J Mol Evol (1999) 49:789–797

JOURNAL OF MOLECULAR EVOLUTION

Amino Acid Reiterations in Yeast Are Overrepresented in Particular Classes of Proteins and Show Evidence of a Slippage-Like Mutational Process

M. Mar Albà,* Mauro F. Santibáñez-Koref, John M. Hancock

Comparative Sequence Analysis Group, Medical Research Council Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

Received: 3 March 1999 / Accepted: 26 July 1999

Abstract. Long amino acid repeats are often observed in eukaryotic proteins. In humans, several neurological disorders are caused by proteins containing abnormally long polyglutamines. However, no systematic analysis has attempted to investigate the relationship between reiterations of particular amino acids and protein function, the possible mechanisms involved in the generation of these regions, or the contribution of selection in restricting their genomic distribution, in a large collection of wild-type proteins. We have used baker's yeast open reading frames to study these questions. The most abundant amino acid repeats found in yeast proteins are repeats of glutamine, asparagine, aspartic acid, glutamic acid, and serine. Different amino acid repeats are concentrated in different classes of proteins. Acidic and polar amino acid repeats are significantly associated with transcription factors and protein kinases, while serine repeats are significantly associated with membrane transporter proteins. In most cases the codon structures encoding the repeats at the gene level show a significant bias toward long tracts of one of the possible codons, suggesting that trinucleotide slippage has played an important role in generating these reiterations. However, many, particularly those encoding serine repeats, do not show evidence of slippage. The distributions of codon repeats within proteins and between coding and noncoding regions of the genome, and of amino acids between proteins with different functions, suggest that repeats of these kinds are subject to strong selection.

Key words: Yeast — Slippage — Amino acid tandem repeats — Homopeptides — Protein function — Genome analysis — Codon composition

Introduction

Amino acid repeats are relatively common in eukaryotes (Green and Wang 1994). The most common are formed by uncharged polar amino acids (such as Gln, Asn, Ser, Pro, and Thr), acidic amino acids (Glu and Asp), or small amino acids (such as Gly and Ala) (Green and Wang 1994; Karlin and Burge 1996). The number of proteins that contain such regions cannot be explained by the frequencies of the individual amino acids. Interest in the functional context of such repeats has been stimulated by the association of several proteins that contain abnormally expanded glutamine tracts with human neurological disorders (reviewed by Reddy and Housman 1997). Homopolymeric amino acid tracts have often been observed in transcription factors (Wharton et al. 1985; Gerber et al. 1994; Hancock 1993; Karlin and Burge 1996; Nakachi et al. 1997). There is some evidence that these homopeptide stretches can mediate or modulate protein–protein interactions (Mitchell and Tjian 1989; Perutz et al. 1994; Kazemi-Esfarjani et al. 1995).

To understand the origins of these repeats it is important to understand the mutational processes that are most important in shaping them. The predominant mode of

**Present address:* Wohl Virion Centre, Windeyer Institute of Medical Sciences, University College London, London W1P 6DB, UK
Correspondence to: John M. Hancock; *e-mail:* jhancock@rpms.ac.uk

mutation of tandemly repeated DNA sequences (microsatellites) is thought to be replication slippage (Levinson and Gutman 1987), a conclusion supported by extensive molecular genetic studies in yeast (reviewed by Sia et al. 1997). However homopeptide regions could also arise by the accumulation of point mutations. The relative contributions of these two processes should be distinguishable because they should give rise to different patterns of codon organization. Slippage should give rise to long runs of single codons, while long runs should not be present if these regions originally arose by point mutation and positive selection.

Baker's yeast (*Saccharomyces cerevisiae*) currently offers a unique opportunity to study both the mutational processes underlying the generation of amino acid repeats and their occurrence in different protein classes (Richard and Dujon 1997). This is because both its complete genome sequence (Goffeau et al. 1996, Mewes et al. 1997) and an extensive functional classification of open reading frames (ORFs) (Hodges et al. 1999) are available. Hancock (1995) and Field and Wills (1996) detected numerous simple sequences in the yeast genome. More formally, many studies have shown that long microsatellite repeats are overrepresented in yeast (Valle 1993; Behe 1995; Richard and Dujon 1996; Dechering et al. 1998; Field and Wills 1998; Rose and Falush 1998; Pupko and Graur 1999; J.M.H. and M.F.S.K., unpublished).

Richard and Dujon (1996, 1997) studied triplet repeats in the yeast genome and showed differential representation of different classes of triplet repeat and frame preferences leading to high representations of repeats of certain amino acids (particularly Gln and Ser). However, their analysis did not allow detection of amino acid repeats that might derive other than by slippage or allow analysis of the proportion of, say, Glu repeats that have arisen predominantly by slippage. Nor did they consider systematically the functions of the proteins containing these repeats. To do this we have analyzed the set of ORFs derived from the yeast genome sequence (Goffeau et al. 1996, Mewes et al. 1997), for which there is currently functional information for more than half (Hodges et al. 1999).

We have set out to investigate three main questions. First, has replication slippage been the predominant mechanism generating homopeptide regions in wild-type yeast proteins, based on the codon structure of the regions encoding them, or has point mutation also played an important role? Second, given the apparent nuclear localization of many repeat-containing proteins in yeast (Richard and Dujon 1997), can previous reports of associations of particular types of amino acid repeat with particular functional groups of proteins (e.g., Wharton et al. 1985; Gerber et al. 1994; Karlin and Burge 1996), which have necessarily been based on incomplete and biased data sets, be borne out in an analysis of a complete genome? If so, which types of proteins from an organism contain long amino acid repeats? Finally, is there evidence of selective constraints that allow particular amino acids reiterations but not others?

We show that a small number of amino acids are highly overrepresented in amino acid repeats in the yeast genome and that they show associations with particular protein functional classes. Analysis of the codons encoding these regions indicates that while, on average, many amino acid repeats show evidence of the action of slippage, many do not, especially regions coding for serine repeats. We also describe strong reading frame preferences for codon repeats and differences in repeat motif frequencies between coding and noncoding regions. These findings indicate the direct influence of selection in regulating the emergence of amino acid repeats by slippage.

Methods

Database Searches

BLASTP (Altschul et al. 1990) at the NCBI was used to find all yeast protein entries with long homopeptides. Polypeptide sequences were obtained from the GenBank database using the *S. cerevisiae* subset. The expected number of repeats was estimated by calculating $N\Sigma_{i=1}^{20} p_i^n$, where N is the total number of codons in yeast coding sequences (Goffeau et al. 1996), p_i the frequency of each amino acid (*S. cerevisiae* Codon Usage Table from the ftp server at Stanford University), and n the size of the repeat. This expectation assumes that adjacent codons are independent, which is not necessarily true, but for such low expectations the error is expected to be very small. Only four or five repeats of six or more amino acids, mainly of the more abundant Leu and Val residues, are expected to occur by chance in the entire yeast genome (see Table 1). We therefore took this length as a general cutoff for further analyses. Proteins with long homopeptides were classified according to the Yeast Protein Database functional categories (Hodges et al. 1999).

Analysis of Codon Composition

As a measure of the contribution of replication slippage to the recent evolutionary history of amino acid repeat regions, we determined the length of the longest run of any single codon within the region. This should be longer than expected by chance if slippage had made a major contribution to generating the region. To illustrate this concept we can use the following example.

Consider an amino acid reiteration of length 10 glutamines and with the following codon composition: $(CAG)_7(CAA)_2(CAG)_1$. The length of the longest CAG run is 7 and that of the longest CAA run 2.

Given the length of the amino acid repeat (g) and the codon frequency in the population of repeats (p), and assuming that the codon coding for a particular residue is independent of the codons encoding other residues in the array, then the probability of the longest run of a single codon being of length l is

$$p(l|g,p) = \sum_{i=0}^{g}\sum_{j=0}^{g} N_s(i,j,l)\left(\binom{g-i-1}{j} + 2\binom{g-i-1}{j-1} + \binom{g-i-1}{j-2}\right)p^i(1-p)^{(g-i)}$$

Table 1. Amino acid reiterants in yeast ORFs

Amino acid	Observed ≥10	Observed ≥8	Observed ≥6	Expected (≥6)
Leu	2	3	6	1.78
Val	0	1	3	0.86
Lys	2	7	23	0.44
Glu	10	30	67	0.27
Ala	0	11	17	0.23
Ser	18	44	122	0.17
Ile	0	0	0	0.16
Asp	16	33	60	0.14
Thr	0	1	10	0.13
Gly	0	0	4	0.11
Asn	27	38	72	0.10
Pro	3	11	25	0.02
Arg	2	2	5	0.02
Phe	1	2	4	0.02
Gln	58	92	147	0.01
Tyr	1	1	1	0
Met	0	0	0	0
His	1	5	15	0
Trp	0	0	0	0
Cys	0	1	2	0
Total	141	282	583	4.46

where

$$N_s(i,j,l) = \sum_{k=1}^{j} (-1)^k \binom{j}{k} \left(\binom{i - lk - 1}{j - 1} - \binom{i - (l-1)k - 1}{j - 1} \right)$$

The deviation of the observed values from this expectation gives an indication of whether the codon distribution within the repeat encoding region is consistent with a random organization of the codons, or, alternatively, is indicative of codon slippage. For the statistical analysis of this deviation we used the normalized deviations from the expected values and calculated $Z_l(l|g,p) = (l - E_l(l|g,p))/\sqrt{V_l(l|g,p)}$, where l is the observed length of the longest run, $E_l(l|g,p) = \Sigma_{j=0}^{g} jp(l|g,p)$ the expected length by a random codon distribution (taking the codon frequency in the collection of repeats), and $V_l(l|g,p) = \Sigma_{j=0}^{g} j^2 p(j|g,p) - (E_l(l|g,p))^2$ the variance. For our example array, the expected repeat lengths are 2.43 for CAG and 3.22 for CAA, for $p_{CAG} = 0.45$ and $p_{CAA} = 0.55$, giving values of Z of 3.39 for CAG and 0.69 for CAA. This indicates a high likelihood that slippage has acted on CAG codons in this repeat but a low likelihood for CAA.

The average Z value for all reiterations gives the general trend in the population of repeats. These values can be calculated for each codon encoding a particular reiterated amino acid. If Z is greater than 0, the observed values are higher than expected by chance, and if lower, they are smaller. If slippage has contributed significantly to the evolution of the region, we would expect the length of the longest run of a single codon to be higher than expected, i.e., $Z_l(l|g,p) > 0$. The significance of any deviation from 0 can be assessed on the basis that the variance of the average of the normalized values is approximately $1/\sqrt{n}$ for large sample size n.

To test whether certain codons might be overrepresented with respect to the general codon frequency in the collection of repeats, we compared the codon frequencies in the repeats, and those in the remainders of the repeat-containing ORFs, excluding the repeats, with those in the yeast codon usage table (*S. cerevisiae* Codon Usage Table from the ftp server at Stanford University). Significance was assessed using χ^2 with $p < 10^{-4}$.

Codon usages for individual genes containing amino acid repeats were obtained from a table (Saccharomyces_cerevisiae.pln) obtained from the CUTG WWW server (Nakamura et al. 1997) at http://www.dna.affrc.go.jp/~nakamura/CUTG.html.

To test the distributions of different classes of triplet repeat within protein coding regions and between coding and noncoding regions, all tandem repeats of length 5 or greater were identified using the program Arrayfinder (Hancock et al. 1999) and classified using sequence annotations. We chose the lower threshold of 5 for this analysis to provide a sufficiently large sample of noncoding repeats for subsequent analysis.

Results

Amino Acid Reiterations in Yeast Proteins

We identified all long yeast homopeptides by performing a search of all repeats of six or more amino acids in the yeast protein database (GenBank subset). This allowed us to determine which amino acids have a higher tendency to form long homopeptides in this organism and the kind of proteins in which they are found.

The total number of yeast proteins with amino acid repeats of this length or greater was 444, which is about 7.6% of the estimated number of proteins in yeast (Goffeau et al. 1996). The chance expectation of repeats of this size is 4.46 assuming a random distribution of the amino acids. The observed repeats are limited to a subset of residues which are mostly polar or charged (Table 1). We observed a high abundance ($n > 50$) of Gln (147), Ser (122), Asn (72), Glu (67), and Asp (60) repeats and a more moderate abundance ($50 > n \geq 10$) of Pro (25), Lys (23), Ala (17), His (15), and Thr (10) repeats. No or few repeats were observed for Cys, Phe, Trp, Ile, Tyr, Leu, Val, Gly, and Met.

Gln and Asn tracts tended to be longer than other types of repeat, with more than one-third (39.5% for Gln and 37.5% for Asn) containing more than 10 residues, compared to 26.7% of Asp repeats, 14.9% of Glu repeats, and 14.8% of Ser repeats (Table 1).

Functional Protein Classes that Contain Amino Acid Reiterations

For subsequent analyses we concentrated on the five most frequently occurring types of amino acid repeat, Asn, Asp, Gln, Glu, and Ser, which made up more than 75% of all repeats six or more amino acids long between them. All of these occurred more than 50 times. We then used the Yeast Protein Database to classify the proteins into functional categories (Hodges et al. 1999) and tested the occurrence of the different functional categories in the groups containing the different kinds of amino acid repeat for statistically significant overrepresentation of

Table 2. Functional protein classes overrepresented among proteins containing long amino acid tandem repeats

		Repeat type[c]				
Function[a]	N(%)[b]	Gln	Asn	Glu	Asp	Ser
ATP-binding cassette	30 (0.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (4.8)
Inhibitor	12 (0.4)	0 (0.0)	**3** **(7.3)**	0 (0.0)	0 (0.0)	0 (0.0)
Protein kinase	123 (3.9)	**13** **(19.1)**	4 (9.8)	2 (7.1)	0 (0.0)	4 (9.3)
Transcription factor	262 (8.2)	**26** **(38.2)**	**11** **(26.8)**	4 (14.3)	**10** **(37.0)**	5 (11.9)
Transporter	97 (3.0)	0 (0.0)	1 (2.4)	0 (0.0)	1 (3.7)	5 (11.9)
Total	3174	68	41	28	27	42

[a] Functions are classified according to YPD (Hodges et al. 1999).
[b] Numbers and percentages (in parentheses) of the total classified yeast ORFs made up by the functional class.
[c] Numbers and percentages (in parentheses) of proteins with the specified class of amino acid repeat in the identified functional classes. Numbers in boldface indicate frequencies expected to occur with a binomial probability of less than 0.05 (observed frequency or greater) given the relative frequency of the functional class in the set of classified yeast ORFs and the number of classified proteins containing that class of repeat, correcting for multiple tests.

Table 3. Frequencies of association of tandem amino acid repeats within yeast proteins

	Ala[a]	Arg	Asn	Asp	Gln	Glu	His	Lys	Pro	Ser	Thr
Ala	1		3		4		1				
Arg		1									
Asn			**10**	4	10		1	1		1	2
Asp				**8**	4	4		2		1	1
Gln					**26**	1	1		2	1	2
Glu						**8**				3	
His							0			1	1
Lys								1			
Pro									4	1	
Ser										**9**	**3**
Thr											0

[a] Frequencies are frequencies of occurrence of the two types of amino acid repeat (length, ≥6 residues) in the same protein. Thus for a protein containing Gln, Asp, and Asn repeats, all three pairs would be counted. Figures in boldface represent frequencies significantly overrepresented at at least the $p = 0.007$ level.

functional categories. We identified six such overrepresentations at the $p < 0.05$ level after correction for multiple tests (Table 2). Transcription factors were significantly overrepresented in Gln, Asn, and Asp repeat-containing proteins, while protein kinases were also overrepresented among Gln repeat-containing proteins and inhibitors among Asn repeat-containing proteins. Although transcription factors and protein kinases were very common among Glu repeat-containing proteins (as were protein kinases in Asn repeat-containing proteins), they did not achieve statistical significance. Transporter proteins were overrepresented among Ser repeat-containing proteins but did not reach statistical significance. A complete listing of yeast proteins containing long amino acid repeats is available over the WWW at http://www.med.ic.ac.uk/dc/GGE/publicdata/ystprots.html. The functional categories of the proteins that contain the five most common repeats are also included in this list.

Fifty yeast proteins that contain long amino acid reiterations (11% of the total) contain more than one, either of the same residue or of a different one (Table 3). The frequency of occurrence of reiterations of at least one other type of amino acid was more than double the expectation for proteins containing repeats of Gln, Asn, or Asp (16–27%). For Glu repeats this frequency was lower (12%) and for Ser repeats it was very close to the expected value (9%). Chi-square analysis of the frequencies of occurrence of second amino acid repeats for the five most common classes of repeat showed a strong tendency for a second repeat of the same amino acid to occur ($p < 0.007$) in all cases. In Ser repeat-containing proteins both second Ser repeats ($p = 0.007$) and Thr repeats ($p = 4.0 \times 10^{-7}$; second Ser repeats excluded) were overrepresented.

Evidence for Slippage in the Generation of Yeast Homopeptides

The sequences encoding the repeats of the five most reiterated amino acids were analyzed to determine the extent to which slippage had contributed to their generation. We predicted that slippage would generate long runs of single codons. A simple way to assess this was to compare the length of the longest pure codon run for each amino acid repeat with the one expected in a homopeptide of the same length (see Methods). The calculation of expected lengths depends on the frequency of the different codons. We assessed three types of frequencies for use in this calculation: the frequencies in the yeast genome as a whole, the frequencies in proteins containing the amino acid repeat concerned, and the frequencies within the repeats themselves. Table 4 shows the codon frequencies in these different classes of sequence. For three amino acids, Asp, Asn, and Glu, we found no significant differences in the codon usages of repeat-coding proteins with respect to genomewide codon usage, either within or outside their repeats (after correction for multiple testing). However, Ser repeats showed a highly significant deviation from the codon usage both in Ser repeat-containing proteins and in the genome as a whole. In regions coding for Ser repeats, TCN codons were overrepresented and AGC/AGT codons underrepresented. Further, Gln repeat-containing proteins showed a highly significant increased usage of CAG codons, which was further accentuated in the Gln repeats themselves. Because of these differences, we used the codon frequencies in the repeats to predict the

Table 4. Analysis of codon frequencies in amino acid repeat proteins

Amino acid[a]	Overrepresentation[b]	Codons[c]	Codon composition[d]			Codon usage correlation[h]
			All[e]	Outside[f]	Inside[g]	
Aspartic acid	1.28	GAT > GAC	0.65	0.66 ($p = 0.064$)	0.63 ($p_g = 0.302$) ($p_r = 0.062$)	0.243 ($p = 0.066$)
Asparagine (Gln 1.21)	1.74	AAT > AAC	0.59	0.60 ($p = 0.126$)	0.63 ($p_g = 0.023$) ($p_r = 0.090$)	0.402 ($p = 0.001$)
Glutamic acid (Asp 1.24)	1.43	GAA > GAG	0.71	0.72 ($p = 0.282$)	0.74 ($p_g = 0.086$) ($p_r = 0.194$)	0.232 ($p = 0.101$)
Glutamine (Asn 1.36) (Pro 1.31)	1.68	CAA > CAG	0.69	0.66 ($p = 1.9 \times 10^{-8}$)	0.55 ($p_g = 1.4 \times 10^{-34}$) ($p_r = 3.5 \times 10^{-19}$)	0.053 ($p = 0.569$)
Serine (Thr 1.30)	1.53	TCA	0.210	0.215	0.249	0.037 ($p = 0.72$)
		TCC	0.160	0.162	0.179	−0.060 ($p = 0.56$)
		TCG	0.097	0.089	0.114	0.313 ($p = 0.002$)
		TCT	0.265	0.268	0.347	0.147 ($p = 0.153$)
		AGC	0.109	0.105	0.061	0.098 ($p = 0.342$)
		AGT	0.160	0.153 ($p = 0.332$)	0.049 ($p_g = 1.6 \times 10^{-26}$) ($p_r = 2.3 \times 10^{-24}$)	0.134 ($p = 0.193$)

[a] Amino acid repeated. Amino acids listed in parentheses are also overrepresented by 20% or more in this set of proteins. Overrepresentations are calculated for the set of proteins containing the specified repeat by comparison with the general yeast codon usage values.

[b] Degree to which the repeated amino acid is overrepresented in the set of proteins containing repeats of length ≥6. The total frequency of the amino acid outside the repeat in the total set of proteins was compared to the predicted frequency for the same total number of amino acids using the YPD codon usage table as a basis.

[c] Codons encoding the specified amoni acid. For Asp through Glu, the more common of the two codons is given on the left-hand side. For Ser all six codons are listed.

[d] For Asp through Glu, the proportion of all codons for the specified amino acid contributed by the more common codon. For Ser the proportions of all six codons are given.

[e] Codon proportions in the yeast genome as a whole, derived from the YPD codon usage table.

[f] Codon proportions in the set of proteins containing repeats but calculated excluding the repeats. p is the probability of obtaining this value by chance (chi-square) assuming the overall genomic ratio.

[g] Codon proportions within the repeats. p_g is the probability of obtaining this value by chance (chi-square) assuming the overall genomic ratio; p_r is the corresponding probability assuming the codon frequencies found outside the repeats in these proteins.

[h] Correlation coefficients of the codon proportion within and outside the repeats for the individual genes in the set. p is the probability of obtaining this value by chance (t test).

expected length distribution of codon repeats. The average values of Z calculated in this way for each class of amino acid repeat are shown in Table 5.

We observed longer than expected expansions of single codons for Gln, Asn, Glu, and Asp repeats but not for Ser. While for Asn and Asp the Z score was significantly high for both codons, for Gln and Glu repeats this was the case for only one of the two codons for each amino acid (CAG for Gln and GAA for Glu). Therefore we found evidence of slippage for six different codons, encoding four different amino acids. This was confirmed by the observation that for the same set of codons there was an excess (significantly more than 5% by χ^2 $p < 10^{-4}$) of repeats for which Z exceeded 1.96 (Table 5).

A possible artifact resulting from using repeat codon usage values would occur if the codon compositions of individual repeats were correlated with those in the rest of the protein. This would result in different expected probabilities of codon arrays of a given length in different proteins and could lead to over- or underestimation of the contribution of slippage to repeat evolution. To test for this we carried out correlation analysis of codon frequencies inside and outside repeats. A significant relationship ($p = 0.001$) was found only for Asn among the amino acids for which codons showed mean Z values greater than 0. Similar results (not shown) were obtained for the relationship between the Z value for each array and the codon usage outside it.

Selective Constraints

In the absence of any selection at the protein level, trinucleotide slippage should occur equally on the two strands and in the three possible reading frames of a gene. However, the fact that only a subset of amino acids is reiterated suggests a bias in both the frame and the orientation in which triplet runs are found (see Richard and Dujon 1996). In order to investigate this aspect formally for the entire yeast genome we identified all perfect tandem repeats of six or more triplets (excluding AAA/TTT and CCC/GGG, which are considered to be

Table 5. Analysis of the length of pure codon runs

Amino acid	Codon	Mean Z[a]	Significance of mean Z	$Z >$ 1.96[b]	$Z \leq$ 1.96[c]
Gln	CAG	0.37	***[d]	**26**[d]	121
	CAA	0.18	NS[e]	16[e]	131
Asn	AAC	0.57	***	**16**	56
	AAT	0.51	***	**16**	56
Glu	GAA	0.62	***	**21**	46
	GAG	−0.01	NS	6	61
Asp	GAC	0.42	***	**10**	50
	GAT	0.53	***	**11**	49
Ser	TCA	0.05	NS	13	109
	TCT	0.04	NS	13	109
	TCG	−0.19	NS	8	114
	TCC	−0.01	NS	8	114
	AGC	−0.07	NS	6	116
	AGT	−0.08	NS	3	119

[a] Average of the standardized values for the length of the longest pure codon repeat in the amino acid reiterants (negative values are observations lower than expected by chance, and positive values higher than expected by chance).
[b] Number of individual arrays reaching a Z score of 1.96 or greater (expected proportion: 0.05).
[c] Number of individual arrays reaching a Z score of $\leq$1.96 (expected proportion: 0.95).
[d] Significant ($p < 10^{-4}$).
[e] Not significant ($p > 10^{-4}$).

mononucleotide repeats) using the program Arrayfinder (Hancock et al, 1999). We compared the frequency of the circular permutations (i.e., all three frames in both orientations) of all coding repeats (Table 6). As expected, the most commonly repeated triplets are generally also the most slippage-prone as indicated by the Z scores (Tables 5 and 6). Two exceptions were tandem CAA repeats, which were more abundant than tandem CAG repeats, and tandem GAT repeats, which outnumbered GAC repeats.

Certain frames were strongly favored. For example, of 48 CAG/GCT repeats, 34 corresponded to CAG (Gln), 3 to AGC (Ser), 2 to GCA (Ala), 8 to GCT (Ala), 1 to CTG (Leu), and none to TGC (Cys). We have observed a similar frame bias for this codon in mammalian exons (E.A. Worthey, M.F.S.K., and J.M.H., unpublished data). Similarly, of 74 tracts involving AAC/GGT, 51 consisted of CAA (Gln) and 23 of AAC (Asn). In the opposite orientation there were none corresponding to Val, Cys, or Leu, in spite of the fact that Leu and Val are the most abundant amino acids in yeast.

As a second indicator of whether selection has acted on triplet repeats in coding regions, we investigated the frequencies of tandem triplet repeats in coding and noncoding regions (Table 6). There was no significant correlation between frequencies of triplet repeats in the two types of regions ($r = 0.320$, $p = 0.367$). Hancock (1995) showed a significant positive correlation between the AT content and the frequency of simple sequence motifs in noncoding, but not coding, regions sampled from a number of genomes, including yeast. A similar correlation could be seen here for noncoding regions ($r = 0.768$, $p = 0.009$) but not for coding regions. Table 6 also shows markedly more repeats in the sense strand that are purine-rich (predominance of A/G over C/T; $N = 263$) than pyrimidine-rich (C/T > A/G; $N = 60$).

Discussion

Despite the compact nature of the yeast genome and its low overall level of sequence repetition (Hancock 1995), these analyses show that amino acid repeats are common in yeast proteins, as suggested by earlier studies of triplet repeats in the yeast genome (Richard and Dujon 1996, 1997). These authors analyzed eight yeast chromosomes for coding tandem repeats and identified 12 Gln, 12 Ser, 9 Asp, 8 Glu, 6 Asn, and 6 Ala encoding arrays. Our more extensive survey identified the same set of highly represented amino acids (Table 1) except that we found a lower representation of Ala repeats. Green and Wang (1994) carried out a broader survey of amino acid repeats in the databases which was biased primarily toward mammalian proteins. They also found Gln repeats to be most common among long repeats. However, we find an underrepresentation of certain amino acids, such as glycine and alanine, and an overrepresentation of others, particularly aspartic acid and asparagine, in yeast compared to Green and Wang's (1994) survey.

We initially asked three questions about the yeast repeats: Are they generated predominantly by slippage? Do they occur in particular types of proteins? and Is there evidence that selection acts upon them?

Role of Slippage

We investigated the role of slippage in generating tandem amino acid repeats by analyzing the codon composition of regions encoding the repeats. On average, six of the eight codons coding for the highly overrepresented polar or acidic amino acids (Asn, Gln, Asp, and Glu) had mean Z values significantly greater than zero, consistent with a significant role for replication slippage in the evolution of these regions. Significantly more than the expected 5% of Z scores greater than 1.96 were also seen for these amino acids. Experimental analysis indicates that replication slippage is the predominant mechanism involved in microsatellite instability in yeast (Henderson and Petes 1992; Petes et al. 1997). These patterns are therefore consistent with slippage playing an important role in generating these amino acid repeats.

Despite the significantly high mean Z scores found for many codons in this analysis, regions encoding Ser repeats, and many regions encoding the other classes of repeat (having $Z \leq 1.96$; Table 5), did not deviate strongly from random expectation in this analysis. Our

Table 6. Frequencies and locations of triplet repeats of length ≥5 in the yeast genome

Motif[a]	NC[b]	Reading frame[c]					
		1	2	3	4	5	6
AAC/ACA/CAA/TTG/TGT/GTT	12	23	0	51	0	0	0
		(N)	(T)	(Q)	(L)	(C)	(V)
AAG/AGA/GAA/TTC/TCT/CTT	5	9	3	54	1	11	2
		(K)	(R)	(E)	(F)	(S)	(L)
AAT/ATA/TAA/TTA/TAT/ATT	36	38	0	0	1	1	0
		(N)	(I)	(∗)	(L)	(Y)	(I)
ACC/CCA/CAC/GTG/TGG/GGT	0	0	4	1	0	0	0
		(T)	(P)	(H)	(V)	(W)	(G)
ACG/CGA/GAC/GTC/TCG/CGT	1	0	0	8	0	2	2
		(T)	(R)	(D)	(V)	(S)	(R)
ACT/CTA/TAC/GTA/TAG/AGT	5	4	2	0	0	0	0
		(T)	(L)	(Y)	(V)	(∗)	(S)
AGC/GCA/CAG/CTG/TGC/GCT	2	3	2	34	1	0	8
		(S)	(A)	(Q)	(L)	(C)	(A)
AGG/GGA/GAG/CTC/TCC/CCT	0	0	0	3	0	5	4
		(R)	(G)	(E)	(L)	(S)	(P)
ATC/TCA/CAT/ATG/TGA/GAT	3	0	9	2	0	0	35
		(I)	(S)	(H)	(M)	(∗)	(D)
CCG/CGC/GCC/GGC/GCG/CGG	0	0	0	0	0	0	0
		(P)	(R)	(A)	(G)	(A)	(R)

[a] All six permutations of each of the 10 triplet motifs are presented. All occurrences in both orientations are pooled in the column of noncoding repeats. For coding repeats, frequencies for the six possible reading frames are presented in the same order as listed in this column. Repeats of A/T and C/G are not included, as these are considered to be mononucleotide repeats.

[b] Frequency of noncoding repeats.

[c] Frequency of repeats in the six possible reading frames. Single-letter codes for the amino acids encoded by each repeat are given below the numbers. ∗ represents a stop codon.

calculations depend on the codon frequencies encoding the repeats. If codon frequencies within the repeats were biased by slippage (i.e., some codons underwent slippage more readily than others), we might underestimate *Z* for the most slippage-prone codons. Codon usage within the repeat region deviated significantly from the genomic codon usage only for Gln and Ser repeats. However, recalculating *Z* for Gln and Ser codons using genome-wide codon usage values did not result in different conclusions (data not shown). Similarly, the correlation between codon usage outside and inside Asn repeats had no significant effect on our conclusions. We therefore conclude that Ser repeats, and indeed many other low-*Z* repeats found for the other amino acids, have not undergone a significant amount of slippage during their recent evolutionary history. Point mutation and subsequent selection may therefore have been an important driving force in the generation of these repeats in yeast proteins.

Proteins containing amino acid repeats were enriched in that same amino acid outside the repeat (Table 4) and tended, weakly, to have a different codon usage from the general codon usage in the yeast genome. We also observed a general overrepresentation of proteins containing two or more amino acid repeats and, among these, proteins containing more than one repeat of the same amino acid. Two models can explain this (see Hancock 1993; Karlin and Burge 1996; Richard and Dujon 1997). The first is that repeat-containing proteins have been enriched in these amino acids during their previous evolutionary history by slippage followed by point mutation, which has obscured the original patterns generated by slippage. This is a similar argument to one used to explain the presence of cryptically simple sequence regions in many genomes (Tautz et al. 1986; Ohno and Epplen 1983). The alternative is that amino acid repeats arise by slippage in proteins already enriched in the amino acid that becomes tandemly repeated, and which have a somewhat biased codon usage. Such an overrepresentation would increase the probability that codons for that amino acid would become tandemly reiterated by chance and act as seeds for subsequent slippage.

Role of Selection and Functional Associations

The action of selection on tandem triplet repeats is best seen by considering their distribution between the different strands and reading frames within ORFs, and between coding and noncoding regions of the genome. Purely neutral structures would be expected to be randomly distributed in both senses within ORFs. Richard and Dujon (1996), in an analysis of eight yeast chromosomes, showed that tandem triplet repeats showed strong reading frame and strand preferences and our whole-genome analysis confirms this (Table 6).

We carried out correlation analysis to determine

whether frequencies of different classes of repeat were similar in coding and noncoding regions (see Hancock 1995). This showed that the two classes of frequency were not significantly correlated, that is, that triplet repeat frequencies within ORFs differ from those in intergenic regions. We also observed that the triplet repeat frequencies in intergenic regions correlate significantly with their base composition. This agrees with a previous analysis of frequencies of significantly simple motifs derived from analysis of cryptically simple regions in a variety of genomes (Hancock 1995), which also showed a significant correlation of the frequencies of simple sequence motifs with their AT-richness in noncoding, but not coding, regions. This was attributed to a higher probability of slippage taking place in AT-rich sequences (because of their lower melting temperatures). We therefore take a strong correlation of motif frequency with base composition as an indication that triplet repeats in noncoding regions of the yeast genome are largely (although not necessarily entirely) neutral structures. The frequencies of repeats in coding regions therefore reflect the action of functional selection.

The abundance of CAG repeats in yeast coding regions parallels its abundance in mammalian exons (Stallings 1994). However, AAT repeats, also very abundant in yeast coding regions, are rare in the exons of mammals (Stallings 1994), and GGC repeats, relatively abundant in mammalian exons (Stallings 1994), are uncommon in yeast genes. This may be because Asn repeats (AAT) are not well tolerated in mammals and that the same is true for Gly repeats (GGC) in yeast. Asn repeats appear to be rare in vertebrates but more common in invertebrate, yeast, and plant proteins (Stallings 1994). Alternatively, these differences could be due to differences in the slippage process between the groups or may reflect the low GC content of the yeast genome (Richard and Dujon 1997).

A further indication that selection has influenced the distribution of amino acid repeats comes from our analysis of the functions of repeat-containing proteins. This analysis represents the first attempt at a whole-genome analysis of amino acid repeats and protein function. Previous analyses of this question have been based on broad database surveys (Wharton et al. 1985; Gerber et al. 1994; Karlin and Burge 1996) and have been broadly consistent with one another in associating Gln repeats with transcription factors. Our analysis showed a number of statistically significant overrepresentations of proteins belonging to particular functional groups, as defined by the YPD classification (Hodges et al. 1999), in proteins containing particular classes of amino acid repeats. These overrepresentations include an association of Gln repeats with transcription factors, which we are able for the first time to put on a secure statistical footing. Further, the finding that Gln repeats are abundant in yeast transcription factors indicates that this is a very widespread characteristic of eukaryotic transcription factors. The associations we observe must reflect the action of selection at some level, either due to a positive contribution of amino acid repeats to the functions of some classes of proteins (particularly transcription factors and protein kinases for the polar and acidic amino acids) or because these groups of proteins have characteristic structures or interactions which permit the incorporation of types of amino acid repeats that cannot be accommodated in other types of protein (Green and Wang 1994).

There is some evidence that changes in length of Gln repeats, in particular, can affect protein function, particularly in the case of the androgen receptor (Kazemi-Esfarjani et al. 1995), and there is some circumstantial evidence that they and other repeated amino acids may participate in protein–protein interactions (Mitchell and Tjian 1989; Perutz et al. 1994; Pinto and Lobe 1996). There is also evidence that frameshifts within regions coding for Gln repeats can be deleterious (Lanz et al. 1995). Expanded Gln repeats can also cause disease in humans (Reddy and Housman 1997). Thus it is clear that changes in repeat length can produce selectable phenotypes at some level. Perhaps the strongest evidence for a positive function for amino acid repeats is the observation that, in plants, insertion of a homopolymeric Gln stretch into a transcription factor gives rise to a higher level of expression of a target gene (Schwechheimer et al. 1998). However, none of these studies demonstrates unambiguously whether amino acid repeats play a positive or a (nearly) neutral role in the functions of proteins. Our data provide intriguing associations between the most common amino acid repeats and cellular components which appear to be part of the cell signaling system, and it is tempting to speculate that changes in the length of repeats in such systems could alter their behavior and therefore contribute to their evolutionary diversification (Hancock 1993; Karlin and Burge 1996; Richard and Dujon 1997), perhaps involving molecular coevolution between proteins (Dover and Flavell 1984; Hancock 1993). Such diversification could be relatively rapid on an evolutionary time scale because of the high mutation rates of microsatellites (see Hancock 1999). However, further experimental analysis will be necessary to test ideas of this kind.

In conclusion, our analyses provide evidence that amino acid repeats are common in yeast (*S. cerevisiae*) proteins and that their locations are subject to selection on the reading frames in which tandemly repeated codons can accumulate, the types of codon repeats that can accumulate in proteins in general, and the types of amino acids that can accumulate in proteins of particular functional types. Slippage appears to have made a significant contribution to the evolution of amino acid repeats in yeast, but many repeats, particularly Ser repeats,

appear to have evolved primarily by the accumulation of point mutations, at least during their recent evolutionary history.

Acknowledgments. We are grateful to Steve Oliver for comments on an early version of the manuscript. We thank the UK Medical Research Council for financial support. M.M.A. was supported by a postdoctoral fellowship from Ministerio de Educación y Cultura, Spanish Government.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Behe MJ (1995) An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. Nucleic Acids Res 23:689–695

Dechering KJ, Cuelenaere K, Konings RNH, Leunissen JAM (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res 26:4056–4062

Dover GA, Flavell RB (1984) Molecular coevolution: DNA divergence and the maintenance of function. Cell 38:622–623

Field D, Wills C (1996) Long, polymorphic microsatellites in simple organisms. Proc R Soc Lond B 263:209–215

Field D, Wills C (1998) Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proc Natl Acad Sci USA 95:1647–1652

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811

Goffeau AB, et al. (1996) Life with 6000 genes. Science 274:563–567

Green H, Wang N (1994) Codon reiteration and the evolution of proteins. Proc Natl Acad Sci USA 91:4298–4302

Hancock JM (1993) Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID). Nucleic Acids Res 21:2823–2830

Hancock JM (1995) The contribution of slippage-like processes to genome evolution. J Mol Evol 41:1038–1047

Hancock JM (1999) Microsatellites and other simple sequences: Genomic context and mutational mechanisms. In: Goldstein DB, Schlötterer C (eds) Microsatellites: Evolution and applications. Oxford University Press, Oxford, pp 1–9

Hancock JM, Shaw PJ, Bonneton F, Dover GA (1999) High sequence turnover in the regulatory regions of the developmental gene *hunchback* in insects. Mol Biol Evol 16:253–265

Henderson ST, Petes TD (1992) Instability of simple sequence DNA in Saccharomyces cerevisiae. Mol Cell Biol 17:3382–3387

Hodges PE, McKee AHZ, Davis BP, Payne WE, Garrels JI (1999) Yeast Protein Database (YPD): A model for the organization and presentation of genome-wide functional data. Nucleic Acids Res 27:69–73

Karlin S, Burge C (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci USA 93:1560–1565

Kazemi-Esfarjani P, Trifiro MA, Pinsky L (1995) Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: Possible pathogenetic relevance for the (CAG)n-expanded neuronopathies. Hum Mol Genet 4:523–527

Lanz RB, Wielands S, Hug M, Rusconi S (1995) A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. Nucleic Acids Res 23:138–145

Levinson G, Gutman GA (1987) Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Mewes HW, et al. (1997) Overview of the yeast genome. Nature 387: 7–65

Mitchell PJ, Tjian R (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science 245: 371–378

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

Nakamura Y, Gojobori T, Ikemura T (1997) Codon usage tabulated from the international DNA sequence databases. Nucleic Acids Res 25:244–245

Ohno S, Epplen JT (1983) The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc Natl Acad Sci USA 80:3391–3395

Perutz MF, Johnson T, Suzuki M, Finch JT (1994) Glutamine repeats as polar zippers: Their role in inherited neurodegenerative disease. Proc Natl Acad Sci USA 91:5335–5358

Petes TD, Greenwell PW, Dominska M (1997) Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. Genetics 146:491–498

Pinto M, Lobe CG (1996) Products of the grg (Groucho-related gene) family can dimerize through the amino-terminal Q domain. J Biol Chem 271:33026–33031

Pupko T, Graur D (1999) Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. J Mol Evol 48:313–316

Reddy PS, Housman DE (1997) The complex pathology of trinucleotide repeats. Curr Opin Cell Biol 9:364–372

Richard G-F, Dujon B (1996) Distribution and variability of trinucleotide repeats in the genome of the yeast Saccharomyces cerevisiae. Gene 174:165–174

Richard G-F, Dujon B (1997) Trinucleotide repeats in yeast. Res Microbiol 148:731–744

Rose O, Falush DA (1998) Threshold size for microsatellite expansion. Mol Biol Evol 15:613–615

Schwechheimer C, Smith C, Bevan MW (1998) The activities of acidic and glutamine-rich transcriptional activation domains in plant cells: Design of modular transcription factors for high-level expression. Plant Mol Biol 36:195–204

Sia EA, Jinks-Robertson S, Petes TD (1997) Genetic control of microsatellite instability. Mutat Res 383:61–70

Stallings RL (1994) Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequences: implications for human genetic diseases. Genomics 21:116–121

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Valle G (1993) TA-repeat microsatellites are closely associated with ARS consensus sequences in yeast chromosome III. Yeast 9:753–759

Wharton KA, Yedvobnick B, Finnerty VG, Artavanis-Tsakonas S (1985) opa: A novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in D. melanogaster. Cell 40:55–62