Comparative Analysis of Amino Acid Repeats in Rodents and Humans

M. Mar Albà[1,3] and Roderic Guigó[1,2]

[1]Grup de Recerca en Informàtica Biomèdica, Departament de Ciències Experimentals i de la Salut, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra and [2]Centre de Regulació Genòmica, 08003 Barcelona, Spain

Amino acid tandem repeats, also called homopolymeric tracts, are extremely abundant in eukaryotic proteins. To gain insight into the genome-wide evolution of these regions in mammals, we analyzed the repeat content in a large data set of rat–mouse–human orthologs. Our results show that human proteins contain more amino acid repeats than rodent proteins and that trinucleotide repeats are also more abundant in human coding sequences. Using the human species as an outgroup, we were able to address differences in repeat loss and repeat gain in the rat and mouse lineages. In this data set, mouse proteins contain substantially more repeats than rat proteins, which can be at least partly attributed to a higher repeat loss in the rat lineage. The data are consistent with a role for trinucleotide slippage in the generation of novel amino acid repeats. We confirm the previously observed functional bias of proteins with repeats, with overrepresentation of transcription factors and DNA-binding proteins. We show that genes encoding amino acid repeats tend to have an unusually high GC content, and that differences in coding GC content among orthologs are directly related to the presence/absence of repeats. We propose that the different GC content isochore structure in rodents and humans may result in an increased amino acid repeat prevalence in the human lineage.

The availability of three mammalian genome sequences, human, mouse, and rat, has provided the opportunity to study genome-wide evolutionary patterns with an unprecedented degree of resolution (Rat Genome Sequencing Project Consortium 2004). A clear picture of the relative importance of various mutational processes in different lineages can be obtained by the comparison of orthologous sequences. In this work, we focused on the occurrence of tandem amino acid repeats, and associated gene features, using a large collection of 7039 rat:mouse:human orthologous protein and coding sequences.

It has long been observed that tandem amino acid repeats, also known as homopolymeric tracts, are a very common feature of eukaryotic proteins (Green and Wang 1994). Annotations such as "alanine tract" or "proline-rich region" are relatively common in databases, but the function and evolution of these regions are in general still poorly understood. Tandem repeats show a high degree of repeat unit length polymorphism, lie outside well defined structural/functional domains (Huntley and Golding 2002), and tend to occur in sequences which are poorly conserved in evolution (Nishizawa et al. 1999). They are often embedded in low-complexity regions, or simple sequences, which also include interrupted, nontandem, repeats. The low degree of conservation, or high turnover, of repeats may be related to a low degree of purifying selection (Hancock et al. 2001) and to the effect of trinucleotide slippage on expanding or shortening the repeats (Levinson and Gutman 1987). At the genomic level, slippage of short DNA motifs (1–6 units) results in the formation of microsatellites, the length distribution of which can be modeled as a balance between two evolutionary forces: slippage and point mutation (Kruglyak et al. 1998). In the case of coding regions, slippage of certain repeat units, such as dinucleotide or tetranucleotides, will have a deleterious effect, as the protein reading frame will be disrupted. However, some types of trinucleotide repeats, resulting in in-frame homopolymeric amino acid runs, are not uncommon in gene coding regions (Stallings 1994; Tóth et al. 2000; Subramanian et al. 2003).

The high polymorphism and wide distribution of amino acid repeats may imply that in many cases they are functionally neutral. However, there is increasing biochemical evidence that in particular proteins some repeat types, such as glutamine, alanine, proline, and glycine runs, can modulate protein-protein interactions and/or regulate transcription (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). In addition, tandem amino acid repeats do not appear in proteins in a random fashion; on the contrary, a significant association of different types of repeats with transcription factors and developmental proteins has been observed (Karlin and Burge 1996; Albà et al. 1999a; Young et al. 2000). Of special interest are repeat expansions that lead to disease, in particular of CAG triplets resulting in abnormally long glutamine tracts (for review, see Reddy and Housman 1997). Recent work on mutant huntingtin, causing Huntington's disease, has suggested that glutamine expansion may lead to aberrant interactions with the transcription factors Sp1 and TAF II, two proteins that make contact through glutamine-rich regions (Dunah et al. 2002; Freiman and Tjian 2002).

Although important differences in amino acid repeat content have been reported in various eukaryotic genomes (Albà et al. 2001; Karlin et al. 2002), to date no large-scale comparisons of mammalian proteins have been performed. At the level of DNA motifs it has been observed that mouse genomic sequences contain more short motif repeats, including trinucleotide repeats, than human ones (Mouse Genome Sequencing Consortium 2002), which has been interpreted as a higher rate of DNA slippage in the former (Kruglyak et al. 1998). It has also been reported that trinucleotide repeat unit microsatellites are more abundant in rodents than in primates (Tóth et al. 2000). However, in the present study we show that, at both the protein and coding region levels, repeats are significantly more common in human sequences than in rodent ones. We also detected a higher degree of repeat deletion in rat versus mouse, and we show that

[3]Corresponding author.
E-MAIL malba@imim.es; FAX 34 93-224-0875.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1925704.

there is a direct relationship between amino acid repeat occurrence and an elevated GC content in the gene coding regions.

RESULTS

Amino Acid Repeat Count and Conservation Rate

We identified all tandem single-amino-acid repeats of size 5 or longer in a human–mouse–rat 7039 orthologous protein sequence data set (1:1:1 orthologs). The percentage of proteins with at least one amino acid tandem repeat was remarkably high in all three mammalian species, and clearly larger in humans (17.6%) than in mice (14.9%) or rats (13.7%). The most prominent repeat types were E (glutamic acid), P (proline), A (alanine), L (leucine), S (serine), and G (glycine), followed by Q (glutamine) and K (lysine; Fig. 1). The frequency of different amino acid repeat types was similar to that found for human proteins by Karlin et al. (2002). The number of amino acid repeats in mouse proteins represented 80% (1482 vs. 1833) of the number found in human ones, and that of the rat 72% (1323 vs. 1833). The repeat abundance difference between humans and rodents extended to practically all amino acid types. Among very long repeats, of size ≥8, there was a clear relative increase in repeats of Q and decrease in repeats of L.

To address repeat conservation among the different lineages, we collected two classes of repeats based on sequence alignments: (1) conserved repeats of size 5 or longer in all species considered, and (2) nonconserved repeats of size 5 or longer in the reference species and complete absence of repeat in the other species sequence/s. The first class of repeats is likely to predate the split of different lineages, whereas the second is likely to represent repeats originated in a given lineage or, alternatively, lost in a given lineage. Of the 1833 human amino acid repeats, 52% were conserved in the mouse and 46.5% in the rat, indicating a higher level of repeat conservation in the mouse branch than in the rat one. The amino acid repeat type with the highest conservation rate in humans and rodents was proline, and the lowest, alanine ($P < 0.02$ for P and $P < 0.005$ for A using a binomial test, comparing their relative frequency among repeats conserved in all three species to their relative frequency among human repeats).

The majority of nonconserved repeats represented expansions or deletions: that is, they completely aligned with gaps in the orthologous sequence/s. This fraction was about 90% of the human repeats not conserved in rodents and about 75% of the repeats conserved in rodents but not in humans. Nonconserved repeats showed other interesting features. Among human-specific repeats there was a significant excess of alanine repeats ($P < 0.001$). Among rodent-specific repeats, on the other hand, the most overrepresented was glutamine ($P = 0.02$). We observed that 218 of the mouse repeats not found in the rat, but only 112 of the rat repeats not found in the mouse, had an equivalent repeat of size ≥2 in the human ortholog. So, in the rat lineage a considerably larger number of ancestral repeats would have been lost.

Associated Protein Functions and Location of Repeats Within Proteins

We analyzed the types of functions most strongly associated with proteins containing amino acid repeats, focusing on the human proteins and using Gene Ontology (GO) annotations (Ashburner et al. 2000). Figure 2 shows overrepresented functions ($P < 0.05$) among the top three most abundant functions in proteins with the seven most frequent amino acid repeat types, using annotations of the 'molecular function' GO classification. 'Transcription factor' was overrepresented in most of the amino acid repeat proteins, 'DNA binding' in proteins containing E, S, and Q repeats, and 'RNA binding' and 'protein binding' in proteins containing P repeats. Proteins containing L repeats showed a distinct pattern, being significantly associated with 'transmembrane receptor.'

We also observed a significant excess of the repeat types L, A, G, and Q in the N-terminus of the mammalian proteins (N-terminal 10%, $P < 0.0001$, except $P = 0.007$ and $P = 0.01$ for Q in rats and mice, respectively). In contrast, other abundant repeats types, such as E, P, and S, did not show this deviation. Interestingly, we consistently observed an inverse correlation between repeat location bias and repeat conservation. For example, 30% of the rat repeats absent in the mouse, but only 17% of the ones conserved in the latter, showed the N-terminal location bias.

Trinucleotide Repeats

In coding regions, expansion of trinucleotides by slippage will result in the generation of tandem amino acid repeats. We scanned all gene regions encoding amino acid tandem repeats in the collection of rat, mouse, and human orthologs to identify

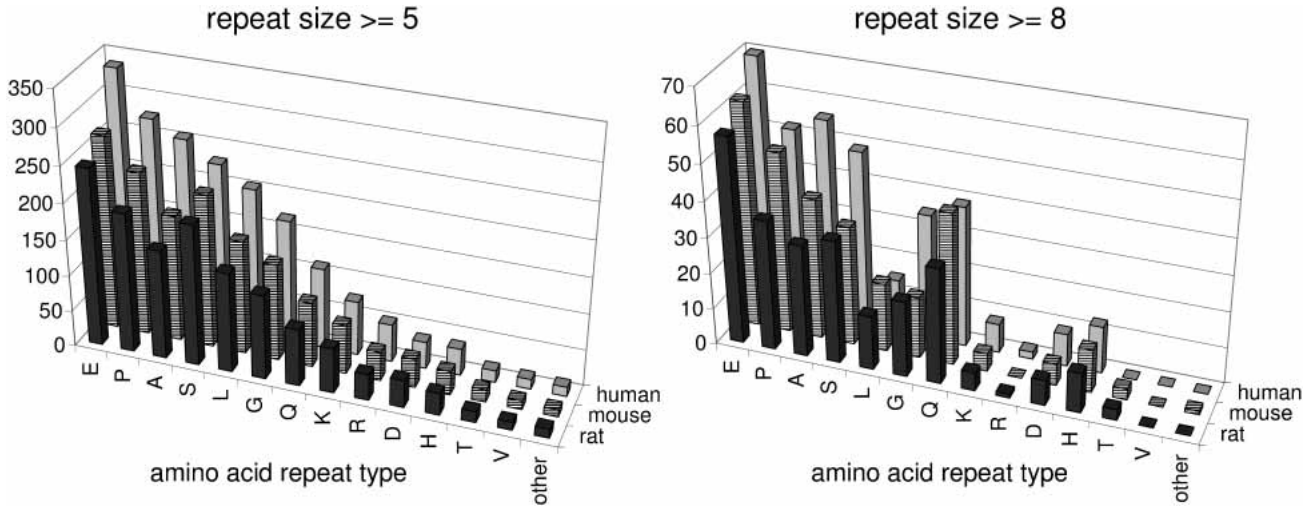

Figure 1 Tandem amino acid repeat counts in 7039 rat–mouse–human orthologous proteins. A repeat size cut-off of at least 5 repeat units, or at least 8 repeat units, was used.

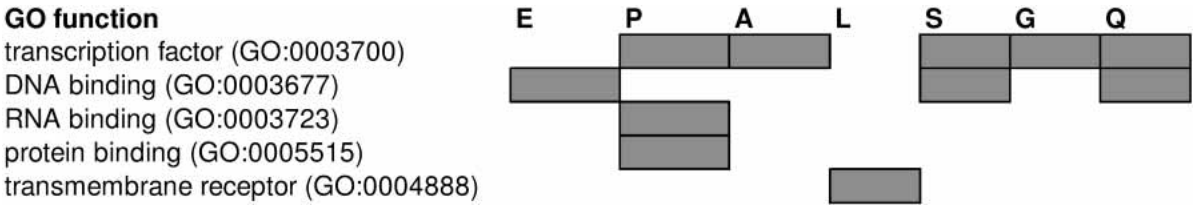

Figure 2 Gene ontology (GO) functions overrepresented in human proteins containing different amino acid repeat types ($P < 0.05$, correcting for multiple tests).

trinucleotide (codon) repeats of size 5 or longer, potentially generated by slippage. In accordance with the difference in the number of amino acid repeats, human coding sequences contained a significantly larger number of trinucleotide repeats than rodent ones (419, in contrast to 267 in the mouse and 266 in the rat). The most frequent codon repeat types were CAG, encoding Q, and GAG, encoding E (Table 1). Repeats of GCG were overrepresented in the human collection: a fourth of the repeats of this codon were in regions encoding human-specific amino acid repeats (cf. 13% for CAG and 11% for GAG). In general, the most prominent codon repeat types tended to be GC-rich, and the regions encoding amino acid repeats showed an average GC content ratio of 0.62–0.64, in marked contrast to 0.52 for all coding regions in the three species.

If trinucleotide slippage plays an important role in the generation of new repeats, one would expect a larger proportion of tandem codon repeats of size 5 or longer among nonconserved repeats than among conserved ones. This was consistently observed in the different lineages. Among the rodents, 26.8% (44 of 164) of the rat-specific repeats but only 16.3% (155 of 948) of the rat repeats conserved in the mouse were encoded by a trinucleotide run of size 5 or longer, and 18.4% (58 of 314) of the mouse-specific repeats but only 15% (142 of 948) of the conserved ones showed this feature. This difference could also be observed over the larger evolutionary distance separating humans and rodents: the percentage of amino acid repeats encoded by a trinucleotide run of size 5 or longer was 29.7% (67 of 225) for human-specific repeats, but only 18.6% (129 of 693) for repeats conserved in all three species.

Table 1. Codon Repeat Counts in the Coding Regions of 7039 Rat-Mouse-Human Orthologs

Codon	AA	Rat	Mouse	Human
CAG	Q	53 (20)	58 (21.7)	75 (18)
GAG	E	56 (21)	52 (19.5)	71 (17)
CTG	L	17 (6.5)	17 (9)	38 (9)
AGC	S	18 (7)	24 (9)	28 (7)
GCG	A	8 (3)	5 (2)	32 (7.6)
GAA	E	3 (1)	8 (3)	17 (4)
CAC	H	13 (5)	9 (3.4)	15 (3.6)
GGC	G	18 (7)	17 (6.4)	24 (5.7)
TCC	S	6 (2.3)	11 (4.2)	11 (2.6)
GCT	A	6 (6)	3 (1.1)	10 (2.4)
CCG	P	5 (2)	8 (3)	18 (4.3)
GAT	D	8 (3)	7 (2.6)	15 (3.6)
GCC	A	7 (2.6)	6 (2.3)	16 (3.8)
CCT	P	1 (0.4)	3 (1.1)	8 (0.9)
AAG	K	14 (5.3)	11 (4.1)	4 (0.9)

Only the most abundant such codon repeat counts, representing 88%–91% of the total, are shown. Relative percentage frequency in each species is shown in parentheses. The order follows codon repeat abundance in human coding regions as a reference.

Relationship Between Coding GC and Repeat Content

The observation that the regions encoding amino acid repeats showed a higher than average GC content pointed towards a more general connection between repeat occurrence and GC-richness. Therefore, we next analyzed the GC content in the coding regions of the genes that encoded amino acid repeats, discarding the regions encoding the repeats, and compared it to the general distribution in the ortholog data set (Fig. 3). In the complete data set, the three species showed a similar average GC ratio (~0.52), although the shape of the distribution varied: In humans a larger number of genes with either a lower or a higher GC content could be observed (Galtier and Mouchiroud 1998). Interestingly, the distribution of GC content in the genes that contained amino acid repeats differed substantially from the general one. There was a clear displacement towards higher GC content ratio values (~0.55). This was in agreement with the significant linear correlation between the GC content in genes containing repeats (excluding the amino acid repeat region/s) and the GC content of the repeat region (rat genes r = 0.35, N = 965, mouse genes r = 0.3, N = 1050, human genes r = 0.34, N = 1241; $P < 10^{-3}$ in all cases). Interestingly, the overall shape of the distribution was similar in the case of mouse and rat genes, with a displacement of the peak from 0.5–0.55 to 0.55–0.6, but clearly changed for human genes, with the appearance of a new peak at a high GC content value (0.6–0.65 interval). In fact, the percentage of human coding regions encoding amino acid repeats with a GC ratio >0.6 was of 32%, in contrast to 18% for the mouse or the rat.

To rule out the possibility that the relationship between amino acid repeat and GC content might be an indirect correlation, due to other properties of the proteins investigated, we examined the relationship between the degree of conservation of repeats and the coding GC content ratio differences among orthologous pairs. We identified those gene pairs that encoded at least one repeat in an equivalent position (conserved), and those in which only one of the two orthologous sequences contained amino acid repeat/s. Figure 4 shows the results of the rat–mouse, rat–human, and mouse–human comparisons. GC content ratio difference clearly correlated with the difference in amino acid repeat content. For example, in the rat–mouse comparison, the proportion of coding regions with higher GC content in the mouse ortholog (Fig. 4A, mouse > rat) increased progressively from the group in which mouse sequences did not contain repeats (A, 'rat') to the group in which both orthologs contained repeats (A, 'conserved') and to that in which rat sequences did not contain repeats (A, 'mouse'). The comparison with humans (Fig. 4B,C) showed the same phenomenon, although in this case similar results were obtained for the groups 'conserved' and 'human,' presumably reflecting the high GC content preponderance in human genes encoding repeats.

DISCUSSION

The analysis of amino acid and trinucleotide repeat content using a large collection of rat, mouse, and human gene orthologs

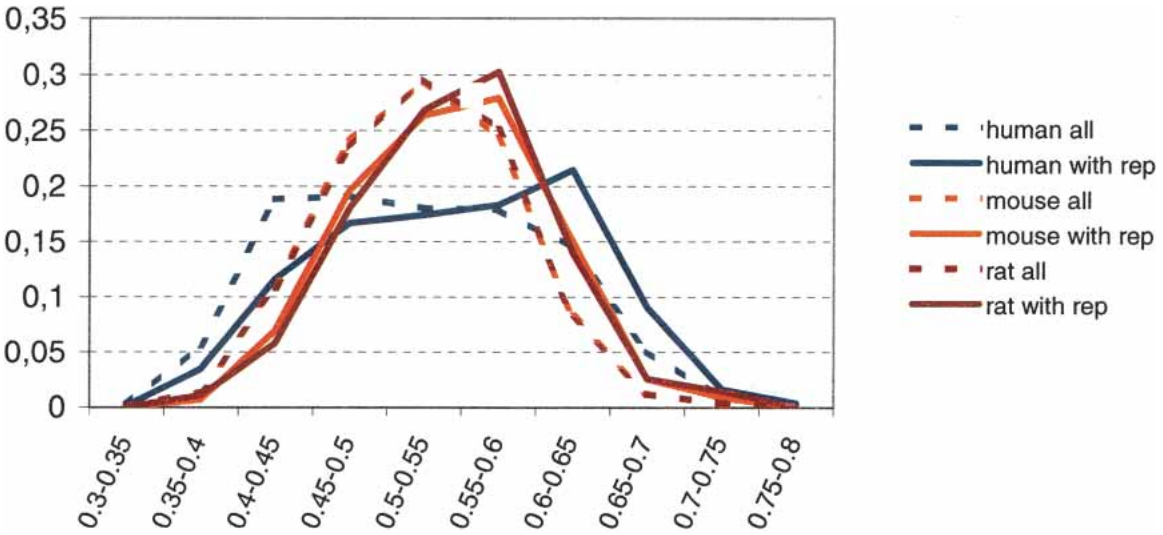

Figure 3 Percentage of coding sequences in different GC content ratio intervals. "all" refers to the totality of orthologous coding regions, and "with rep" refers to coding regions encoding amino acid repeats, after discarding the regions encoding repeats.

has provided new data to understand repeat generation dynamics in mammalian proteins. Comparison of the trinucleotide repeat content in the regions encoding conserved and nonconserved repeats is consistent with the involvement of slippage in the generation of amino acid repeats, as previously observed for glutamine repeats (Albà et al. 1999b). Here we have reported on a bias towards GC richness in the regions encoding the repeats as well as in the coding regions outside the repeats. This generalizes the observation that among vertebrate class III POU transcription factor genes, those with A, G, or P repeats have a higher GC content than those without repeats (Nakachi et al. 1997). It also agrees with the significantly high gene GC content observed in a collection of human and mouse genes containing glutamine repeats (Hancock et al. 2001). The availability of a large data set of orthologous sequences from two species, rat and mouse, with a similar GC distribution background has been particularly valuable in establishing a direct relationship between repeat occurrence and GC content. The results point to a predisposition to repeat generation in particularly GC-rich contexts and, at the same time, as repeats tend to be encoded by GC-rich codons, suggest that the formation of repeats could in turn lead to an increase in the overall gene GC content.

Human sequences show a larger number of amino acid repeats, and of coding trinucleotide repeats, than rodent sequences, which may indicate an overall stronger tendency towards repeat generation (slippage) and/or repeat preservation (lower mutation rate). This is accompanied by a more pronounced bias towards high GC content in the coding regions of genes containing repeats in this species. It thus seems plausible that the isochore differences among the two lineages, with stronger clusters of GC-rich genes in humans, have an influence on the degree of repeat occurrence in proteins. In previous work it was reported that mouse genomic sequences contained more trinucleotide repeats compared to human ones (Mouse Genome Sequencing Consortium 2002). Similarly, we have found that, on average, the frequency of trinucleotide repeats of size 5 or longer in genomic sequences is 29/Mb in rat, 32/Mb in mouse, and only 13/Mb in human sequences. Our results in coding sequences appear to contradict data from a previous study in which trinucleotide microsatellites were found to be more abundant in rodent exons than in primate exons (Tóth et al. 2000). This discrepancy may be affected by the fact that we used a data set approximately three times larger and did not consider nonhuman primate sequences. The differences found between coding and genomic sequences in the mammalian lineages may indicate that a different balance between slippage and point mutation exists in cod-

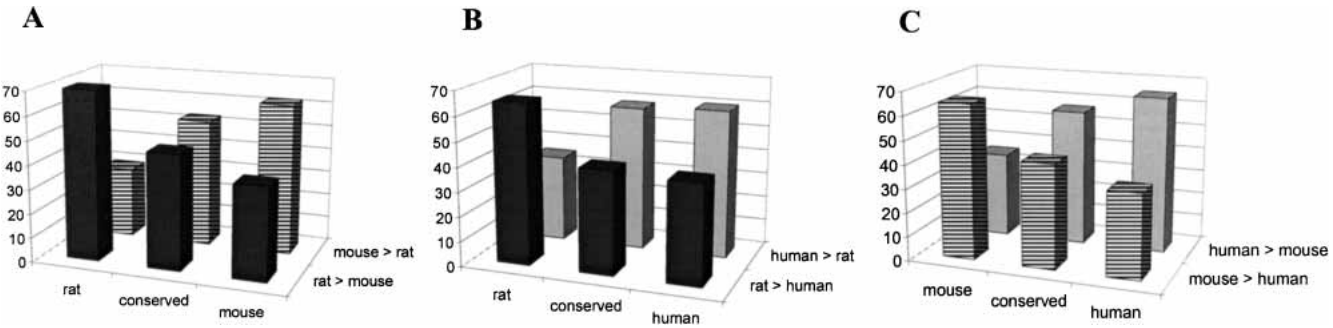

Figure 4 Comparison of coding GC content ratio content between orthologous coding sequences encoding amino acid repeats. (*A*) Rat–mouse comparison, (*B*) rat–human comparison, (*C*) mouse–human comparison. In the calculation of the GC content, all regions encoding repeats were eliminated. Two types of data sets were used: (1) only one species ortholog contained repeat/s (e.g., "rat" in *A* corresponds to rat–mouse orthologous pairs in which only the rat gene encodes amino acid repeats), and (2) both orthologs encoded at least one repeat in an equivalent position (conserved). In each data set, the fraction of pairs in which GC content ratio was superior in one of the two species (e.g., mouse > rat) was calculated. The number of sequence pairs ranged from 65 in *C* (mouse) to 721 in *A* (conserved).

ing and noncoding regions. Interestingly, we also observed that there is twice the number of trinucleotides composed exclusively of a mixture of G and C in human genomic sequences (~0.2/Mb) compared to rat or mouse genomic sequences (~0.1/Mb). This agrees with the relative overrepresentation of GCG repeats in human coding regions, resulting in alanine runs, a type or repeat that is in clear excess among human amino acid repeats not conserved in rodents.

The rat and mouse lineages diverged 12–24 Myr ago (Adkins et al. 2001). Despite this relatively short divergence time, we detected significant differences in amino acid repeat content between the two species in the data set studied. There is an overall excess of repeats of size 5 or longer in mouse proteins, which, using the human as an outgroup species, can be related to a higher rate of repeat loss in the rat lineage. Consistent with this, the ratio between small deletion and insertion events in rodent sequences has been found to be greater in the rat than in the mouse (Taylor et al. 2004; Cooper et al. 2004). However, when species-specific repeats are compared, the percentage of repeats encoded by trinucleotide runs is larger in the rat than in the mouse (26.8% vs. 16.3%), which may indicate that slippage is as active or more so in rat coding sequences than in mouse coding sequences, and that repeat contraction phenomena affect the former in a stronger manner.

Some repeat types (L, A, P, and Q) show a peculiar location within proteins, with a preference for the N-terminus, which in the case of L repeats may be explained by their role as part of peptide signal sequences (Karlin et al. 2002). In general, repeats may be better tolerated at the protein N-terminus, as interference with structured protein domains is reduced, but we do not observe a similar excess in the C-terminal part of proteins. The bias is particularly strong in species-specific repeats, indicating that the 5′ gene regions may be more dynamic in the generation of repeats, although the reasons for this are at present unclear. Nonconserved glutamine repeats between humans and mouse are associated with a low purifying selection environment (Hancock et al. 2001). Thus, many novel repeats may be nearly neutral, whereas a functional role may be more likely for repeats conserved in all three species. Repeats are abundant in proteins that regulate gene expression, and, in some cases, the size of the repeat may have an effect on the binding properties of the protein (Gerber et al. 1994; Lanz et al. 1995; Freiman and Tjian 2002). As these are rapidly evolving regions, these changes may result, in a relatively short time, in changes in gene expression regulatory networks. Repeat turnover may thus be a mechanism for rapid functional diversification. A better understanding of how this complex process may take place, and how it may be affected by genomic properties, is a challenging goal for future studies in the area.

METHODS

Sequences

Human, mouse, and rat protein and cDNA sequences were obtained from the Ensembl database (Clamp et al. 2003) using the Martview facility (http:/www.ensembl.org). We used a list of bona fide 1:1:1 rat–mouse–human orthologs (Rat Genome Sequencing Project Consortium 2004). The cDNA coding regions were extracted by exact match with the corresponding protein sequences. The final data set consisted of 7039 1:1:1 orthologous protein and coding sequences.

Repeat Count and Mapping

Tandem amino acid repeats and tandem trinucleotide repeats, all of size 5 repeat units or longer, were identified in the orthologous protein and coding sequences. The size cut-off was chosen because of its significantly low probability of occurrence by chance (Karlin et al. 2002). We detected a total of 1833 human repeats, 1482 mouse repeats, and 1323 rat repeats. To map equivalent repeats, the orthologous proteins were automatically aligned with CLUSTALW using default parameters (Thompson et al. 1994). Overlap of the same amino acid repeat type in aligned orthologous sequences was taken as the criterion to identify equivalent repeats. Mapping of repeats in sequences allowed us to discriminate between "conserved" and "nonconserved" repeats. "Conserved repeats" were defined as those of size 5 or longer in an equivalent position in the species considered, and "nonconserved repeats" were defined as those of size 5 or longer in the reference sequence/s but absent in any size in the other sequence/s (repeat size of 0).
Examples:

1.
```
seq1: PSL--LLLLQ
seq2: SLLLLLLLLQ
```

conserved repeat, of size 5 in seq1 and of size 8 in seq2.

2.
```
seq1: PS-------Q
seq2: SLLLLLLLLQ
```

nonconserved repeat.

Tests of overrepresentation of particular repeat types in different subsets were performed using the binomial distribution. Computations were performed with in-house Perl programs.

GC Content Comparison

GC content ratio in 0.05 intervals was calculated for all orthologous coding regions and for coding regions encoding repeats after eliminating the regions with repeats. We also computed the difference in coding GC content among orthologous gene pairs, and then classified the pairs in different classes according to the degree of amino acid repeat conservation: (1) repeat/s of size 5 or longer in only one of the two orthologs, and (2) at least one repeat of size 5 or longer in an equivalent position in the two orthologs (conserved).

Protein Function

Gene Ontology (GO) annotations (Ashburner et al. 2000) for Ensembl proteins were retrieved from the Ensembl server (Clamp et al. 2003). The frequency of occurrence of different GO identifiers was calculated for the complete collection of orthologous proteins and for proteins with different amino acid repeat types. For the latter we took the top three 'molecular function' functional annotations and tested the hypothesis of whether their frequency of occurrence was higher than expected according to the general function distribution.

ACKNOWLEDGMENTS

We thank Leo Goodstadt and Abel Ureta-Vidal for the orthologous sequence data set, José Castresana for critical reading of the manuscript, and Simon Jones for linguistic advice. We acknowledge support by the program Ramon y Cajal (M.A.) and the Spanish Ministry of Science and Technology, grant BIO2002-04426-C02-01, partly funded by FEDER.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adkins, R.M., Gelke, E.L., Rowe, D., and Honeycutt, R.L. 2001. Molecular phylogeny and divergence time estimates for major rodent groups: Evidence from multiple genes. *Mol. Biol. Evol.* **18:** 777–791.

Albà, M.M., Santibáñez-Koref, M.F., and Hancock, J.M. 1999a. Amino acid reiterations in yeast are over-represented in particular classes of proteins and show evidence of a slippage-like mutational process. *J. Mol. Evol.* **49:** 789–797.

Albà, M.M., Santibáñez-Koref, M.F., and Hancock, J.M. 1999b.

Conservation of polyglutamine tract size between mouse and human depends on codon interruption. *Mol. Biol. Evol.* **16:** 1641–1644.
Albà, M.M., Santibáñez-Koref, M.F., and Hancock, J.M. 2001. The comparative genomics of polyglutamine repeats: Extreme difference in the codon organization of repeat-encoding regions between mammals and *Drosophila*. *J. Mol. Evol.* **52:** 249–259.
Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.
Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., et al. 2003. Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res.* **31:** 38–42.
Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., and Sidow, A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* (this issue).
Dunah, A.W., Jeong, H., Griffin, A., Kim, Y.M., Standaert, D.G., Hersch, S.M., Mouradian, M.M., Young, A.B., Tanese, N., and Krainc, D. 2002. Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. *Science* **296:** 2238–2243.
Emili, A., Greenblatt J., and Ingles, C.J. 1994. Species-specific interaction of the glutamine-rich activation domains of Sp1 with the TATA box-binding protein. *Mol. Cell Biol.* **14:** 1582–1593.
Freiman, R.N. and Tjian, R. 2002. Neurodegeneration. A glutamine-rich trail leads to transcription factors. *Science* **296:** 2149–2150.
Galtier, N. and Mouchiroud, D. 1998. Isochore evolution in mammals: A human-like ancestral structure. *Genetics* **150:** 1577–1584.
Gerber, H.P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., and Schaffner, W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263:** 808–811.
Green, H. and Wang, N.1994. Codon reiteration and the evolution of proteins. *Proc. Nat. Acad. Sci.* **91:** 4298–4302.
Hancock, J.M., Worthey, E.A., and Santibáñez-Koref, M.F. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol. Biol. Evol.* **18:** 1014–1023.
Huntley, M.A. and Golding, G.B. 2002. Simple sequences are rare in the Protein Data Bank. *Proteins* **48:** 134–140.
Imafuku, I., Waragai, M., Takeuchi, S., Kanazawa, I., Kawabata, M., Mouradian, M.M., and Okazawa, H. 1998. Polar amino acid-rich sequences bind to polyglutamine tracts. *Biochem. Biophys Res. Commun.* **253:** 16–20.
Karlin, S. and Burge, C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Nat. Acad. Sci.* **93:** 1560–1565.
Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., and Gentles, A.J. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Nat. Acad. Sci.* **99:** 333–338.
Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Nat. Acad. Sci.* **95:** 10774–10778.
Lanz, R.B., Wieland, S., Hug, M., and Rusconi, S. 1995. A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. *Nucleic Acids Res.* **23:** 138–145.
Levinson, G. and Gutman, G.A. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4:** 203–221.
Mitchell, P.J. and Tjian, R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245:** 371–378.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Nakachi, Y., Hayakawa, T., Oota, H., Sumiyama, K., Wang, L., and Ueda, S. 1997. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. *Mol. Biol. Evol.* **14:** 1042–1049.
Nishizawa, M., Nishizawa, K., and Kim, K.S. 1999. Tendency for local repetitiveness in amino acid usage in modern proteins. *J. Mol. Biol.* **294:** 937–953.
Perutz, M. 1994. Polar zippers: Their role in human disease. *Protein Sci.* **3:** 1629–1637.
Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature* (in press).
Reddy, P.S. and Housman, D.E. 1997. The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol.* **9:** 364–372.
Stallings, R.L. 1994. Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: Implications for human genetic diseases. *Genomics* **21:** 116–121.
Subramanian, S., Madgula, V.M., George, R., Mishra, R.K., Pandit, M.W., Kumar, C.S., and Singh, L. 2003. Triplet repeats in human genome: Distribution and their association with genes and other genomic regions. *Bioinformatics* **19:** 549–552.
Taylor, M.S., Ponting, C.P., and Copley, R.R. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* (this issue).
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 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.
Tóth, G., Gáspári, Z., and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10:** 967–981.
Wilkins, R.C. and Lis, J.T. 1999. DNA distortion and multimerization: Novel functions of the glutamine-rich domain of GAGA factor. *J. Mol. Biol.* **285:** 515–525.
Xiao, H. and Jeang, K.T. 1998. Glutamine-rich domains activate transcription in yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273:** 22873–22876.
Young, E.T., Sloan, J.S. and Van Riper, K. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* **154:** 1053–1068.

Received September 1, 2003; accepted in revised form November 17, 2003.