A Genomic Basis for the Evolution of Vertebrate Transcription Factors Containing Amino Acid Runs

Sandrine Caburet, Daniel Vaiman and Reiner A. Veitia

INSERM E0021 Génomique et Développement, IFR Alfred Jost, Hôpital Cochin, 75014 Paris, France

Manuscript received January 2, 2004
Accepted for publication May 3, 2004

ABSTRACT

We have previously shown that polyAla (A) tract-containing proteins frequently present runs of glycine (G), proline (P), and histidine (H) and that, in their ORFs, GC content at all codon positions is higher than that in the rest of the genome. In this study, we present new analyses of these human proteins/ORFs. We detected striking differences in codon usage for A, G, and P in and out of runs. After dividing the ORFs, we found that 5’ halves were richer in runs than 3’ halves. Afterward, when removing the runs, we observed that the run-rich halves (grouped irrespectively of their 5’ or 3’ position) had a marked statistical tendency to have more homo- and hetero-dicodons for A, G, P, and H than the run-poor halves. This suggests that, in addition to the necessary GC-rich genomic background, a specific codon organization is probably required to generate these coding repeats. Homo-dicodons may indeed provide primers for run formation through polymerase slippage. The compositional analysis of human HOX genes, the most polyAla-rich family, and their comparison with their zebrafish homologs, support these hypotheses and suggest possible effects of genomic environment on ORF evolution and organismal diversification.

Many proteins have runs of single amino acids. In the case of humans, the most frequently encountered amino acids in homopolymeric tracts are glutamine (Gln/Q), leucine (Leu/L), proline (Pro/P), alanine (Ala/A), and glycine (Gly/G; Karlin et al. 2002). Expansions of monotonic tracts can be pathogenic, the most notable examples being polyQ expansions leading to neurodegenerative disorders (Ross 1995). Polyalanes account for 16.9% of runs in human proteins and they appear very frequently in association with runs of Pro, Gly, and His. These amino acids are encoded by GC-rich codons (Karlin et al. 2002; Cocquet et al. 2003). In at least nine genes polyAla expansions exceeding a critical threshold cause human disease (Mendelian inheritance in man database entries 110100, 142989, 600211, 602279, 603073, 142959, 300382, 313430, 603851; Brown and Brown 2004). This points to the fact that the polyAla domains undergo strong structural and/or functional constraints. Two mechanisms leading to polyAla run-length changes have been proposed: expansions/contractions of stretches of GCN codons by polymerase slippage (i.e., in PABPN1, ARX, MICA, GPX1, TGFBR1, RPL14, and FOXL1) and duplications/deletions of mixed (GCN) sequences due to unequal crossing over during meiosis (i.e., in PABPN1, HOXD13, HOXA13, RUNX2, ZIC2, FOXL2, ARX, SOX3, PHOX2B, and MSH3; Lavoie et al. 2003). However, we have recently shown that imperfect repeats can also arise as a result of polymerase slippage in regions of strong secondary structure (De Baere et al. 2002, 2003).

In a previous study, we carried out a preliminary compositional analysis of human proteins containing at least one run of seven or more Ala and of their open reading frames (ORFs). As we decided to focus on proteins involved in information transfer processes, this threshold was retained since 75% of proteins with runs of seven or more Ala are RNA/DNA-binding factors, while this fraction drops to 64% when proteins with runs of five Ala are considered (Cocquet et al. 2003; Lavoie et al. 2003). Specifically, by comparing a set of human polyAla proteins with a reference gene set representing the human genome, we demonstrated that the GC content at all codon positions (GC1, GC2, and GC3) of the ORFs was higher than that in the reference gene set (Cocquet et al. 2003). AGPH content (i.e., the frequency of A + G + P + H) correlated with GC3 in both the polyAla and reference genes. However, correlation was much stronger in the former, suggesting that the compositional specificity of the polyAla proteins is dictated mainly by the evolution of their ORFs. This evidence strongly supports the connection between high GC content and the presence of AGPH runs. Two pioneering works on a specific family of transcription factors have reported findings in line with ours (Sumiyama et al. 1996; Nakachi et al. 1997). These authors proposed 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. Interestingly, several reports focusing on amino acid runs have shown that the proteins carrying them are involved mainly in transcriptional regulation or developmental processes. In particular, over 80% of polyAla-containing proteins in humans are transcription factors or interact with RNA (Karlin et al. 2002; Cocquet et al. 2003; Lavoie et al. 2003).

In this study, we have carried out an in-depth statistical exploration of 75 human polyAla proteins having at least one run of seven or more Ala. These proteins are essentially independent: only 5 of them had close paralogs in the data set. In most other cases, polyAla domains likely arose by convergence even within the context of paralogs (see detailed discussion in Comparative genomics of amino acid runs in Hox evolution). We present insights about the mechanisms that generate the amino acid runs encoded by GC-rich codons and potential sources of GC enrichment in these ORFs. Furthermore, we explore the evolutionary implications of run accumulation, as we and others have shown that run-rich proteins are more likely to play roles in DNA/RNA binding. That is why we present also a case study of Hox genes, the most polyAla-rich class of transcription factors in mammals.

MATERIALS AND METHODS

The sample of 75 human polyAla proteins and ORFs was assembled and analyzed with COMPSEQ as described in Cocquet et al. (2003; see http://www.genetics.org/supplemental/). Concerning Hox genes, from GenBank we retrieved the sequences of well-characterized orthologs (human/zebrafish) described in the evolutionary analysis of Amores et al. (1998). To assess the correlation of GC3 vs. GC of the 50 kb surrounding an ORF, we retrieved 25 kb of gDNA sequence on each side of each relevant gene. GC contents were determined using COMPSEQ. Statistical tests were performed at http://faculty.vassar.edu/lowy/VassarStats.html or using the StatView software (Abacus, Berkeley, CA).

RESULTS AND DISCUSSION

Global compositional characteristics of polyAla proteins and ORFs: The set of 75 run-rich proteins that we analyzed contained 246 A, G, P, and H runs in total. From these, 151 were polyAla (9.1 ± 4.2 residues/run), 55 were polyGly (5.7 ± 2.3 residues/run), 30 were polyPro (5.4 ± 1.6 residues/run), and 10 were polyHis (7.3 ± 2.4 residues/run). The predominance of polyAla reflects the recruitment bias (querying GenBank for “(Ala)”), Run lengths ranged from 4 to 20 residues for Ala, to 15 for Gly, 12 for Pro, and 11 for His. The threshold of 4 for defining a run was chosen since, on average, a repeat of three amino acids will occur by chance about once in a 300-amino-acid protein.

We calculated the mean run length (MRL) as the total number of amino acids in runs divided by the number of runs for each ORF. We found a negative correlation between the number of runs and the MRL ($R = -0.48; P < 10^{-5}$), suggesting that the more runs that a protein has, the shorter they are. Neither the number of runs nor the MRL correlated with ORF length after discarding the sequences coding for runs longer than three amino acids. Thus, increasing protein length does not translate directly into an increase in the likelihood of developing runs.

Run density (RD), defined as the number of runs per length unit of the protein (X100), correlated strongly with GC3 ($R = 0.45, P < 4 	imes 10^{-5}$, Figure 1). In addition, the proportion of amino acids in runs with respect to the total length of the protein displayed a strong correlation with GC3 ($R = 0.56, P < 10^{-7}$). In line with this, Montoya-Burgos et al. (2003) have shown that GC richness clearly favors run development as they found a strong correlation between genomic GC content and minisatellite density (i.e., per megabase). It is also known that high GC is associated with shorter ORFs (Montoya-Burgos et al. 2003 and references therein). Indeed, our polyAla set showed a weak, although significant, negative correlation between total ORF length (runs

![Figure 1](http://www.genetics.org/supplemental/) — GC content at the third codon position (GC3) correlates with run density (75 ORFs). Two ORF subsets are highlighted: 10 ORF with very low RD (≈0.2, solid circles) and 10 ORFs with high RD (≥1.3, solid squares) over all the spectrum of GC3 (see text for details). Note the non-linear relationship between RD and GC3. Fitting data to $y = ax^b$ shows that the best estimate of $b$ is 1.89 ($P = 0.0015$ to reject the null hypothesis $b = 1$). $P$-value and $r^2$ are shown in the figure.
counted) and GC3 ($R = -0.26$, $P < 0.02$), which remained true when runs longer than three amino acids were excluded ($R = -0.28$, $P = 0.015$).

We noted that single runs were longer on average than multiple runs. As shown in Figure 2, it appears clearly that in our protein/ORF set high GC3 content translates into a higher number of shorter runs of various amino acid compositions. As shown in Figures 1 and 2b, GC3 correlated with an increase in the number of runs. Moreover, nonlinear curves generally fit data better than a straight line. This nonlinear behavior suggests the existence of a facilitating effect of high GC content, mediated probably by a peculiar codon organization. Thus, a high GC content background favors trinucleotide expansion, but the main driving element could be the organization of C and G into favorable “seeds.” The first clue of this is provided, at the protein level, by the existence of a correlation between the frequency of $A + G$, $A + G + P$, or $A + G + P + H$ in runs with their frequency out of the runs ($R_{AG} = 0.23$, $P < 0.04$; $R_{AGP} = 0.37$, $P < 0.001$; $R_{AGPH} = 0.35$, $P < 0.002$). This suggests that a peculiar amino acid/codon composition out of runs modulates the probability of run existence. To further explore this possibility, we looked in more detail at the codon usage and organization in the run-rich proteins/ORFs.

**Run-rich proteins/ORFs bear specific signatures:** We computed the GC content for the set of run-rich ORFs from which we removed the sequences encoding the runs ($A/G/P/H$ in runs greater than three). Wilcoxon matched-pairs signed-ranks tests showed that GC1 + 2 was smaller than in the original ORFs, as expected from the fact that runs are encoded by GC-rich codons ($P < 0.0001$; Table 1). In contrast, no difference was detected for GC3. However, these indexes were all higher than those in the reference gene set (Mann-Whitney U-test, $P < 0.0001$).

The set of run-containing genes was analyzed for usage of AGPH codons. Specifically, for each of the 75 ORFs, the number of codons for $A$ (GCN), $G$ (GGN), $P$ (CCN), or $H$ (CAC/T) was counted inside and outside runs. The first noticeable feature was the differential

---

**Figure 2.** GC3 correlates with a diversification of runs in polyAla proteins. (a) The proportion of $A$ (shaded squares) or GPH (solid triangles) in runs over total run length for each ORF are plotted vs. GC3 for the 75 polyAla proteins/orfs. Fitting data to $y = ax^b$ shows that the best estimate of $b$ is 2.48 ($P = 0.02$). (b) The ratios of $A$ in run (shaded squares) or GPH in runs (solid triangles) over total protein length are plotted vs. GC3. The first ratio behaved almost linearly (exponent $b = 0.93$) while behavior of the second ratio was obviously nonlinear ($b = 5.35$, $P = 0.005$). $P$-value and $r^2$ are shown in the figure.
usage of codons in and out of the runs and highly significant differences for all amino acids except His were detected (Table 2). Although codons ending by GC were predominant, there was a clear avoidance of GGG (Gly) and CCC (Pro) in runs as they may lead to frequent frameshifts by polymerase slippage. Moreover, a principal component analysis performed on codon usage in and out of runs indicated a clear segregation (details in supplemental data at http://www.genetics.org/supplemental/).

As shown in Figure 1, ORFs having similar GC3 may have different RD, which could be explained by a different codon organization. Thus, we defined two ORF subsets with very low (~0.2) or high RD (>1.3) for a wide range of GC3. We computed the frequencies of dicodons, for all combinations of AGPH, of only AA, PP, GG, HH and of homo-dicodons in proteins/ORF from which we removed the runs (length greater than three). The two sets were different with regard to those indexes (Mann-Whitney: \( P < 0.02 \), \( P < 0.002 \), and \( P < 0.008 \), respectively, Table 3), showing that a strong RD is associated with a high usage of homo- or hetero-dicodons, which may constitute run nucleation centers (seeds).

We next explored whether such differences in codon organization were visible within the ORFs. First of all, we tested a biased distribution of runs in N- or C-terminal halves. For this, we used the ORFs, from which runs were removed. Among the 75 ORFs, 45 had most runs in the 5' halves, 25 ORFs had the runs concentrated in the 3' halves of the 75 proteins together against the 3' halves paired \( P = 0.8 \), \( P < 10^{-18} \), with the slope of the regression line significantly smaller than
Thus, at least in these run-rich genes, 5' halves seem GC3 richer than those in the rest of the ORF. In line with this idea, 5'-UTR sequences are on average GC richer than 3'-UTRs (Pesole et al. 1997).

Finally, we evaluated the potential predominance of hetero/homo-dicodons in run-rich halves. Indeed, the run-rich halves (run removed) had a marked statistical tendency to contain more hetero/homo-dicodons (Table 3). This provides clues to understand the generation of coding microsatellites. A GC-rich background is necessary but not sufficient, as the 3' halves carry fewer runs than the 5' halves do, although they are GC richer than the reference set. The run-rich halves are naturally rich in AGPH codons and their combinations, in line with the general correlation between AG, AGP, and AGPH in and out of the runs. Many of these codons are just one mutation step away from one another (i.e., A/G, P/A, P/H). Thus, homo-dicodons, the "primers" of runs, can be easily generated by point mutation. The intrinsic nature of the run-rich ORF/proteins is thought to favor homo-dicodon production by point mutation and run expansion by slippage and illegitimate recombination.

Comparative genomics of amino acid runs in Hox evolution: To illustrate our previous conclusions about the genomic impact on polyAla generation, we focused on homeobox proteins since they constitute the largest single functional group of polyAla proteins in humans (49/494; Lavoie et al. 2003). In addition, this family is involved in major developmental processes. Human Hox are distributed in four paralogous clusters (A–D; Ruddle et al. 1994; Gehring et al. 1994; Finnerty and Martinez 1998) and carry polyAla (and other repeats) that have appeared independently in various paralogs (Lavoie et al. 2003). The appearance of a polyAla domain may underlie changes in Hox activity. Indeed, the presence of a repressive polyAla domain in the carboxy-terminal region of Ultrabithorax in insects leads to suppression of abdominal limbs (Galant and Carroll 2002; Ronshaugen et al. 2002). PolyAla-rich sequences have been found in repression domains in other transcription factors (Han and Manley 1993a,b).

To conduct our analysis, we gathered a sample of 33 Hox genes for which the human and zebrafish orthologs were available. The latter were included as they basically lack runs. This characteristic is not specific to Hox proteins (Cocquet et al. 2003; Lavoie et al. 2003) and could be driven by forces acting on the genome and not on the protein structure itself at first place. The human HOX set included 20 ORFs containing AGPH runs while 13 lacked any run. GC1 + 2 and GC3 contents in the 33 human HOXs were significantly higher than those of the reference set (RefSet, Mann-Whitney U-test, P < 0.0001) and similar to those of the run-rich ORFs described above (Table 4). The strong AGPH vs. GC3 correlation was also found in the HOXs (R = 0.60, P < 0.01).
TABLE 3

<table>
<thead>
<tr>
<th>Protein run</th>
<th>GC3</th>
<th>Protein run</th>
<th>GC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/20</td>
<td>0.7</td>
<td>1/20</td>
<td>0.5</td>
</tr>
<tr>
<td>2/20</td>
<td>1.5</td>
<td>2/20</td>
<td>1.0</td>
</tr>
<tr>
<td>3/20</td>
<td>1.5</td>
<td>3/20</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Dicoiodons for all combinations of di-amino acids (i.e., AG, PH, GP), dicoiodons for homodi-amino acids (i.e., AA, GG), and homo-dicoiodons (i.e., GCG, GCG for AA) were considered.**

The GC indexes between the 13 naturally run-less HOXs from our set and the 20 HOXs (run removed) were similar. This similarity translated also into a similar usage of hetero- and homo-dicoiodons for AGPH (data not shown), which suggests that even the naturally run-less HOXs have a potential to develop runs. So, why have some HOX developed runs and some have not? Apart from selectionistic issues, a random component might help to trigger run formation. This random element can well be (i) point mutations generating homo-dicoiodons from interconvertible hetero-dicoiodons and/or (ii) an initial event of polymerase slippage on a preexisting homo-dicoiodon. Consistently, pairs of paralogs HoxA11/D11 and A13/D13 bear Ala runs in different regions of the ORFs, while D8 presents a run lacking in its very close paralog C8. This is also valid for tracts of other amino acids such as G, P, and H. For instance, B3 has a long polyGly that is absent in A3 (and D3); D8 has a long polyPro lacking in C8 and A9 carries a polyHis absent in the other paralogs.

As shown above, GC richness is directly linked to the presence of runs. The zebrafish Hox, in line with their lack of runs, had GC1 + 2 contents similar to those of the zebrafish reference genes, while the GC3 was found to be even lower (see Table 4 and Veitia 2004). Accordingly, amino acid runs observed in the set of human proteins can be viewed as acquisitions due to the effect of genomic constraints on ORF evolution. Evolutionary convergence in run formation might be driven by a pressure increasing GC content and acting locally on mammalian HOXs.

The isochore/mosaic structure (Bernardi 2000) of the human genome leads to the existence of a strong correlation between the GC3 of the ORFs and the GC content of the region in which the genes are embedded (GCg; Eyre-Walker and Hurst 2001). Such a statistical analysis for the 33 human HOX genes failed to produce any significant correlation of GC3 vs. GCg (not even within subsets after ranking by total GC or GC3). Absence of correlation can be attributed to clustering and paralogy. In contrast, a set of 39 truly independent genes
from the reference set (having GC contents similar to those of HOX genes) displayed a significant correlation between GC3 and GCg ($R = 0.41$, $P < 0.01$). Base composition is less structured in cold-blooded species such as zebrafish (Eyre-Walker and Hurst 2001). Consistently, GC3 vs. GCg did not correlate for the 33 zebrafish Hox genes. However, after ranking them by GC3 content, the subset of 18 genes with heavier GC3 showed a significant correlation ($R = 0.59$, $P < 0.02$), suggesting that a certain degree of genomic compartmentalization does exist in the zebrafish, as already suggested (Veitia 2004 and references therein).

As recombination associated with biased gene conversion (favoring GC against AT) seems to be a primary determinant of local GC enrichment and isochore generation (Montoya-Burgos et al. 2003), we compared GC3 of 13 homologous genes located on both X and Y chromosomes (in the Y region excluded from inter- and intrachromosomal recombination; Rozen et al. 2003). A Wilcoxon matched-pairs signed-ranks test showed significant differences between X and Y GC3 contents for these genes (GC3X = 0.51 ± 0.18 vs. GC3Y = 0.45 ± 0.13, $P < 0.0008$; one-tailed test). Moreover, a linear regression of GC3Y vs. GC3X showed a covariation ($R = 0.96$, $P < 2 	imes 10^{-7}$) but the slope of the line was significantly smaller than 1 (slope = 0.73, standard error = 0.064, $P = 0.02$). Notably, the best fit was obtained with curves showing some degree of “saturation” (data not shown). These results show that recombination is associated with higher GC content that leads to a biased codon and dicodon usage. Homo-dicodons, preexisting or generated by point mutations, facilitate initial events of polymerase slippage. Then high recombination rates increase the probability of illegitimate events. This closes an autocatalytic loop that may explain the series of nonlinear behaviors shown in this study (Figures 1 and 2).

An exploration of a more direct link between recombination and generation of coding repeats is underway.

**Concluding remarks and speculations:** Compositional transitions differentiating the genomes of warm- and cold-blooded vertebrates essentially concern the present-day GC-rich genes (such as polyAla and Hox; Bernardi 2000). Accordingly, Hox genes belong to different compartments in humans and zebrafish and evolve with different “tempi,” accumulating runs in humans that potentially introduce functional changes. This idea can be extended to their regulatory elements that may have modified their pattern of expression (Aissani and Bernardi 1991a,b). Changes in such genes, encoding mainly transcription factors, would have affected important processes, thus playing an underestimated role in the evolution from amniotes to warm-blooded amniotes.

A possible functional role of polyAla is suggested by the existence of alternative protein isoforms differing only by a domain containing the repeat (for instance, HSP90 in X51346 and EST CD673431; TBX1C/NM_080647 vs. TXB1A/NM_080646; and TBX1B/NM_005992, mFOMXE1/NM_183298, and, potentially, XM_143748). Moreover, the interruption of the polyAla run by a single Val residue in FOXE1 impairs transactivation, pointing to an essential functional/structural role of this domain (Hishinuma et al. 2001). On the other hand, polyAla tracts display a threshold length beyond which deleterious effects may appear (Cocquet et al. 2003; Brown and Brown 2004). This may be due to the formation of toxic intranuclear aggregates that are sensitive to the effect of chaperones (HSP 40, 70, and 90). A recent work using the polyAla protein PABP2 fused to the green fluorescent protein has shown that fluorescence was not uniformly distributed in the nucleus, even for the normal protein (Berciano et al. 2004). PolyAla runs might serve a general function, such as regulation of the intranuclear concentrations of active factor by establishing a chaperone-dependent equilibrium between inactive/aggregated and active forms. On the other hand, it is known that when Hsp90 is inhibited, cryptic phenotypic variation is uncovered (Quetsch et al. 2002). These variants, expressed when the buffering effect of Hsp90 is impaired, can be selected and continuously expressed. This provides a mechanism for promoting evolutionary change in developmental processes. Given the strong dependence of protein aggregation upon the polyAla length, one can expect a role for the chaperones under normal conditions that might explain the persistence of polyAla length polymorphisms documented by Lavoie et al. (2003), but also their association with disease in case of deficient buffering. Namely, deletion/insertion polymorphisms (deletions being predominant) in PMX2B display an association with schizophrenia and more specifically with schizophrenia manifesting strabism.

**TABLE 4**

<table>
<thead>
<tr>
<th>Genes</th>
<th>Human HOXs</th>
<th>Zebrafish HOXs</th>
<th>Zebrafish RefSet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run+</td>
<td>Run−</td>
<td>Run less</td>
<td></td>
</tr>
<tr>
<td>GC1 +2</td>
<td>0.58 ± 0.04</td>
<td>0.56 ± 0.04</td>
<td>0.54 ± 0.04</td>
</tr>
<tr>
<td>GC3</td>
<td>0.77 ± 0.06</td>
<td>0.77 ± 0.06</td>
<td>0.76 ± 0.06</td>
</tr>
</tbody>
</table>

We removed the sequences coding for the runs longer than three AGP or H residues in the run-containing human HOX ORFs (run+) to obtain the base composition in the sequence out of the runs (run−) and compare it to the composition of the naturally run-less HOX ORFs.

PolyAla Runs and Transcription Factors

1819
mus (Toyota et al. 2004). Moreover, a deletion of 10 of 14 residues from the polyAla of FOXL2 has been associated with premature ovarian failure, while the change was inherited from the apparently normal mother (Harris et al. 2002). Another possibility worth exploring comes from the fact that polyAla proteins are also rich in Ser and Thr. Thus, differential phosphorylation of these sites could also alter the structure and function of the relevant factors.

The authors thank E. De Baere and J. Cocquet for help and discussions. R.A.V. is funded by the Université Paris VII and the INSERM.

LITERATURE CITED

Han, K., and J. L. Manley, 1993b Transcriptional repression by the Drosophila even-skipped protein: definition of a minimal repression domain. Genes Dev. 7: 491–503.

Communicating editor: K. W. Broman