Haematology and Blood Transfusion Vol. 29
Modern Trends in Human Leukemia VI
Edited by Neth, Gallo, Greaves, Janka
© Springer-Verlag Berlin Heidelberg 1985

Cellular Oncogenes as the Ancestors of Endocrine and Paracrine Growth Factors and Their Evolutionary Relic Status in Vertebrates

S. Ohno[1]

A. Introduction

The generally held belief that any gene whose expression is precisely regulated in development ought to perform an indispensable function to the host organism is not quite correct and, in fact, has no solid foundation. It should be recalled that the DNA replication mechanism of modern organisms has developed a high degree of precision (the error rate is 10^{-10} per base pair per replication or thereabout) and that whatever damage sustained by DNA is effectively rectified by multitudes of DNA repair mechanisms. Accordingly, even dispensable peptide chains such as fibrinopeptides A and B do not change their amino acid sequences very rapidly, a 1% change in their amino acid sequences taking roughly 1.0–1.6 million years [1]. Under these circumstances, a gene which has lost its usefulness to the host organism does not disappear quite so readily. The half-life of an enzyme gene that has become redundant has been estimated as 50 million years [2].

B. Redundant and Useless Genes May Persist for 50 Million Years or More

By becoming tetraploid, an organism initially gains four alleles at every gene locus. Subsequently, each set of four homologous chromosomes differentiates into two pairs, thus completing the process of diploidization. At this stage, freshly diploidized tetraploid species are endowed with twice the number of gene loci when compared with their diploid counterparts. This is the stage, at which trout and salmon of the teleost family Salmonidae, whitefish of the family Coregonidae, and grayling of the family Thymallidae find themselves [3]. A few of those duplicated, and therefore redundant, genes manage to acquire a new role; e.g., of all the vertebrates, only diploidized tetraploid teleost fish are endowed with liver-specific lactate dehydrogenase (LDH), in addition to the customary skeletal muscle and heart LDH. The mechanism of gene duplication as the means to acquire new genes with previously nonexistent functions, however, is very inefficient, having a very low success ratio: the phrase Salvandrum paucitas, dammnundrum multitudo gives ample testimony to its high failure ratio. Accordingly, older diploidized tetraploids of the teleost family Cyprinidae as well as Catostomidae have lost progressively larger numbers of these redundant, duplicated loci by silencing mutations. Since the fossil record gives the origin of these diploidized tetraploids, Ferris and Whitt [2] were able to calculate the average half-life of enzyme loci that became redundant as 50 million years. It should be noted here that this half-life refers to the average time needed for half of the redundant enzyme loci to lose their assigned functions. After losing their assigned functions, these redundant enzyme loci may continue to code for functionless polypeptide chains. The case in

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

point is the murine Slp locus, situated in the middle of the major histocompatibility (MHC) antigen gene complex region of the mouse genome. While the neighboring Ss locus specifies C4 (complement 4 of antibody-mediated lysis), a protein specified by the Slp locus has already lost its assigned function as C4 owing to accumulated mutations. Yet, this Slp locus was androgen dependent in most mouse strains, and operator constitutive mutation of this locus was found in wild mice [4]. It would thus appear that a substantial portion of the redundant gene loci may continue to specify functionless proteins even after 100 million years of independence from natural selection.

Although mammal-like reptiles were already an independent lineage at the time of the dinosaurs, mammals as we know them came into being only 70 million years ago. This should make us realize the unreality of the statement that any gene loci with precisely regulated expression must be indispensable to the host organism. The fact is that genes that have outlived their usefulness may linger on for 50–100 million years.

C. Most Oncogenes are Evolutionary Relics of the Cell Autonomous Stage of Development

Although multitudes of cellular oncogenes perform divergent functions (some of their products are found in the nucleus, while others are found inside the plasma membrane), it is clear that all of them function as intracellular cell growth factors. In unicellular eukaryotes such as baker's yeast as well as in many of the multicellular eukaryotes with underdeveloped circulatory systems such as insects, these intracellular growth factors have apparently played a vital role, for it should be recalled that embryonic development of insects is still largely a cell autonomous process as discussed in detail elsewhere [5]. With the advent of the cardiovascular system, development of vertebrates became a centrally controlled affair via multitudes of peptide and steroid hormones and cellular autonomy was suppressed. While intracellular growth factors of earlier times served as ancestors of these peptide hormones as well as of their plasma membrane receptors, they themselves largely became evolutionary relics whose functions have become redundant [5].

D. Near Immortality of Certain Oncogenes Conferred on Them by Their Original Construction

If cellular oncogenes became redundant at the onset of vertebrate evolution, most of them should have become silent by now, in spite of their long estimated half-life of 50 million years, for primitive vertebrates were already in evidence more than 300 million years ago. However, the continued performance of essential functions need not be invoked to explain this persistence for more than 300 million years.

The view first expressed in 1981 [6] that all the coding sequences originally were repeats of base oligomers has found increasing support from independent sources [7–9]. Provided that the number of bases in the oligomeric unit is not a multiple of three, coding sequences made of oligomeric repeats are inherently impervious to normally very damaging base substitutions, deletions, and insertions, thus possessing a near immortality. In this kind of oligomeric repeat, three consecutive copies of the oligomeric unit translated in three different reading frames gives the unit periodicity to their polypeptide chains: while nonameric repeats, the unit sequence being a multiple of three, can give only tripeptide periodicity to their peptide chains, three consecutive copies of decameric repeats encode the decapeptidic periodicity to its polypeptide chain. Thus, if one reading frame of this kind of oligomeric repeat is open, the other two are automatically open as well. It follows then that the potentially most damaging base substitution that changes an amino acid-specifying codon to the chain terminator (e.g., Trp codon TGG to chain-terminating TAG or TGA) merely silences one of the three open reading frames. Deletions or insertions of bases that are not multiples of three are usually as damaging, for resulting frame shifts alter downstream amino acid sequences and most often result in premature chain terminations. In this type of oligomeric repeat, such insertions and deletions are of no consequence either, for downstream amino acid sequences are not at all affected by frame shifts.

In a previous paper [10], we have analyzed the published coding sequence of human c-*myc* gene [11] in detail. Within the 5′ half of c-*myc* coding sequence, we identified one each of recurring base tetradecamer, duodecamer, and two monodecamers. The significance of this becomes clear once it is realized that if c-*myc* is a unique sequence *sensu stricto,* even a given base decamer is expected to recur only once every 1 048 576 bases. Yet, here we found a recurring base tetradecamer within a mere 687-base 5′ half c-*myc* coding sequence. Furthermore, recurring duodecamer and monodecamers were found to represent slightly modified parts of the tetradecameric sequence GGCCGCCGCCTCCT. Thus, it was concluded that the entire 5′ half of the c-*myc* coding sequence originated from repeats of the previously noted base tetradecamer. Since 14 is not a multiple of 3, three consecutive copies of it translated in three different reading frames would have given the following tetradecapeptidic periodicity to the original c-*myc* polypeptide chain, at least the amino terminal half of it

Gly Arg Arg Leu Leu
GGC CGC CGC CTC CT/G

Ala Ala Ala Ser Trp
GCC GCC GCC TCC T/GG

Pro Pro Pro Pro
CCG CCG CCT CCT

Indeed, the human c-*myc* coding sequence, at least the 5′ half of it, apparently inherited a measure of immortality from its original construction, for we found two long, alternative open reading frames, one covering the first 301 bases and the other from the 599th to 952nd bases. When this region of human c-*myc* coding sequence [11] was compared with the corresponding region of v-*myc* coding sequence of avian retrovirus MC29 [12], we found that the two differed from each other not so much by amino acid substitutions as by five stretches of insertions and two stretches of deletions. Thus, c-*myc*'s inherent imperviousness to deletions and insertions was shown [10].

E. Resurrection of a Silenced v-*src* Gene by Utilization of its Alternative Open Reading Frame

A measure of immortality inherited by some of the oncogenes from their original construction was indeed shown by the following experiment of Mardon and Varmus [13]. First, they established the rat cell line that was transformed by the integration into the genome of a single copy of strain B77 Rous sarcoma virus v-*src* coding sequence. One of the defective mutations sustained by the integrated v-*src* that deprived from the rat cell line of the transformed phenotype was identified as an insertion of a single base A between 146th Glu codon GAA and 147th Glu codon GAG. A resulting frame shift created a new chain terminator 51 bases further downstream, thus, silencing a mutated v-*src* [13]. The surprise was the second mutation that resurrected a silenced v-*src* as a transforming gene. This second event was an insertion of a duplicated 242-base segment into the position between T and GG of the 148th Trp codon in the original reading frame. This 242-base segment started from GAT representing the 68th Asp codon in the original reading frame and ended in T of 148th Trp also in the original reading frame, thus including a previously inserted A. Since the inserted segment is now translated in an alternative reading frame, the resulting double frame shifts restored the original reading frame, starting from GAG of the 147th Glu of the wild-type v-*src* and downward which in the resurrected v-*src* became the 228th Glu.

Such restoration of function by an insertion in the midst of the polypeptide chain of an 81-residue new amino acid sequence is hardly believable, unless a new sequence specified by a repeated coding segment translated in an alternative open reading frame resembles parts of the preexisting amino acid sequence. Such a resemblance, in turn, is expected only if the coding sequence itself still maintains a sufficient vestige of the ancestral construction;

i.e., the coding sequence originating from repeats of a base oligomer in which the number of bases in the oligomeric unit was not a multiple of three. Indeed, the existence of so long an alternative open reading frame itself is a reflection of the v-*src* coding sequence's ultimate derivation from oligomeric repeats, the number of bases in the oligomeric unit not being a multiple of three. As might be expected, when translated in an alternative open reading frame, amino acid residues 1–8 encoded by a duplicated 242-base segment were Thr-Pro-Se-Arg-Arg-Arg-Ser-Val. In the standard amino acid sequence of Rous sarcoma v-*src*, the very similar nonapeptide, Thr-Pro-Ser-(Gln)-Arg-Arg-Arg-Ser-Leu customarily occupies positions 10–18 [5].

F. Summary

Contrary to the popularly held view, genes that have lost their usefulness to the host organism may continue to encode proteins for 50 million years or longer. Accordingly, precisely regulated expression of genes can not be taken as proof of their indispensability. My view is that multitudes of oncogenes of vertebrates are evolutionary relics harking back to the days of invertebrate ancestors in which embryogenesis was still a cell autonomous process. Parts of certain oncogene coding sequences originated from repeats of base oligomers whose numbers of bases were not multiples of three. Thus, these segments are still endowed with a measure of immortality in that they are impervious to normally very deleterious base substitutions, insertions, and deletions.

References

1. Dayhoff MO (ed) (1972) Atlas of protein sequences and structure. National biomedical research foundation, Silver Springs, Maryland
2. Ferris SD, Whitt GS (1977) Loss of duplicated gene expression after polyploidization. Nature 265:258–260
3. Ohno S (1970) Evolution by gene duplication. Springer, Heidelberg Berlin New York
4. Hansen TH, Shreffler DC (1976) Characterization of a constitutive variant of the murine serum protein allotype, *Slp*. J Immunol 117:1507–1513
5. Ohno S (1984) Repeats of base oligomers as the primordial coding sequence of the primeval earth and their vestiges in modern genes. J Mol Evol 20:313–321 (1984)
6. Ohno S (1981) Original domain for the serum albumin family arose from repeated sequences. Proc Natl Acad Sci USA 79:1999–2002
7. Blake C (1983) Exon – present from the beginning? Nature 306:535–537
8. Gō M (1983) Modular structural units, exons and functions in chicken lysozyme. Proc Natl Acad Sci USA 80:1964–1968
9. Alexander F, Young PR, Tilghman SM (1984) Evolution of the albumin:α-fetoprotein ancestral gene from the amplification of a 27 nucleotide sequence. J Mol Biol 173:159–174
10. Ohno S, Yazaki A (1983) Simple construction of human c-*myc* gene implicated in B-cell neoplasms and its relationship with avian v-*myc* and human lymphokines. Scand J Immunol 18:373–388
11. Watt R, Stanton LW, Marcu KB, Gallo RC, Croce CN, Tovera G (1983) Nucleotide sequence of cloned cDNA of human c-*myc* oncogene. Nature 303:725–727
12. Alitalo K, Bishop MJ, Smith DH, Chen E, Colby WW, Levinson AD (1983) Nucleotide sequence of the v-*myc* oncogene of avian retrovirus MC29. Proc Natl Acad Sci USA 80:100–105
13. Mardon G, Vermus HE (1983) Frameshift and intragenic suppressor mutations in a Rous sarcoma provirus suggest *SRC* encodes two proteins. Cell 32:871–879