Nucleotide Compositional Constraints on Genomes Generate Alanine-, Glycine-, and Proline-rich Structures in Transcription Factors

Yutaka Nakachi, Toshiyuki Hayakawa, Hiroki Oota, Kenta Sumiyama, Li Wang, and Shintaroh Ueda

Department of Biological Sciences, Graduate School of Science, University of Tokyo

Correlation between amino acid composition and nucleotide composition is examined. Class III POU transcription factors having higher third GC contents showed higher contents of alanine, glycine, and proline residues encoded by GC-rich nucleotides, and vice versa. This correlation was observed even among various types of transcription factors from vertebrates and invertebrates regardless of functional and structural constraints inherent to each protein. Furthermore, reptile class III POU sequences revealed no evolutionary directionality increasing the GC contents from cold- to warm-blooded vertebrates.

Introduction

Proteins have unique amino acid sequences that are specified by their corresponding nucleotide sequences. Since there are 20 amino acids and their corresponding 61 codons, many amino acids are encoded by two or more codons. Such degeneracy of the genetic code permits nucleotide sequences to vary without altering amino acid sequences. As a significant portion of the nucleotide changes at the third codon position are synonymous (silent), the third codon position is considered to directly reflect the degree of nucleotide compositional constraints onto DNAs harboring the sequences.

We have recently discovered a significant correlation between the GC content at the third codon position (the third GC content) and the homopolymeric amino acid repeat ratio of the mammalian class III POU transcription factor genes (Sumiyama et al. 1996): the mammalian Brain-1, Brain-2, and Scip genes have homopolymeric amino acid repeats (sequences without interruptions in the run of a single amino acid residue) including alanine, glycine, and proline, whereas most or all of these repeats are absent from their homologs in nonmammals (amphibians and fish). These characteristic amino acid repeats are well conserved in both position and repeat number among mammals. In contrast, the mammalian Brain-4 gene, like its nonmammalian homolog, has no homopolymeric amino acid repeats. The mammalian Brain-1, Brain-2, and Scip genes containing the homopolymeric amino acid repeats have a higher third GC content. In contrast, the respective nonmammalian homologs lacking the homopolymeric amino acid repeats have a lower third GC content. However, the mammalian Brain-4 gene, like its nonmammalian homolog, has a lower third GC content. There was a clear positive correlation between the homopolymeric amino acid repeat ratio and the third GC content. The amino acids of these characteristic homopolymeric repeats were encoded mainly by codons with a relatively high GC content. These findings indicate that nucleotide compositional constraints increasing the GC contents (GC pressure) have facilitated the generation of homopolymeric amino acid repeats in mammalian class III POU transcription factors.

To understand how protein sequences have evolved by compositional constraints on genomes, we first determine the class III POU gene structures in reptiles. We then examine correlation between the amino acid composition of proteins and the third GC content using not only the vertebrate class III POU genes, but also various transcription factors from vertebrates and invertebrates.

Key words: GC pressure, POU, transcription factor, reptile.

Address for correspondence and reprints: Shintaroh Ueda, Laboratory of Molecular Biology and Evolution, Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan. E-mail: sueda@uts2.s.u-tokyo.ac.jp.

Mol. Biol. Evol. 14(10):1042–1049. 1997

Materials and Methods

Reptile Class III POU Genes

Recombinant phage clones containing the reptile class III POU genes were isolated from the green anole (*Anolis carolinensis*) genomic library using the POU domain of the chicken Brain-1 gene as a probe. As washing condition in hybridization was lowly stringent (four times for 30 min each in 0.15 M sodium chloride and 0.015 M sodium citrate at 55°C), these clones were identified by hybridization using DNA fragments upstream of the POU domain as probes and partially determining their nucleotide sequences. Nucleotide sequences of the class III POU genes were determined by the dideoxynucleotide chain termination method on both strands. The nucleotide sequence data reported in this paper will appear in the DDBJ/EMBL/GenBank international nucleotide sequence database with the accession numbers AB001868–AB001870.

Sequence Analysis

Transcription factors employed here were as follows; Brain-1 (AB001835, M88299, AB001868, D13045), Brain-2 (L37868, M88300, L27663, AB001869, X64835, Y07905), Brain-4 (X82324, M88301, M84645, AB001870, U17654), and Scip (L26494, X54628, M35205/M72711, X59056, X96422) for the vertebrate class III POU; Evx-1 (D10455, X54239, X60655), Gbx-2 (L39770, L47990), goosecoid (L03395, M81481, M85271, X70471), hairy (L04527, L19314, U36194), HNF-3β (L10409, L25637), HoxA7 (M17192, M24752), HoxB4 (M26884, M36654), HoxB7 (M16937, X06592), HoxC6 (X12499, X16510), Mash-1 (L11871, M98272, U14587, X53725), Mox2

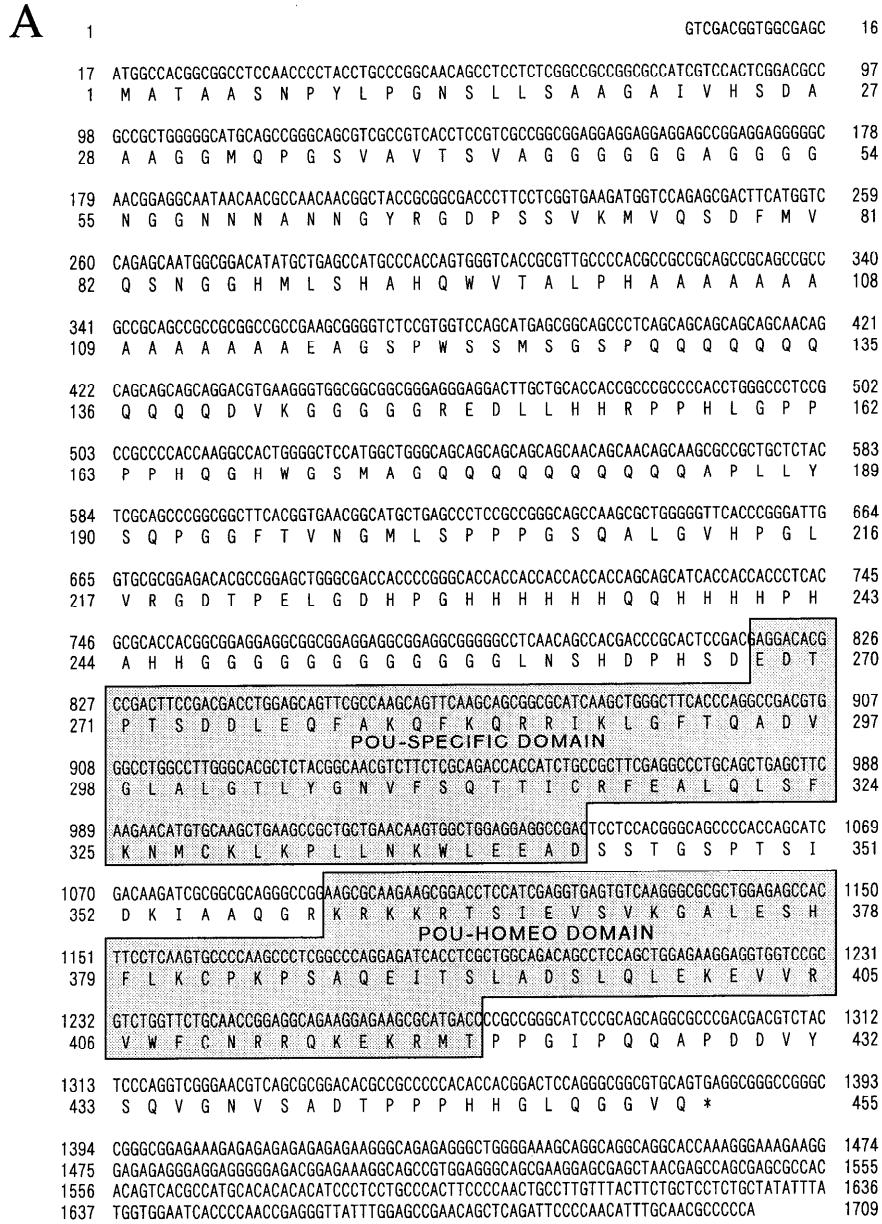

FIG. 1.—Nucleotide sequences of the green anole (*A*) Brain-1, (*B*) Brain-2, and (*C*) Brain-4 genes and their deduced amino acid sequences. The POU-specific and POU-homeo domains are indicated with shadowed boxes. An asterisk indicates a stop codon. We used the notation of the sequences deposited in the DDBJ/EMBL/GenBank international nucleotide sequence database for numbering nucleotides presented here.

(L20432, Z16406), neuroD (U28067, U50822), Nkx-2.5 (L25600, S70708), Otx1 (L32602, U14591), Pax6 (M93650, X63183, X63963), and Oct1 (M29972, X13403, X17190) for vertebrates; and achaete-scute AS-T3 and AS-T8 (X12549/Y00846, X12550/Y00846), apterous (M92841), atonal (L36646), Cf1a (M81959), daughterless (Y00221), distalless (S47947), empty spiracle (X51653), engrailed (M10017), even-skipped (X05138), extramacrochaetae (M31902), fushitarazu (M62856), goosecoid (X95420), hairy (M87885), orthodenticle (X58983), runt (U22357), sloppy paired slp1 and slp2 (X66095, X669098), and tailless (M34639) for invertebrates, respectively. For multiple alignment of the amino acid sequences, we initially used a multiple alignment program in CLUSTAL W (Thompson, Higgins, and Gibson 1994) and further visually adjusted to increase similarity.

Results and Discussion

Genomes of warm-blooded vertebrates are reported to be mosaics of very long (more than 200 kb) DNA segments called isochores, whereas those of cold-blooded vertebrates are compositionally uniform; there is a wide variation in the GC content among genes of warm-blooded vertebrates, but cold-blooded vertebrate genes show the relatively lower GC content within a narrow range. Bernardi proposed that directional nucleotide substitutions have occurred during evolution from cold- to warm-blooded vertebrates (Bernardi et al. 1985; Ber-

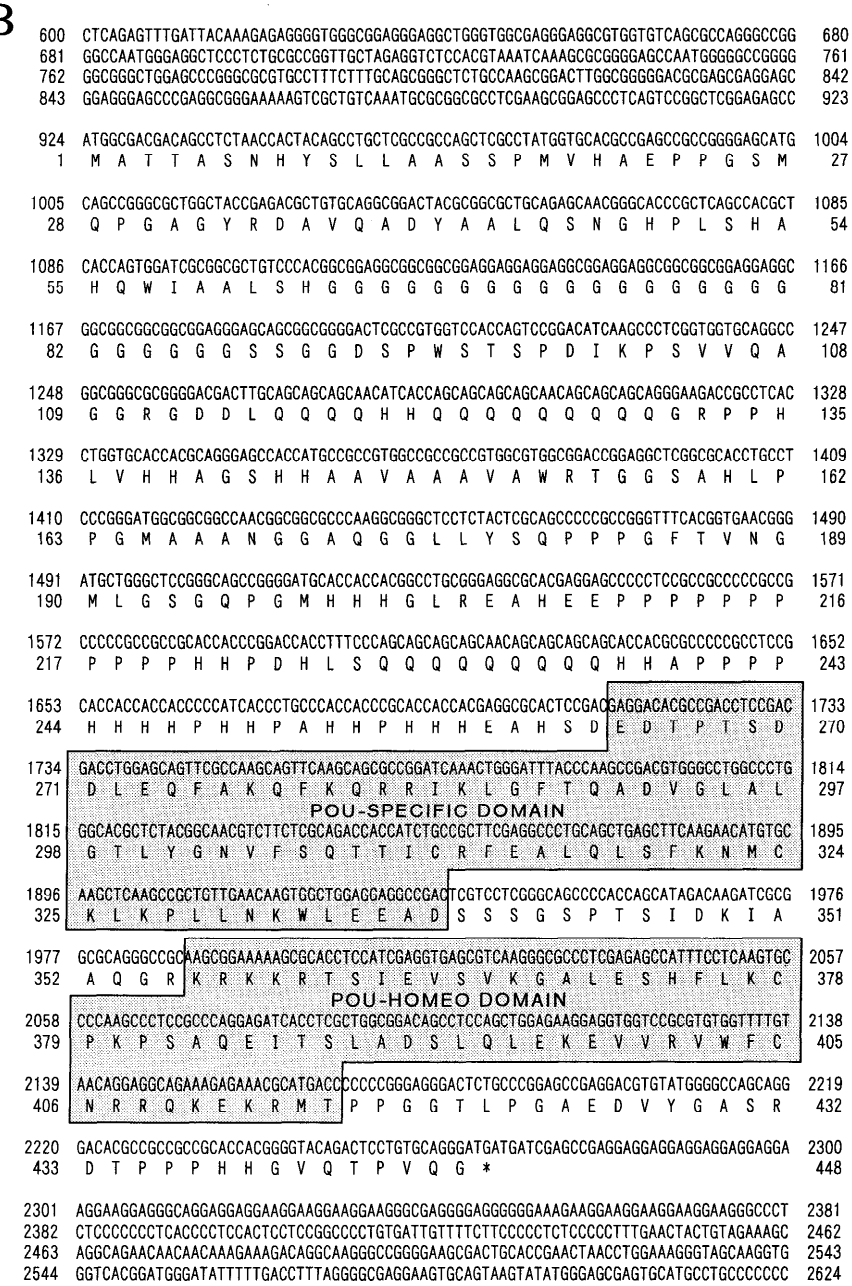

FIG. 1 *(Continued)*

nardi and Bernardi 1991; Kadi et al. 1993; Bernardi 1995). A mammalian genome is constituted by a mosaic of very long DNAs with different GC contents, in each of which the GC content is fairly homogeneous (Ikemura and Aota 1988). We have recently discovered a significant correlation between the third GC content and the homopolymeric amino acid repeat ratio of the vertebrate class III POU genes (Sumiyama et al. 1996). The class III POU genes from *Xenopus* and zebrafish with few homopolymeric amino acid repeats had low third GC contents, whereas those from mammals (human, mouse, and rat) showed two phases opposite to each other: genes containing several homopolymeric amino acid repeats such as Brain-1, Brain-2, and Scip showed higher third GC contents, while those containing very few homopolymeric amino acid repeats such as Brain-4 showed lower third GC contents. Therefore, genes of reptiles belonging to cold-blooded vertebrates are expected to have no or few homopolymeric amino acid repeats like amphibians and fish. We obtained complete nucleotide sequences of the reptile Brain-1, Brain-2, and Brain-4 genes. Contrary to our expectation, all of the reptile class III POU genes had homopolymeric amino acid repeats (see fig. 1).

The reptile Brain-1 gene had a number of homopolymeric amino acid repeats nearly equal to that of the mammalian Brain-1 gene (fig. 2*A*), but the taxon-specific homopolymeric amino acid repeats were found in addition to the common repeats between mammals and reptiles. Mammalian-specific repeats were alanine-re-

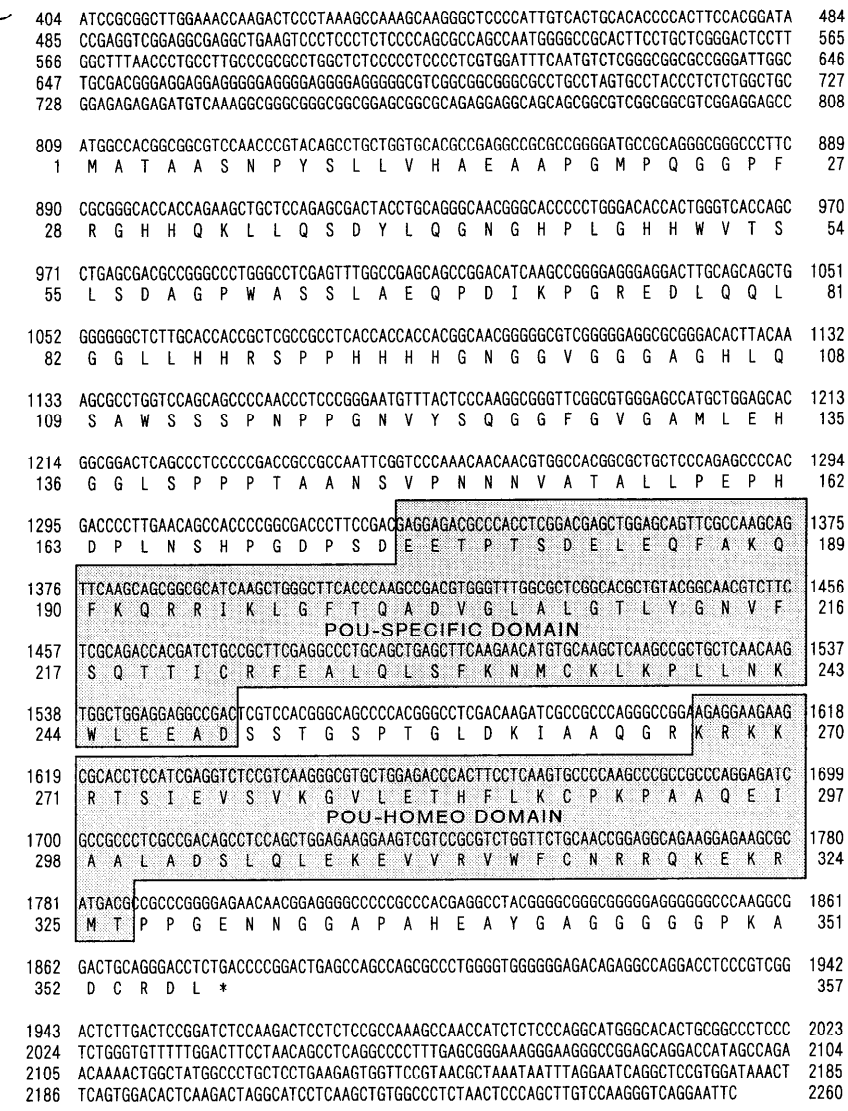

FIG. 1 *(Continued)*

peats in positions 214–229, glycine repeats in positions 31–44, 46–51, and 271–276, and proline repeats in positions 163–168 and 199–204, whereas reptile-specific repeats were glycine repeats in positions 64–69 and 175–179 and glutamine repeats in positions 158–168 and 238–247. Furthermore, there was a wide variation in repeat number of the homopolymeric amino acid repeats common to mammals and reptiles, although all of the amino acid repeats were well conserved among mammals not only in position but also in repeat number. In the Brain-2 gene, the situation was the same as that in the Brain-1 gene. In addition to taxon-specific repeats, there were common repeats including glycine repeats in positions 69–92 and glutamine repeats in positions 139–161 between mammals and reptiles, but a wide variation in repeat number was observed as in the Brain-1 gene (fig. 2*B*). Surprisingly, the reptile Brain-4 gene also had homopolymeric amino acid repeats, glycine repeats in positions 366–370 (fig. 2*C*), different from both the mammalian and amphibian homologs with no characteristic amino acid repeats.

Present data on the reptile class III POU genes were consistent with our previous conclusion (Sumiyama et al. 1996): genes having homopolymeric amino acid repeats showed higher third GC contents. The reptile Brain-4, however, did not show a good fit to the plot between the homopolymeric amino acid ratio and the third GC content (see fig. 3*A*). Moreover, it had another remarkable feature: less amino acid sequence similarity to the mammalian homologs than to that of the amphibian homolog, although mammals are more closely related to reptiles than to amphibians. The phylogenetic tree using the well-conserved POU domain sequence showed the cluster of the mammalian and reptile Brain-4 genes, excluding the possibility that the reptile Brain-4 obtained here was not a homolog to the mammalian and amphibian Brain-4 (data not shown).

Through overall comparison with the other vertebrate homologs, we noticed extraordinary contents of alanine (A), glycine (G), and proline (P) residues in the reptile Brain-4. The codons for these amino acids are GC-rich (GCN, GGN, and CCN, respectively). Con-

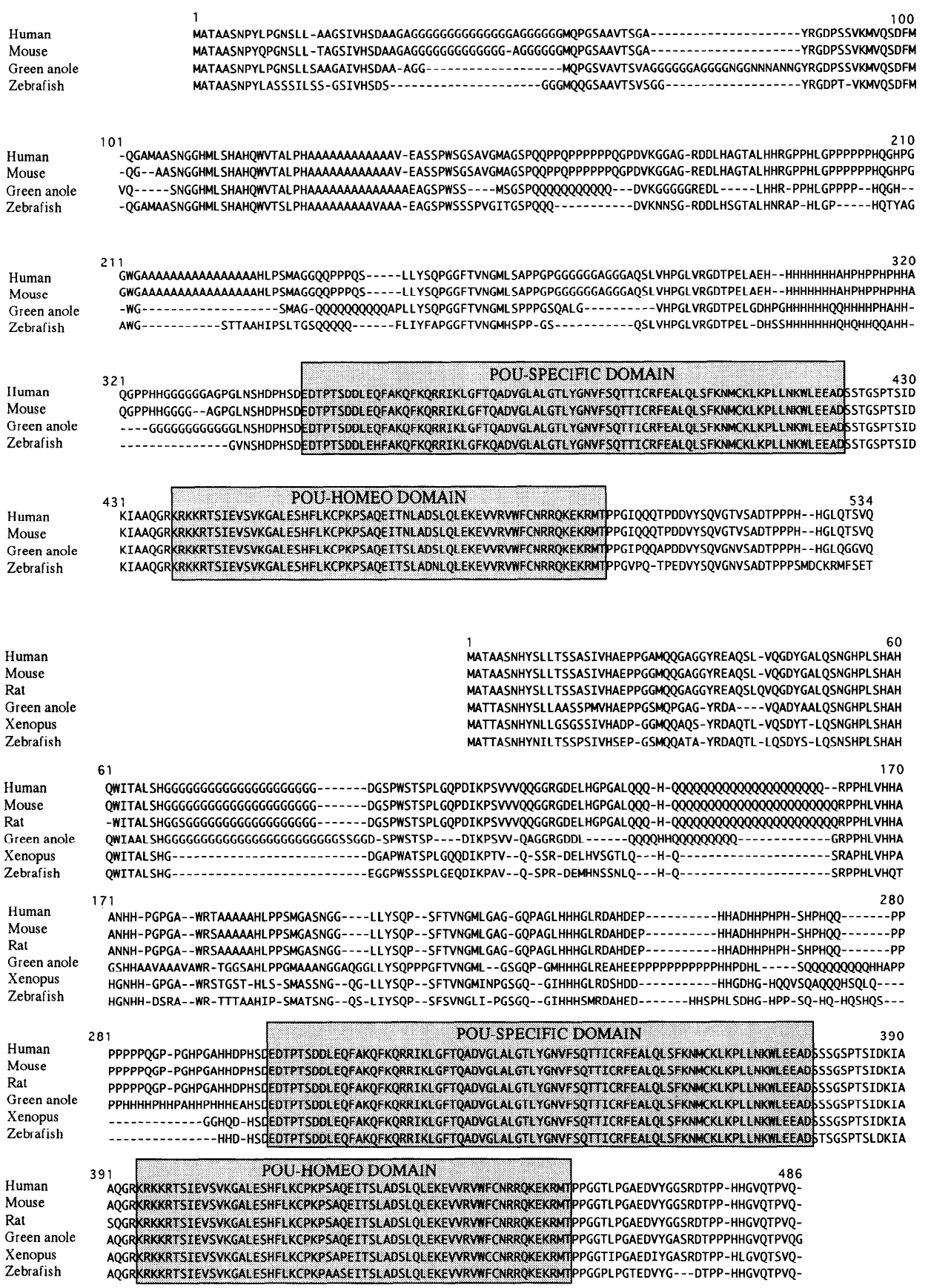

FIG. 2.—Sequence alignment within the homologs of the class III POU genes. *A*, Brain-1. *B*, Brain-2. *C*, Brain-4. The POU-specific and POU-homeo domains are indicated with shadowed boxes. Gaps are represented by hyphens.

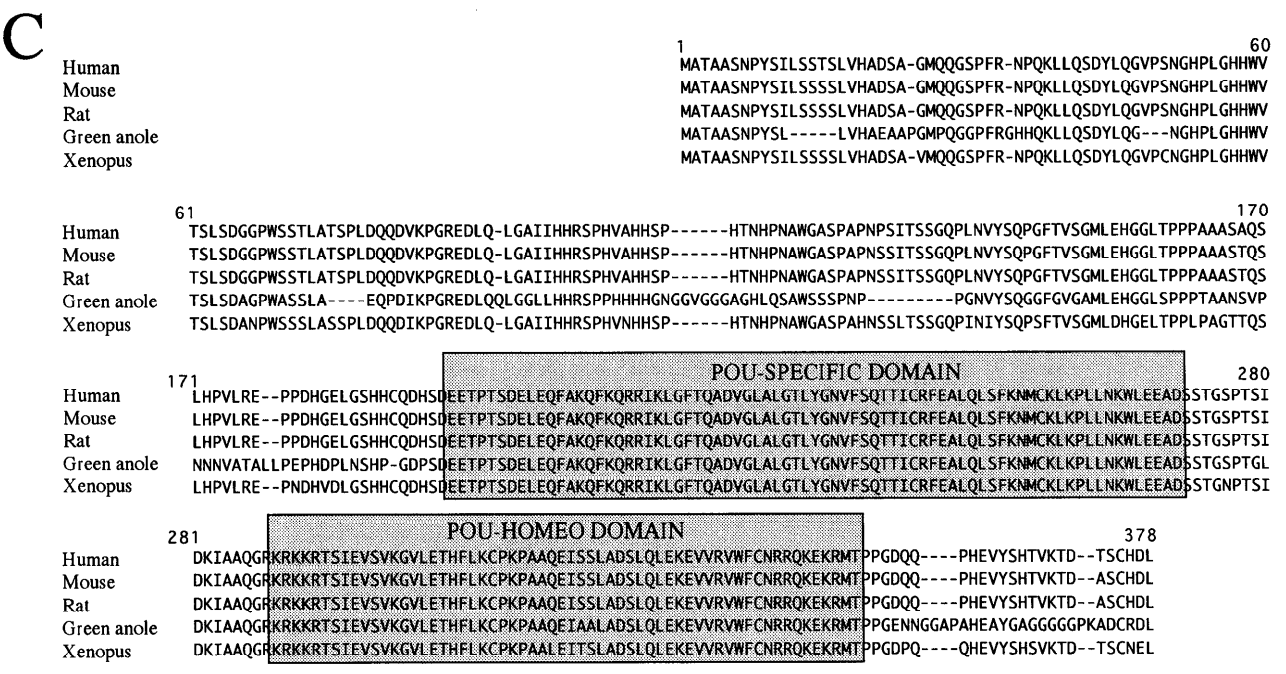

FIG. 2 *(Continued)*

versely, the number of arginine residues, the remaining amino acid encoded by GC-rich codons, did not vary among the vertebrate Brain-4 genes (14 residues for all the vertebrate Brain-4 genes). The Brain-2 gene also showed no variation in number of arginine residues (16 for all the vertebrates). We found some variations in arginine number in both Brain-1 and Scip genes, but their difference is only one residue (14 and 15 for human/mouse/green anole and zebrafish Brain-1, respectively, and 15 and 16 for human/mouse/rat/*Xenopus* and zebrafish Scip, respectively). We thus used the total content of A, G, and P residues (AGP content) to investigate a correlation between the third GC content and the amino acid composition among the vertebrate class III POU genes. As the third GC content of the entire region is shown to be nearly equal to that of the POU domain with no homopolymeric amino acid repeats (Sumiyama et al. 1996), we used the third GC content of the entire region so as to make comparisons among a wide variety of transcription factors possible, as mentioned below. We found a significant correlation between the third GC and AGP contents (correlation coefficient was 0.82), as shown in figure 3*B*.

Phylogenetic analysis indicates that Brain-1, Brain-2, Brain-4, and Scip genes already existed in the common ancestor of vertebrates (Sumiyama et al. 1996). These genes are not tandemly located on the same chromosome, at least in mammals; the mammalian Brain-4 gene is located on the evolutionarily well-conserved X chromosome, while the other class III POU genes are on autosomes (Avraham et al. 1993; Xia et al. 1993; Atanasoski et al. 1995; de Kok et al. 1995). The present data thus suggest that (1) the ancestral class III POU gene possesses no enrichment of A, G, or P residues, (2) enrichment of AGP residues including generation of the homopolymeric amino acid repeats occurred independently both in particular lineages and in particular class III POU genes (AGP enrichment in the Brain-1 and Brain-2 genes occurred both in the common ancestor of amniotes and independently in each lineage of amniotes, whereas that in the Brain-4 gene occurred only in reptiles). Moreover, we conclude that the hypothesis based on evolutionary directionality increasing the GC contents from cold- to warm-blooded vertebrates is incorrectly drawn due to sparse sequence data on cold-blooded vertebrates except for amphibians. In fact, there are only five reptile genes used in figure 3 of Bernardi and Bernardi (1991), and the GC contents of those appear to be nearly equal to those of the mammalian homologs.

A similar situation holds for other transcription factors of vertebrates: a wide variation of AGP content and a nearly equal arginine content. We found a clear tendency for the AGP content to increase relative to the third GC content in each transcription factor (data not shown). We thus plotted the AGP content against the respective third GC content for all the vertebrate transcription factors studied (fig. 4*A*). To our surprise, there was a clear positive correlation (correlation coefficient was 0.72) regardless of functional and structural constraints inherent in each protein. We also analyzed data from both vertebrates and invertebrates (fig. 4*B*) and again found a positive correlation (correlation coefficient was 0.71). No arginine-rich region was found in any of the transcription factors examined. This observation is compatible with the frequency distribution of homopolymeric amino acid repeats of proteins in general (Green and Wang 1994).

Present results provide a general picture for protein structure and its evolution: amino acid compositions are under profound influence of nucleotide compositional constraints on genome DNAs harboring coding sequenc-

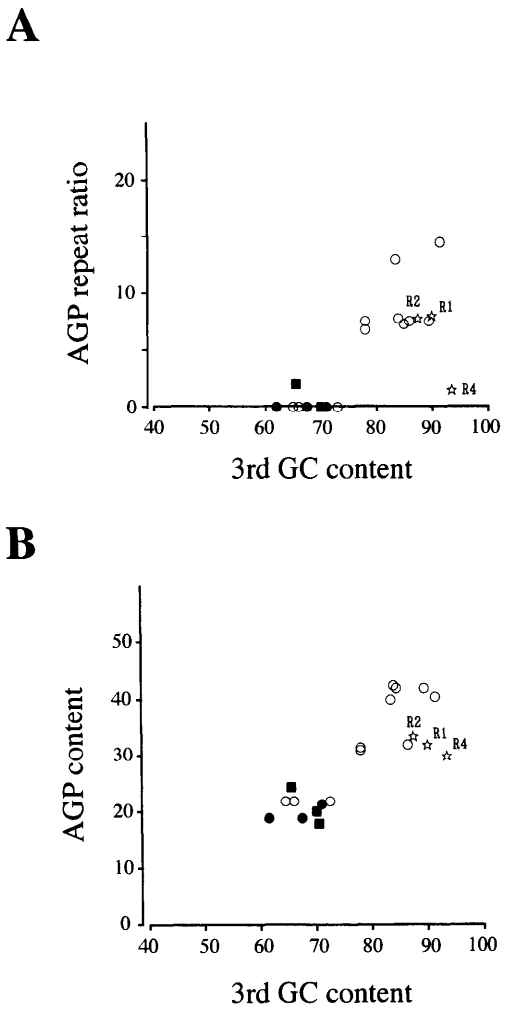

FIG. 3.—*A*, Plot of the homopolymeric alanine/glycine/proline amino acid repeat ratio (AGP repeat ratio) of the vertebrate class III POU transcription factors against the respective third GC content. *B*, Plot of the alanine/glycine/proline amino acid content (AGP content) of the vertebrate class III POU transcription factors against the respective third GC content. ○: mammals, ☆: reptiles, ●: amphibians, ■: fish. R1, R2, and R4 represent the green anole Brain-1, Brain-2, and Brain-4 genes, respectively. Homopolymeric amino acid repeats are defined as sequences consisting of more than four consecutive identical amino acid residues without interruptions.

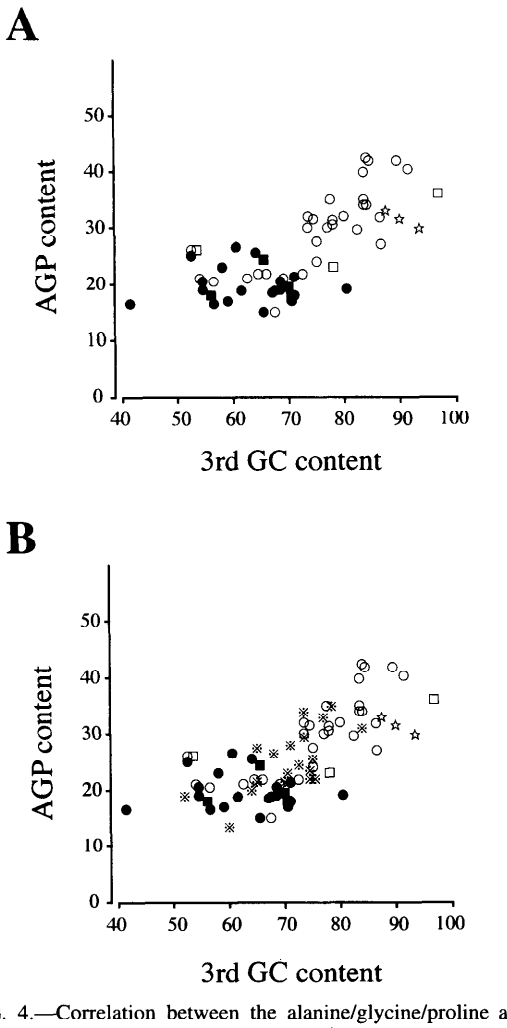

FIG. 4.—Correlation between the alanine/glycine/proline amino acid content (AGP content) of transcription factors and the respective third GC content. *A*, Vertebrates. *B*, Vertebrates plus invertebrates. ○: mammals, □: avians, ☆: reptiles, ●: amphibians, ■: fish, ※: Drosophila. Correlation coefficient for invertebrates alone was 0.67.

es. As a result, the ratio of A, G, and P residues linearly correlates with the degree of nucleotide compositional constraints increasing the GC contents, and changes in nucleotide compositional constraints have caused concomitant alterations in amino acid compositions through evolution.

A-, G-, and P-rich sequences are identified as transcriptional activation domains of transcription factors (Mermod et al. 1989; Mitchell and Tjian 1989; Licht et al. 1990; Tanaka, Clouston, and Herr 1994; Catron et al. 1995). In fact, a transcription factor artificially fused with homopolymeric proline repeats significantly modulates its transcriptional activation (Gerber et al. 1994). We therefore suggest that enrichment of A, G, and P residues in transcription factors caused by GC pressure should have a profound influence on diversification of gene regulation mechanisms (fig. 5). A possible functional difference in transcription factors caused by nucleotide compositional constraints will be the subject of forthcoming studies.

[step 1] changes in nucleotide compositional constraints

↓

[step 2] concomitant alterations in amino acid compositions (generation of homopolymeric amino acid repeats)

↓

[step 3] modulation of transcriptional activation

↓

[step 4] diversification at organismal levels

FIG. 5.—Proposed hypothesis on molecular mechanisms for diversification at the organismal level caused by genomic evolution. Note that this hypothesis is applicable not only to species diversification caused by functional differentiation among orthologous genes, but also to functional differentiation among paralogous (duplicated) genes.

Acknowledgments

We thank S. Yokoyama for the reptile library. We also gratefully acknowledge the thoughtful criticism of N. Saitou and the anonymous reviewers. This study was supported by grants from the Ministry of Education, Science, Sports, and Culture of Japan.

LITERATURE CITED

ATANASOSKI, S., S. S. TOLDO, U. MALIPIERO, E. SCHREIBER, R. FRIES, and A. FONTANA. 1995. Isolation of the human genomic Brain-2/N-Oct 3 gene (POUF3) and assignment to chromosome 6q16. Genomics **26**:272–280.

AVRAHAM, K. B., B. C. CHO, D. GILBERT et al. (15 co-authors). 1993. Murine chromosomal location of four class III POU transcription factors. Genomics **18**:131–133.

BERNARDI, G. 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. **29**:445–476.

BERNARDI, G., and G. BERNARDI. 1991. Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. Evol. **33**:57–67.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

CATRON, K. M., H. ZHANG, S. C. MARSHALL, J. A. INOSTROZA, J. M. WILSON, and C. ABATE. 1995. Transcriptional repression by Msx-1 does not require homeodomain DNA-binding sites. Mol. Cell. Biol. **15**:861–871.

DE KOK, Y. J. M., S. M. VAN DER MAAREL, M. BITNER-GLINDZICZ, I. HUBER, A. P. MONACO, S. MALCOLM, M. E. PEMBREY, H. H. ROPERS, and F. P. M. CREMERS. 1995. Association between X-linked mixed deafness and mutations in the POU domain gene *POU3F4*. Science **267**:685–688.

GERBER, H., K. SEIPEL, O. GEORGIEV, M. HÖFFERER, M. HUG, S. RUSCONI, and W. SCHAFFNER. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science **263**:808–811.

GREEN, H., and N. WANG. 1994. Codon reiteration and the evolution of proteins. Proc. Natl. Acad. Sci. USA **91**:4298–4302.

IKEMURA, T., and S. AOTA. 1988. Global variation in G+C content along vertebrate genome DNA. J. Mol. Biol. **203**: 1–13.

KADI, F., D. MOUCHIROUD, G. SABEUR, and G. BERNARDI. 1993. The compositional patterns of the avian genomes and their evolutionary implications. J. Mol. Evol. **37**:544–551.

LICHT, J. D., M. J. GROSSEL, J. FIGGE, and U. M. HANSEN. 1990. Drosophila Krüppel protein is a transcriptional repressor. Nature **346**:76–79.

MERMOD, N., E. A. O'NEILL, T. J. KELLY, and R. TJIAN. 1989. The proline-rich transcriptional activator of CTF/NF-1 is distinct from the replication and DNA binding domain. Cell **58**:741–753.

MITCHELL, P. J., and R. TJIAN. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science **245**:371–378.

SUMIYAMA, K., K. WASHIO-WATANABE, N. SAITOU, T. HAYAKAWA, and S. UEDA. 1996. Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. J. Mol. Evol. **43**:170–178.

TANAKA, M., W. M. CLOUSTON, and W. HERR. 1994. The Oct-2 glutamine-rich and proline-rich activation domains can synergize with each other or duplicates of themselves to activate transcription. Mol. Cell. Biol. **14**:6046–6055.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

XIA, Y. R., B. ANDERSON, M. MEHRABIAN, A. T. DIEP, C. H. WARDEN, T. MOHANDAS, R. J. MCEVILLY, M. G. ROSENFELD, and A. J. LUSIS. 1993. Chromosomal organization of mammalian POU domain factors. Genomics **18**:126–130.

NARUYA SAITOU, reviewing editor

Accepted June 30, 1997