)
k = tempo constant (>0), higher number = slower tempo.

3.0 Programming and realization

All of the preliminary programming is scripted in Hypercard. The scripts prepare all the necessary data, that is, the table of codons, and the genomes as collections for the MAX object code language (Copyright by IRCAM and Opcode Systems). The initial table data contains the index number followed by the codon, amino acid, pK(a), amino acid class number, and the molecular weight of the amino acid.

Codons 11 - 16 from the table:
11, TGT, CYS, 1.900, 6, 121;
12, CCG, PRO, 1.952, 7, 155;
13, CCC, PRO, 1.952, 7, 155;
14, CCT, PRO, 1.952, 7, 155;
15, CCA, PRO, 1.952, 7, 155;
16, ACG, THR, 2.088, 1, 120;

Each codon is transformed into a list in a collection. The list specifies the address, MIDI pitch, velocity, channel number, pitch bend, and duration of the event for the corresponding codon.

MAX collection**:**
11, 26 60 6 24 23;
12, 38 120 7 1 31;
13, 35 108 7 19 30;
14, 32 78 7 26 30;
15, 32 96 7 26 30;
16, 34 108 1 28 25;

A second series of algorithms reads the raw DNA strings for a genome, searches for the start and stop codons, and then forms the three-letter codon sequences. Uncoded filler, ubiquitous extraneous material bearing no significance to amino acid production, is ignored. In addition, each codon from the genome is checked in the look-up table and its index number is put into another collection.

Beta Globin sequence:
ATG ATG ATG GTG CAC CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT GCC CTG TGG GGC AAG GTG AAC GTG GAT GAA . . . . etc.

Corresponding MAX collection:
1,32;2,32;3,32;4,33;5,56;6,42;7,17;8,14;9,23;10,23;11,64;12,24;13,46;14,36;15,17;16,46;17,42;18,54;19,52;20,64;21,33;22,30;23,33;24,20; . . . . etc.

Actual genomes of human or bacterial proteins, or complete viruses can then be scanned by a MAX patch so that each of the codons is culled from the data- base table and then played in real-time linear sequence as MIDI events. This process is analogous to the scanning of the mRNA by the ribosomes as it adds amino acids sequentially to make proteins—a process not unlike several cars (ribosomes) on a roller coaster negotiating the identical track (mRNA), but at different locations, speeds, and spacings. Polyphonic voices can occur just as multiple ribosomes run along a single strand of mRNA. At this point in our work, the computer performs the music on a Yamaha TX802 digital synthesizer according to a duration constant (the greater the constant, the longer each relative MIDI event).

4 Future work

Thus far, we have generated musical compositions for blood and liver cells, the polio virus, botulinin toxin (botulism), measles, rubella, four distinct common cold viruses, and the HIV virus (we have presently avoided most human proteins because of large amounts of uncoded filler found in between sequences). The next major goal is to realize the Smallpox (Variola) Virus (now extinct save for two vials in Atlanta and Moscow respectively). Because of its many distinct sequences and extreme length (20,000 base pairs), the MAX patch presently being used will require some modifications. Future plans also include the investigation of replacing MIDI events with real-time synthesis programming.