

The Musical Gene: Generating Harmonic Patterns from Sequenced DNA of *E.coli* Bacteria to Compose Music

Frederic Bertino^{*}, Ching-Hua Chuan[†] and Jared Peroune[‡]

^{*}Department of Biology, [‡]Department of Music

[†]Department of Mathematics and Computer Science

Barry University

Miami Shores, USA

Frederic.Bertino@mymail.barry.edu, chchuan@mail.barry.edu

Abstract— Our research concerns the blending of three major areas of computer science, biology, and music, to encourage learning about natural patterns and algorithms in living and natural systems. Our goal was to reveal aspects of the amino acid patterns that cannot be as easily detected through other means, and to reveal some parallel between the structures of music and Deoxyribonucleic acid (DNA). We created a computer program to generate the backbone chord progression for a piece of tonal music based on the amino acid sequence of a DNA strand. The chord choices produced by the program reflect the inherent characteristics of the biological sequence. The computer program took the data from a translated nucleotide sequence (LysA gene of the *E. coli* bacterium) and converted the amino acid derivative into a chord progression, which then was arranged in the spectrum of time and orchestrated. The fundamental theory of diatonic chords, frequency, and cadence were taken into account when assigning the amino acids to their respective chord. An atonal piano concerto was generated and orchestrated from the harmonic progression derived from the amino acid sequence. This indicates that genetic material possesses a similar backbone to that of music and can be translated as such.

[keywords]: *DNA Music, Amino Acid Sequence, Western Tonal Music Theory, DNA Piano Concerto, Lysine A, E. coli.*

I. INTRODUCTION

The difficulty that laypeople have in understanding the complex nature and vital function of DNA inspired this study. In contrast, many researchers have observed that most people are very sensitive to tonality and structural information in music, even if they have no formal musical training. This research was designed to aid biologists in discovering the possibility of reoccurring patterns and algorithms within DNA to reveal patterns behind what was believed to be random assimilation of the molecules when DNA is generated within the cell. Because of humankind's sensitivity to tonal music, it is believed that tracking trends in harmonic patterns derived from an amino acid backbone would facilitate the discovery of similarities in various parts of a DNA sequence.

DNA is a complex molecular arrangement of four nucleotide bases: Adenine (A), Cytosine (C), Thymine (T), and Guanine (G). The nucleotide sequences of these bases are very long, with some gene sequences containing thousands of base pairs. When DNA is transcribed to RNA (A single strand nucleic acid) it can then be translated to proteins by the construction of an amino acid sequence. Three nucleotides (in any arrangement) give rise to an amino acid codon, a three-letter code for any one of the 20 naturally occurring amino acids in living systems. Codons of different nucleotides can sometimes code for the same amino acid, while other codons are designed to start and stop the translating process. Amino acid sequences code for proteins. The correlative observations made in this study were that the general rules of Western Tonal Theory in music had similar facts and figures to those of DNA. There are 24 triads (major/minor) and similarly, 20 amino acids that could have chords easily assigned to them. A chord triad within the major/minor spectrum consists of three pitches, just as amino acid codons consist of three nucleotides. These similarities led us to coordinate the two disciplines through a technological medium.

We took the DNA sequence of the gene *LysA* in *E.coli* bacteria and searched for recurring patterns in its amino acid sequence after codon translation. After translating the DNA sequence, a computer program was written in MATLAB to assign a chord triad to a respective amino acid based on the amino acid's frequency of occurrence within the translated DNA sequence. Because a codon or "triad" classifies amino acids, it would be appropriate to group these amino acids to a chord triad. Amino acids vary based on the order and contents of the nitrogenous bases involved, similar to musical chords (triads) and the pitches that form them. Musical triad assignment was based on a logical theoretical foundation within the study of music theory such that the most frequently occurring amino acid would be assigned the tonic chord, or root chord of the piece's intended key. This was followed by the dominant and sub-dominant chords being more common than chords that would not normally be found in the piece's key.

In western music theory, harmony provides the backbone by which music obtains its character. Elements of orchestration, melody, and time are facets, which the artist interprets to make the piece desirable. The amino acid sequence generated the harmonic progression (musical backbone) that was later orchestrated. Inspiration for the concerto's orchestration, rhythmic progression, and melody (that are built upon the progression) came from a variety of famous composers, such as Gershwin, Liszt, Chopin, Prokofiev, Rachmaninoff, and Beethoven, to name a few. However, the harmonies within the piece came only from the amino acid sequence without any deviation from the scientific data obtained. For a piece of this magnitude, and its origins in the molecular nature of life, we wrote a concerto for full orchestra and piano, and captured the "whimsical" nature of the wonders of science (as well as the helical structure of DNA). These contained constantly ascending and descending moments of tension and release, and ascending and descending arpeggiated patterns.

II. RELATED WORK

Hayashi and Munakata [4] assigned the nucleotide bases to pitches in the interval of a fifth according to the thermal stability of the nitrogenous bases. It was found that research subjects could recall DNA sequences and specific patterns were more easily after having heard the sequence. Dunn and Clark [2] have been assigning amino acids to pitches based upon chemical principles such as molecular weight, thermal stability, molecular volume, and biochemical category and have generated music. These two research projects differ from this study in that the composers have used the element of melody (by assigning amino acid or nitrogenous base) to a specific pitch, whereas this research couples harmonic property with amino acid property, as the harmony of a musical composition defines its character better than variables of pitch. In a set harmony, pitch can be varied and different melodies can be obtained. Should harmony be a variable, the melody must conform to the changes in the fundamental musical backbone.

III. MATERIALS AND METHODS

A. Deoxyribonucleic Acid

DNA is the backbone of living systems in that it contains genetic blueprints that promote the construction of proteins and facilitate biochemical processes in living organisms. DNA consists of sugar-phosphate spines spiraled about, and connected by base pairs at the center. The base Adenine (A) binds to Thymine (T), and Cytosine (C) binds to Guanine (G) forming the double-helical structure of the molecule. Base pairing remains consistent (A, T and C, G); however, the ordering and varying of these pairs is infinitely variable. DNA is translated into genetic codes of Amino acids based on its genetic makeup within living cells. A group of three base pairs in any order constitutes an amino acid, which is used to create life proteins. Amino acids, though comprised of three nucleotides, are often abbreviated with one letter.

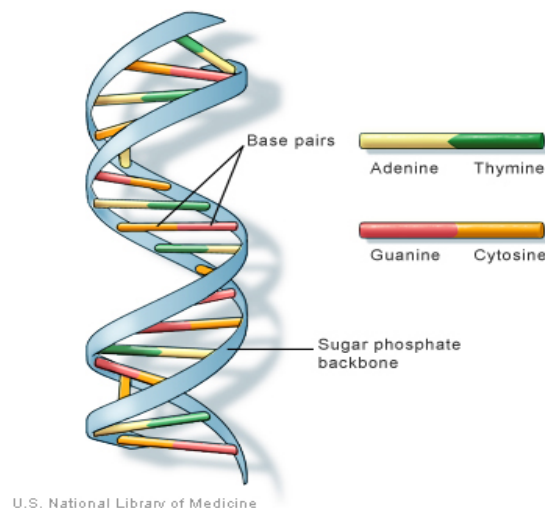


Figure 1. DNA with base pairing A-T, C-G..

Musical chord triads, consisting of three tones spaced from each other by an interval of a third, have a structure parallel to amino-acid codons. These chords within a musical scale, all have properties in relation to a tonal center, or the tonic note or chord of the respective key in which the chord lies. We call these chords diatonic in that they reside within a specific scale. Chords within a scale are labeled in a roman numerical system to identify their function within a key. Capital roman numerals are major chords where lower-case roman numerals are minor chords. The corresponding number has to do with the scale degree, where, in relation to the tonic, this chord lies.

B. Mapping Amino Acids into Triads

In an effort to search for naturally occurring patterns within a DNA sequence, we translated the sequence consisting of amino acids into a harmonic chord progression such that the musical role of chords mimics the biological patterns. During the process of mapping, we focus on the frequency of occurrence of amino acids since this frequency is considered one of the most important features for describing a DNA sequence [1]. We first analyzed the DNA sequence by counting each type of amino acid appearing in the sequence. In this study we chose the DNA sequence of the *Lysine A* protein in *Escherichia Coli*, due to its availability in the free online genetic database, ECOGENE. The amino acid frequency of occurrence in *Lysine A* is shown in Figure 2 where each amino acid is labeled using one-letter symbol.

After obtaining the frequency of occurrence, we mapped each amino acid to a triad as follows. The more frequently occurring amino acid is represented by a triad that is considered more important in tonal music. Figure 3 illustrates the *Tonnetz* (or Harmonic Network) [5] where closely related chords are geographically adjacent to each other. The mapping scheme starts from the center of the *Tonnetz*, using the triad tonic to represent the most frequently

occurring amino acid. The relative minor triad (vi) is chosen for the second-most frequent one, and dominant (V) and subdominant (IV) are used for the third and fourth. The mapping scheme next travels on *Tonnetz* in two directions, moving the center to the dominant first and then the subdominant, and selects their relative minor and dominant/subdominant for the next frequently occurring amino acids.

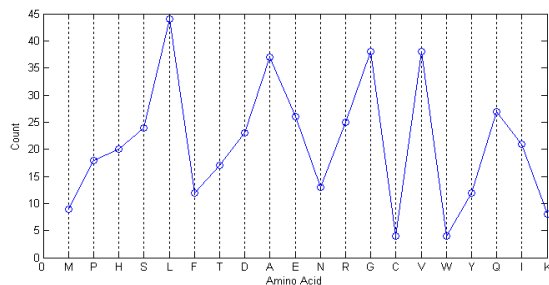


Figure 2. Amino Acid frequency of occurrence in *Lys A*.

The mapping scheme is summarized as a tree structure shown in Figure 4. To retrieve the triads in order, the tree can be read from the top to the bottom first, and then from the left to the right. For instance, the triad I is used for the most frequently occurring amino acid, followed by vi, V, IV, iii, ii, II, VII^b, etc. In this way, the more frequently occurring amino acids are assigned chords closer to the tonic, and the less frequent ones are mapped to chords that are less used according to the tonal center. The first 112 chords generated using the mapping scheme are shown in Figure 5.

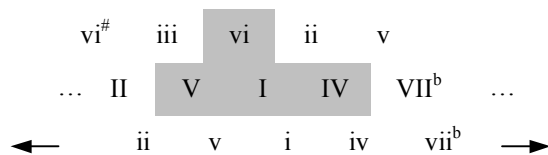


Figure 3. A subset of *Tonnetz*: closely related chords are located adjacently.

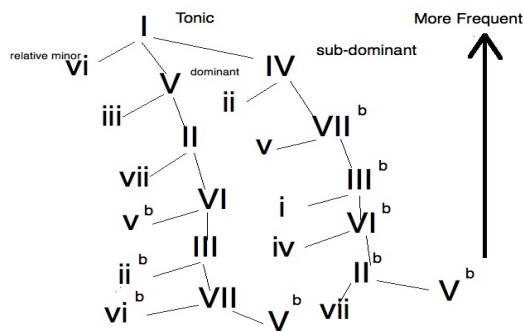


Figure 4. The tree structure used for mapping amino acids into triads based on the frequency of occurrence.

C. From the harmonic backbone to a piano concerto

The orchestrated product was derived from the harmonic backbone by manipulation of orchestral, melodic, and musicological themes and influences by the scientist-composers. The translated amino acid sequence yielded only the harmonic chord progression. To create an orchestrated work, this was easily interpreted artistically using the chord progression as the musical backbone for further analysis and composition.

Interpreting the harmonic backbone into an orchestral piece relied on drawing musical influence from major classical composers of the 19th and 20th centuries combined with prior knowledge of orchestral techniques, musicality, and theory. This prevented the melodic themes and overtones from skewing the harmonic backbone of the genetic information in any way.

After the score's composition in Sibelius® Music Composition Software, brief harmonic analysis was applied as proof reading to verify that chords from the genetic sequence remained unaltered.

Rhythmic practices were independent of the genetic material, though the orchestral interpretation of the piece limited the chord changes to one chord per measure.

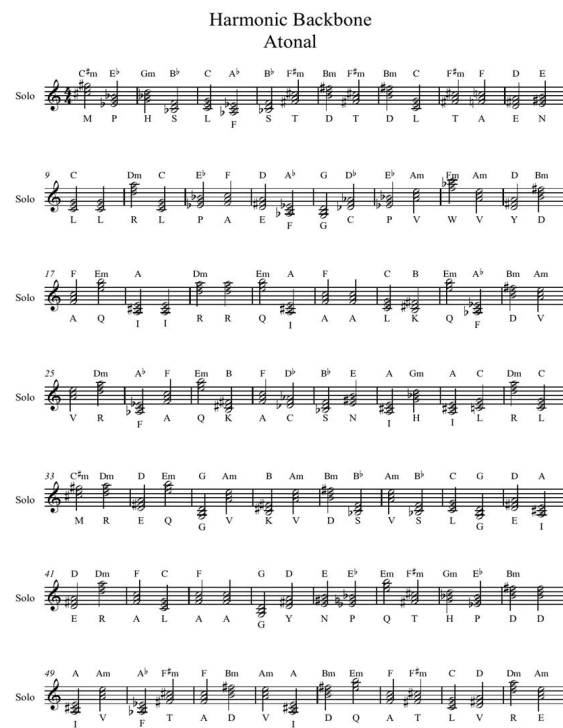


Figure 5. The first 112 chords in the harmonic backbone of the composition.

IV. RESULTS AND DISSCUTIONS

Musically, the results of the study proved very conducive to a theoretically sound piece of music. From the composer's standpoint, the composition contains a very apparent element

of tension and release, in that the genetic harmonic backbone established a tonal center (in this case, the amino acid Lysine was tonic) and the supporting chords served in moving the music to and from the tonal center as the piece progressed. The piece is to be considered atonal, though not in the same respect as Schoenberg's 12-tone row concept, but atonal in that every chord within the chromatic scale, major and minor, was employed into the work. Assigning a chord's function based upon its frequency of occurrence within the amino acid sequence gave the finished work a level of solidity with a hierarchal progression through diatonic and chromatic chords. This quality is just as one would see in a piece by any classical composer of the modern era—all with influences and lessons learned from previous classical masters. The piece is, unlike some classical works, very easy to listen to because the harmonic principle of locating and adhering to a tonal center was strictly followed.

The resulting chord progression was nothing more than what the genetic information possessed. The chord sequence is simply a musical representation of the amino acid sequence.

V. CONCLUSIONS AND FUTURE WORK

The DNA Concerto derived from the Lysine A Gene within *E.coli* bacterium was conceived only from the genome itself. While human involvement rendered the final piece's character, the harmonic backbone derived from the amino acid sequence was kept unaltered. The work's harmonic progression could be altered if the mapping procedure (done by frequency of amino acid occurrence

within the genome) was reassessed to map in a different way. The DNA concerto suggests that DNA, through its somehow random nature, has elements of patterns and algorithms within it, as observed by its harmonic progression within the framework of an orchestral piece.

To better assess the algorithmic possibilities of genetic material, the mapping system used will be re-evaluated to map based on a more biologically relevant plane than the frequency of amino acid occurrence, taking chemical and biological principles into a higher account, than previous research has done.

REFERENCES

- [1] R. Brooker, *Genetics: Analysis and Principles*, McGraw-Hill Science/Engineering/Math, 3rd edition, 2008.
- [2] J. Dunn and M. A. Clark, "Life Music," *The Sonification of Proteins*, Leonardo, 32(1), pp. 25-32, 1999.
- [3] T. Eerola and P. Toivainen, "MIDI Toolbox: Matlab Tools for Music Research," University of Jyväskylä: Kopijyvä, Jyväskylä, Finland. 2004
- [4] K. Hayashi and N. Munakata, "Basically Musical", *Nature*, 310(96), 1984.
- [5] H. C. Longuet-Higgins and M. J. Steedman, "On Interpreting Bach," *Machine Intelligence*, pp. 6-221, 1971.
- [6] B. Pierce, *Genetics: A Conceptual Approach*, 3rd edition, W. H. Freeman, 2007.
- [7] S. Kostka and D. Payne, *Tonal Harmony*, 5th edition, McGraw Hill, 2004.
- [8] R. Takashai, J. H. Miller, and F. Pettie, gene2music, http://www.mimc.ucla.edu/faculty/miller_jh/gene2music/