A Physiological Approach to DNA Music

Peter Gena, Ph.D.
School of the Art Institute of Chicago

Charles Strom, M.D., Ph.D.
Medical Director, Biochemical and Molecular Genetics Laboratories
Nichols Institute / Quest Diagnostics
San Juan Capistrano, CA

Abstract*:* *As a consequence of the Human Genome Project, there has been an explosion of primary DNA sequencing data available on the internet. Five years ago, we envisioned a type of computer-generated music that would take cues for its musical parameters directly from the physiological ones present in DNA. The first paper,* ***Musical Synthesis of DNA Sequences,*** *was presented and published at the Sixth International Symposium on Electronic Art, 1995 in Montreal; and XI Colloquio di Informatica Musicale, 1995 in Bologna.*[1] *This updates our recent work.*

1. Introduction: DNA and RNA

With the exception of Prions, all known life forms on the planet use nucleic acid molecules (either DNA or RNA) to store genetic information. In eukaryotes, protozoans, yeast, and bacteria, the genetic material is invariably DNA, whereas some viruses use RNA as their genetic material. DNA molecules are comprised of long chains consisting of four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). In RNA, the thymine is replaced by uridine (U). The bases are linked to each other by phosphodiester covalent bonds to form the final genetic material. A DNA molecule contains two such chains wound around each other in a structure known as the double helix. In the double helix the base on one strand exactly determines the corresponding base on the opposite strand. Whenever a T residue is on one strand, an A residue will be exactly opposite to it on the other. When a G is on one strand, there will be a C on the complementary. The G-C and A-T pairs stabilize the double helix by forming hydrogen bonds with each other, thus keeping the double helix together. G-C pairs contain three hydrogen bonds and A-T pairs only two, making G-C pairs more stable than A-T pairs. The order of these bases contains the complete genetic blueprint for a given organism. Within a gene, the sequence of bases will specify exactly the amino acid sequence of a protein chain or RNA species. The exact ordering of amino acids in any protein chain is designated as the primary sequence.

Genetic words (codons) consist of a sequence of three base pairs, i.e. AAA specifies the amino acid lysine and GGG specifies glycine. In order to synthesize a protein the sequence on the DNA is read by a protein known as RNA polymerase and a special RNA molecule called messenger RNA (mRNA) is synthesized. The mRNA will have a complimentary sequence to the DNA template on which it was formed. For example, if the DNA contains the codon CCC the corresponding mRNA will have the sequence GGG. The genetic code (Figure 2) is the dictionary that translates the 64 possible three-base combinations in mRNA into their corresponding amino acids. The genetic code uses U instead of T because the codons refer to the mRNA sequence and not the DNA sequence. The DNA template, located in the nucleus of each cell, acts as a blueprint that directs the production of proteins. After the DNA is translated into mRNA, the mRNA is then serially scanned by ribosomes, organelles located in the cell's cytoplasm (Figure 1). Ribosomes use the mRNA as a template to direct the synthesis of proteins.

[1] Peter Gena and Charles Strom. *Musical Synthesis of DNA Sequences* . Sixth International Symposium on Electronic Art, Montreal, 1995: 83-85. XI Colloquio di Informatica Musicale, Univeristà di Bologna, 1995: 203-204.

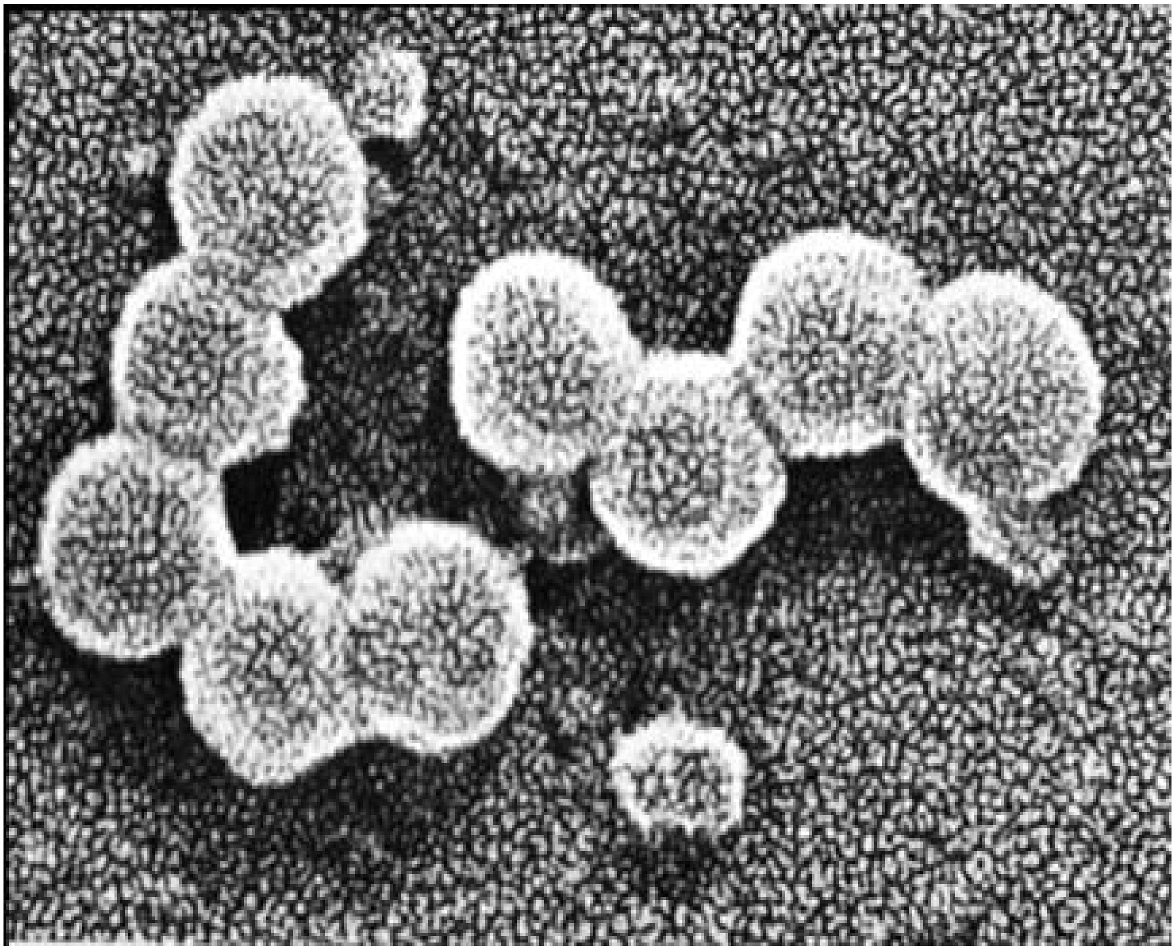


Figure 1: **Ribosomes** (electron microscope image)

As there are 64 possible codons (permutations of four bases taken three at a time) and only 20 possible amino acids, the genetic code is degenerate, meaning that more than one codon sequence can specify the same amino acid. For example there are four different codons that specify the amino acid leucine (GUU, GUC, GUA, GUG).

There are also four special codons. AUG codes for the amino acid methionine, but also acts as the "start" signal that alerts the cellular machinery to begin reading the code; and three codons, UAA, UAG and UGA are the termination codons that signal the machinery that it has reached the end of the gene and to halt chain elongation. The genetic code is also universal; all known life forms use the same genetic code, which is compelling evidence that all life on earth evolved from a common ancestor.

Although the primary amino acid sequence of a protein determines its chemical structure, the function of a protein is also determined by its three-dimensional conformation. Proteins fold up on themselves and have areas, or domains, that may anchor them in a cell membrane or cause them to be excreted. This is called secondary structure. Secondary structure is dependent on the chemical properties of each individual amino acid and how it relates to chemical properties of its neighbors. The most important contributors to the secondary structure of a protein are the physical properties of the amino acids with respect to their solubility in water or lipid (the cytoplasm is filled with water, and cell membranes are comprised of lipids). Amino acids that are more soluble in water are hydrophyllic (water loving) and those less soluble in water but more so in lipid are known as hydrophobic residues. Proteins fold so that the hydrophobic amino acids are on the inside, hidden from the water of the cytoplasm, and the hydrophyllic residues are on the outside. A membrane bound protein will have sequences containing several hydrophobic residues in succession. This hydrophobic domain will actually pass through a cell membrane so that the hydrophobic groups are dissolved in the lipid of the cell membrane and the hydrophobic groups will be exposed on the inner or outer surface of the cell. The extent to which an amino acid is hydrophobic or hydrophyllic is based on the side chain of the particular amino acid. If the side chain is charged (either positively or negatively) the residue will be hydrophyllic, and if the side chain is neutral the residue will be hydrophobic. Thus the 20 amino acids can be categorized as follows:

Aromatic (very hydrophobic)
Phenylalanine, Tyrosine, Tryptophan
Aliphatic (hydrophobic)
Glycine, Alanine, Valine, Leucine, Isoleucine
Aliphatic hydroxyl (less hydrophobic)
Serine, Threonine
Imino (less hydrophobic)
Proline
Sulfur containing (less hydrophobic)
Cysteine, Methionine
Amide (less hydrophyllic)
Asparagine, Glutamine
Basic (hydrophyllic)
Lysine, Arginine, Histidine
Acidic (hydrophyllic)
Aspartic Acid, Glutamic Acid

A further determination of secondary structure is the dissociation constant, or Pk(a), of the amino acid. This will determine the overall charge of a protein and of particular domains, and influence folding as well as enzyme activity. Pk(a) is a measure of the acidity of each residue and the pH at which the amino acid will release its proton into an aqueous solution. Each amino acid contains a different number of atoms, and therefore has a specific molecular weight. The molecular weight of any molecule is that of one molecular equivalent (Mole) of the substance. There are 6.02 X 1023 molecules in every Mole (Avogadro's number).

Any algorithm whose purpose is to convey information in a meaningful way regarding structure and function of proteins must take into account both the primary (amino acid sequence) and secondary (chemical properties of the amino acids, hydrophobic/hydrophyllic nature, Pk(a), and molecular weight) structures. We implemented an additional level of complexity by using the physical nature of the codons themselves. The codons that are more G-C rich will be more difficult to pull apart than A-T rich codons. This can be measured as the melting temperature of the codon and can be approximated by the formula of 4 (G + C) + 2 (A + T) in degrees Celsius.

2. Physio-musical conversion

An algorithm was designed to convert the list of sixty-four codons into distinct musical events according to the above-mentioned 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, serves as a database (Figure 2). There are eight different basic timbres at any time; one for each of the eight classes of amino acids.

2.1 Pitch/frequency

Each of the nineteen amino acids has a distinct Pk(a) [p] that helps define pitch or frequency [f & f_1]. Additional modifications involve physical properties of the molecular bonding occurring in the codon itself, independent of the amino acid for which it codes [$G = \sum G + C$ per codon; $T = \sum A + T$ per codon; k = Hydrogen Bonding constant]. 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 program makes a binary choice with each selection.

$$f = ((\, p\, (4G + 2T) + 12) + k \qquad (2.1)$$

$$f_1 = (([p - 7.0])\, (4G + 2T) + 12) + k \qquad (2.2)$$

2.2 Intensity

Intensities [I and I_1] are also adjusted according to the hydrogen bonding [H] occurring in each codon. As with pitch, there are two corresponding equations for each codon:

$$I = 6H \tag{2.3}$$
$$I_1 = 109 - I \tag{2.4}$$

2. 3 Duration

The Pk(a) [p] and atomic weights of the amino acids [M] and sum of the hydrogen bonding [S] times a tempo constant [k] determine duration in clock ticks (D):

$$D = 0.01pM + 0.1Sk \tag{2.5}$$

3.0 Realization

A preliminary program prepares all the necessary data, that is, the table of codons (Figure 2), and the genomes as collections for the MAX/MSP object code and DSP language (Copyright by IRCAM, Opcode Systems, and Cycling '74). 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.

```
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;
```

Figure 2: **Codons 11 - 16 from the look-up table**

Each codon is transformed into a list in a collection (Figure 3). The list specifies the address, pitch (frequency) number, intensity, timbre (amino acid class), and duration of the event for the corresponding codon.

```
11, 26 60 6 23;
12, 38 120 7 31;
13, 35 108 7 30;
14, 32 78 7 30;
15, 32 96 7 30;
16, 34 108 1 25;
```

Figure 3: **Corresponding MAX collection**

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 codon index number is put into another collection as follows:

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/MSP 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.

Complete genomes of human or bacterial proteins, or viruses are then scanned by a MAX/MSP patch, *DNA Mixer*, so that each of the codons is culled from the database table (Figure 3) and then played in real-time linear sequence. Each "channel" of the mixer is made up of a sub-patch that plays the sequence via the look-up (database) table. The event scheduling for each "track" is handled by an embedded sub-patch, which in turn uses yet another, a rhythm processor, to monitor duration. The present DNA mixer can execute up to six individual sequences at different starting points. This process is analogous to the scanning of the mRNA by the ribosomes [Figure 1] 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. This creation of polyphony is analogous to the way multiple ribosomes run along a single strand of mRNA.

4.0 Recent Music

From the onset, I believed that the musical reading of DNA ought to be rendered literally. As the sequences represent life of many sorts, I am reluctant to tamper with the "score." The DNA mixer can realize sequences as digital sound and/or print them out in musical notation. Ideally, performances of the synthesized pieces should be done live from the computer, where the ribosome simulations can be set spontaneously before playing. Notes on the early work based on human proteins and enzymes (red blood cells, liver proteins, elastin, etc.) and viruses (HIV, polio, measles, rubella, common colds, botulism, etc.), and works with instruments (***Collagen and Bass Clarinet*** and ***Chopin's Catarrh***—using the cystic fibrosis gene as a cantus firmus sought in the Nocturnes) can be found in YLEM (Sept/Oct. '99) at http://ylem.org/NewSite/archive/issuethmbs/newsletters/SeptOct99/article2.html.

As the first sequences were initially conceived for concert performance, two of the early pieces (***Red Blood Cells,*** 1995, and ***Polio Virus,*** 1998) were placed in the **Collage Jukebox 2.1**, installed at the Galerie ERSEP/Université Lille 3, in Tourcoing, France, April 27 to May 20, 2000. The actual hi-fidelity jukebox contained selections from 307 artists worldwide.

The first use of the DNA mixer in an installation was for ***Genesis*** (1999), an interactive piece by Eduardo Kac with DNA manipulation by Dr. Strom. The premiere took place at *Ars Electronica* (Linz), on September 4 -19, 1999, and subsequently in São Paulo and New York City. The mixer used for Genesis utilizes a mutation factor (changing a different DNA signifier after approximately each 100,000 base pairs) during the two-week exhibition. Three sequences: the genesis gene (a short text from Genesis, translated into DNA code by Kac and Strom, and synthesized by Strom into plasmid), the cyan and the yellow plasmids can loop infinitely. Timbral changes were made whenever a website user switched on the UV light over the projected live plasmid, which sped up mutations (Figure 4). In addition, as participants controlled the light from the website, the tempo of the sequence gradually increased to a maximum, then worked its way down again (see http://www.ekac.org/dnamusic.html).

There were two recent instrumental DNA works premiered in 2000. ***Progesterone/Testsoterone*** combined the electronic playback of the two hormones with live instrumental improvisation (winds, piano, bass, and percussion). ***La Peste per oboe d'amore*** (and computer-generated Yersinia Pestis sequences), was performed at the Talloires International Composers Festival in Annecy, France in June. The oboist who commissioned the work, from Chicago's CUBE Ensemble, specifically requested the bubonic plague. Though the Yersinia Pestis sequence is rather large, consisting of some 35,800 base pairs, it conveniently sub-divided into six parts. Thus, one of the six was outputted as musical notation (properly transposed) for the oboe d'amore, and the remaining five "channels" were realized and played electronically by the DNA mixer.

Him, Himself and He, is a recent installation commissioned for **From Steel to Flesh**, an exhibit held from February 5 to March 2, 2001 on the occasion of the first **Miss USA Pageant,** under the auspices of the Trump Casino in Gary, Indiana. The show, which took place in the Contemporary Art Gallery, adjacent to the facilities used by the fifty beauty contestants at Indiana University Northwest, had *Masculinity* as its theme. For ***Him, Himself and He,*** I chose five male genes: the SrY (sex-determining region of the Y chromosome) the anti-Müllerian hormone, the Androgen Receptor, Dihydrotestosterone (5 Alpha Reductase Enzyme) and the female pseudo-hermaphrodite gene (steroid 21-hydroxylase). The Mixer automatically chooses among the five genes, or four other fixed combinations of them. These combinations reflect the natural flow of the genes in action. The viewer/listener has the opportunity to override the program and choose among the nine entries, or “User Mix” from the menu (Figure 5). Selecting the latter allows a choice of mixing up to eight of the five genes together, while starting the sequences at varying points, and choosing a preferred tempo (in beats per minute). There are a total of 37 timbres that the mixer has at its disposal, selecting eight new ones each time a fixed sequence or user mix begins. In ***DNA Music Installation,*** like the version presented at CADE 2001, the process and DNA mixer used is the same as in ***Him, Himself and He***. However, it usually offers from 18 to 25 sequences, and a palette of up to 64 timbres.

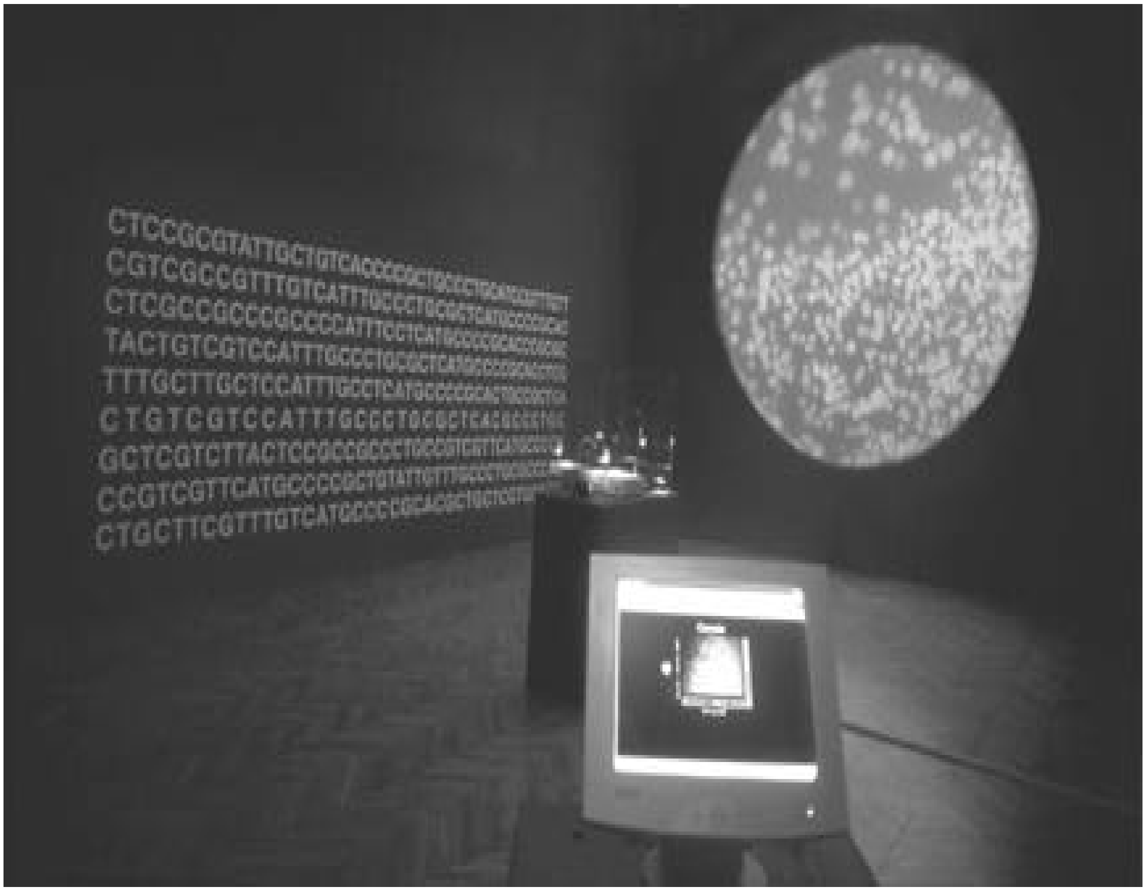
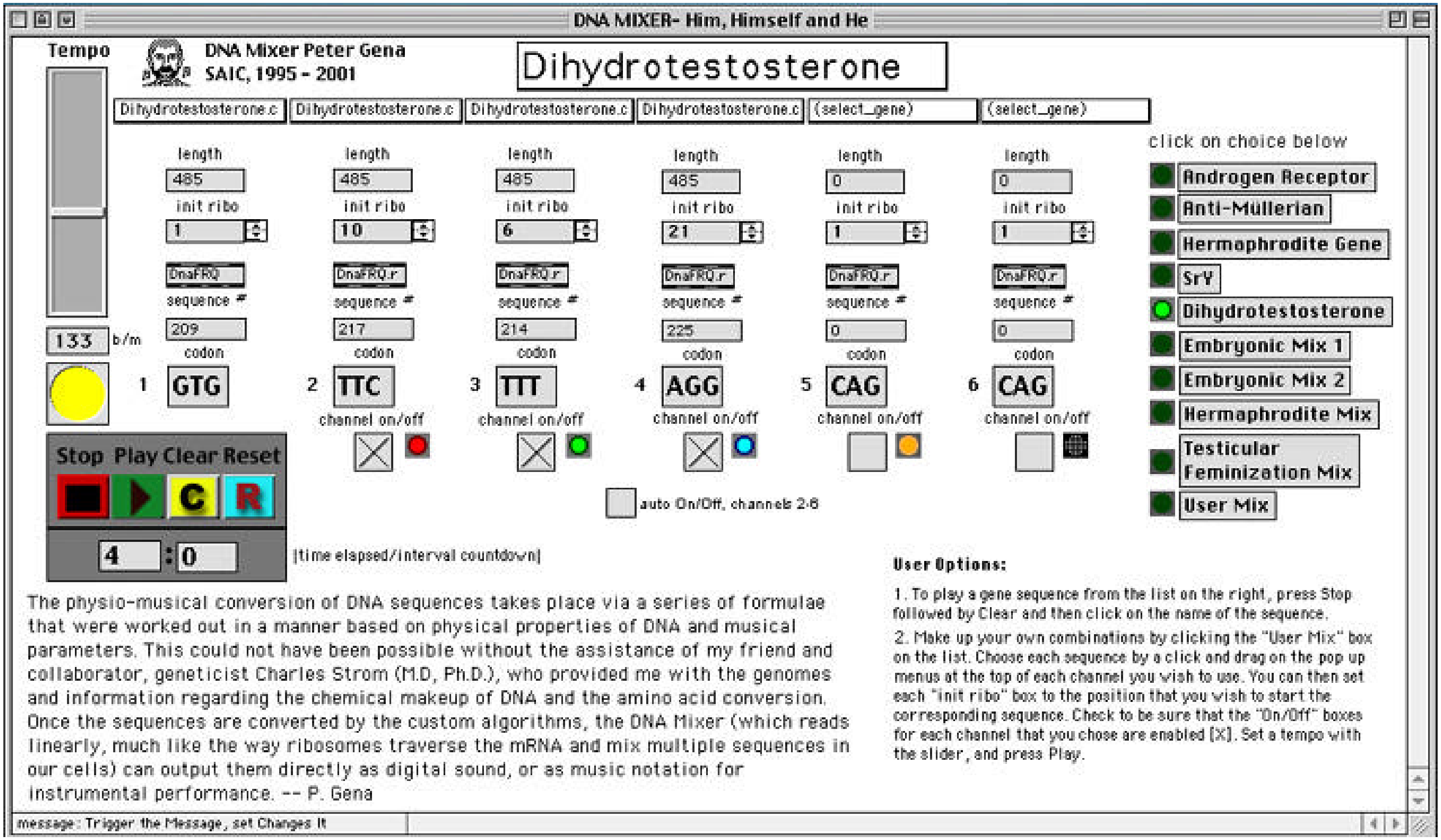

CTCCGCGTATTGCTGTCACCCCGCTGCCCTGCATCGTTGCT
CGTCGCCGTTTGTCATTTGCCCTGCGCTGATGCCCCGC
CTCGCCGCCCGCCCCATTTCCTCATGCCCCGCACCGCGC
TACTGTCGTCCATTTGCCCTGCGCTCATGCCCGCACCTG
TTTGCTTGCTCCATTTGCCTCATGCCCCGCACTGCCCTGA
CTGTCGTCCATTTGCCCTGCGCTCACGCCCTGC
GCTCGTCTTACTCCGCCGCCCTGCCGTCGTTCATGCCG
CCGTCGTTCATGCCCCGCTGTATTGTTTGCCCTGCGCC
CTGCTTCGTTTGTCATGCCCCGCACGCTGCTCGT

Figure 4: ***Genesis*** (1999) by Eduardo Kac, with DNA manipulation by Charles Strom, and music by Peter Gena (O.K. Centrum für Geganwartskunst Oberösterreich, 1999)

Figure 5: DNA Mixer: *Him, Himself and He* (2001) – next page



Dihydrotestosterone

133 b/m

	1	2	3	4	5	6
	Dihydrotestosterone.c	Dihydrotestosterone.c	Dihydrotestosterone.c	Dihydrotestosterone.c	(select_gene)	(select_gene)
length	485	485	485	485	0	0
init ribo	1	10	6	21	1	1
	DnaFRQ	DnaFRQ.r	DnaFRQ.r	DnaFRQ.r	DnaFRQ.r	DnaFRQ.r
sequence #	209	217	214	225	0	0
codon	GTG	TTC	TTT	AGG	CAG	CAG
channel on/off		☒	☒	☒	□	□

□ auto On/Off, channels 2-6

4 : 0 |time elapsed/interval countdown|

The physio-musical conversion of DNA sequences takes place via a series of formulae that were worked out in a manner based on physical properties of DNA and musical parameters. This could not have been possible without the assistance of my friend and collaborator, geneticist Charles Strom (M.D, Ph.D.), who provided me with the genomes and information regarding the chemical makeup of DNA and the amino acid conversion. Once the sequences are converted by the custom algorithms, the DNA Mixer (which reads linearly, much like the way ribosomes traverse the mRNA and mix multiple sequences in our cells) can output them directly as digital sound, or as music notation for instrumental performance. -- P. Gena

User Options:

1. To play a gene sequence from the list on the right, press Stop followed by Clear and then click on the name of the sequence.
2. Make up your own combinations by clicking the "User Mix" box on the list. Choose each sequence by a click and drag on the pop up menus at the top of each channel you wish to use. You can then set each "init ribo" box to the position that you wish to start the corresponding sequence. Check to be sure that the "On/Off" boxes for each channel that you chose are enabled [X]. Set a tempo with the slider, and press Play.