Haematology and Blood Transfusion Vol. 31
Modern Trends in Human Leukemia VII
Edited by Neth, Gallo, Greaves, and Kabisch

Repetition as the Essence of Life on this Earth: Music and Genes

S. Ohno [1]

A. Introduction

While it is believed that life on this earth started as long ago as a few billion or more years ago, a number of true innovations in evolution appears to have been rather dismally small. Most of the successful adaptive radiation of living organisms have apparently been accomplished by extensive plagiarization of those preciously few innovations via the mechanism of gene duplication [1]. Furthermore, it appears that most of these true innovations have occurred at the very beginning, before the division of prokaryotes from eukaryotes. For example, nearly all the sugar-metabolizing enzymes appear to have achieved their inviolable functional competence at the above-noted early date. Natural selection has since been spinning wheels in the air.

B. The Story of Glyceraldehyde 3-Phosphate Dehydrogenase

It would be noted in Fig. 1 that the 332-residue-long glyceraldehyde 3-phosphate dehydrogenase of the pig differs from the lobster enzyme only at 86 positions. Inasmuch as vertebrates, or rather chordates diverged from crustaceans roughly 500 million years ago, one can conclude from the above and similar data on additional species that this enzyme has been undergoing 1% amino acid sequence divergence every 20 million years, thus accumulating 26% amino acid sequence difference in 500 million years. If such a rate calculation can be extended indefinitely, however, even at this snail's pace one still expects this enzyme to have undergone 100% amino acid sequence divergence in 2 billion years. Now 2 billion years ago would have been about the time prokaryotes diverged from eukaryotes. Yet the bacterial amino acid sequence from *Bacillus stearothermophilis*, also shown in Fig. 1, still maintains 177 out of the 332 sites (53%) homology with the pig enzyme, and similar 180 out of 332 sites homology with the lobster enzyme. In fact, there are 19 segments (tripeptidic or longer), comprised of 92 residues in total, that remain invariant in all three species. The longest conserved segment, tridecapeptidic in its length, occupying 144th to 156th position, represents the most critical of the substrate binding sites, 149th Cys forming the thiol linkage with substrate intermediates [2]. Indeed, after achieving the appropriate degree of functional competence 2 billion or more years ago, glyceraldehyde 3-phosphate dehydrogenase has not changed in its essence; evolutionary compatible amino acid substitutions that accompanied successive diversification and speciation merely symbolizing futile spinning of the wheel. Such a futility is also evident in Fig. 1, for at the 14 positions, a eukaryote (the pig) and a prokaryote (*Bacillus stearothermophilis*) share the identical residues, while the other eukaryote (the lobster) is left out as an oddball; e.g., the third position of the pig and the bacillus is Val, while that of the lobster is Ile. At these and many other positions, the game of musical chairs

[1] Beckman Research Institute of the City of Hope Duarte, California 91010, USA

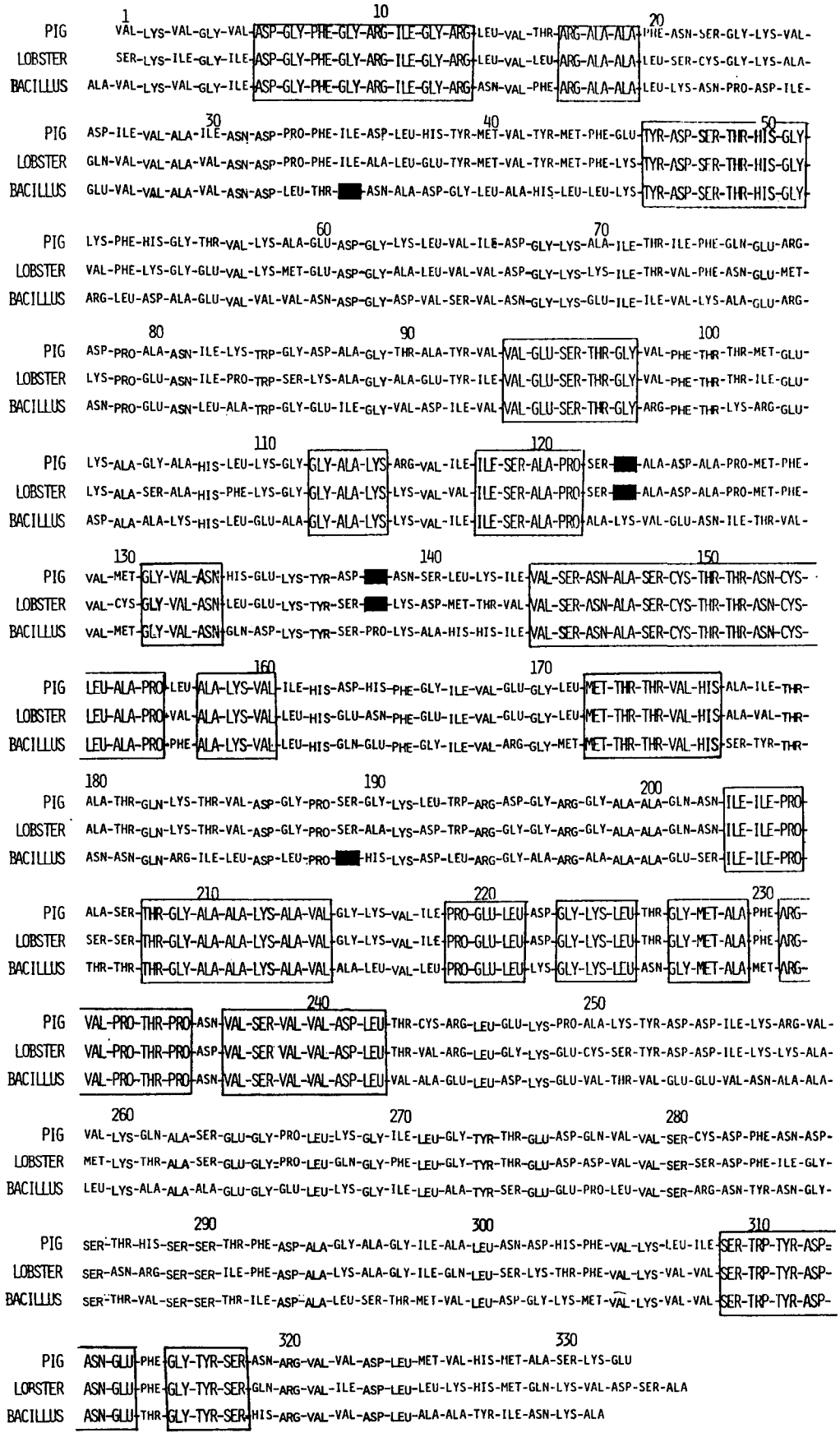

Fig. 1. The amino acid sequences of glyceraldehyde 3-phosphate dehydrogenases from three divergent species are compared. Bacillus refers to *Bacillus stearothermophilis*. Discordant and identical residues are shown slightly displaced from each other; discordant ones are placed little above identical ones. Amino acid residues of tripeptidic or longer conserved segments are shown in *large capital letters* and segments are *boxed in*. Deleted residues are identified as *black boxes*

have apparently been in play among a limited number of functionally compatible amino acids.

Analogous situations have been found with regard to other sugar metabolizing enzymes, e.g., phosphoglycerate kinase, triose isomerase etc. Furthermore, all these sugar-metabolizing enzymes are constructed of the same mould. The amino terminal half and the carboxyl terminal half forming two distinct domains, a cleft between the two accommodating the substrate and the coenzyme. The amino terminal half is for the coenzyme binding and the carboxyl terminal half is for the substrate binding. Furthermore, Rossman [3], among others, has pointed out that in the case of kinases, the mononucleotide (e.g., ATP) binding site of the amino terminal half is comprised of three β-sheet-forming segments and two α-helix-forming segments in the following order from the amino terminus; $\beta\alpha\beta\alpha\beta$. The dinucleotide (NAD or NADP) binding site of dehydrogenases, on the other hand, evolved from the above by duplication; thus, it can be expressed as $2 \times \beta\alpha\beta\alpha\beta$. Inasmuch as the most critical portion of the substrate binding site evolved within the last segment of the duplicate (e.g., 144th to 156th tridecapeptide of Fig. 1), this intrusion of the substrate binding active site into the dinucleotide binding domain froze the dinucleotide binding domain of each enzyme as uniquely its own. Thus, there is no more than 20% amino acid sequence homology between dinucleotide binding sites of different enzymes in spite of the fact that all are made of the same $2 \times \beta\alpha\beta\alpha\beta$ mould. It would be recalled that within the same enzyme, conservation of greater than 50% homology is the rule for the whole enzyme, therefore, the dinucleotide binding amino terminal half.

At any rate, two notable facts emerge from the above. First, coding sequences for sugar-metabolizing enzymes and probably for many other enzymes (e.g., proteases) have already achieved the appropriate degree of functional competence before the division of prokaryotes from eukaryotes. Second, repetitions were the rule of the game from the very onset of life on this earth; the dinucleotide binding site evolving from the mononucleotide binding site by duplication, and that the mononucleotide binding site itself likely to have evolved by 2.5 times duplication of the one $\beta\alpha$ or $\alpha\beta$ unit.

C. Ingeniousness Embodied in the First Set of Coding Sequences that Were Repeats of Base Oligomers

Orgel's group [4] has shown that in the presence of Zn ion, nonenzymatic synthesis of nucleic acids occurs in the proper 3′- to 5′ linkage, provided that there is a template. Thus, it would appear that what was in short supply in the prebiotic world, before the emergence of life on this earth was long templates from which copies can be made. Put it more succinctly, the first primordial question is: "How did oligonucleotides manage to extend themselves to become worthy coding sequences?" There is one simple answer: One tandem duplication of the preexisted oligomer assures indefinite extension of that template, as illustrated at the top of Fig. 2. What if the heptameric template CAGCCTG duplicated to become tetradecamer? After completion of its complementary strand, the two might pair in the manner shown; second copy pairing with the first copy of the complementary strand. The paired portion would now serve as the primer for the next round of nucleic acid synthesis. At the completion of the second round, the 14-mer template now becomes 21-mer. In this way, the indefinite extension of the primer is assured a priori, a paired segment always serving as a primer for the next round of nucleic acid synthesis. The above then is the first reason for believing that the first set of coding sequences, or rather all nucleic acids in the prebiotic world that presaged the emergence of life, on this earth were all repeats of various base oligomers.

How accurate was a copying function of the nonenzymatic nucleic acid replication? Of various nucleic acid polymerases known, the most error prone appear to be reverse transcriptase of retroviruses, for their error rate has been estimated as of the order of 10^{-3}/base pair/year [5]. This is one million times higher error rate compared to DNA polymerases of vertebrates, and at this rate, there would be 100% base sequence change every one thousand years. The inherent error rate of prebiotic, therefore, nonen-

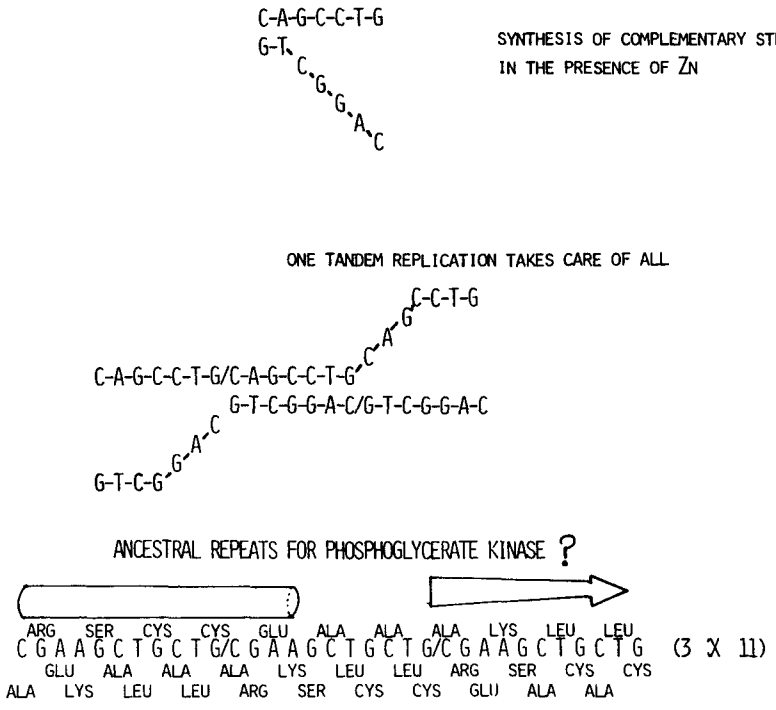

Fig. 2. Replication of nucleic acids is based upon the inherent complementarity that exists between two purine-pyrimidine pairs; A pairs with T or U, while G pairs with C. Accordingly, provided that there is a template (the heptamer CAGCCTG shown at the top), mononucleotides would readily assemble themselves in the 3′, 5′ linkage to form a complementary strand in the presence of Zn [4] as shown at the top. What was in short supply in the prebiotic world then were templates of substantial lengths. What if the above noted haptamer repeated itself in tandem or some of the base oligomers were by chance tandem repeats (two copies of the shorter oligomer) to begin with. It and its complementary strand can pair unequally in the manner depicted at the middle. As a paired segment now functions as a primer for the next round of nucleic acid synthesis, infinite extension of templates is now assured. All it takes to start this process is the one tandem duplication.

Of long oligomeric repeats thus formed, those that evolved to be the first set of coding sequences likely started from oligomeric units whose numbers of bases were not multiples of three. There were two distinct advantages: (1) They gave longer periodicities to polypeptide chains; e.g., repeats of the base octamer would have given octapeptidic periodicity while repeat of the base nonamer would have only the tripeptidic periodicity. (2) They would have encoded polypeptide chains of identical periodicity in all three reading frames. Within the periodic unit such repeats could have given both α-helical segment and β-sheet forming segment as shown at the bottom. Such alternating α,β structures gave rise to the mononucleotide binding site (3) which would have been utilized immediately as parts of the primitive nucleic acid polymerase. Later they gave rise to ATP and NAD, NADP binding sites of many enzymes as discussed in the text

zymatic nucleic acid replication is expected to be higher than the above-noted 10^{-3}; as error prone as they are, reverse transcriptases are, after all, the enzyme of a sort. Prebiotic coding sequences had to contend with this very high replication error rate and should still have been able to encode polypeptide chains of potential function. Provided that the number of bases in the oligomeric unit was not a multiple of three, repeats of the base oligomer would have been very stable under this mostly trying circumstance of constant base substitutions, deletions, and insertions. This is also illustrated at the bottom of Fig. 2. Since the monodecamer CGAAGCTGCTG cannot be divided by 3, three consecutive copies of it translated in three different reading frames gives the monodecapeptidic periodicity to a polypeptide chain. Contrast the above to repeats of the base dodecamer, which can give only the tetrapeptidic periodicity to the polypeptide chain. Furthermore, since within a given reading frame three consecutive copies of the monodecamer are to be translated in all three reading frames, such

repeats encode polypeptide chains of the identical periodicity in all three reading frames. This openness of all three reading frames give them a great deal of imperviousness to base substitutions, deletions, and insertions. Repeats of the monodecamer shown at the bottom of Fig. 2 encode both potentially α-helix-forming segment and potentially β-sheet-forming segment within one monodecapeptidic unit. In fact, sugar-metabolizing enzymes in general and phosphoglycerate kinase in particular might have originally been encoded by repeats of such a monodecamer, for AAGCTGCTG portion of the monodecameric unit recur in many variations in the modern coding sequence (e.g., of man) for phosphoglycerate kinase as already noted in our previous paper [6].

D. Repetition as the Essence of Coding Sequences and Musical Compositions

Earth on which life has evolved has always been governed by the hierarchy of periodicities. First, earth rotates on its own axis to create days, while the moon's revolution around the earth gives months, with neap tides and spring tides to be topped by years, reflecting the earth's travel around the sun. It is small wonder if life itself was born out of periodicities embodied in repetition of unit base oligomers. Just as man eventually devised seconds, minutes, and hours as arbitrary units of time measurement, one of the periodicities embodied in polypeptide chains encoded by the first set of codeing sequences that were oligomeric repeats must soon have been chosen as the arbitrary time-measuring unit by the ancestral biological clock. It now appears that this arbitrarily chosen unit was the simplest dipeptidic periodicity. The polypeptide chain encoded by *per* locus of *Drosophila merlanogaster,* fundamentally involved in the expression of biological rhythms such as cicardian behaviors and 55-s rhythm of courtship song, is largely comprised of the Gly-Thr dipeptidic repeats interspersed with short stretches of its deviant Gly-Ser dipeptidic repeats, and that the homologous gene encoding the polypeptide chain of the above-noted dipeptidic periodicity is conserved in the mouse as well [7]. Observing the *per* locus coding sequence, one notices that there have been numerous neutral base substitutions, e.g., free base substitutions at the redundant 3rd base position of glycine codons. Thus, it would appear that the time-keeping was done from the beginning at the polypeptide level rather than at the level of coding sequences, although the initial periodicity of that polypeptide chain had to be the consequence of its coding sequence being repeats of unit base oligomers.

Now we come to the origin shrouded in mist, of the prehistory of musical compositions. Inasmuch as songs of canaries and skylarks are as pleasing to our ears as they must be to their mates as well as to themselves, it is clear that melodies as such are no human invention. Furthermore, the vocal cord and other sound-making apparatuses of our immediate relatives (e.g., *Homo neanderthalensis*) appear to have been rather underdeveloped. Accordingly, I wonder if early *Homo sapiens* were capable even of imitating beautiful bird songs noted above even if they wanted to. I would rather believe that music as such were invented by primitive man as purely rhythmic timekeeping device. For example, a hunting party intent on bringing down a mammoth or two would have to coordinate activities of several cohorts spread over a wide arc surrounding the herd of mammothes. This, I suspect, was done by rhythmic beatings of hollowed tree trunks for example; fast repetitions of a given rhythm conveying an urgent need to close in whereas slow repetitions of the same rhythm meaning cautious approach. It would thus appear that music, too were initally born out of repetitious rendition.

Even today of wonderous melodies, music is still used as a time keeping device, as in dancing and military parades. Rhythm of the latter, marching music are essentially that of our heart beat. Our heart beats slow in slumber and contemplation, while it beats uncontrollably fast in fright. Rhythm of marching music should be somewhere in between to indicate willingness either to go forth against formidable adversaries or to defend against adversaries until death. Because of this homage to the periodicity inherent both in coding sequence construction and musical composition, the way was sought to interconvert the two. The solution

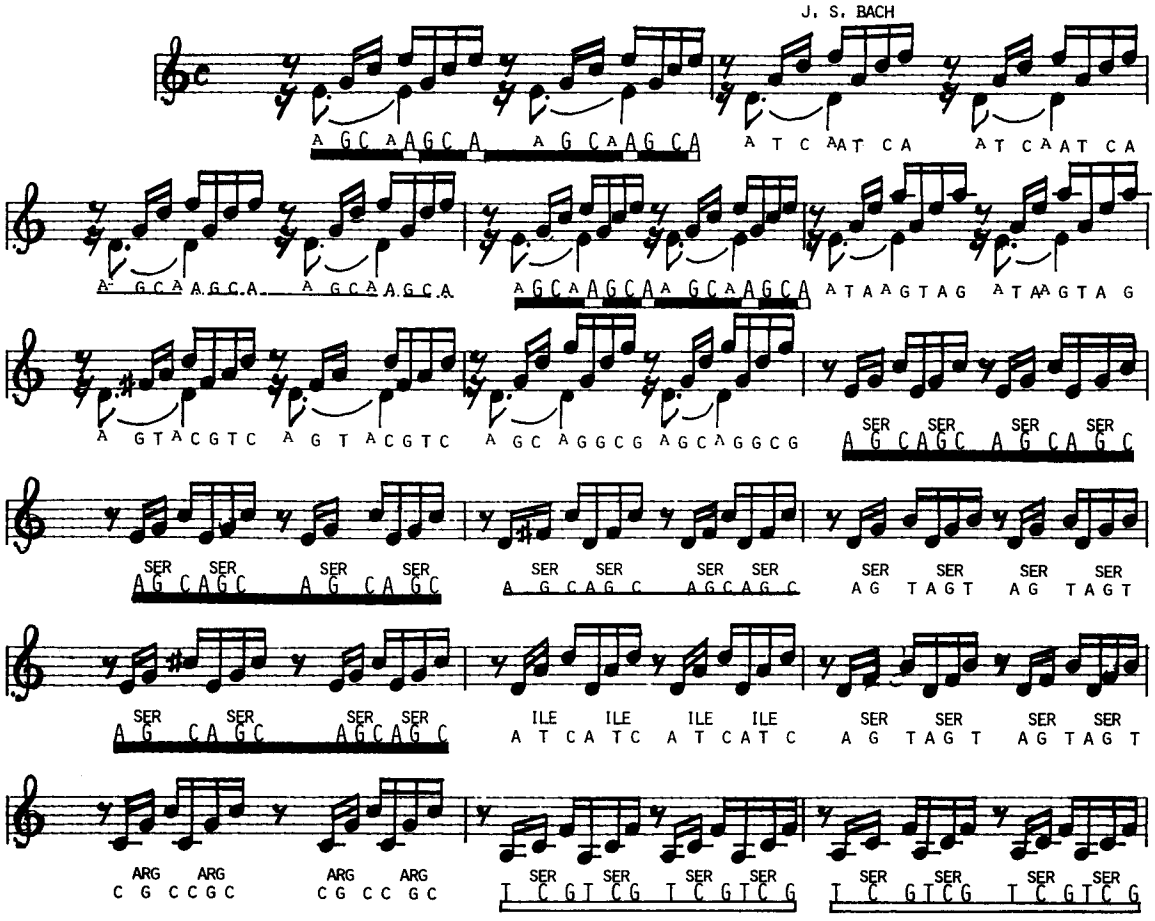

Fig. 3. An initial part of the treble-clef musical score of Prelude No. 1 from well-tempered clavichord by J. S. Bach, accompanied by the base sequence and the amino acid sequence transcribable from that base sequence

that we arrived at is to assign a space and a line on the octave scale to each base in the ascending order of A, G, T, C in such a way so that the classical middle-C position would be occupied by C on the line, A in the space occupying the position immediately above [6].

In Fig. 3, the treble-clef musical score of Prelude No. 1 from well-tempered clavichord by J. S. Bach, the great master of the early Baroque, is accompanied by the base sequence transcribed from it according to the rule stated above. It would be noted that with regard to every 4/4th or 8/8th time signature unit, the second half is the exact repeat of the first half. Furthermore, until the 3rd line, each half is repeats of four notes, the four-note subunit consisting of one 3/16th note and three 1/16th notes followed by one 1/4th note and four 1/16th notes. Translated to base sequence, the first time signature unit is comprised of four exact copies of the AGCA tetramer followed by four copies of a single-base substituted deviant of the above-noted tetramer ATCA. The AGCA recurrs again 8 times. Since 4 is not a multiple of three, these tetrameric repeats are capable of giving the tetrapeptidic periodicity to a polypeptide chain, but alas. chain terminators TAA and TAG come in pairs at the extreme right of 2nd line. From the 4th line onward, one 3/16th note and a quarter note are relegated to the base clef; therefore, the treble-clef score becomes trimeric repeats. When translated, this portion yields polyserine interspersed with teterailsoleucyne and tetraarginine.

In general, I found musical compositions of the early Baroque period to be repeats of short base oligomers, these oligomers being single-base substituted variants of each other. Indeed, their resemblance to what I conceived as the first set of coding sequences at the very beginning of life on this earth is uncanny (see Fig. 2). Most of the coding sequences possessed by modern organisms

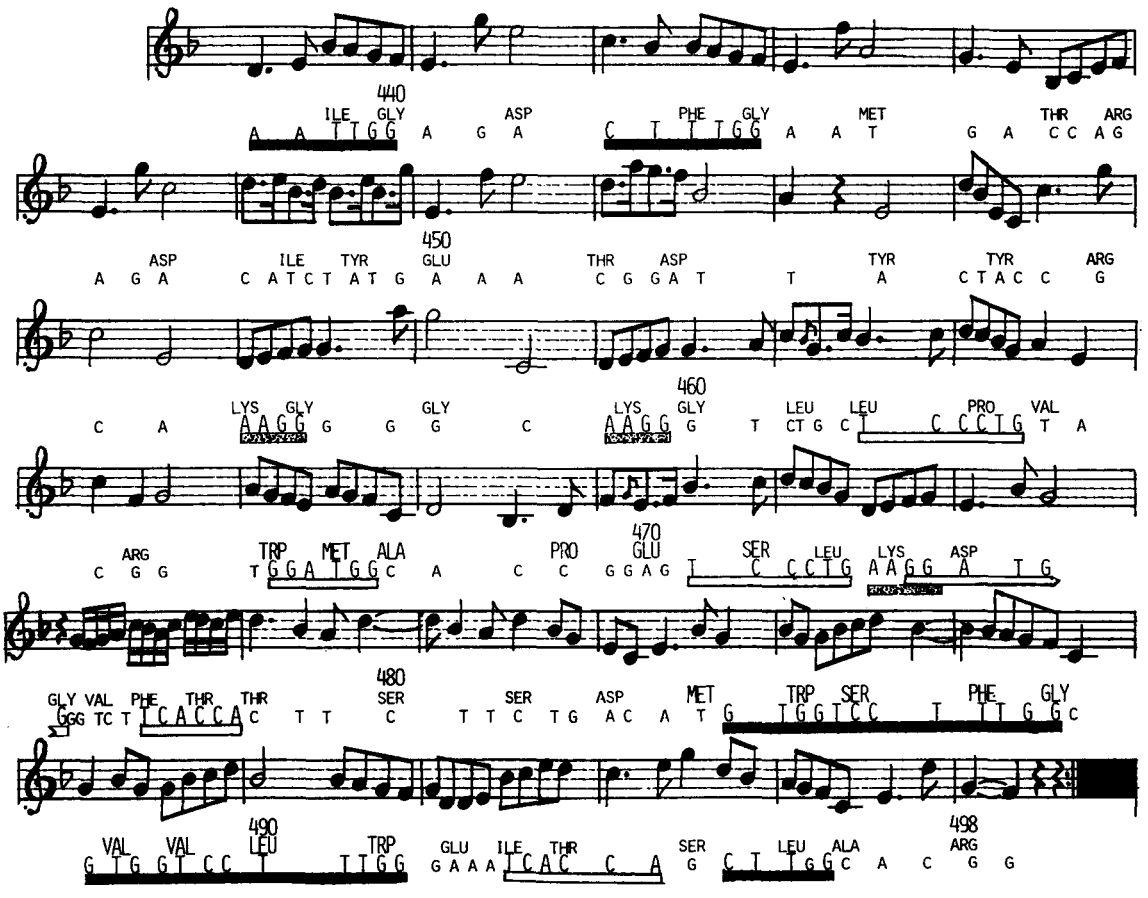

Fig. 4. The heart of the coding segment for tyrosine kinase domain of the human insulin receptor β-chain (8). Amino acid residues of the two active site segments are shown in *large capital letters*. This musical transformation for violin of the coding segment is in E minor, 4/4th or 8/8th time signature

have endured for hundreds of millions of years. In the case of those for sugar-metabolizing enzymes, 2 billion years or more as already noted. Thus, their original periodicities are obvious only for discerning eyes. Not surprisingly, musical compositions of the late Romantic period resemble these coding sequences. We have previously shown that Frédéric Chopin's nocturne Opus 55, No. 1, resembled the last exon for the largest subunit of RNA polymerase II [6]. In Fig. 4, the musical transformation for violin of the most functionally critical part of the tyrosine kinase domain of the human insulin receptor β-chain [8] is shown. This segment includes two active site segments most critical for the assigned function of tyrosine kinase. Amino acid residues of these two active site oligopeptides are shown in large capital letters. It would be noted that nearly all of the second active site is encoded by tandem repeats of the dodecamer GTGGTCCTTTGG, thickly underlined by solid bars (2nd from the last line of Fig. 4). Its two truncated derivatives at the top line of Fig. 4 are also underlined by solid bars. Other, more musically pertinent repeats are also underlined by open bars and shaded bars; e.g., the hexamer TCCCTG in 3rd and 4th lines of Fig. 4.

E. Summary

In prebiotic nucleic acid replication, templates appear to have been in short supply. A single round of tandem duplication of existing oligomers assured progressive extension of templates to the length adequate for encoding of polypeptide chains. Thus, the first set of coding sequences had to be repeats of base oligomers encoding polypeptide chains of various periodicities. On one hand, the readiness of these periodical polypeptide chains to assume α-helical and/or β-sheet secondary structures contributed to the extremely rapid initial functional diversification of these polypeptide chains. It

would be recalled that most, if not all, of the sugar-metabolizing enzymes had already achieved the inviolable functional competence before the division of prokaryotes from eukaryotes. On the other hand, a certain (dipeptidic?) of the peptidic periodicities was apparently chosen as the timekeeping unit by the biological clock. Musical compositions too apparently evolved originally as a timekeeping device. Accordingly, repetitiousness is evident in all musical compositions. Evolution of musical compositions from the early Baroque to the late Romantic parallels that of coding sequences from rather exact repeats of base oligomers to more complex modern coding sequences in which repetitious elements are less conspicuous and more varied.

Inasmuch as the earth is governed by the hierarchy of periodicities (days, months and years), such reliance on periodicities is rather expected.

References

1. Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin Heidelberg New York
2. Dayhoff MO (ed) (1972) Atlas of protein sequences and structure. National biomedical research foundation, Silver Springs, Maryland
3. Rossman MG (1981) Evolution of glycolytic enzymes. Philos Trans R Soc Lond [Biol] B293:191–203
4. Bridson PK, Orgel LE (1980) Catalysis of accurate poly (C)-directed synthesis of 3′-5′-linked oligoguanylates by Zn^{+2}. J Mol Biol 144:567–577
5. Gojobori T, Yokoyama S (1985) Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. Proc Natl Acad Sci USA 82:4198–4201
6. Ohno S, Ohno M (1985) The all-pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition. J Immunogenet 24:71–78
7. Shin H-S, Bargiello TA, Clark BT, Jackson FR, Young MW (1985) An unusual coding sequence from a *Drosophila* clock gene is conserved in vertebrates. Nature 317:445–451
8. Ulrich A, Bell JR, Chen EY, Herrera R, Petruzzelli LM, Dull TJ, Gray A, Coussens L, Kiao Y-C, Tsubokawa M, Mason A, Seeburg PH, Gunfeld C, Rosen OM, Ramachandran J (1985) Human insulin receptor and its relationship to the tyrosine kinase family of oncogenes. Nature 313:756–761

Translation to Human Temparaments
Session in "De Emmenhof"

Fotos Regina Völz