Immunogenetics 24: 71–78, 1986

Immunogenetics

The All Pervasive Principle of Repetitious Recurrence Governs Not Only Coding Sequence Construction But Also Human Endeavor in Musical Composition

Susumu Ohno and Midori Ohno

Beckman Research Institute of The City of Hope, Duarte, California 91010

Abstract. Organisms which have evolved on this earth are governed by multitudes of periodicities; tomorrow is another today, and the next year is going to be much like this year. Accordingly, the principle of repetitious recurrence pervades every aspect of life on this earth. Thus, individual genes in the genome have been duplicated and triplicated often to the point of redundancy, and each coding sequence consists of numerous variously truncated as well as variously base-substituted copies of the original primordial building block base oligomers and their allies. This principle even appears to govern the manifestations of human intellect; musical compositions also rely on this principle of repetitious recurrence. Accordingly, coding base sequences can be transformed into musical scores using one set rule. Conversely, musical scores can be transcribed to coding base sequences of long open reading frames.

"Whereas ordinary mortals are content to mimic others, creative geniuses are condemned to plagiarize themselves" is my shorter, albeit inarticulate, version of what Van Veen said in *Ada* by Vladimir Nobokov. Indeed, it seems that vaunted geniuses seldom invented more than one modus operandi during their lifetimes, and even civilization has largely been dependent upon plagiarizing a small number of creative works; e. g., the multitudes of Gothic churches can be viewed as pan European plagiarism of the abbey church of St. Denis and/or the cathedral at Sens. This is not surprising for new genes sensu stricto have seldom been invented. Evolution rather relies on plagiarizing an old and tested theme; the mechanism of evolution by gene duplication (Ohno 1970). For example, the adaptive immune system of vertebrates has apparently evolved by plagiarizing one ancestral gene. This gene encoded a 90 or so residue long polypeptide chain which folded itself to form two β-barrel structures held together by one intrachain disulfide bridge. Included in this superfamily of genes are not only those responsible for antigen-binding immunoglobulins (Igs) and T-cell receptors but also those responsible for class I and class II major histocompatibility antigens and even for the transmembrane receptor for poly-IgM and IgA (Mostov et al. 1984). Inasmuch as reliance on repetitions of tested themes has been the hallmark of all lives evolved on this earth, the redundancy resulting from this has become very prominent in mammalian genomes. Most of the nine or more factors involved in blood clotting that finally convert fibrinogen to fibrin are inert serine proteases that become activated by cleavage induced by a preceding activated serine protease; why should there not be one or two instead of nearly ten, and why should factor VIII be so enormous, being comprised of 2332 amino acid residues (Gitschier et al. 1984)? Similarly, the tyrosine kinase domain is an integral component of membrane receptors for various growth factors such as insulin and the epidermal growth factor. In addition, however, new *c-onc* gene loci for tyrosine kinase are being discovered at an alarming rate (Martin-Zaca et al. 1985). Interspecific comparison also revealed different redundancies involving various gene loci for class I as well as class II major histocompatibility antigens (Klein and Figueroa, in press).

In this paper, we shall show that this principle of repetitious recurrence pervades both the construction of coding sequences in the genome, which can be regarded as being representative of nature, and musical composition which can be regarded as the most abstract and therefore the most intellectual expression of nurture.

Historical development of coding base sequences and musical composition. The principle of repetitious recurrence also applies to the construction of individual coding sequences. Available evidence indicates that all the coding base sequences were—at their inception eons ago—repeats of base oligomers. They therefore encode polypeptide chains of exact periodicities (Ohno and Epplen 1983, Ohno

Address correspondence to: Dr. S. Ohno, Beckman Research Institute of the City of Hope, 1450 East Duarte Road, Duarte, CA 91010-0269, USA

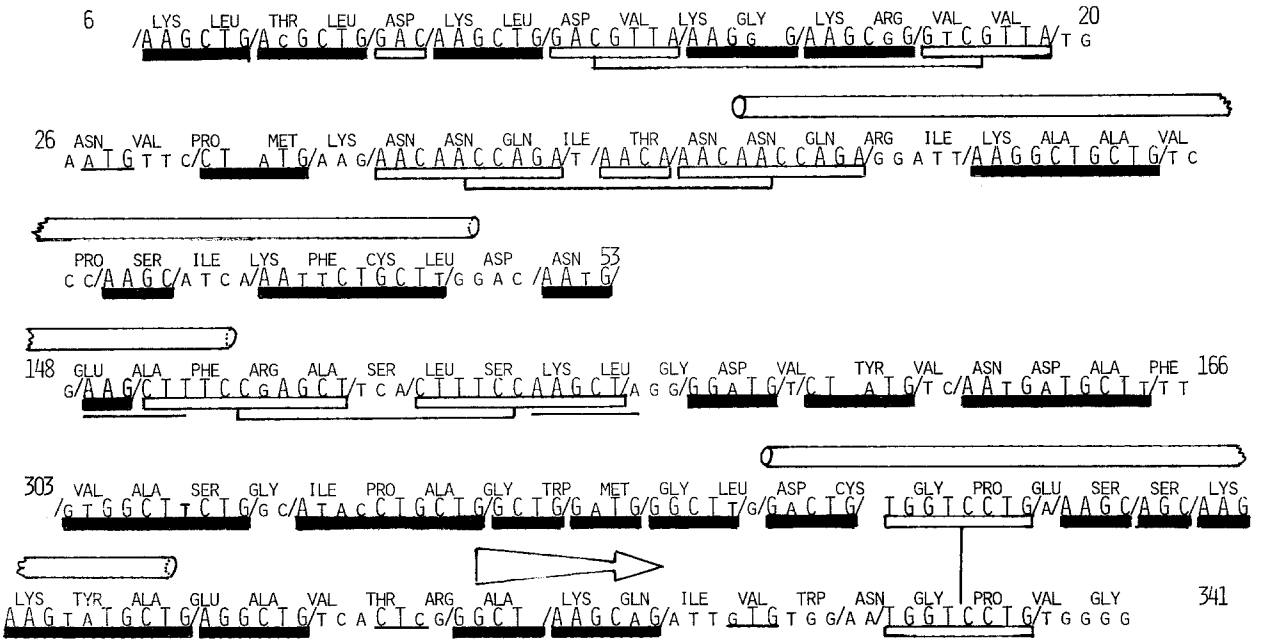

Fig. 1. Four widely spaced representative segments of the 419 codon long, human X-linked phosphoglycerate kinase coding sequence (Singer-Sam et al. 1983) accompanied by amino acid residues encoded by them are shown. The first and last codons of each segment are numbered to indicate the segment's position in the entire coding sequence. The primordial building block hexamer A A G C T G (*first row, extreme left,* for example) and its parental decamer A A G G C T G C T G (*second row, extreme right),* shown in *large capital letters,* and their variously truncated and/or base-substituted copies (deviant bases are shown in *small capital letters*) are underscored by *thick solid bars.* Tandemly recurring base oligomers not directly related to the above-mentioned primordial building blocks are underscored by *open bars* and connected to each other. α-helix-forming regions are designated by *cylinders* and β-sheet-forming regions by *arrows* above the sequences (Watson et al. 1982)

1984, 1985). Even today, rare coding sequences that arose de novo in divergent organisms fit the above description; e. g., antifreeze proteins of cods and flounders inhabiting the Arctic and Antarctic oceans (DeVries 1982) and circumsporozoite antigens of malarial protozoa belonging to the genus *Plasmodium* (Zavala et al. 1985). These original periodicities have not readily disappeared from modern coding sequences in spite of the hundreds of millions of years that have elapsed since their inception. This point is illustrated in Figure 1 using three representative segments of the human X-linked, 419 codon long coding sequence for one of the sugar metabolizing enzymes, phosphoglycerate kinase (PGK; Singer-Sam et al. 1983). It appears that such sugar metabolizing enzymes were well established very early in evolution, shortly after the beginning of life on this earth in fact, and they have changed very little since then. Accordingly, the amino acid sequence of this human enzyme has retained an almost 70% homology with the corresponding enzyme of baker's yeast, and the two are indistinguishable in their tertiary conformation (Watson et al. 1982). In Figure 1, the primordial building block base decamer A A G G C T G C T G (Fig. 2, second row at the extreme right) and its truncated hexameric derivative A A G C T G (two in the first row of Fig. 1) and the variously base-substituted as well as truncated derivatives of both are underscored by solid bars. Derivatives of these two primordial building block base oligomers occur indiscriminately in regions encoding α-helical peptide segments (designated by cylinders above the sequence), in those encoding β-sheet forming segments (designated by arrows above the sequence), as well as in regions in between. This attests to the antiquity of these primordial building block base oligomers which date back to the time of inception of this coding sequence eons ago. Nevertheless, the monotony created by the endless recurrence of these decamers, hexamers, and their derivatives is broken by refreshing appearances of tandemly recurring base oligomers that are not directly related to the two primordial building blocks mentioned above; e. g., the A A C A A C C A G A decamer tandemly recurs in the second row of Figure 1, and its tetrameric portion A A C A is recapitulated in between the two copies. By this time, those familiar with music will have been struck by the

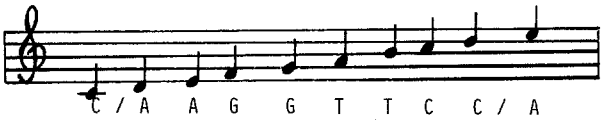

Fig. 2. The assignment of two consecutive positions each in the octave scale to four bases in the ascending order of A, G, T, and C. This is the inviolable rule to abide by in all musical transformations of coding base sequences, as well as in the transcription of existing musical scores to base sequences

striking similarity between the representative coding sequence presented as in Figure 1 and musical composition. In musical terms, the primordial A A G G C T G C T G decamer and its truncated hexamer A A G C T G are principal subjects (melodies) that are developed in endless variations (variously base-substituted and truncated derivatives), while tandemly recurring base oligomers not directly related to the primordial ones such as the A A C A A C C A G A decamer are secondary subjects (melodies) that introduce the desired complexity into musical composition.

Accordingly, the inviolate rule for musical transformation of coding sequences was invented as shown in Figure 2. In order to fill the octave scale, two consecutive positions were assigned to each of the four bases in the ascending order of A, G, T, and C. Inasmuch as adenine (A) and guanine (G) are purines with heavier molecular masses than pyrimidines, it seemed natural that A and G should occupy the lower end of the octave scale. Since A preceded G, again it seemed natural that of the two pyrimidines, T (which is complementary to A) should precede C. Going up the octave scale, the next A can be placed above C, and going down the scale C can occupy two consecutive positions below A. In order to transform the human X-linked PGK coding sequence into a musical score based on the rule established in Figure 2, various melodies were assigned to the primordial decamer A A G G C T G C T G and its truncated hexamer A A G C T G and tried. The melody for A A G C T G shown in the first and second rows of Figure 3, and that for A A G G C T G C T G shown in the fourth row of Figure 3 appeared to be most appropriate. The choice of these two principal subjects automatically decided the key as well as the time signature of this musical transformation, thus at the same time dictating the secondary subjects to be assigned to the tandemly recurring heptamer G A C G T T A (Fig. 3, second row) and the decamer A A C A A C C A G A (Fig. 3, third and fourth rows). In Figure 3, the musical transformation into D minor with a time signature of $^{9}/_{8}$ of the human X-linked PGK coding sequence up to the 52nd codon is shown. If played on a violin, this transformation is hauntingly melancholy, as though reflecting the Weltschmerz of the gene that persevered for hundreds of millions of years.

Although we might regard musical composition as the most abstract and therefore probably the most intellectual form of human artistic endeavor, melodies are clearly not human inventions; the songs of skylarks, canaries, and certain other songbirds are as pleasing to our ears as they must be to themselves, as well as to their prospective mates. It is likely that in the early stages of development of our own species, the invention of rhythms preceded that of melodies.

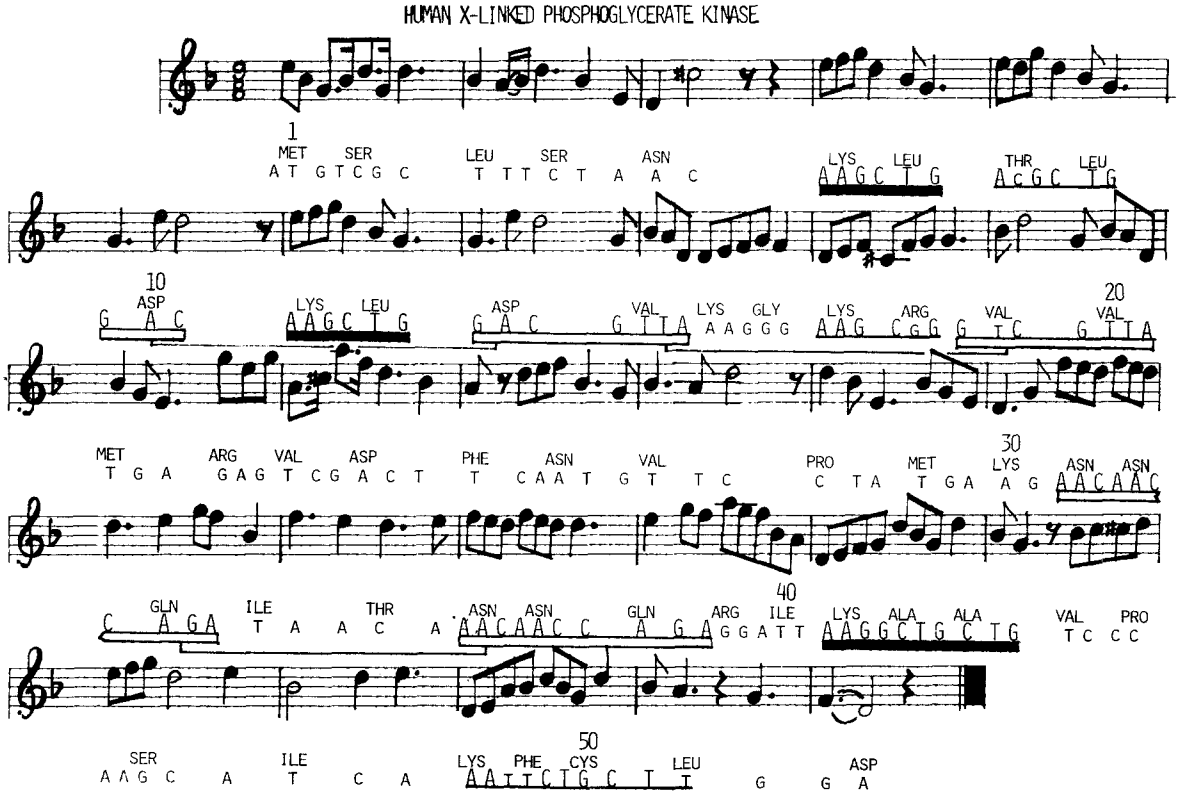

Fig. 3. Musical transformation of the first 52 codons of the human X-linked phosphoglycerate kinase coding sequence (Singer-Sam et al. 1983) in D minor and with a time signature of $^{9}/_{8}$. As the piece was written for the violin, only the treble clef musical score is given

Rhythmic sounds created by, for example, beating a hollowed tree trunk with wooden sticks might have originally served as a time-keeping device in ancestral societies of hunters and gatherers. It should be recalled that even today, music is often utilized for this purpose in dances and military parades. At the very beginning, each composition was no doubt an endless repetition of a chosen rhythm, thus resembling coding sequences which, at their inception, were repeats of base oligomers. Even after progressive refinement in the Renaissance, Baroque, and Romantic periods, repetitious recurrence is still the hallmark of musical composition. This is illustrated in Figure 4 with a section from Frederic Chopin's Nocturne, opus 55, no. 1. The treble clef musical score of Chopin's Nocturne is accompanied by a base sequence transcribed from it in accordance with the unambiguous rule set out in Figure 2. Transcribed are not only every grace note but also every note in each cord from the top to the bottom. A transcribed base sequence of Figure 4 is also accompanied by an amino acid sequence encodable by one long open reading frame. It should be noted that the segment contained in the sixth, seventh, and eighth rows of Figure 4, corresponding to the 51st to 81st codons of the transcribed base sequence, is a near exact repeat of the preceding segment corresponding to the 20th to 50th codons occupying the third, fourth, and fifth rows of Figure 4. The third repetition of the same segment begins at the extreme right of the eighth row starting with the 82nd codon. Thus, the all pervasive principle of repetitious recurrence is very evident, even in this relatively modern musical composition. Within the initial segment which ends at the extreme right of the fifth row and corresponds to the 50th codon, the principal subject transcribable to the primordial building block nonamer C A A C C T C C C (underscored by a thick solid bar) recurs thrice. Furthermore, there are five derivatives of this principal subject (underscored by thin solid bars). These derivatives sustained the insertion of a grace note or notes transcribable to A or two or more Ts and/or base substitutions; three times C to T, once each C to A and T to C. One short segment underscored by an open bar that begins with the triad cord transcribable to T A C also recurs first at the center right of the second row and a second time at the extreme right of the third row. This recurring segment, which is transcribable to the heptamer T A C G G T G, may be regarded as one of the secondary subjects.

As already mentioned, the segment occupying the sixth, seventh, and eighth rows of Figure 4, corresponding to the 51st to 81st codons of the transcribed base sequence, is a near exact recapitulation of the preceding segment corresponding to the 20th to 50th codons. It differs from the former only by two independent insertions of a grace note which are both transcribable to A, as seen in the eighth row. However, this recapitulation, which is evident in the transcribed base sequence, is not reflected in the amino acid sequence translated from it. This is because the second segment is translated into a different reading frame from the first. For many of the modern coding sequences of considerable antiquity, a reading frame shift usually results in a premature chain termination. Nevertheless, provided that the number of bases in the primordial building block oligomer was not a multiple of three, coding sequences were at their inception provided with three long open reading frames encoding polypeptide chains of identical periodicities. Long unused open reading frames persist to this day in certain kinds of coding sequences, notably those of bacterial plasmids, retroviruses, and *c-onc* genes in the vertebrate genome. Accordingly, these coding sequences are rather impervious to reading frame shifts, thus retaining a measure of immortality that was originally endowed by their ultimate ancestors (Ohno 1985). In this context, it is of great interest to note that in spite of the fact that the aforementioned 31 codon long, tandemly recurring segments can be translated into two different reading frames, the base sequence transcribed from the treble clef musical score of Chopin's Nocture retained one 160 codon long, open reading frame, although the sequence is only shown up to the 86th codon in Figure 4. The resemblance between coding base sequences and musical composition is indeed more than skin deep.

Homology between the Nocturne and the last exon of the largest subunit of mouse RNA polymerase II. It should be noted in Figure 5a and b that the primordial building block base oligomer of the last exon of the largest subunit of mouse RNA polymerase II (Corden et al. 1985) is the nonamer C A A C C T C T C which recurs four times in Figure 5a and b. This is nothing more than a single-base deviant of the principal subject of Chopin's Nocturne shown in Figure 4. Accordingly, the first 106 codons of this last exon were transformed into the musical score for the piano according to the modus operandi of Chopin's Nocturne. Both treble and base clef scores were transcribed, as shown in Figure 5a and b. When played on the piano, most listeners readily identified Chopin as its composer, but the similarity to his Nocturne was not noticed. This composition has a lively dance cadence, which is not altogether surprising because RNA polymerase is not a nocturnal creature, having instead to engage in transcriptional activity day and night. There is another difference. It should be recalled that in the case of the transcribed base sequence of Chopin's Nocturne, tandem repetition of the 31 codon long segment was not reflected in the amino acid sequence encodable by it, because two copies were to be translated into different reading frames. In the case of the last exon of the largest subunit of mouse RNA polymerase II, the exact heptapeptidic periodicity Tyr-Ser-Pro-Thr-Ser-Pro-Ser which begins at the extreme right of the fifth row of Figure 5a was not clearly reflected in the base sequence as an exact repetition of the 21 base long unit. This was because of multiple samesense base substitutions. For example, Tyr composed of six hep-

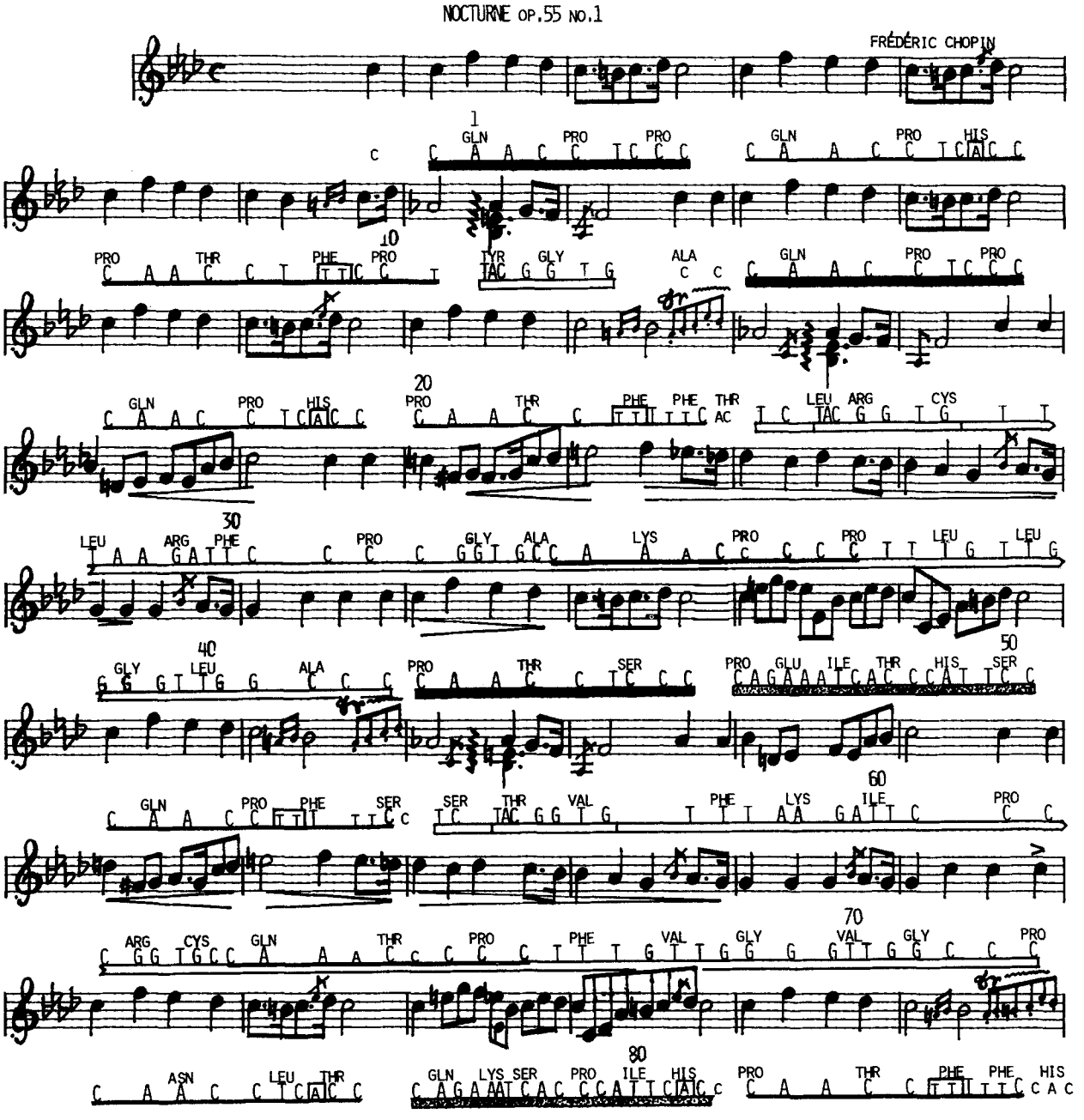

Fig. 4. Transcription of the initial portion of the treble clef musical score of the Nocturne, opus 55, no. 1 by Frederic Chopin to the base sequence according to the dictate specified in Figure 2. The amino acid sequence encodable by the longest open reading frame is also shown. The principal subject transcribable to the base nonamer C A A C C T C C C that recurs three times is underscored by *thick solid bars*, while nine variously base-inserted and/or base-substituted copies of this primordial building block are underscored by *thin solid bars*. Inserted bases are encased in *boxes*. Heptamer T A C G G T G, which is not directly related to the primordial nonamer, recurs thrice in Figure 4. These three copies are underscored by *open bars*. In actual fact, the segment shown in the sixth, seventh, and eighth rows corresponding to the 51st to 81st codons is the exact recapitulation, except for two inserted grace notes, of the preceding segment corresponding to the 20th to 50th codon. This recurring segment is divided into four sections: (1) a derivative of the primordial building block; (2) a 53 base long section underscored by an *open bar* that contained one copy of the heptamer T A C G G T G and one 2 base-substituted copy of the primordial building block; (3) the primordial building block or its single base inserted copy; (4) 17 or 18 base long segment underscored by a *shaded bar*

Fig. 5a and b. The musical transformation for the piano of the first 105 codons of the last exon of the largest subunit of mouse RNA polymerase II (Corden ▶ et al. 1985). Inasmuch as the primordial building block of this exon proved to be the nonamer C A A C C T C T C (which recurs four times within Figure 5a and b and is underscored by *thick solid bars*), and since this is only a single-base deviant of the primordial building block of Chopin's Nocturne shown in Figure 4, its musical transformation with regard not only to the treble clef score but also to the base clef score corresponds to the modus operandi of Chopin in general and to his Nocturne in particular. Nevertheless, the composition is very distinct from the Nocturne, having a livelier dance cadence

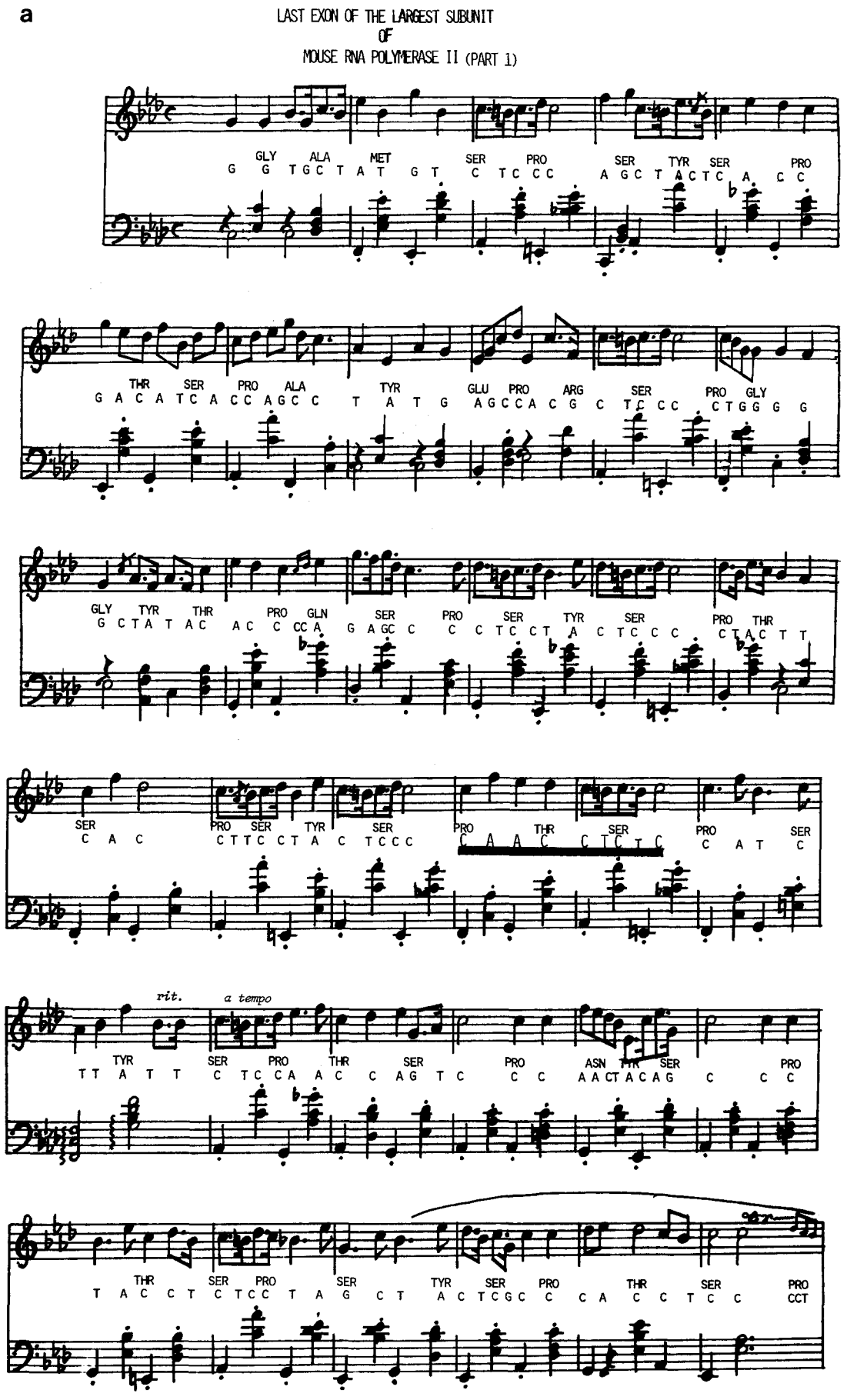

Fig. 5a. (Legend, see page 75)

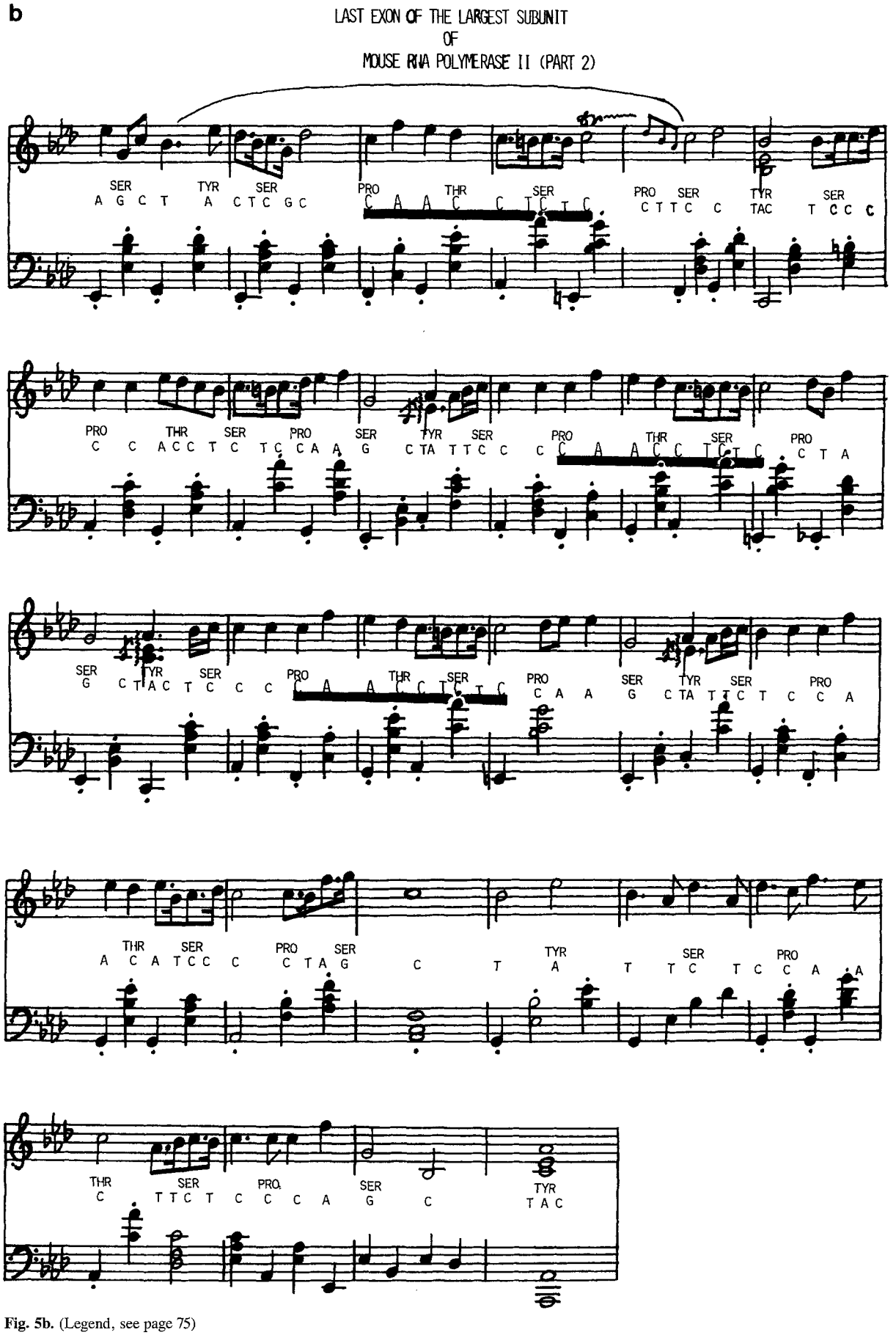

Fig. 5b. (Legend, see page 75)

tapeptidic units in Figure 5b was encoded three times each by T A C and T A T. Indeed, in Figure 5b, the primordial nonamer C A A C C T C T C recurred only thrice instead of six times. The heptapeptidic periodicity in this instance is not merely the accidental remains of the original periodicity. Rather, it has been actively maintained by natural selection because of its critical role in the proper functioning of RNA polymerase II.

References

Corden, J. L., Cadena, D. L., Ahearn, Jr., J. M., and Dahmus, M. E.: A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proc. Natl. Acad. Sci. U.S.A. 82:* 7934–7938, 1985

DeVries, A. L.: Biological antifreeze agents in cold water fishes. *Comp. Biochem. Biophysiol. 73A:* 627–640, 1982

Gitschier, J., Wood, W. I., Goralka, R. M., Wion, K. L., Chen, E. Y., Eaton, D. H., Vehar, G. A., Capon, D. J., and Lawn, R. M.: Characterization of human factor VIII gene. *Nature 312:* 326–330, 1984

Klein, J. and Figueroa, F.: Evolution of the major histocompatibility complex. *CRC Crit. Rev. Immunol.*, in press, 1986

Martin-Zaca, D., Hughes, S. H., and Barbacid, M.: A human oncogene formed by the fusion of truncated tropomyosin protein kinase sequences. *Nature 319:* 743–748, 1985

Mostov, K. E., Friedlander, M., and Blobel, G.: The receptor for transepithelial transport of IgA and IgM contains multiple immunoglobulin-like domains. *Nature 308:* 37–43, 1984

Ohno, S.: *Evolution by Gene Duplication,* Springer-Verlag, Heidelberg, 1970

Ohno, S.: Repeats of base oligomers as the primordial coding sequences of the primitive earth and their vestiges in modern genes. *J. Mol. Evol. 20:* 313–321, 1984

Ohno, S.: Immortal genes. *Trends Genet. 1:* 196–200, 1985

Ohno, S. and Epplen, J.: The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. *Proc. Natl. Acad. Sci. U.S.A. 80:* 3391–3395, 1983

Singer-Sam, J., Simmer, R. L., Keith, D. H., Shively, L., Teplitz, M., Itakura, K., Gartler, S. M., and Riggs, A. D.: Isolation of a cDNA clone for human X-linked 3-phosphoglycerate kinase by use of a mixture of synthetic oligodeoxyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A. 80:* 802–806, 1983

Watson, H. D., Walker, N. P. C., Shaw, P. J., Bryant, T. N., Wendell, P. L., Fothergill, L. A., Perkins, R. E., Conroy, S. C., Dobson, M. J., Tuite, M. F., Kingsman, A. J., and Kingsman, S. M.: Sequence and structure of yeast phosphoglycerate kinase. *EMBO J. 1:* 1635–1640, 1982

Zavala, F., Tam, J. P., Cochrane, A. H., Quakyi, I., Nassenzwieg, R. S., and Nassenzwieg, V.: Synthetic oligopeptide immunization against *Plasmodium falsiparum. Science 228:* 1436–1440, 1985

Received April 8, 1986