Note

Compositional Biases and Polyalanine Runs in Humans

Julie Cocquet,* Elfride De Baere,† Sandrine Caburet* and Reiner A. Veitia*,[1]

**INSERM E0021 and U361, Reproduction et Physiopathologie Obstétricale, Hôpital Cochin, 75014 Paris, France and*
†*Department of Medical Genetics, Ghent University Hospital, B-9000 Ghent, Belgium*

Manuscript received May 8, 2003
Accepted for publication July 8, 2003

ABSTRACT

Human proteins containing polyalanine tracts tend to have runs of other amino acids and their open reading frames (ORFs) display a biased codon usage. Their alanine, glycine, proline, and histidine content strongly correlates with the GC content of the third codon base, suggesting that the compositional specificity of these proteins is dictated to a great extent by the evolution of their ORFs.

MANY proteins with runs of single amino acids have been described. In human the amino acids most frequently encountered in repetitive tracts are, in decreasing incidence, glutamine (Gln, Q), leucine (Leu, L), proline (Pro, P), alanine (Ala, A), and glycine (Gly, G) (Karlin *et al.* 2002). PolyGln and polyAla expansions are associated with human disease (Cummings and Zoghbi 2000). We have recently demonstrated that 30% of all mutations identified in the forkhead transcription factor gene FOXL2 lead to polyAla expansions, resulting mainly in BPES type II (De Baere *et al.* 2003). In addition, in at least eight other genes polyAla expansions exceeding a critical threshold have recently been shown to cause human disease: mutations of HOXD13 in synpolydactyly, of RUNX2 in cleidocranial dysplasia, of PABP2 in oculopharyngeal muscular dystrophy, of ZIC2 in holoprosencephaly, of HOXA13 in hand-foot-genital syndrome, of ARX in X-linked mental retardation, of SOX3 in X-linked mental retardation with growth hormone deficiency, and of PHOX2B in congenital central hypoventilation syndrome. This motivated us (i) to carry out a compositional analysis of human proteins containing at least one polyAla run, (ii) to explore the evolutionary "stability" (presence or absence) of polyAla tracts, and finally, (iii) to gain insights about the mechanisms underlying Ala run evolution.

We assembled a sample of 78 genes containing one or more polyAla repeats (TestSet) by querying GenBank with BLAST, searching for matches of the word A_n. We retained polyAla proteins containing a run of at least seven Ala residues whose expression was supported by expressed sequence tag data but irrespective of their functional annotation or their implication in pathology. We set the threshold arbitrarily to 7 to avoid blurring potentially interesting correlations with the inclusion of proteins with smaller runs, which have a higher probability of appearing by chance. The TestSet contained 132.7 kb [mean open reading frame (ORF) length, 1.68 ± 1.50 kb]. To compare this sequence set with a nonredundant reference sample of the human genome, we collected 223 ORFs (RefSet) by querying the on-line Mendelian inheritance of man (OMIM) database with the word "gene" and retrieving the coding sequences using the links with GenBank. This type of query allows retrieval of genes irrespective of their potential implication in pathology. The RefSet represented 439 kb (mean ORF length, 1.97 ± 1.41 kb). ORF lengths in both sequence sets (Test and Ref) were statistically similar ($P \sim 0.1$).

The compositional properties of both DNA sequence sets (Test- and RefSet), and of the proteins they potentially encode, were analyzed using COMPSEQ (http://bioweb.pasteur.fr/seqanal/interfaces/compseq.html). This program counts the frequencies of all words (monomers, dimers, etc.) that occur in a sequence, using a sliding window. This window moves up by the length of the "word" (1, 2, etc.) each time, skipping over the intervening words. In the case of DNA it can count only those words that occur in a specified frame.

The frequencies of the four nucleotides were not the same in the TestSet and in the RefSet. This translates into statistically different GC contents: 0.60 (length-weighted mean) for the TestSet, *vs.* 0.53 for the RefSet (Mann-Whitney test; $P < 0.0001$). RefSet GC content was in perfect agreement with the weighted mean recalculated from data in Zhang (1998) and the statistics of the codon usage database (http://www.kazusa.or.jp/codon/) for man (*i.e.*, 0.53). When we removed the GC-rich regions encoding runs of four or more Ala residues

[1]*Corresponding author:* INSERM U361, Reproduction et Physiopathologie Obstétricale, Hôpital Cochin, Pavillon Baudelocque, 123 Bd. de Port Royal, 75014 Paris, France. E-mail: veitia@cochin.inserm.fr

TABLE 1

Synthesis of some compositional properties of the polyAla proteins and their corresponding ORFs

Amino acids (AA)	AA odds ratio polyAla/Ref	Preferred codons in polyAla set	Odds ratio codons	$(AA)_2$ odds ratio polyAla/Ref	$(AA)_4$ odds ratio polyAla/Ref
Ala A	1.66	*GCG*/GCC	*4.02*/1.60	5.90	59.84
Pro P	1.44	*CCG*/CCC	*3.24*/1.51	2.21	5.37
Gly G	1.23	GGC/GGG	1.66/1.42	2.82	14.03
His H	1.22	CAC	1.47	3.66	16.68
Ser S	1.12	*TCG*/AGC	*2.79*/1.37	1.56	3.59
Gln Q	1.09	CAG	1.14	2.46	3.92
Arg R	1.06	CGC/CGG	1.45/1.28	1.86	4.96

The downward order of the amino acids reflects their decreasing tendency to overrepresentation in the TestSet with respect to the reference but not the absolute representations. For instance, Ser is represented about three times more than His in the TestSet. However, the former is also more abundant than the latter in the RefSet. $(AA)_2$ and $(AA)_4$ odds ratios were included *only* as an indication of a tendency to form runs (statistical significance tests lack them). Ratios are exaggerated by the fact that COMPSEQ counts $n - p + 1$ times a p-peptide belonging to a run of length n. AA, amino acid; AA odds ratio polyAla/Ref, fx(Test)/fx(Ref); f, frequency of x; x, any amino acid; odds ratio codons, $fxyz$(Test)/$fxyz$(Ref); $fxyz$, observed frequency of in-frame trimers. $x, y, z \in${A, C, G, T}; $(AA)_n$ odds ratio polyAla/Ref, fx_n(Test)/fx_n(Ref); x, any amino acid; odds ratio >1 (<1) → *tendency* toward over- (under)representation; in italics, least preferred codons in the RefSet.

in the TestSet (*i.e.*, some proteins had several Ala runs), the GC content of this polyAla(−) set dropped to 0.59 but was still different from that of the RefSet ($P < 0.0001$). The GC content of third codon positions (GC3), which are under less selective pressure, was different: 0.67 for the TestSet and 0.61 for the RefSet ($P < 0.0003$, GC3 in the codon usage database = 0.59). GC1 and GC2 were also different ($P < 0.0001$).

We analyzed also possible departures of the dinucleotide frequencies in the sequence sets from what is expected from the random assortment of the nucleotides. We used the well-known measure $\rho^*xy = f^*xy/f^*xf^*y$, where f^*xy denotes the frequency of dinucleotide xy and f^*x and f^*y are the *a priori* frequencies of the mononucleotides, x, y. ρ* analysis was applied to a concatemer of ORFs concatenated with their inverted complementary sequence. $\rho^* \geq 1.23$ or $\rho^* \leq 0.78$ translate into extreme over- or underrepresentation of the dinucleotide in question ($P < 0.001$; Karlin and Burge 1995). ρ* values of our data set were statistically similar to those found by Karlin and Burge (1995), and both the Test- and the RefSets reflected the known pervasive underrepresentation of dinucleotides CG and TA. However, in the case of our RefSet $\rho^*_{CG} = 0.50$, while in the polyAla TestSet $\rho^*_{CG} = 0.74$. In the polyAla(−) TestSet ρ^*_{CG} was 0.71. Thus, CG usage in the TestSet was closer to the normal range and the polyAla-encoding regions affected this quantity, but in a minor way. Other compositional biases such as Pro and Gly richness can explain part of this CG contribution. Namely, one-fourth of Pro residues in the polyAla proteins are encoded by a CG-containing codon and some GlyGly dicodons (*i.e.*, GGCGGN) can generate a CG doublet. Other GC-rich codons or their combinations can also make a contribution. In line with the evoked trend, dinucleotide CG was represented twice as much in the TestSet as in the RefSet, having the highest score of all dinucleotides. Dinucleotide AC is at the boundary of underrepresentation in the TestSet ($\rho^*_{AC} = 0.79$ *vs.* 0.83 in the RefSet) while GT was specifically underrepresented ($\rho^*_{GT} = 0.74$ *vs.* 0.80 in the RefSet).

The codon frequencies of the RefSet were in agreement with those reported in the codon usage database for man and with the length-weighted values recalculated from Zhang's (1998) data. This and the fact that Karlin's ρ^*_{xy} for the RefSet are similar to those reported in the literature (Karlin and Burge, 1995) further suggest that this set contained a representative sample of the ORFs in the genome (online Table 1 at http://www.genetics.org/supplemental/). On the other hand, we detected many departures of the codon usage between the Test and RefSets. In most cases, irrespective of the representation of the amino acid, the most biased codons were the most GC rich, as intuitively expected given the GC richness of the ORFs in the TestSet. In some instances the most deviant codons in the TestSet were the least-preferred codons in the "general" genome (Table 1 and online Table 1). Specifically, GCG and CCG encoded ~25% of Ala and Pro, respectively (and only ~11% in the RefSet) whereas TCG encoded ~13% of Ser residues (compared to ~5% in the RefSet) (online Table 1). We also studied Ala codon usage in the ORF regions specifically encoding runs of four or more Ala residues. From these 1433 codons, 35.9% corresponded to GCC and 32.2% to GCG. High GCC usage was expected (*i.e.*, 38.9% in TestSet and 40.5% in RefSet), whereas the bias favoring the usage of GCG was unexpectedly high, as noted above.

Di-, tri- and homotetrapeptides of Pro, Gly, His, Ser, and Gln also showed a tendency (in decreasing order)

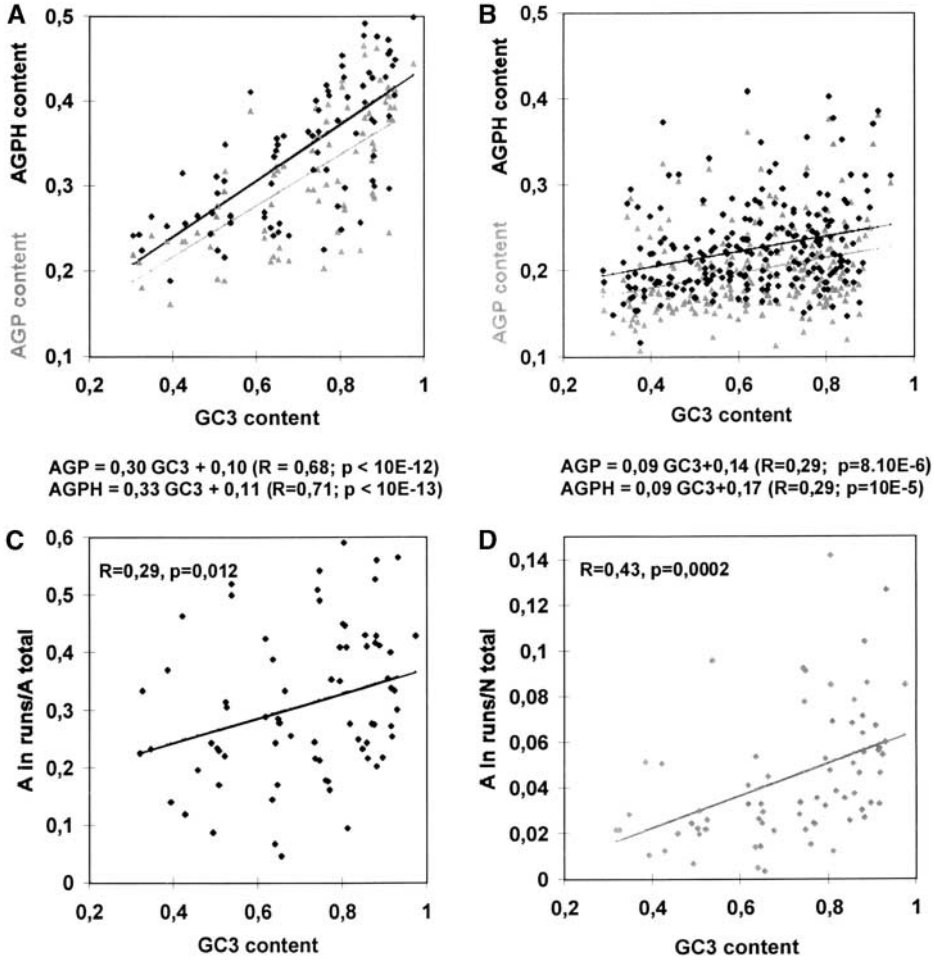

FIGURE 1.—Statistical analyses of polyAla proteins and their ORFs. (A and B) Linear regression and correlation analyses of the AGP and AGPH content of the polyAla proteins (more than seven Ala's/runs) and the GC3 content of their corresponding ORFs. The parameters of the regression lines are shown. In both cases evidence for significant correlation was obtained for the TestSet (A). A similar analysis involving 223 genes from the RefSet showed a much weaker correlation (B). This correlation is expected because GC1 + GC2 and GC3 are correlated (D'ONOFRIO *et al.* 1999). The statistical comparison of the slopes of the corresponding lines AGP(H) *vs.* GC3 in the Test- and RefSets confirmed the intuitive expectation that the slopes are greater in the former set than in the latter ($P < 0.0001$). This suggests that the composition of the proteins containing homopolymeric runs in humans better reflects the compositional constraints acting upon their ORFs than other proteins might do. (C and D) Linear regression and correlation analyses of the proportion of Ala residues in runs/total Ala (or total amino acid counts) *vs.* GC3 content (N total = number of residues in the protein). Statistical significance was obtained in both cases. After removing all AGPH codons and correcting GC3 accordingly, the correlations in A remained highly significant ($R > 0.6$; $P < 10E\text{-}9$) and the slopes of the regression in the Test and RefSet different ($P < 0.0001$). After removing Ala codons, correlations in C and D also remained significant ($R > 0.25$; $P < 0.015$). Thus, bias favoring G/C at third codon positions in AGPH does not drive the observed correlations.

toward overrepresentation in the TestSet. This suggests that these amino acids tend to be organized in runs in the polyAla-rich proteins (Table 1). This corroborates a previous suggestion that GC-rich genomes (*i.e.*, human and fly) have favored runs of GC-rich codons such as those coding for Ala, Pro, Gly, and His, whereas GC-poor genomes (worm, yeast, and weed; GC < 40%) have favored runs encoded by AT-rich codons (KARLIN *et al.* 2002).

To explore the relationship between GC content and protein primary sequence we analyzed the behavior of the content in the amino acids Ala, Gly, and Pro (AGP) *vs.* the GC content of the third base of the codons for each ORF of the TestSet. Our analysis is similar to that described by SUMIYAMA *et al.* (1996) and NAKACHI *et al.* (1997) for several transcription factors from vertebrates and invertebrates. However, our study is the first to focus on an intragenomic comparison between a subset of human proteins (containing polyAla tracts and their ORFs) and a representative sample of human ORFs. Here, we have found a highly significant positive correlation between AGP content and GC3 content with a Pearson correlation coefficient $R = 0.68$, which increased to 0.71 when we included His in the analysis (*i.e.*, AGPH content). For the RefSet, on the contrary, we found a weaker though statistically significant correlation (Figure 1, A and B). This latter result is not unexpected, taking into account the "universal" correlation between GC1 + GC2, the codon positions that essentially dictate the identity of the amino acids, and GC3 documented by D'ONOFRIO *et al.* (1999; and previous references therein). Our results suggest that the amino acid composition of the proteins containing homopolymeric runs in human reflects to a great extent the compositional constraints of the underlying ORFs. Specifically, the impact of the evolution of the ORFs on amino acid composition is stronger for polyAla proteins than for reference proteins. SUMIYAMA *et al.* (1996) and NAKACHI *et al.* (1997) attributed their significant correlation to variable intra- and intergenomic GC pressure. Following the current knowledge, it is clear that polyAla-encoding ORFs should bear the compositional signatures of the specific genomic compartments (*i.e.*, isochores) in which they lie (*i.e.*, GC3 and local GC strongly

correlate; Eyre-Walker and Hurst 2001). We extended our correlation analysis to the proportion of Ala residues in runs with respect to total Ala or to total amino acid counts *vs.* GC3 content. Significant correlations were found in both analyses (Figure 1, C and D). This further suggests that the GC richness of the ORFs plays a role in Ala runlength variation.

Departures of the observed homodicodon frequencies from random expectation were observed in the regions of the TestSet coding for the polyAla tracts (1433 codons). All homodicodons showed a tendency to overrepresentation with odds ratios >1.5. The most frequent dicodons, $(GCC)_2$ and $(GCG)_2$, had odds ratios of 1.55 and 1.75, respectively. Standard frequency comparison showed also highly significant differences with random expectations ($P < 0.0001$). This is in line with the notion that the initial mechanism to alter the length of amino acid runs is polymerase slippage during replication of repeated units. A further array homogenization process cannot be excluded on the basis of our analysis (Bouzekri *et al.* 1998).

Both the least-preferred Pro codon CCG and its homodicodon $(CCG)_2$ are more represented in the TestSet. Interestingly, $(CCG)_n$ becomes $(GCC)_n$ and would encode polyAla when read in another frame. Similarly, GGC, the most frequent Gly codon in both sequence sets, shows a tendency to form homodicodons in the TestSet. Again, a run of GGC read in another frame is interpreted as $(GCG)_n$ (a run of Ala codons, the least preferred in the rest of the genome but strongly represented in the polyAla proteins). This suggests that there might have been interconversion between $Gly_n \leftrightarrow Ala_n$ or $Pro_n \leftrightarrow Ala_n$, although not many examples could be found in GenBank. If this hypothetical "framesliding" occurred, its result has persisted within mammals and only statistical relicts remain. (See L09550 *vs.* NM_00523.)

In several transcription factors, alanine-rich regions have been shown to be responsible for repression of target genes (Han and Manley 1993). A previous work has shown that Ala, Gly, and Pro repeats present in the mammalian *Brn*-1 and -2 genes are absent in the nonmammalian homologs (Nakachi *et al.* 1997), leading the authors to propose that changes in the nucleotide compositional constraints of the genomes during evolution resulted in a concomitant generation of amino acid runs. This would have in turn modified transcriptional activity producing organismal diversification. Karlin and Burge (1995) have also proposed that polymeric runs might serve to tune the activity of the transcription factors. Here, a comparison of the 78 polyAla protein sequences against the nonredundant division of GenBank provided strong evidence for absence or shortening of polyAla tracts in one or more nonmammalian vertebrate species (same trend for Gly and Pro; online Table 2 at http://www.genetics.org/supplemental/). This casts doubts about a possible universal role of polyAla tracts as mediators of repression. Note, in online Table 2, that the primary sequences of the factors under comparison were extremely close (very low BLAST *E*-values), which argues in favor of a preservation of function. However, note that, in spite of the rather unbiased way in which the polyAla proteins (seven or more Alas/run) were collected, over 75% of those displaying functional annotation are transcription factors. Besides, the proteins without clear annotation contain a DNA- or RNA-binding domain in most cases. This is in line with the ideas of Nakachi *et al.* (1997). Namely, an alteration of functional (or physical) properties of a large body of transcription factors would explain, at least in part, the differences among vertebrate taxa. One can imagine a scenario in which Ala run shortening or formation/growth by polymerase slippage and unequal crossing-over (as is likely to be the case in mammals) are a mere consequence of the local compositional properties of the genes. This may alter the properties of many transcription factors. PolyAla tracts form α-helices with multiple equilibrated isoforms, leading to the existence of a threshold length beyond which deleterious effects appear (Cocquet *et al.* 2002). Ala runs might serve a general function, such as regulation of the intracellular/intranuclear concentrations of the active factor by establishing a chaperone-dependent equilibrium between inactive aggregated or misfolded and active forms (Sakahira *et al.* 2002).

In conclusion, our results show that the ORFs encoding polyAla proteins have higher GC and GC3 contents than reference human ORFs do. The stronger correlation between AGP (AGPH) and GC3 contents in the TestSet suggests that constraints operating on these ORFs leave a stronger imprint on amino acid composition than those operating on other ORFs in the rest of the genome. The fact that most polyAla proteins in our TestSet are transcription factors raises the question of the impact of the evolution of a set of ORFs, influenced by their genomic contexts, on major evolutionary transitions (organismal diversification).

The authors thank Kenta Sumiyama for interesting discussions about a previous version of the manuscript and Alex Fedorov, a reviewer of this article, for his helpful comments. R.A.V. is funded by the Université Denis Diderot/Paris VII and the Institut National de la Santé et de la Recherche Médicale.

LITERATURE CITED

Bouzekri, N., P. G. Taylor, M. F. Hammer and M. A. Jobling, 1998 Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. Hum. Mol. Genet. **7:** 655–659.

Cocquet, J., E. Pailhoux, F. Jaubert, N. Servel, X. Xia *et al.*, 2002 Evolution and expression of FOXL2. J. Med. Genet. **39:** 916–921.

Cummings, C. J., and H. Y. Zoghbi, 2000 Fourteen and counting: unraveling trinucleotide repeat diseases. Hum. Mol. Genet. **9:** 909–916.

De Baere, E., D. Beysen, C. Oley, B. Lorenz, J. Cocquet *et al.*, 2003 FOXL2 and BPES: mutational hotspots, phenotypic variability, and revision of the genotype-phenotype correlation. Am. J. Hum. Genet. **72:** 478–487.

D'Onofrio, G., K Jabbari, H. Musto and G. Bernardi, 1999 The correlation of protein hydropathy with the base composition of coding sequences. Gene **238:** 3–14.

Eyre-Walker, A., and L. D. Hurst, 2001 The evolution of isochores. Nat. Rev. Genet. **2:** 549–555.

Han, K., and J. L. Manley, 1993 Functional domains of the Drosophila Engrailed protein. EMBO J. **12:** 2723–2733.

Karlin, S., and C. Burge, 1995 Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11:** 283–290.

Karlin, S., L. Brocchieri, A. Bergman, J. Mrazek and A. J. Gentles, 2002 Amino acid runs in eukaryotic proteomes and disease associations. Proc. Natl. Acad. Sci. USA **99:** 333–338.

Nakachi, Y., T. Hayakawa, H. Oota, K. Sumiyama, L. Wangand *et al.*, 1997 Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol. Biol. Evol. **14:** 1042–1049.

Sakahira, H., P. Breuer, M. K. Hayer-Hartl and F. U. Hartl, 2002 Molecular chaperones as modulators of polyglutamine protein aggregation and toxicity. Proc. Natl. Acad. Sci. USA **99:** 16412–16418.

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.

Zhang, M. Q., 1998 Statistical features of human exons and their flanking regions. Hum. Mol. Genet. **7:** 919–932.

Communicating editor: M. Noor