Correspondence

Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns

Rie Takahashi and Jeffrey H Miller

Address: Department of Microbiology, Immunology and Molecular Genetics and the Molecular Biology Institute, University of California, Los Angeles, CA 90095-1489, USA.

Correspondence: Rie Takahashi. Email: gene2music@gmail.com

Published: 3 May 2007

Genome ***Biology*** 2007, **8:**405 (doi:10.1186/gb-2007-8-5-405)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/5/405

Abstract

We have converted genome-encoded protein sequences into musical notes to reveal auditory patterns without compromising musicality. We derived a reduced range of 13 base notes by pairing similar amino acids and distinguishing them using variations of three-note chords and codon distribution to dictate rhythm. The conversion will help make genomic coding sequences more approachable for the general public, young children, and vision-impaired scientists.

In an effort to make science appealing to a wider audience, interdisciplinary groups have combined efforts to initiate novel approaches for the presentation and perspective of scientific material. An example is that of Victor Wong, a blind meteorology graduate student studying at Cornell University. He developed a computer program that translates different colors of a weather map into 88 distinct piano notes. With the use of a stylus to scan across a weather map, Wong was able to hear a gradation of colors ranging from blue to red with respect to electron density [1]. Another example of an interdisciplinary approach involves Japanese biologists at the RIKEN Center for Developmental Biology in Kobe, who have incorporated basic concepts of developmental biology into card games based on manga characters like Pokémon to interest young people. Aside from the amusing, colorful characters, the creators hope to preempt the introverted, asocial stereotypes of scientists before they "take root" [2]. Also, the *Biochemist's Songbook* by Harold Baum describes scientific concepts with lyrics and song [3].

In the context of basic research, a conversion from genomic sequences to music would open a door for the vision-impaired to study genomic biology. An auditory presentation could also be a means of exposing students to the concepts of DNA sequences and protein sequences at an earlier age through the use of auditory characteristics such as length, tempo, and dynamics. Some studies have attempted to transpose DNA sequences directly to music [4]. This approach suffers from a limited number of notes based on nucleotides composed of only four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Although the DNA could be read as a note for every two or three consecutive bases, this would focus the melodies more on DNA sequence organization and be less informative than looking at the coded information *per se*. Moreover, the result creates a string of notes that has no recognizable theme or musical depth as a composition. Other attempts to convert DNA sequences to music have used mathematical derivations based on the physical properties of the individual nucleotides in codons to create a set of equations for translating DNA sequences to musical notes [5,6]. A number of studies have dealt with pure protein sequences [7-10]. For example, Dunn and Clark used algorithms and the folding patterns of proteins to translate amino-acid sequences into musical themes [9]. Such an assignment creates a range that spans two to four octaves. Notes spanning such large ranges typically yield scores that lack musicality. They also examined a nine-note scale, but without distinguishing among amino acids having the same note value [10].

The goal of our work is to find a mode of converting genomic sequences

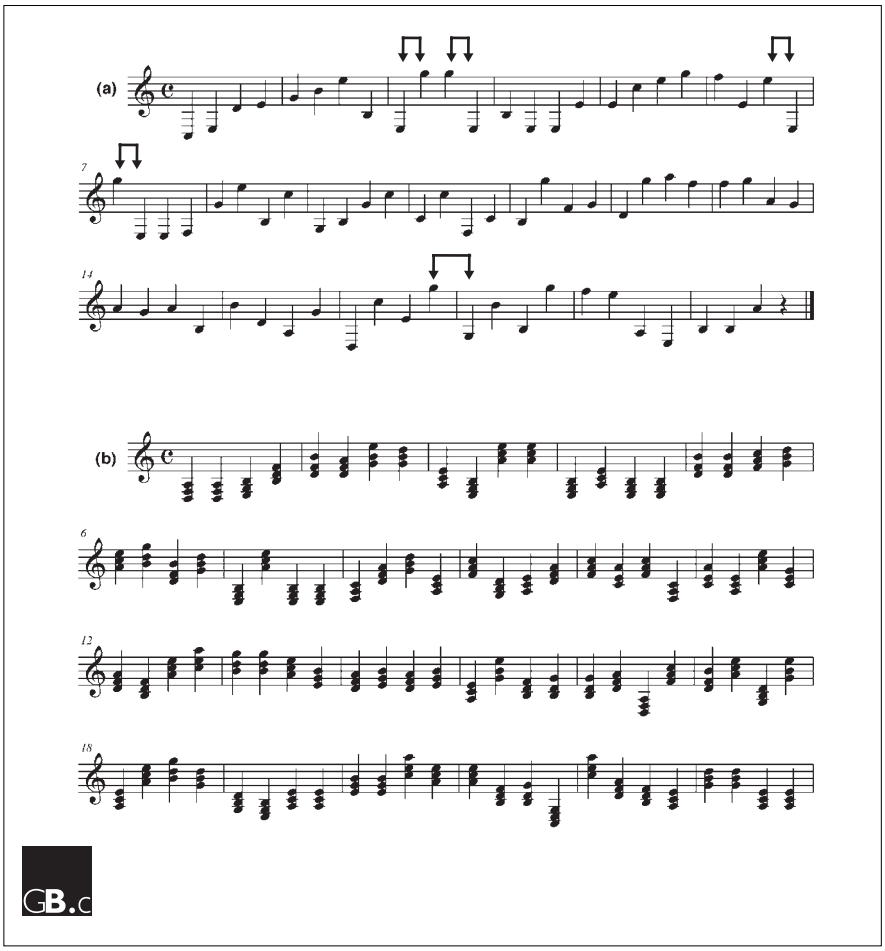

Figure 1
Human thymidylate synthase A (ThyA) protein sequence converted into single notes based on a 20-note range. **(a)** Amino acids were assigned a musical note starting an octave below middle C and based primarily on the hydrophobicity of the particular amino acid (Trp-C, Met-D, Pro-E, His-F, Tyr-G, Phe-A, Leu-B, Ile-C, Val-D, Ala-E, Cys-F, Gly-G, Thr-A, Ser-B, Gln-C, Asn-D, Glu-E, Asp-F, Arg-G, Lys-A). Having a one-to-one ratio of amino-acid assignment to musical notes results in a range that spans 2.5 octaves. Though this code may initially be the most obvious assignment, the approach leaves large jumps between consecutive notes as pointed out by the arrows. The large intervals occur sporadically and tend to interrupt any cohesive melody that may be heard. The 20-note range assignment also limits musicality and the ability to create a memorable theme. **(b)** Partial human ThyA protein sequence converted into chords based on a reduced-note assignment. Certain similar amino acids were paired and assigned a three-note chord (triad) starting an octave below middle C. Each member of the amino-acid pair was distinguished from the other by using different variations of the same fundamental triad, namely the root position (RP) and first inversion (FI) chord. Amino acids were assigned to the following notes: Trp-C, Met-D, Pro-E, His-F, {Tyr-G (RP), Phe-G (FI)}, {Leu-A (RP), Ile-A (FI)}, {Val-B (RP), Ala-B (FI)}, Cys-C, Gly-D, {Thr-E (RP), Ser-E (FI)}, {Gln-F (RP), Asn-F (FI)}, {Glu-G (RP), Asp-G (FI)}, {Arg-A (RP), Lys-A (FI)}. The result is a reduced, 13-base note range that minimizes the interval jumps between consecutive notes and produces a fuller sound with the use of the triads based on a particular key signature. For example, tyrosine is represented by a G major, root position triad.

(including coding and, eventually, non-coding) to piano notes that sound reasonable to a musician's ear while remaining faithful to the science of the protein sequences. The classic problem to overcome is the jump between consecutive notes as a consequence of the 20-note range when each amino acid is represented by a unique note. The wide range of the notes results in melodies that have many large, sporadic jumps, making them difficult to follow musically. A second problem is the question of how to incorporate rhythm into the sequence of notes. We describe here several innovations in coding assignments that generate a reduced note range and that also introduce rhythm into the sequence of notes.

Our pilot study focused on the amino-acid sequence of the human thymidylate synthase A (ThyA) protein. We used numerous assignments, including the chromatic scale, before finalizing our coding assignment based on a diatonic scale. Figure 1a shows the beginning portion of this sequence fixed to a 20-note range (2.5 octaves), where each amino acid was initially assigned to a unique note. One way to improve the musicality is to express each amino acid as a chord, rather than a single note. We then devised a reduced note range using chords, in which similar amino acids were paired initially. Thus, aspartic acid and glutamic acid were paired, as were leucine and isoleucine, tyrosine and phenylalanine, valine and alanine, threonine and serine, glutamine and asparagine, and arginine and lysine. The paired amino acids were assigned the same fundamental single note, but distinguished by being given a different version of their respective chord. For example, tyrosine and phenylalanine are both assigned a G major chord. The paired amino acids are distinguished from each other by either being assigned to a root position or first inversion chord of the same key signature. Tyrosine is assigned a G major root position (RP) chord and phenylalanine is assigned to a G major first inversion (FI) chord. The initial 13 base notes, assigned roughly according to hydrophobicity, yielded the music for ThyA shown in Figure 1b (see legend). Although the complete range of notes included in the chords spans more than 13 notes, the use of triads modulates the sound of the large jumps and range in addition to increasing the complexity of the music.

The next step was to add rhythm, which we did by referring to the coding sequence shown for humans and assigning one of

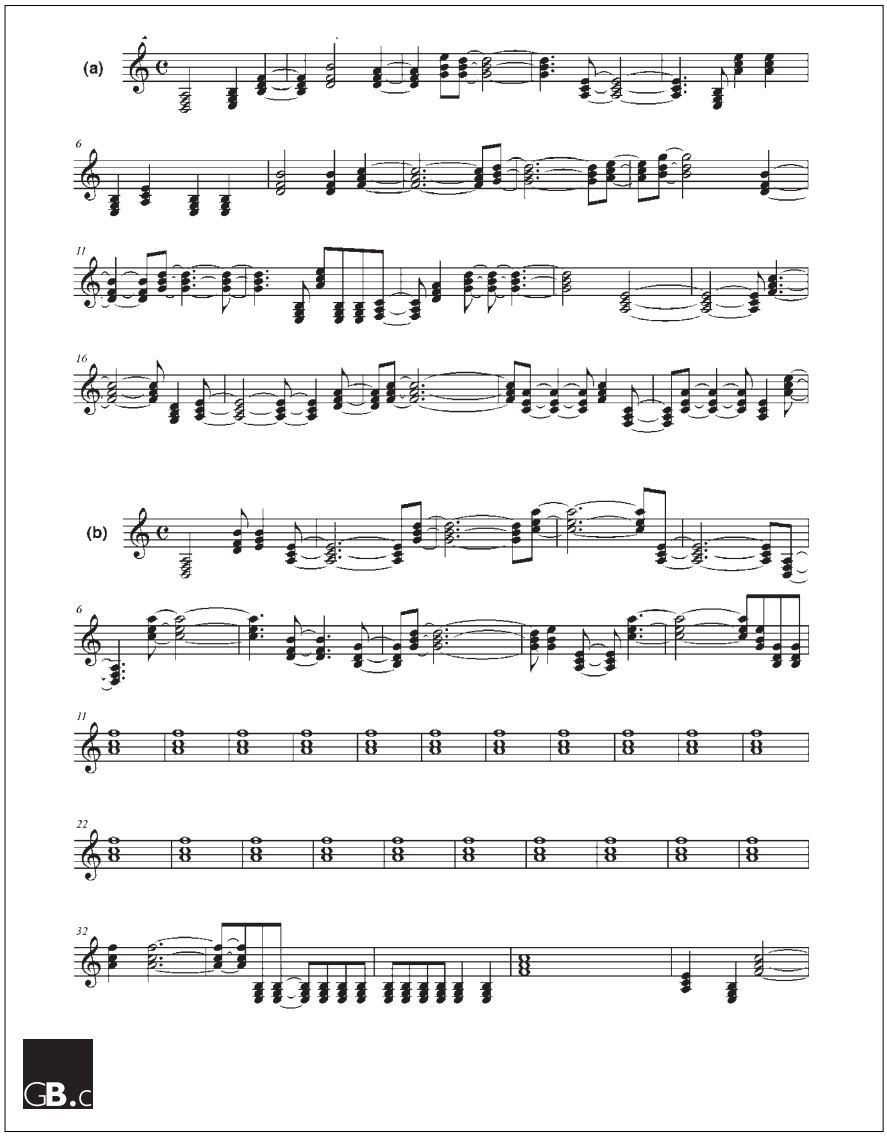

Figure 2
Partial human ThyA protein sequence with rhythm based on the human codon distribution. **(a)** Four different note lengths (eighth, quarter, half, whole note) were each assigned to a particular codon usage range based on frequency per 1,000. Zero to 10 (per 1,000) was assigned the eighth note, 11 to 20 the quarter note, 21 to 30 a half note, and a codon usage greater than 30 was assigned the whole note. The more frequently a particular codon is used, the longer the note length that represents such a codon. **(b)** Huntingtin protein translated into musical notes based on the reduced-note range and human codon distribution. The wild-type huntingtin protein contains 21 glutamines in the beginning portion of the sequence. The protein also contains proline-rich regions. The repetition in these regions can be distinctly heard in the musical translation.

four note durations to each amino-acid codon based on the codon usage (frequency per 1,000 occurrences). The more abundant the codon is for a particular organism, the longer the note duration. One can see the new rhythmic adjustments in Figure 2a, where the reduced note range assignment is used for the human ThyA protein. The resulting music addresses the issues of musicality such as large interval jumps and rhythm, which makes the musical translation more pleasing to listen to and maintains the integrity of the protein sequence within the music. Figure 2b illustrates the difference that can be recognized when various protein motifs are scored. Here, we transposed the beginning segment of the huntingtin protein involved in Huntington's disease [11]. A clear auditory pattern emanates from both repetitive glutamines (21 in this normal individual) and polyproline stretches. The repeated notes are distinctly set apart from the rest of the sequence, allowing one to recognize this region by ear.

By converting genomic sequences into music, we hope to achieve several goals, which include investigating sequences by the vision impaired. Another aim is to attract young people into molecular genetics by using the multidisciplinary approach of fusing music and science. There are strong associations between music and perception. Heightened interest in a historically known condition called synesthesia (or synaesthesia) has also spanned multiple fields of study including science, music, and history [12]. The condition has prompted a collaborative approach among various disciplines aimed at developing a more comprehensive picture of this syndrome. Synesthesia is an involuntary perception produced by stimulation of another sense. Commonly one hears a certain pitch that consistently evokes a particular color. Synesthesia is considered an unusually strong cross-modal association in the brain and has been observed in children and adults [12]. Another example of a collaborative, cross-disciplinary effort includes research pertaining to sound-induced photisms. Sound-induced photisms have been recorded where a startled reaction to a sound (soft or loud) evokes colors ranging from flashes of white light to a colorful flame [13]. A separate study confirms that lighter colors 'fit together' with higher pitches of sound and darker stimuli are better fitted to lower pitches [14].

In future studies, we will use a recently created program (F. Pettit, unpublished work), now in its testing stages, which implements the translation rules we have formulated. Use of this program will enable very rapid translation of large segments of genomes into music. Furthermore, different instruments can

be assigned to unique parts of the genome, such as regulatory, intergenic, and promoter/operator sequences, in order to use the obvious distinction as a teaching tool for introducing the function of the genome and its parts. Finally, each protein provides a theme that can be used as a source to make variations that would involve improvisation and elaboration, which would allow the investigator/author to contribute an artistic component to the original melody. For further examples of protein music and references to previous work, go to our website gene2music [15]. Also, browse this website to access our computer program in order to convert your own gene of interest to music.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a music clip of the human ThyA protein based on the single note assignment of one amino acid per musical note. Additional data file 2 is a music clip of the human ThyA protein derived from the reduced 13-base note chord assignment. Additional data file 3 is a music clip of the human ThyA protein based on our final coding assignment, which includes rhythm. Additional data file 4 is a music clip of the huntingtin protein based on our final coding assignment.

References

1. Oberst T: **Blind graduate student 'reads' maps using CU software that converts color into sound**. *Cornell Chronicle* 2005, **36:**5.
2. Cyranoski D: **Japan plays trump card to get kids into science.** *Nature* 2005, **435:** 726.
3. Miller JN: ***The Biochemists' Songbook* by Harold Baum.** *J Pharm Biomed Anal* 1983, **1:**379.
4. Ohno S, Ohno M: **The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition.** *Immunogenetics* 1986, **24:**71-78.
5. Gena P, Strom C: **Musical synthesis of DNA sequences**. In: *XI Colloquio di Informatica Musicale*: November 1995; Bologna: Universita di Bologna; 1995: 203-204.
6. Gena P, Strom C: **A physiological approach to DNA music**. In: *CADE 2001*. Glasgow, UK: Glasgow School of Art Press; 2001:129-134.
7. Hance BD: **Art exhibit to showcase musical works based on genetic sequences.** *Arizona Daily Wildcat* 1996, Jan 31.
8. Jensen E, Rusay R: **Musical representations of the Fibonacci string and proteins using Mathematica**. *Mathematica J* 2001, **8:**55.
9. Dunn J, Clak MA: **Life music: the sonification of proteins.** *Leonardo* 1999, **32:**25-32.
10. **A protein primer: a musical introduction to protein structure** [http://www.whozoo.org/mac/Music/Primer/Primer_index.htm]
11. The Huntington's Disease Collaborative Research Group: **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.** *Cell* 1993, **72:**971-983.
12. Ione A, Tyler C: **Neuroscience, history and the arts. Synesthesia: is F-sharp colored violet?** *J Hist Neurosci* 2004, **13:** 58-65.
13. Jacobs L, Karpik A, Bozian D, Gothgen S: **Auditory-visual synesthesia: sound-induced photisms.** *Arch Neurol* 1981, **38:** 211-216.
14. Hubbard T: **Synesthesia-like mappings of lightness, pitch, and melodic interval.** *Am J Psychol* 1996, **109:**219-238.
15. **gene2music** [http://www.mimg.ucla.edu/faculty/miller_jh/gene2music/home.html]