Elevated Evolutionary Rate in Genes with Homopolymeric Amino Acid Repeats Constituting Nondisordered Structure

Jun Gojobori1 and Shintaroh Ueda*,2

1School of Advanced Studies, Graduate University for Advanced Studies, Hayama, Kanagawa, Japan
2Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

Abstract

Homopolymeric amino acid repeats are tandem repeats of single amino acids. About 650 genes are known to have repeats of this kind comprising seven residues or more in the human genome. According to the evolutionary conservativeness, we classified the repeats into three categories: those whose length is conserved among mammals (CM), those whose length differs among nonprimate mammals but is conserved among primates (CP), and those whose length differs among primates (VP). The frequency of each repeat, especially Ala, Leu, Pro, and Glu repeats, varies greatly in each category. The 3D structure of homopolymeric amino acid repeats is considered to be intrinsically disordered. As expected, a large proportion of the repeats had a disordered structure, and nearly half of the repeats were predicted as completely disordered. However, a number of the repeats predicted to have nondisordered structure: 13% and 25% of the repeats for categories CM and VP, respectively. Comparison of the substitution rates showed a higher $K_a/K_s$ ratio for the genes with not disordered repeats than the genes with disordered repeats. These results indicate that amino acid substitution rates have been elevated in the genes with nondisordered repeats.

Key words: homopolymeric amino acid repeats, intrinsically disordered region, mammal, primate.

Introduction

Tandem repeats of single amino acids in eukaryotic proteins are called homopolymeric amino acid repeats. This type of repeats is observed in wide range of eukaryotic species (Golding 1999; Huntley and Golding 2000; Huntley and Clark 2007). It is shown that 21% of all yeast (Saccharomyces cerevisiae) proteins contain the regions that have significant similarity to either poly-serine, poly-glutamic acid, poly-aspartic acid, poly-glutamine or poly-asparagine (Golding 1999). For cellular slime molds (Dictyostelium discoideum), more than 32% of all its proteins show similarity to a single repeat, and thale cress (Arabidopsis thaliana) has 5.4% of its proteins similar to repeats (Huntley and Golding 2000). In the study of 12 Drosophila species genomes, the percentage of proteins containing at least one repeat within their genes ranges from 16% in D. sechellia to 30% in D. grimshawi (Huntley and Clark 2007). In the human genome, approximately 650 genes are known to contain one or more homopolymeric amino acid repeat whose length is seven residues or longer, and many of the genes play roles in transcription, translation, or DNA/RNA binding (Albà and Guigó 2004; Faux et al. 2005).

It has been shown that there is divergence in the length of repeats among closely related species (Djian et al. 1996; Sumiyama et al. 1996; Nakachi et al. 1997; Albà and Guigó 2004; Mularoni et al. 2007). In addition, at least 21% of amino acid repeats show polymorphism of length in humans (Mularoni et al. 2006). DNA strand slippage during replication or unequal crossing over is considered to cause the repeat length changes (Albrecht and Mundlos 2005; Pearson et al. 2005; Huntley and Golding 2006; Mirkin 2006, 2007). Our previous study showed a positive correlation between amino acid composition and nucleotide composition; transcription factors with higher third GC contents (the proportion of nucleotide G or C of third position in codons) had more alanine, glycine, and proline residues, which are encoded by GC-rich nucleotides. It is likely that the enrichment of these amino acid residues is accomplished by the generation of homopolymeric amino acid repeats (Nakachi et al. 1997). For particular homopolymeric amino acid repeats, such as glutamine (Gln, Q) and alanine (Ala, A) repeats, their expansion is known to cause disorders, triplet repeat diseases; glutamine repeats encoded by CAG repeats can cause neurodegenerative diseases such as Huntington disease and spinocerebellar ataxia, and CGN repeats can cause several predominantly congenital malformation syndromes (Orr et al. 1993; Muragaki et al. 1996; Albrecht and Mundlos 2005; Orr and Zoghbi 2007). Proline (Pro, P) and glycine (Gly, G) repeats are shown to have the function of regulating transcription or modulating protein–protein interactions (Mitchell and Tjian 1989; Emili et al. 1994; Gerber et al. 1994; Perutz 1994; Imafuku et al. 1998; Xiao and Jeang 1998; Wilkins and Lis 1999).

The traditional view of protein structure–function relationships is that an amino acid sequence defines a 3D structure that is essential for protein function. Recent studies, however, have revealed the native and functional state to be disordered or unstructured for many proteins and protein domains (Wright and Dyson 1999; Ward et al. 2004; Fink 2005; Fukuchi et al. 2009). Homopolymeric amino acid
repeats are suggested to form an intrinsically disordered structure (Huntley and Golding 2002; Simon and Hancock 2009). “Intrinsically disordered” refers to proteins or regions that do not form stable structures like alpha-helices or beta-sheets. They are involved in protein–protein interaction leading to the formation of polymers or involved in molecular recognitions by binding to kinases, transcription factors, translation inhibitors, and DNA/RNA (Dunker et al. 2002). In eukaryote genomes, about 30% of proteins are predicted to have intrinsically disordered regions (Ward et al. 2004). This number is much higher than in archaea (2%) and bacteria (4.2%) and thought to reflect the complexity of signaling and regulatory processes within eukaryotes (Ward et al. 2004; Tompa et al. 2006; Lobley et al. 2007). Several proteins undergo disordered-to-ordered transition upon binding to their substrate and engage in molecular switching. The structural role of the homopolymeric amino acid repeats as a disordered region is suggested to be as a flexible spacer between functional domains in molecules that mediate protein–DNA/RNA or protein–protein interaction (Faux et al. 2005). Tompa (2003) found the percentage of proteins with amino acid repeats to be much higher in intrinsically disordered proteins than in other proteins.

To investigate the evolutionary meaning of homopolymeric amino acid repeats, we first performed a search in silico for mammalian orthologues of the human genes with such repeats and identified the repeats. Next, we sequenced the repeats and their flanking regions in the orthologues for nonhuman primates (apes, Old World monkeys, New World monkeys, and prosimians), which are closely related to humans and whose phylogeny and divergence times are well known, and compared repeat lengths among species based on categories classified according to the pattern of change in length. We also predicted the likelihood of the repeats being in intrinsically disordered regions and addressed the association among the evolutionary pattern of the repeats, the likelihood of being intrinsically disordered, and the amino acid substitution pattern outside the repeats.

**Materials and Methods**

**Blast Research**

Nucleotide sequences of 861 homopolymeric amino acid repeats of the 658 genes reported by Faux et al. (2005) were obtained from GenBank (Benson et al. 2009). In order to obtain orthologous genes of mammals, we performed a Blast search against the “nr” database of National Center for Biotechnology Information (NCBI) using the deduced amino acid sequences of the human genes (McGinnis and Madden 2004). The mammalian orthologues obtained were verified by a reciprocal Blast search to obtain the human genes used as queries in the first Blast search, and their nucleotide and deduced amino acid sequences were aligned by the ClustalW program (Thompson et al. 1994). Our data set is mainly based on genes from mice (Mus musculus) and rats (Rattus norvegicus), with only a small number of genes from other mammals, such as dogs (Canis familiaris) and cows (Bos taurus). The genes of dogs and cows do not cause any bias because our data show the same results were obtained after these genes were excluded.

**Genomic Polymerase Chain Reaction and Sequencing**

We examined genomic DNA of the following nonhuman primates: orangutans (Pongo pygmaeus), white-handed gibbons (Hylobates lar), agile gibbons (H. agilis), Japanese macaques (Macaca fuscata), owl monkeys (Aotus trivirgatus), and greater galagos (Otolemur crassicaudatus). DNA samples were prepared from blood specimens by conventional methods. We designed universal primers for polymerase chain reaction (PCR) to amplify the regions containing the homopolymeric amino acid repeats for all the primate orthologues. Reactions were performed in a 50 μl volume containing 1× PCR buffer for KOD-plus+, 200 μM of dNTPs, 5% dimethyl sulfoxide (DMSO), 1.5 mM of MgSO4, 16 pmoles each of the primers, 10–100 ng of genomic DNA, and 0.5 units of KOD-plus DNA polymerase (TOYOBO, Japan). The PCR conditions included an initial denaturing at 94 °C for 4 min, followed by 30 cycles at 94 °C for 30 s, 61.5 °C for 30 s, and 68 °C for 4 min and an additional extension at 68 °C for 10 min. The amplified fragments were purified using NucleoSpin Extract II (MACHEREY-NAGEL, Germany), and directly sequenced with a BigDye Terminator version 3.1 Cycle Sequencing kit and an ABI prism 3100 Genetic Analyzer (Applied Biosystems, Carlsbad, CA) We determined nucleotide sequences of two individuals for each primate species. When more than one allele were observed, the allele with the longest repeat was used for further analysis. The primers used in this study are shown in supplementary table S1, Supplementary Material online.

**Prediction of Intrinsically Disordered Regions**

Intrinsically disordered or naturally unstructured regions of protein were predicted by the software Spritz (Vullo et al. 2006). We used a “short predictor” because the least length of homopolymeric amino acid repeat is seven in our study. For each homopolymeric amino acid repeat, the ratio of the number of amino acid residues in the repeat predicted as “disordered” to the length (total number of amino acid residues) of the repeat was defined as a Disorder Index: A value of 1 means that every amino acid in the repeat is predicted as disordered, whereas 0 means that every amino acid is not predicted as disordered. We confirmed the results using the other software package, Poodle (Shimizu et al. 2007).

**Result**

**Genes with Homopolymeric Amino Acid Repeats**

We first obtained nucleotide sequences of 861 homopolymeric amino acid repeat of the 658 genes reported by Faux et al. (2005) from GenBank. We checked their authenticity and excluded ambiguous repeats from subsequent analysis.
We excluded Lys (K) repeats if the poly (A) tails of mRNA were regarded as Lys repeats due to an annotation error (19 repeats). Then, using the deduced amino acid sequences of the remaining 842 repeats of 639 human genes as queries for a Blast search against the nr database of NCBI, we collected their orthologous genes in mammals. As expected, a majority of the genes obtained were from the chimpanzee, rhesus macaque, mouse, and rat because whole-genome sequencing projects have been conducted for these species. After verification by a reciprocal Blast search, we aligned the amino acid sequences for each gene, focusing on the homopolymeric amino acid repeats. We found 28 repeats in 23 genes to show differences in length among primates. To further elucidate the degree of change, we attempted to newly determine nucleotide sequences of genes containing these repeats for 6 primate species belonging to 5 genera of 5 families of 3 infraorders of 2 semiorders; orangutans (P. pygmaeus) belonging to great apes, lar and agile gibbons (H. lar and H. agilis) belonging to lesser apes, Japanese macaques (M. fuscata) belonging to Old World monkeys, owl monkeys (A. trivirgatus) belonging to New World monkeys, and great galagos (O. crassicaudatus) belonging to prosimians. The primate classification was based on Goodman et al. (1998). We newly identified nucleotide sequences of 27 repeats in 22 genes.

Based on evolutionary conservativeness, we classified the 805 repeats of 615 genes as follows:

- **Category CM**: Homopolymeric amino acid repeats whose lengths are conserved among mammals.
- **Category CP**: Homopolymeric amino acid repeats whose lengths are different among nonprimate mammals but conserved among primate species.
- **Category VP**: Homopolymeric amino acid repeats whose lengths differ among primates.

In total, 686, 91, and 28 repeats were classified into categories CM, CP, and VP, respectively. Figure 1 illustrates the procedure described above. The genes with the repeats classified into each category are listed in supplementary tables S2–S4, Supplementary Material online.

**Amino Acids Constituting Homopolymeric Repeats**

The distribution of amino acid types in homopolymeric amino acid repeats is shown in figure 2. No repeats were found for cysteine (Cys, C), isoleucine (Ile, I), valine (Val, V), tryptophan (Trp, W), or tyrosine (Tyr, Y). Aspartic acid (Asp, D) repeats were distributed at low frequencies in any of the three categories. Phenylalanine (Phe, F), methionine (Met, M), and asparagine (Asn, N) repeats were extremely rare, whereas glutamic acid (Glu, E), glycine (Gly, G), and serine (Ser, S) repeats were observed in all three categories. Glutamine (Gln, Q) repeats were the most abundant in category VP. They were mainly encoded by CAG codons, which are well studied (Orr et al. 1993; Orr and Zoghbi 2007). Glu repeats were the second most abundant amino acid in category VP and were mainly encoded by GAG codons. Alanine (Ala, A) and proline (Pro, P) repeats had higher proportion in category CP. However, Ala

![Figure 1](https://example.com/fig1.png)

**Fig. 1** Classification of homopolymeric amino acid repeats into three categories, CM, CP, and VP (see text for details).

![Figure 2](https://example.com/fig2.png)

**Fig. 2** Frequency distributions of homopolymeric amino acid repeats for categories CM, CP, and VP. Note that no repeats were observed for Cys (C), Ile (I), Val (V), Trp (W), or Tyr (Y).
repeats were less frequent in category VP, and neither Pro nor histidine (His, H) repeats were found in the category VP. A strong functional constraint seems to have been operating on homopolymeric Ala, His, and Pro repeats lengths during primate evolution. Similar to Ala repeats, leucine (Leu, L) repeats were relatively frequent in category CM but rare in categories CP and VP. This suggests that the length of Leu repeats has had less chance to change during mammalian evolution. Lysine (Lys, K), arginine (Arg, R), and threonine (Thr, T) repeats were observed in category CM alone, though they were less frequent.

**Codon Homogeneity and Repeat Lengths**

Previous studies showed a correlation between codon homogeneity and the length of homopolymeric amino acid repeats, with repeats with greater codon homogeneity more likely to have slippage and become longer (Albà et al. 1999; Albà and Guigo’ 2004; Mularoni et al. 2006). As shown in figure 3, we found no correlation between repeat length and codon homogeneity ($r = 0.06$, $P = 0.08$), although the repeats with less codon homogeneity tended to be shorter while the repeats with greater homogeneity had longer repeat lengths. Codon homogeneity seems to be correlated with length variability; the greater the codon homogeneity, the more variable the length. Such a tendency was found for all three categories, CM, CP, and VP. The probability of DNA strand slippage during replication is likely higher for repeats with greater codon homogeneity.

**Intrinsically Disordered Structure of Homopolymeric Amino Acid Repeats**

Intrinsically disordered regions were predicted using Spritz software (Vullo et al. 2006) for 805 homopolymeric amino acid repeats in 615 human genes. The likelihood of an intrinsically disordered region (Disorder Index) was evaluated by counting those residues predicted to be disordered relative to the entire length of the repeat. The result was essentially the same as that obtained with another software package Poodle (Shimizu et al. 2007). As expected from previous studies (Huntley and Golding 2002; Simon and Hancock 2009), the majority of the repeats were predicted as a disordered structure: The Disorder Index was 1 for nearly half of the repeats (table 1 and supplementary fig. S1, Supplementary Material online). Surprisingly, however, a number of the repeats had a Disorder Index of 0. For category CM, a little more than 10% of the repeats had a Disorder Index of $\leq 0.1$, whereas more than 50% had a Disorder Index of $0.9$. For category CP, more than 60% of amino acid repeats had a Disorder Index of $0.9$ and less than 4% had a Disorder Index of $\leq 0.1$. Interestingly, about 35% of the repeats had a Disorder Index of $0.9$, whereas more than 25% had a Disorder Index of $\leq 0.1$ for category VP. The proportion of amino acid repeats which was not predicted as disordered was greater in category VP than category CM. We evaluated the statistical significance of these tendencies using permutation test. Category VP had smaller proportion of repeats with Disorder Index of $0.9$ than the randomly expected ($P < 0.05$). On the other hand, categories CM and CP did not show the deviation from the randomly expected for repeats with Disorder Index of $0.9$. Categories VP and CM had significantly larger proportion of repeats with Disorder Index of $0.9$ and less than 4%. For category CP, more than 60% of amino acid repeats had a Disorder Index of $0.9$, whereas Category CP had significantly smaller proportion of the repeats than the randomly expected ($P < 0.05$ for each comparison). In table 1, we did not take into account the repeats with Disorder Index $>0.1$ or $\leq 0.9$. This is because that for this type of repeats, the disordered/nondisordered boundary was predicted in the middle of homopolymeric amino acid repeats. To predict the exact boundary between disordered and nondisordered regions is difficult and might be inaccurate. Figure 4 shows the average Disorder Index for each amino acid residue of the homopolymeric amino acid repeats. ANOVA tests showed that the average Disorder Index was different among amino acids ($P < 0.001$). When we removed Phe, Leu, Thr, Met, Ala, Glu, and Gln (the lowest seven amino acids for the Disorder Index), the ANOVA did not show any significance. This indicates one group of amino acids (Phe, Leu, Thr, Met, Ala, Glu, and Gln) to have a lower Disorder Index and another group (Asp, Ser, Lys, Arg, Gly, His, Pro, and Asn) to have a higher Disorder Index. Notably, Phe and Leu repeats had an extraordinary low Disorder Index.

**Substitution Rates for the Genes with Homopolymeric Amino Acid Repeats**

To address the evolutionarily permissive nature of proteins with homopolymeric amino acid repeats whose Disorder Index was 0, we estimated the synonymous ($K_s$) and nonsynonymous ($K_a$) substitution rates and their ratio ($K_a/K_s$) for the genes with repeats whose Disorder Index was 0 and 1. These substitution rates were evaluated from the entire gene coding region using the method of Li (1993). When

**Table 1. Distribution of Disorder Index Values for All Repeats and for Each Category.**

<table>
<thead>
<tr>
<th>Disorder Index</th>
<th>CM (%)</th>
<th>CP (%)</th>
<th>VP (%)</th>
<th>All (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–0.1</td>
<td>13.1</td>
<td>3.3</td>
<td>25.0</td>
<td>12.4</td>
</tr>
<tr>
<td>0.9–1.0</td>
<td>52.8</td>
<td>60.4</td>
<td>35.7</td>
<td>53.0</td>
</tr>
</tbody>
</table>

*Fig. 3* Scatter plot of the length of homopolymeric amino acid repeats against the codon homogeneity of the nucleotide sequence. Homogeneity was defined as the ratio of the number of codons that appear most frequently in the repeat region to the total codon number in the repeat.

546

![Scatter plot of the length of homopolymeric amino acid repeats against the codon homogeneity of the nucleotide sequence.](https://example.com/figure3.png)
a gene had at least one repeat with a Disorder Index of 0, it was classified as “Disorder Index = 0,” and when the gene had at least one repeat with a Disorder Index of 1, it was classified as “Disorder Index = 1.” Genes with both types of repeats were excluded from the calculation. The Ka and Ks values and the Ka/Ks ratio between human and mouse were significantly higher for the genes with repeat of Disorder Index = 0 than for the genes with repeat of Disorder Index = 1 (P < 0.001, see table 2). We also obtained average Ka/Ks ratios for each of the three categories, CM, CP, and VP. Statistically significant differences between the genes with repeat of Disorder Index = 0 and 1 were observed in categories CM and VP (P < 0.001).

**Discussion**

As shown in figure 2, the distribution of amino acid residues constituting homopolymeric amino acid repeats is different among the categories classified based on evolutionary conservativeness. No repeats were found for Cys, Ile, Val, Trp, and Tyr, whereas repeats were extremely rare for Phe, Met, and Asn. In addition to the Phe, Met, and Asn repeats, repeats of Lys, Arg, and Thr were observed in category CM alone. This indicates the length of homopolymeric amino acid repeats to have been highly conserved during mammalian evolution. His and Pro repeats were observed in categories CM and CP but not in VP, showing lineage-specific conservativeness. Namely, His and Pro repeats have been conserved commonly in mammals or uniquely in the lineage of primates. Gln repeats are the most abundant in categories VP and CP. Gln repeats show variations in length in humans and their extraordinary expansion causes neurodegenerative diseases (Orr et al. 1993; Orr and Zoghbi 2007). Gln repeats with length variation in humans are mainly encoded by CAG repeats and considered to be more likely to change their length by forming a hairpin structure on the chromosome during DNA replication (Pearson et al. 2005; Mirkin 2006, 2007). Many of the Gln repeats in categories CP and VP were also encoded by CAG. Ala repeats show less variation in length among primates even though they are relatively abundant in category CP. Leu repeats were among most abundant in category CM but rare in categories CP and VP. Leu repeats have very strong toxicity when long (Dorsman et al. 2002).

The present study shows that Phe, Leu, Thr, Met, Ala, Glu, and Gln are less likely to form disordered repeats, whereas Asp, Ser, Lys, Arg, Gly, His, Pro, and Asn are more likely. Accordingly, we divided the amino acids into lower and the higher Disorder Index groups. This classification does not correlate with the physicochemical properties of the amino acids (Zhang 2000). For example, Ala and Glu are both lower Disorder Index amino acid, but Ala is hydrophobic and nonpolar, whereas Glu is hydrophilic and polar. Leu had an extremely low Disorder Index, consistent with a recent study (Simon and Hancock 2009). It is known that longer Leu repeats are very toxic (Dorsman et al. 2002). The extremely low Disorder Index of Leu repeats may be associated with this toxicity. The ranking of likelihood of being disordered region for each amino acid are roughly consistent with previous studies (Dunker et al. 2001; Weathers et al. 2004). We obtained Pearson’s correlation coefficient between the ranking of Disorder Index and the ranking of likelihood of being disordered region for the data from Dunker et al. (2001) and that from Weathers et al. (2004), respectively. The correlation coefficients were both significant for the two comparisons (P < 0.05). Therefore, our prediction of disordered regions is appropriate compared with previous studies.

When average Ka/Ks ratios were compared between the genes with repeat of Disorder Index = 0 and 1, they were found to be significantly higher in the former group. Especially, we observed a larger difference for the Ka value than for the Ks value. This means that the genes with repeat of Disorder Index = 0 could change their amino acid sequences faster than the genes with repeat of Disorder Index = 1. When estimated for each of the categories, statistical significance was observed for categories CM and VP, also showing a higher Ka/Ks ratio of the genes with repeat of Disorder Index = 0. On the other hand, the Ka/Ks ratio of the genes with repeat of Disorder Index = 1 was greater than that with repeat of Disorder Index = 0 in category CP, although the difference was marginally insignificant. In category CM, more than 10% of the repeats are Disorder Index = 0 genes, whereas more than 20% of the repeats are in category VP. The amino acids frequently repeated in

**Table 2.** Average Substitution Rates in Humans and Mice of the Genes with Repeats Whose Disorder Index Was 0 and 1.

<table>
<thead>
<tr>
<th></th>
<th>Disorder Index = 0</th>
<th>Disorder Index = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ka/Ks</td>
<td>0.207</td>
<td>0.111</td>
</tr>
<tr>
<td>Ks</td>
<td>0.530</td>
<td>0.426</td>
</tr>
<tr>
<td>Ka</td>
<td>0.121</td>
<td>0.052</td>
</tr>
<tr>
<td>Ka/Ks (CM)</td>
<td>0.205</td>
<td>0.112</td>
</tr>
<tr>
<td>Ka/Ks (CP)</td>
<td>0.089</td>
<td>0.113</td>
</tr>
<tr>
<td>Ka/Ks (VP)</td>
<td>0.333</td>
<td>0.073</td>
</tr>
</tbody>
</table>

Asterisks indicate that the difference between Disorder Index = 0 and Disorder Index = 1 is significant. (*P < 0.001).
category CM are Ala, Glu, Gly, Leu, Pro, and Ser (see fig. 2). Among them, Ala and Leu belong to the lower Disorder Index group (see fig. 4). As previously discussed, these repeats, especially Leu repeats, are highly conserved among mammals. Then, it is possible that homopolymeric amino acid repeats characteristic to mammals such as Ala, Leu, and Glu repeats help to generate the repeat of Disorder Index \(= 0\). In category VP, the frequent amino acid repeats are Glu, Gly, and Gln repeats among which Glu and Gln belong to the lower Disorder Index group. In fact, the majority of the repeats with Disorder Index \(= 0\) are Ala, Leu, and Gln repeats for category CM and Glu, and Glu repeats for category VP. Highly proportional distribution of the genes with repeat of Disorder Index \(= 0\) in categories CM and VP is well concordant with that the categories CM and VP alone have higher nonsynonymous substitution rates (see tables 1 and 2). This means the occurrence of elevated amino acid substitution rate outside the repeats with the lower Disorder Index.

It is generally considered that the globular structural domain is well conserved during evolution because one can detect this kind of region by homology search of protein sequence among distantly related organisms, even across different kingdom (Orengo and Thornton 2005). On the other hand, intrinsically disordered region of protein is known to be less conserved than structural domains (Brown 2005). The repeats with Disorder Index \(= 0\) are potentially structural domains because they are not predicted as disordered domains. Our result shows that the gene including homopolymeric amino acid repeats with Disorder Index \(= 0\) evolves faster than that with Disorder Index \(= 1\). This result is quite different from what is expected by the traditional view of the evolution of protein structure that the intrinsically disordered regions evolve faster than non-disordered regions. One possibility is that the length change in homopolymeric amino acid repeats with Disorder Index \(= 0\) affected evolutionary rate of the flanking regions in a way to change their protein sequences in compensatory manner for the repeat length change (note that “the flaking sequences” means that the amino acid sequences other than homopolymeric amino acid repeats). This type of repeat length change should influence the nucleotide substitution rates for particular region of a protein not for the whole coding region. An alternative explanation is that the repeat length change weakened functional constraint operating on the genes with Disorder Index \(= 0\) and more nucleotide substitutions were introduced in their whole coding regions. It is likely that homopolymeric amino acid repeats with Disorder Index \(= 0\) have accumulated in categories CM and VP through different mechanisms. The category CM is rich with Ala, Glu, and Leu repeats, whereas the category VP is with Glu and Gln repeats. The different composition of amino acids between categories CM and VP can lead to the different evolutionary property. Further investigation on category CM would help to discuss evolutionary implications of the elevated evolutionary rate because there is no repeat length change for the genes in this category.

We have previously proposed that the acquisition of homopolymeric amino acid repeats has caused functional diversification in gene regulatory systems (Nakachi et al. 1997). Direct evidence of this was provided by a recent study using a knock-in mouse in which the wildtype gene having homopolymeric amino acid repeats was replaced with one lacking the repeats (Anan et al. 2007). The present study newly disclosed the possibility that not only the acquisition/expansion of homopolymeric amino acid repeats but also the accompanying accelerated amino acid changes in their flanking sequences have caused the diversification of gene function during mammalian (and probably primate) evolution. Comparative studies show that Gln repeats in various genes have been acquired during primate evolution toward humans (Rubinsztein et al. 1995; Choong et al. 1998; Andrés et al. 2003; Kurosaki et al. 2006). In the case of the spinocerebellar ataxia type 1 gene, ATXN1, the common ancestor of primates has no repetitive structure, and the corresponding region is made up of several Gln and Pro residues. After the divergence of prosimians and simians, the common ancestor of simians acquired an increase in Gln residues. A further increase in the total number of Gln repeats occurred in the common lineage of great apes and humans (Kurosaki et al. 2006). It would be worth searching for amino acid substitutions unique to the human lineage, focusing on the genes that have acquired Glu and Gln homopolymeric amino acid repeats during primate evolution.

Other eukaryotic species such as S. cerevisiae, A. thaliana, or Drosophila species also have homopolymeric amino acid repeats in their genes. Therefore, investigation using these species, distantly related to mammals, can verify our conclusion that evolutionary rate has been elevated in genes with homopolymeric amino acid repeats constituting non-disordered structure. If we obtain negative conclusion, it could evoke different property among species in evolution of genes with homopolymeric amino acid repeats, like that homopolymeric amino acid repeats constituting nondisordered structure are different between categories CM and VP. And a clue for deciphering evolutionary implications in acquisition of homopolymeric amino acid repeats might be given.

**Supplementary Material**

Supplementary tables S1–S4 and figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We thank Fumihiro Niwa, Hiroshi Ishii, and Hitomi Iwasaki for helping with PCR direct sequencing. We also thank Prof. Fuyuhiko Inagaki for helpful comments about disordered regions and Prof. Li Wang for helpful discussion and encouragement during the preparation of this paper, respectively. We thank two anonymous reviewers for helpful comments. This work was supported by a Grant-in-Aid for Scientific Research.
References


