MUSICAL SYNTHESIS OF DNA SEQUENCES

By Peter Gena

pgena@artic.edu

&

By Charles Strom

DNAguy@aol.com

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.

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.

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 codons and their corresponding amino acid types, followed by the dissociation constant or pK(a) and molecular weight, was constructed as a data-base (Figure 1, below). There are eight

ISEA95 montréal isea95@er.uqam.ca T : (514) 990-0229

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. Pitch bend commands for each note place the music in just intonation.

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

$$f_1 = (([p - 7.0])\ (4G + 2T) + 12) + k$$

where:
f & f_1 = MIDI pitches
p = pK(a)
$G = \Sigma G + C$ per codon
$T = \Sigma A + T$ per codon
k = Constant (Hydrogen Bonds):
[AA = -2, TT = -1, CC = +1, CG = +2, GC = +3, GG = +4]

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.

$$I = 6H$$

$$I_1 = 109 - I$$

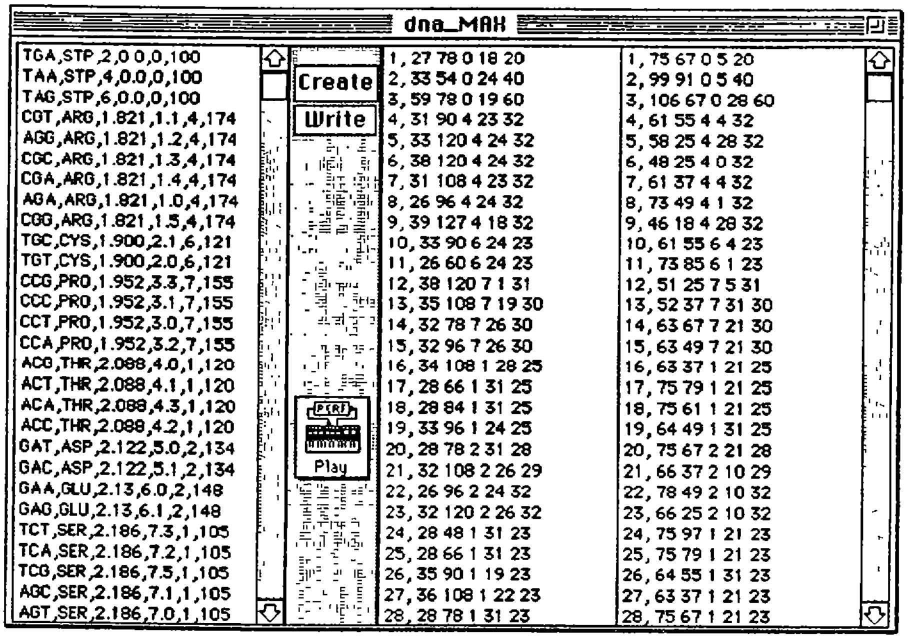
where:
I and I_1 are MIDI velocity levels
H is proportional to codon melting temperature and Hydrogen-bond strength per codon.
(each G = +8, C = +6, A = +4, T = +1)

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

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 codon, followed by its amino acid, pK(a), amino acid class numbers, and the molecular weight of the amino acid (Figure 1, first column). Each codon is transformed into a list in a collection (Figure 1, second and third columns). The list specifies the address, MIDI pitch, velocity, channel number, pitch bend, and duration of the event for the corresponding codon. 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 (Figure 2, left column). 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 codon index number is put into another collection (Figure 2).

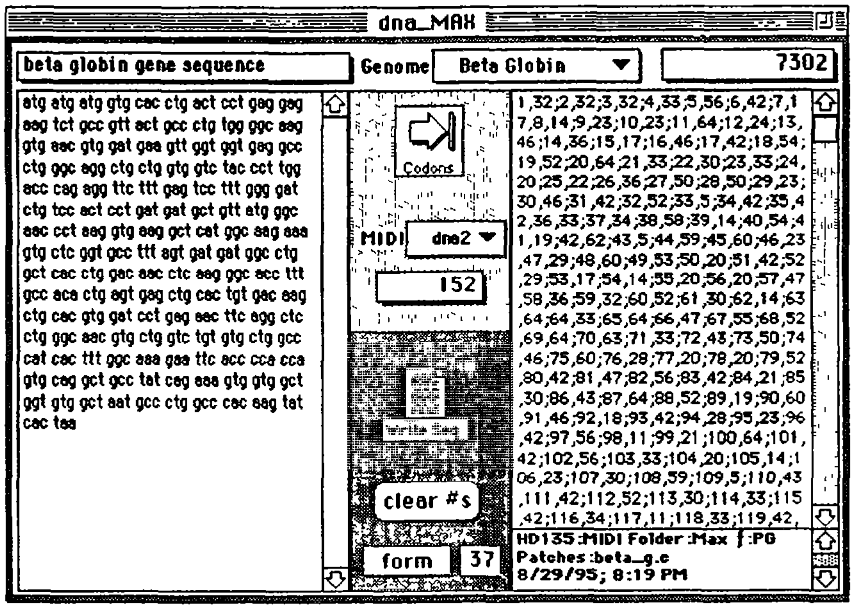


```
TGA,STP,2,0 0,0,100
TAA,STP,4,0.0,0,100
TAG,STP,6,0.0,0,100
CGT,ARG,1.821,1.1,4,174
AGG,ARG,1.821,1.2,4,174
CGC,ARG,1.821,1.3,4,174
CGA,ARG,1.821,1.4,4,174
AGA,ARG,1.821,1.0,4,174
CGG,ARG,1.821,1.5,4,174
TGC,CYS,1.900,2.1,6,121
TGT,CYS,1.900,2.0,6,121
CCG,PRO,1.952,3.3,7,155
CCC,PRO,1.952,3.1,7,155
CCT,PRO,1.952,3.0,7,155
CCA,PRO,1.952,3.2,7,155
ACG,THR,2.088,4.0,1,120
ACT,THR,2.088,4.1,1,120
ACA,THR,2.088,4.3,1,120
ACC,THR,2.088,4.2,1,120
GAT,ASP,2.122,5.0,2,134
GAC,ASP,2.122,5.1,2,134
GAA,GLU,2.13,6.0,2,148
GAG,GLU,2.13,6.1,2,148
TCT,SER,2.186,7.3,1,105
TCA,SER,2.186,7.2,1,105
TCG,SER,2.186,7.5,1,105
AGC,SER,2.186,7.1,1,105
AGT,SER,2.186,7.0,1,105
```

```
1, 27 78 0 18 20
2, 33 54 0 24 40
3, 59 78 0 19 60
4, 31 90 4 23 32
5, 33 120 4 24 32
6, 38 120 4 24 32
7, 31 108 4 23 32
8, 26 96 4 24 32
9, 39 127 4 18 32
10, 33 90 6 24 23
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
17, 28 66 1 31 25
18, 28 84 1 31 25
19, 33 96 1 24 25
20, 28 78 2 31 28
21, 32 108 2 26 29
22, 26 96 2 24 32
23, 32 120 2 26 32
24, 28 48 1 31 23
25, 28 66 1 31 23
26, 35 90 1 19 23
27, 36 108 1 22 23
28, 28 78 1 31 23
```

```
1, 75 67 0 5 20
2, 99 91 0 5 40
3, 106 67 0 28 60
4, 61 55 4 4 32
5, 58 25 4 28 32
6, 48 25 4 0 32
7, 61 37 4 4 32
8, 73 49 4 1 32
9, 46 18 4 28 32
10, 61 55 6 4 23
11, 73 85 6 1 23
12, 51 25 7 5 31
13, 52 37 7 31 30
14, 63 67 7 21 30
15, 63 49 7 21 30
16, 63 37 1 21 25
17, 75 79 1 21 25
18, 75 61 1 21 25
19, 64 49 1 31 25
20, 75 67 2 21 28
21, 66 37 2 10 29
22, 78 49 2 10 32
23, 66 25 2 10 32
24, 75 97 1 21 23
25, 75 79 1 21 23
26, 64 55 1 31 23
27, 63 37 1 21 23
28, 75 67 1 21 23
```

Figure 1: Partial Table of Codons



beta globin gene sequence

Beta Globin ▼

7302

```
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 gtt ggt ggt gag gcc
ctg ggc agg ctg ctg gtg gtc tac cct tgg
acc cag agg ttc ttt gag tcc ttt ggg gat
ctg tcc act cct gat gat gct gtt atg ggc
aac cct aag gtg aag gct cat ggc aag aaa
gtg ctc ggt gcc ttt agt gat gat ggc ctg
gct cac ctg gac aac ctc aag ggc acc ttt
gcc aca ctg agt gag ctg cac tgt gac aag
ctg cac gtg gat cct gag aac ttc agg ctc
ctg ggc aac gtg ctg gtc tgt gtg ctg gcc
cat cac ttt ggc aaa gaa ttc acc cca cca
gtg cag gct gcc tat cag aaa gtg gtg gct
ggt gtg gct aat gcc ctg gcc cac aag tat
cac taa
```

dna2 ▼

152

37

```
1,32;2,32;3,32;4,33;5,56;6,42;7,1
7,8,14;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;25,22;26,36;27,50;28,50;29,23;
30,46;31,42;32,52;33,5;34,42;35,4
2,36,33;37,34;38,58;39,14;40,54;4
1,19;42,62;43,5;44,59;45,60;46,23
,47,29;48,60;49,53;50,20;51,42;52
,29;53,17;54,14;55,20;56,20;57,47
,58,36;59,32;60,52;61,30;62,14;63
,64;64,33;65,64;66,47;67,55;68,52
,69,64;70,63;71,33;72,43;73,50;74
,46;75,60;76,28;77,20;78,20;79,52
,80,42;81,47;82,56;83,42;84,21;85
,30;86,43;87,64;88,52;89,19;90,60
,91,46;92,18;93,42;94,28;95,23;96
,42;97,56;98,11;99,21;100,64;101,
42;102,56;103,33;104,20;105,14;1
06,23;107,30;108,59;109,5;110,43
,111,42;112,52;113,30;114,33;115
,42;116,34;117,11;118,33;119,42,
```

HD135:MIDI Folder:Max ƒ:PG
Patches:beta_g.c
8/29/95; 8:19 PM

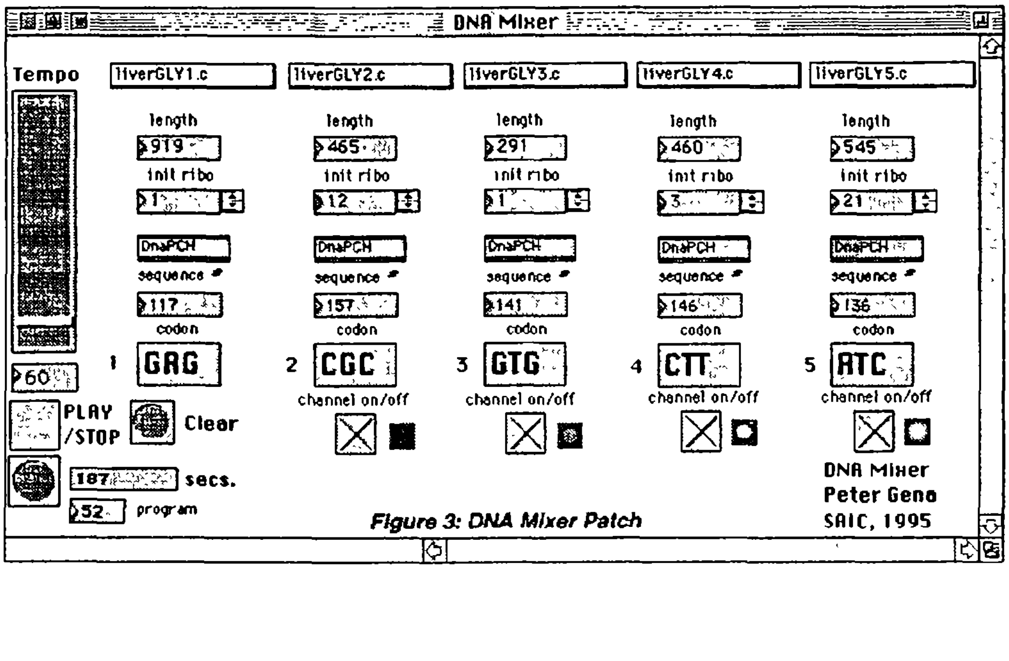
Figure 2: Genome for Beta Globin

Actual genomes of human or bacterial proteins, or complete viruses can then be scanned by the DNA Mixer patch (Figure 3) so that each of the codons is culled from the data-base table and then played in real-time linear sequence as MIDI events. The mixer can play up to five individual sequences at different starting points. 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).

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.

© Peter Gena 1995
School of the Art Institute of Chicago
Chicago, IL 60603

© Charles Strom 1995
Section of Medical Genetics
Illinois Masonic Medical Center
Chicago, IL



	1	2	3	4	5
	liverGLY1.c	liverGLY2.c	liverGLY3.c	liverGLY4.c	liverGLY5.c
length	919	465	291	460	545
init ribo	1	12	1	3	21
	DnaPCH	DnaPCH	DnaPCH	DnaPCH	DnaPCH
sequence #	117	157	141	146	136
codon	GAG	CGC	GTG	CTT	ATC
channel on/off	☒	☒	☒	☒	☒

Tempo 60

187 secs.

52 program

DNA Mixer
Peter Gena
SAIC, 1995

Figure 3: DNA Mixer Patch