Neurological Proteins Are Not Enriched For Repetitive Sequences

Melanie A. Huntley and G. Brian Golding[1]

Department of Biology, McMaster University, Hamilton, Ontario L8S 4K1, Canada

Manuscript received August 19, 2003
Accepted for publication December 11, 2003

ABSTRACT

Proteins associated with disease and development of the nervous system are thought to contain repetitive, simple sequences. However, genome-wide surveys for simple sequences within proteins have revealed that repetitive peptide sequences are the most frequent shared peptide segments among eukaryotic proteins, including those of *Saccharomyces cerevisiae*, which has few to no specialized developmental and neurological proteins. It is therefore of interest to determine if these specialized proteins have an excess of simple sequences when compared to other sets of compositionally similar proteins. We have determined the relative abundance of simple sequences within neurological proteins and find no excess of repetitive simple sequence within this class. In fact, polyglutamine repeats that are associated with many neurodegenerative diseases are no more abundant within neurological specialized proteins than within nonneurological collections of proteins. We also examined the codon composition of serine homopolymers to determine what forces may play a role in the evolution of extended homopolymers. Codon type homogeneity tends to be favored, suggesting replicative slippage instead of selection as the main force responsible for producing these homopolymers.

THE presence and abundance of simple repetitive sequences within nucleotide sequences are well known. Microsatellites and other tandemly repeated sequences within DNA are well characterized; however, similarly repetitive sequences within proteins are less well acknowledged and understood. Nevertheless, such repeats within eukaryotic proteins are abundant. They vary in composition from a simple reiteration of a single amino acid to long tracts of sequence that are predominated by the presence of one or only a few amino acids.

Genome-wide surveys for simple sequences have shown that these low-complexity sequences are the most commonly shared peptide fragments in eukaryotic proteomes (Golding 1999; Huntley and Golding 2000). The prominence of these regions in proteins is a eukaryotic phenomenon, as they are not as common or as highly repetitive in prokaryotes (Marcotte *et al.* 1999; Huntley and Golding 2000). Not enough is known about the structure and function of these highly repetitive, low-complexity regions despite their abundance in eukaryotic proteins.

Only a few functions have been ascribed to these unusual regions. One of the first described and perhaps best known are the *opa* and *opa-like* repeats found in essential developmental proteins in insects (Wharton *et al.* 1985). These repeats are stably located repetitive elements that typically encode a stretch of up to ~30 glutamines, with interspersed histidine residues.

[1]*Corresponding author:* Department of Biology, McMaster University, 1280 Main St. W., Hamilton, Ontario L8S 4K1, Canada.
E-mail: golding@mcmaster.ca

In some prokaryotes, reversible mutations within regions of repetitive simple sequence DNA are involved in phase variation (Stern *et al.* 1986; Hood *et al.* 1996; Saunders *et al.* 2000). This mechanism allows bacterial populations to adapt to changing environments and is important in bacterial virulence (Moxon *et al.* 1994). Functional studies have shown that acidic, glutamine-rich, and proline-rich regions comprise three types of activation domains (Mitchell and Tjian 1989; Triezenberg 1995), while Mar Alba *et al.* (1999) found that transporter proteins were overrepresented among proteins containing serine repeats.

Other well-known repetitive regions in proteins are thought to be the cause of several human neurodegenerative diseases. These are associated with proteins containing extended regions of tandemly repeated glutamine residues. These proteins and others involved in nervous system disease and development contain multiple long homopeptides within their sequence (Karlin and Burge 1996). But not all of the homopeptide tracts are composed of glutamine residues; other residues such as proline, serine, glycine, and glutamic acid form extended homopolymers in these proteins as well.

Huntington's disease was one of the first disorders characterized to be due to homopeptides. This disease is associated with neural cell death, progressive chorea, dementia, and seizures. It is believed to be caused by an increase in the length of a CAG triplet repeat within the *huntingtin* gene. The age of onset is inversely correlated with the length of CAG repeats (Snell *et al.* 1993; Duyao *et al.* 1993; Kieburtz *et al.* 1994). In normal individuals, the repeat length is typically between 9 and 30, while affected individuals tend to have 40 to 121

copies. The triplet repeat encodes a polyglutamine tract, which can form cross-links within and between proteins. This increased cross-linking may induce the formation of aggregates within the cell and consequent neuronal death (Cariello *et al.* 1996).

Kennedy's disease, also known as spinal and bulbar muscular atrophy (SBMA), is an X-linked disease that causes late onset lower-motor and primary-sensory neuropathy. Clinical symptoms include muscular atrophy, twitching, tremors, and androgen deficiency. The primary cause of this disease is an expanded CAG triplet repeat within the androgen receptor (AR) gene (La Spada *et al.* 1991). Like Huntington's disease, the triplet repeat encodes a polyglutamine tract that may have increasingly toxic effects on neuronal cells as the repeat expands.

Dentatorubral-pallidoluysian atrophy (DRPLA) is phenotypically similar to Huntington's disease, including late onset dementia, cerebellar ataxia, myoclonic seizures, and choreic and athetoid movements. Again an expanded CAG repeat, encoding polyglutamine, is responsible for the pathology of this disease (Li *et al.* 1993; Koide *et al.* 1994; Nagafuchi *et al.* 1994). Haw River syndrome is also caused by this same CAG expansion in the DRPLA gene (Burke *et al.* 1994).

Other neurological diseases that fall into this category are spinocerebellar ataxia (SCA) 1, 2, 3 (Machado-Joseph disease), 6, 7, and 17. All are caused by expansions of a polyglutamine tract in separate proteins (Banfi *et al.* 1994; Kawaguchi *et al.* 1994; Pulst *et al.* 1996; David *et al.* 1997; Zhuchenko *et al.* 1997; Nakamura *et al.* 2001; Silveira *et al.* 2002).

Studies of synthetic homopolymers, including glutamine repeats, have shown that some can form stable structures (Krull *et al.* 1965; Perutz *et al.* 1994; Rohl *et al.* 1999). Glutamine repeats have been shown to link pairs of β-strands by hydrogen bonds, forming polar zippers (Perutz *et al.* 1994). This action can result in rigid, irreversible aggregates of proteins within the cell. This has been used as an explanation of how the extended glutamine repeats in some human neurological proteins induce their associated neurodegenerative diseases (Perutz and Windle 2001). However, prion proteins within the yeast *Saccharomyces cerevisiae* also form aggregates, but lack homopeptide sequence within the prion-determining domain (Lindquist *et al.* 2001). Instead these domains tend to be enriched with polar amino acids, such as glutamine and asparagine.

Large numbers of short and long homopeptides are more frequent in developmental proteins than in other classes of proteins (Karlin and Burge 1996). We therefore expect that this may also be true for highly repetitive, low-complexity regions. In this study we collected all developmental and neurological proteins available from the human and *Drosophila melanogaster* proteomes and compared each to similar, but mutually exclusive, data sets to determine whether developmental and neurological proteins do indeed contain more highly repetitive, low-complexity regions than other classes of proteins. We confirm previous results for developmental proteins that they are enriched for homopolymers and, in addition, show that they are enriched for low-complexity sequence regions. But this is not the pattern observed in neurological proteins. As a class of proteins, neurological proteins do not have excess of regions highly enriched for glutamine.

In most of the neurodegenerative proteins, polyglutamine results from a triplet repeat expansion of the CAG codon. It is generally believed that these simple sequences arise as a byproduct of replicative slippage at the DNA level, similar to the process occurring in microsatellite expansion. However, not all repeats follow this pattern. Serine reiterations in yeast do not show bias toward long tracts of one of the possible codons (Mar Alba *et al.* 1999). This suggests that some repeats may have evolved via selection and not slippage.

In this study extended serine homopolymer tracts are used to show that the length of the tract does not affect the mixture of codon types but that the relative position of the codons within a tract does affect codon composition, indicating that these tracts are likely the result of slippage.

MATERIALS AND METHODS

Neurological proteins: Human and Drosophila neurological and kinase proteins were collected from the National Center for Biotechnology Information (NCBI) using the ENTREZ query system. To search for neurological proteins, the key words neural, neuro, nerve, and axon were used. To search for developmental proteins we used the key words development, morphogen, homeotic, differentiation, embryo, larva, and determination. These key words were based on the key words used in the gene ontology database (http://www.godatabase.org/dev/database/). Kinase proteins were collected by searching for the key word kinase. All key words (or modifications of the key word's roots) had to be present on the definition line of the GenPept files. All key word matches were screened to eliminate matches that did not fit into their respective categories, such as homeostasis, which matched to the root of homeotic. These databases are not exclusive, but this method is unbiased, explicit, and easily repeatable. All sequences targeted to the mitochondria were removed.

Many coding sequences within a genome are redundant duplicates, isozymes, or ancient duplications. Additionally, sequence databases can contain redundant sets of sequences. To construct a database of, for example, neurological proteins, such duplicates had to be filtered. First a BLAST search (Altschul *et al.* 1997) was done to screen for similar proteins within the genome. All proteins that had a BLAST expect value <0.75 were then pairwise aligned, using ALIGN (Myers and Miller 1988). The smaller of any two sequences that had a percentage identity $>20\%$ (*e.g.*, the percentage identity between hemoglobin and myoglobin) was thrown away as it was considered to be too recently evolutionarily related. In this way we retained the larger protein of any related pair of sequences. A nonredundant, human neurological database was then constructed, resulting in 433 sequences, equaling a 60% reduction. A nonredundant developmental protein database

containing 242 sequences was similarly constructed by discarding 75% of the sequences. Kinase proteins were collected as a control group and after filtering out 72% comprised 982 nonredundant sequences. From the 982 nonredundant kinase sequences, two more kinase databases were constructed to be comparable to the neurological database. The two kinase databases were each constructed by sampling from the 982 nonredundant sequences. These databases may have a small amount of overlap. The first sampling was designed to be comparable to the neurological database by being within 5% of its protein lengths and contained 422 sequences. The second was within 10% of the neurological sequence lengths and had 429 sequences. In this way, we not only had a full collection of nonredundant kinase sequences for which we could compare the neurological data set, but also had collections of kinase sequences that were compositionally similar to the neurological sequences and thus more directly comparable.

Databases from Drosophila protein sequences were constructed, resulting in 77 neurological proteins (a 45% reduction), 139 developmental proteins (a 56% reduction), and 128 kinases (a 65% reduction). The kinase database within 5% of the neurological lengths had 52 proteins, while the one within 10% of the neurological lengths had 64.

In total, we constructed five types of databases each for human and Drosophila proteins: neurological, developmental, kinase, kinase within 5% of neurological lengths, and kinase within 10% of neurological lengths. To analyze these databases we constructed comparison databases that were similar in composition to the original databases, while excluding neurological, developmental, and kinase proteins, respectively. Each database was used as a basis to sample sequences from the NCBI and to construct 100 random comparison databases. For instance, for human neurological proteins, 50 databases were constructed to contain human sequences that were not neurological, but otherwise randomly chosen from the NCBI and within 5% of the lengths of the neurological proteins. Another 50 databases were constructed to be within 10% of the lengths of the neurological proteins. Therefore, each protein within the neurological database had a protein of similar length within each of the comparison databases. In this way, each of the 100 comparison databases is mutually exclusive to the human neurological database, but is similar in protein length composition.

To determine how common highly repetitive, simple sequences were in these databases, BLAST searches were performed, using 100-residue-long homopolymers of each amino acid. The number of BLAST hits with expect values ≤ 0.01 were compared to those found from the 100 comparison databases and the corresponding percentiles were recorded.

This analysis was also performed on the redundant databases, to examine how the analysis was affected by making the databases nonredundant.

To ensure that these results were robust, we also performed the same analysis using BLAST with 50-amino-acid-long homopolymers and using two entirely distinct algorithms, SIMPLE (Alba *et al.* 2002) and SEG (Wootton and Federhen 1993).

Of these methods, the SIMPLE algorithm has the most rigid window length to search for cryptically simple sequences. During various trials we used total window lengths ranging from 40 to 100 and searched for monomeric-like simple sequences.

For analysis using the SEG algorithm, we chose a window length, L, of 40 and a complexity cutoff value, $K2(1)$, of 2.6. All low-complexity segments were sorted into amino acid categories on the basis of the composition of the segment. If two or more amino acids each had frequencies of 30% or higher, that segment was counted toward each of those categories. This was done to search for highly repetitive, low-complexity regions.

In addition, we analyzed the percentage of low complexity per sequence and the number of low-complexity regions per sequence. We did this using two different sets of SEG parameters: an L of 15 with $K2(1)$ of 1.9 and an L of 40 with $K2(1)$ of 2.6.

This entire analysis was also performed on the proteins from *Caenorhabditis elegans* to determine how widespread the resulting patterns were.

Homopolymer tracts: Analysis similar to a previous study (Karlin and Burge 1996) was performed on nonredundant protein sets for both humans and Drosophila. Following this previous study, we excluded proteins with extremely biased amino acid content if an amino acid had >20% frequency and searched for proteins with three or more homopeptides of lengths ≥5 residues whose combined lengths totaled no less than 20. In sequences with extreme bias in composition long homopeptides are expected to occur more often by chance. Karlin and Burge also screened for proteins containing at least one homopeptide of length ≥10 residues and at least one other of length ≥5 residues. We used the additional requirement that at least one homopeptide within a protein had a length of 15 residues or more to emphasize more extended homopolymers. The protein descriptions, their accession numbers, lengths, and the homopeptide lengths were recorded. Proteins with any known neurological function were grouped in the "neurological" category. Any of the remaining proteins with known developmental function were grouped under the "developmental" category. All other proteins with some known function were termed "other" and any remaining proteins were put in the "unknown or hypothetical" category. We further selected the serine homopeptides within these proteins and analyzed their codon content.

Serine is unique among the amino acid residues as it has two types of codons (TCN and AGY) that are at least two mutational events apart. Because of the mutational distance between the two codon types, studying the codon composition of serine homopolymers allows for a stronger distinction between the two hypotheses for their mechanism of evolution: replicative slippage or selection at the protein level. The TCN codons (TCA, TCC, TCG, and TCT) are more frequent than the AGY codons (AGT and AGC). If the homopeptide was simply the product of DNA slippage during replication, we would expect little mixture of the two codon types. For example, a polyserine tract that was created via strand slippage should be composed of only TCN codons or only AGY codons, but seldom a mixture of both. If, however, other forces, such as selection, are acting to create these homopeptides, then a mixture of the codon types might be more common.

We determined whether the length of the homopolymer tract influenced the mixture of the two codon types, using a likelihood-ratio test, $\chi^2 = -2 \ln (L_0/L)$, where L_0 is the likelihood of the null model and L is the likelihood of the model being tested.

Given genomic codon usage frequencies (f_{AGY} and f_{TCN}) and N polyserine tracts of length $n_i = x_i + y_i$, where x_i is the number of AGY codons and y_i is the number of TCN codons in the ith tract, the likelihood model can be summarized as

$$L = \prod_{i=1}^{N} ([(a + bn_i)f_{\mathrm{AGY}}]^{x_i}[1 - (a + bn_i)f_{\mathrm{AGY}}]^{y_i}). \qquad (1)$$

This model assumes a linear relationship between the length of the tract and codon composition. The parameters a and b were adjusted to maximize the likelihood, L. The null model, L_0, which is a random choice according to the frequencies, is the likelihood obtained with $a = 1$ and $b = 0$.

We used a second model to see if the position of a codon within a homopolymer tract influenced the type of codon found. For instance, if a codon position is flanked by AGY

TABLE 1

The number of significant BLAST hits within a neurological database compared to 100 nonneurological databases

Amino acid homopolymer	*Homo sapiens* (433 sequences)			*D. melanogaster* (77 sequences)		
	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases
A	8	>100	>100	8	100	>100
C	1	58–88	46–82	0	2–84	2–96
D	3	44–72	40–64	3	20–36	40–56
E	27	100	>100	4	8–14	10–12
F	0	2–94	2–96	0	2–100	2–100
G	8	82–90	74–84	6	58–68	46–66
H	6	>100	>100	8	>100	>100
I	0	2–100	2–100	0	2–98	2–98
K	6	18–34	20–34	1	8–32	10–36
L	0	2–96	2–98	0	2–92	2–96
M	0	2–100	2–98	0	2–100	2–100
N	0	2–82	2–76	5	60–82	52–72
P	23	100	>100	3	10–30	12–30
Q	4	14–28	26–34	11	72–82	58–74
R	0	*2–8*	2–12	0	2–44	2–44
S	4	16–20	14–26	5	*4–10*	12–14
T	0	2–30	2–14	0	2–100	*<0*
V	0	2–100	2–100	0	2–100	2–98
W	0	2–100	2–100	0	2–100	2–100
Y	0	2–96	2–94	0	2–98	2–98

Underlines indicate that the number of significant BLAST hits was in the 90th percentile or higher. ">100" indicates that the number of significant BLAST hits was higher and completely outside the distribution. Italics indicate that the number of significant BLAST hits was in the 10th percentile or lower. "<0" indicates that the number of significant BLAST hits was lower and completely outside the distribution.

codons, is that position more likely to be occupied by an AGY or a TCN codon? Given N polyserine tracts each with length n_i, where X_j denotes the codon at position j within the homopolymer tract, we calculated the likelihood as

$$L = \prod_{i=1}^{N} \left(\prod_{j=2}^{n_i - 1} l_j \right)$$

$$l_j = \begin{cases} -\ln P_1 & \text{if } X_{j-1} = \text{AGY},\ X_j = \text{AGY},\ X_{j+1} = \text{AGY} \\ 1 + \ln P_1 & \text{if } X_{j-1} = \text{AGY},\ X_j = \text{TCN},\ X_{j+1} = \text{AGY} \\ -\ln P_2 & \text{if } X_{j-1} = \text{TCN},\ X_j = \text{TCN},\ X_{j+1} = \text{TCN} \\ 1 + \ln P_2 & \text{if } X_{j-1} = \text{TCN},\ X_j = \text{AGY},\ X_{j+1} = \text{TCN} \\ -\ln P_3 & \text{if } X_{j-1} = \text{AGY},\ X_j = \text{AGY},\ X_{j+1} = \text{TCN} \\ 1 + \ln P_3 & \text{if } X_{j-1} = \text{AGY},\ X_j = \text{TCN},\ X_{j+1} = \text{TCN} \\ -\ln P_4 & \text{if } X_{j-1} = \text{TCN},\ X_j = \text{TCN},\ X_{j+1} = \text{AGY} \\ 1 + \ln P_4 & \text{if } X_{j-1} = \text{TCN},\ X_j = \text{AGY},\ X_{j+1} = \text{AGY}. \end{cases} \quad (2)$$

The null model suggests no dependence on neighboring codons. This situation is achieved when $P_1 = P_3 = e^{-f_{\text{AGY}}}$ and $P_2 = P_4 = e^{-f_{\text{TCN}}}$. Otherwise the parameters P_1, P_2, P_3, and P_4 can range from $1/e$ to 1. This results in a logarithmic decay function, bounded between zero and one. The parameters P_1 and P_2 are a measure of how likely the middle codon position will be occupied by the same codon type as the two surrounding codons, given that the two surrounding codons are of the same type. Thus, smaller values of P_1 and P_2 translate to increased probabilities of codon type homogeneity. P_3 and P_4 measure the bias of the middle codon position toward the left or the right codon position when they are not occupied by the same codon type. Therefore, smaller values of P_3 and P_4 mean an increase in the probability of the X_j codon being of the same type as the X_{j-1} codon only, while larger values of P_3 and P_4 correspond to an increase in the probability of being the same type as the X_{j+1} codon.

RESULTS

Neurological proteins: Table 1 shows that the human neurological database contained eight proteins with significant similarity to polyalanine. This number of BLAST hits was larger than that found for any of the 100 human nonneurological databases (matched to be within 5 and 10% of the neurological sequence lengths). The Drosophila neurological database also contained eight significant hits, which were in the 100th percentile of the number of significant BLAST hits from each of the 50 Drosophila nonneurological databases (matched

TABLE 2

The number of significant BLAST hits within a developmental database compared to 100 nondevelopmental databases

Amino acid homopolymer	*H. sapiens* (242 sequences)			*D. melanogaster* (139 sequences)		
	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases
A	4	94–98	100	12	>100	>100
C	0	2–70	2–80	0	2–82	2–90
D	4	94–100	78–88	5	44–64	52–60
E	18	>100	>100	15	84–90	92–100
F	0	2–96	2–100	0	2–100	2–100
G	9	>100	>100	17	>100	>100
H	2	78–94	86–98	13	>100	>100
I	0	2–100	2–100	0	2–100	2–100
K	8	86–94	84–92	1	*2–10*	*2–4*
L	0	2–100	2–98	0	2–90	2–92
M	0	2–100	2–100	0	2–94	2–100
N	1	86–100	82–98	23	>100	>100
P	18	>100	>100	10	>100	>100
Q	6	90–98	84–92	31	>100	>100
R	0	2–24	2–28	1	28–72	34–80
S	11	>100	>100	22	>100	>100
T	2	86–94	68–96	4	52–64	54–74
V	0	2–100	2–100	0	2–98	2–98
W	0	2–100	2–100	0	2–100	2–100
Y	0	2–90	2–88	1	96–100	94–100

Underlines indicate that the number of significant BLAST hits was in the 90th percentile or higher. ">100" indicates that the number of significant BLAST hits was higher and completely outside the distribution. Italics indicate that the number of significant BLAST hits was in the 10th percentile or lower. "<0" indicates that the number of significant BLAST hits was lower and completely outside the distribution.

to be within 5% of the neurological sequence lengths) and larger than that found for any of the 50 nonneurological databases (matched to be within 10% of the neurological sequence lengths).

Table 2 shows that developmental proteins seem to be enriched with alanine (A), glycine (G), proline (P), and serine (S) in comparison to nondevelopmental proteins equally numerous and matched for sequence length. Also, glutamic acid (E) and glutamine (Q) seem to be more common in developmental proteins; however, this result is not as consistent as for A, G, P, and S. It is interesting to note that lysine (K) shows a rather large discrepancy between human and Drosophila. In human developmental proteins, the number of significant BLAST hits to poly(K) was in the 84th to 94th percentile, but in Drosophila it was only in the 2nd to 10th percentile.

Neurological proteins are consistently enriched for alanine (A) and histidine (H) as shown in Table 1. Glutamine, which is associated with many neurodegenerative diseases, is not found to be overrepresented in neurological proteins. There are also large discrepancies between the species for glutamic acid (E) and proline (P).

The kinase proteins in Table 3 show that none of the amino acids are consistently enriched in both species, compared to nonkinase proteins. Kinase databases constructed to be of similar lengths to the neurological proteins (Tables 4 and 5) show no consistent enrichment and an increase in species-to-species discrepancies.

Neurological proteins have much less enrichment compared to developmental proteins. With the exception of alanine and histidine, neurological proteins are not consistently enriched for repetitive protein sequence.

We performed the BLAST analysis on the redundant data sets to investigate the effect of using nonredundant databases. We found no significant difference except for all of the kinases and for the neurological proteins from *D. melanogaster*. In these cases, the nonredundant databases were found to have significantly more BLAST hits per sequence than the redundant databases (data not shown).

Using the SIMPLE algorithm we obtained broadly similar results for neurological proteins. However, in many cases the windows detected as significantly simple were not as enriched for a predominant amino acid as those regions detected by BLAST. Another difference is that SIMPLE is not constructed to recognize residues with similar properties and misses such enriched regions as a result. In a parallel analysis using SEG, again the

TABLE 3

The number of significant BLAST hits within a kinase database compared to 100 nonkinase databases

	H. sapiens (982 sequences)			*D. melanogaster* (128 sequences)		
Amino acid homopolymer	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases
A	11	100	96–98	2	6–100	10–18
C	0	2–20	2–14	0	2–74	2–88
D	2	2–100	*<0*	6	76–88	84–94
E	50	68–72	58–62	7	18–26	20–38
F	0	2–90	2–96	0	2–100	2–100
G	3	*<0*	*<0*	12	94–96	96–100
H	4	62–70	54–72	5	44–64	50–70
I	0	2–100	2–100	0	2–98	2–100
K	30	>100	96–100	2	14–24	10–28
L	0	2–96	2–100	0	2–98	2–98
M	0	2–100	2–100	0	2–100	2–100
N	0	2–56	2–44	7	66–78	68–84
P	33	16–20	26–28	6	58–78	64–76
Q	15	28–44	42–48	8	4–12	*2–6*
R	31	>100	>100	1	32–58	18–66
S	23	92–96	94–96	11	82–84	78–82
T	5	70–80	64–80	5	70–76	74–90
V	0	2–98	2–100	0	2–96	2–100
W	0	2–100	2–100	0	2–100	2–100
Y	0	2–96	2–96	0	2–98	2–96

Underlines indicate that the number of significant BLAST hits was in the 90th percentile or higher. ">100" indicates that the number of significant BLAST hits was higher and completely outside the distribution. Italics indicate that the number of significant BLAST hits was in the 10th percentile or lower. "<0" indicates that the number of significant BLAST hits was lower and completely outside the distribution.

results were consistent with our BLAST analysis, but with more variability found within the Drosophila results (results not shown).

The patterns we obtained using BLAST with 50-amino-acid-long homopolymers were nearly identical to those found using the 100-residue-long homopolymers. However, the Drosophila results, like those from SEG, were more variable.

The parameter space for SEG is very large with numerous parameter sets possible for identifying different types of repetitive low-complexity sequences. Different parameter sets can give rise to dissimilar SEG results. The SEG analysis examining the percentage of low complexity and the number of low-complexity segments per sequence was highly inconsistent between the SEG parameters employed (data not shown).

The proteins of *C. elegans* yielded similar results to those of humans and Drosphila (data not shown). Again, the neurological proteins had no significant enrichment compared to the nonneurological databases. The developmental proteins had the greatest enrichment, while the kinase proteins had enrichment patterns like those found in humans and Drosophila.

Homopolymer tracts: Table 6 shows the lengths of the longest homopolymer tracts for each amino acid. This table does not reflect homopolymer frequency or the average lengths of such tracts. Only the individual extreme cases are listed. In humans, many of the longest tracts for neurological proteins are longer than those for the developmental proteins. In Drosophila the opposite is true. Also, for nine amino acids, in both humans and Drosophila, kinase proteins have homopolymers as long as or longer than those of the developmental and neurological proteins.

The APPENDIX lists proteins with multiple homopolymers containing at least one homopeptide of length 15 or more. This is composed of 29 human proteins (Table A1) and 74 Drosophila proteins (Table A2). While such proteins are more numerous in Drosophila, they also contain significantly ($P < 0.05$) more homopeptides per protein than do the human sequences. There are 559 homopeptide tracts for the Drosophila proteins and only 133 for humans. While KARLIN *et al.* (2002) found 192 human proteins with multiple-amino-acid runs, our altered criteria of at least one homopolymer of length ≥15 residues and our nonredundant database account for this difference.

In both humans and Drosophila, poly(Q) is the most frequent homopolymer tract. However, poly(Q) accounts for only 24.1% of the human homopolymers, while accounting for 53.1% of the Drosophila homopolymers. Another discrepancy between the two species is found in the abundance of poly(E), which accounts for 18.8% of the human homopolymers, but only 0.5%

TABLE 4

The number of significant BLAST hits within a kinase database (containing sequences within 5% of the length of neurological proteins) compared to 100 nonkinase databases

	H. sapiens (422 sequences)			*D. melanogaster* (52 sequences)		
Amino acid homopolymer	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases
A	5	92–94	100	0	*2–8*	*2–8*
C	0	2–42	2–46	0	2–86	2–94
D	1	6–14	6–16	2	36–68	38–70
E	10	8–12	*4–6*	1	*2–8*	*6–12*
F	*0*	*2–96*	*2–96*	*0*	*2–100*	*2–100*
G	*1*	*<0*	2–100	2	12–36	24–36
H	0	2–40	2–24	3	68–86	66–72
I	0	2–100	2–98	0	2–100	2–100
K	7	42–64	34–52	1	18–50	14–48
L	0	2–98	2–96	0	2–98	2–96
M	0	2–100	2–100	0	2–100	2–100
N	0	2–86	2–78	1	10–24	10–26
P	11	10–16	22–32	1	6–28	8–36
Q	2	6–12	6–16	2	*<0*	*6–8*
R	6	>100	>100	0	2–66	2–54
S	4	8–28	18–36	6	92–98	84–92
T	2	62–80	50–80	2	46–74	54–76
V	0	2–98	2–100	0	2–100	2–100
W	0	2–100	2–100	0	2–98	2–100
Y	0	2–96	2–92	0	2–100	2–100

Underlines indicate that the number of significant BLAST hits was in the 90th percentile or higher. ">100" indicates that the number of significant BLAST hits was higher and completely outside the distribution. Italics indicate that the number of significant BLAST hits was in the 10th percentile or lower. "<0" indicates that the number of significant BLAST hits was lower and completely outside the distribution.

of the Drosophila homopolymers. As well, poly(G) and poly(P) are more than double in humans (15.0% *vs.* 7.3% and 10.5% *vs.* 4.7%, respectively).

These interspecies discrepancies are largely consistent with previous results (Karlin *et al.* 2002). However, the lack of polyleucine within the human homopolymers was not found previously. Karlin *et al.* (2002) found 19% of human proteins with at least one homopolymer of length five or more residues contained polyleucine, and only 14 of 192 proteins with multiple homopolymers contained polyleucine. Because of the longer criteria we used to consider homopolymers, only 2 of these 14 proteins were present in our data.

Of the 11 polyserine tracts in human, 7 had absolutely no mixture of the codon types. Of the 56 Drosophila polyserine tracts, 26 had no mixture. From the analysis of the first model, which was used to determine if the length of a homopolymer tract influenced the underlying codon mixture, the likelihood-ratio test gave χ^2 values of 22.83 for humans and 57.18 for Drosophila. Using 2 d.f. these values corresponded to probabilities <0.001 of occurring by chance alone. The likelihood model suggests that longer polyserine tracts did not have significantly less mixture of codon types. In fact, the parameter b is small in both cases and did not have a consistent direction between the two species. However, the maximum-likelihood estimate of a took on a fractional value. This indicated that maximum-likelihood codon frequencies within the homopolymers were different from the genomic codon frequencies.

For the second model, which examines the influence of codon position within a homopolymer tract, we found that P_1, P_2, and P_4 were smaller than the null model values. For humans, P_3 was slightly greater than the null model value, but for Drosophila P_3 was less than the null model value. Likelihood-ratio tests gave χ^2 values of 81.29 for humans and 65.56 for Drosophila. Using 4 d.f., this translates to probabilities $\ll 0.001$. Indeed, being surrounded by one type of codon significantly increases the likelihood of the middle position also being that same codon type. Also, if the two neighboring codons are of different types, the middle position (X_j) will tend to be occupied by a codon type that matches the left-hand (X_{j-1}) site.

DISCUSSION

These results confirm previous reports, showing developmental proteins to be enriched for simple sequences composed primarily of alanine, glutamic acid, glycine, proline, glutamine, or serine. However, unexpectedly,

TABLE 5

The number of significant BLAST hits within a kinase database (containing sequences within 10% of the length of neurological proteins) compared to 100 nonkinase databases

	H. sapiens (429 sequences)			*D. melanogaster* (64 sequences)		
Amino acid homopolymer	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases	No. of significant BLAST hits	Percentile of 50 5% databases	Percentile of 50 10% databases
A	3	68–82	62–88	0	*2–4*	*2–6*
C	0	2–42	2–44	0	2–82	2–94
D	1	6–14	*6–8*	2	40–68	24–48
E	14	20–30	24–26	1	*2–6*	*<0*
F	0	2–96	2–96	0	2–100	2–100
G	1	2–100	*4–10*	3	22–52	38–50
H	1	20–60	28–64	1	4–18	12–36
I	0	2–100	2–98	0	2–100	2–100
K	10	82–86	86–88	2	28–60	40–58
L	0	2–100	2–98	0	2–94	2–90
M	0	2–96	2–100	0	2–100	2–100
N	0	2–74	2–84	1	8–18	6–16
P	17	70–76	80–84	2	34–58	16–48
Q	4	16–32	30–38	3	*<0*	*2–10*
R	11	>100	>100	1	44–68	50–88
S	11	100	100	4	34–50	28–42
T	3	74–92	86–98	2	48–66	48–60
V	0	2–100	2–98	0	2–100	2–100
W	0	2–100	2–100	0	2–100	2–100
Y	0	2–92	2–96	0	2–100	2–100

Underlines indicate that the number of significant BLAST hits was in the 90th percentile or higher. ">100" indicates that the number of significant BLAST hits was higher and completely outside the distribution. Italics indicate that the number of significant BLAST hits was in the 10th percentile or lower. "<0" indicates that the number of significant BLAST hits was lower and completely outside the distribution.

neurological proteins are only slightly enriched for alanine or histidine.

Neurological proteins have been thought to be enriched for repeats. These results show that as a class they do not have an excess of glutamine-enriched regions. Yet many neurological disorders are linked with extended polyglutamine tracts and proteins enriched with glutamine residues. There is evidence that many of these disorders result from protein aggregation, triggered by tracts of polyglutamine forming polar zippers (Yanagisawa *et al.* 2000; Perutz *et al.* 2002). As a result, polyglutamine may well be the best-characterized amino acid repeat to date.

In contrast, these results confirm the well-known example of simple sequence protein repeats, the *opa* and *opa-like* repeats originally found in insects (Wharton *et al.* 1985). The *opa* repeats are typically polyglutamine and are thought to be characteristic of developmentally regulated genes (Wharton *et al.* 1985). Polyglutamine was found to have the greatest number of significant hits within the Drosophila developmental database (Table 2). However, for both humans and Drosophila, significant BLAST hits to poly(A), -(G), -(P), and -(S) are more consistently abundant.

When we look at only the proteins containing multiple homopolymer tracts (Tables A1 and A2), we again find a rather large discrepancy between the two species. Although both species have poly(Q) as the most frequent homopolymer tract, it is far more frequent in the Drosophila proteins, representing over half of the homopolymers, while comprising less than a quarter of the human homopolymers. Poly(E) and poly(P) are much more abundant in humans than in Drosophila.

The amino acids that are found to be overrepresented as repeats within these proteins have diverse properties and thus a variety of implications for the structures of the proteins in which they are embedded.

However, overall, little is known about the types of protein structures extended amino acid repeats can form. A survey of eukaryotic proteins within the structural database revealed that low-complexity protein repeats are underrepresented and rarely structurally characterized (Huntley and Golding 2002). One explanation for their absence in the structural databases is that they are disordered and do not form consistent structures. The relationship between intrinsic structural disorder and sequence complexity in proteins has been well studied (Romero *et al.* 1999, 2001). Interestingly, all of the amino acids found to be enriched within the simple sequences of developmental and neurological proteins

TABLE 6

Length of the longest homopolymer tracts

Amino acid	*H. sapiens*			*D. melanogaster*		
	Neurological	Developmental	Kinase	Neurological	Developmental	Kinase
A	20	*16*	10	13	14	9
C	3	3	6	2	2	3
D	5	4	5	5	6	5
E	10	*15*	13	4	5	8
F	4	3	4	3	3	3
G	21	21	13	13	13	14
H	14	8	13	10	5	9
I	4	3	4	3	4	4
K	6	4	6	3	4	7
L	8	8	8	5	4	6
M	2	3	3	3	4	2
N	4	4	4	7	11	9
P	10	12	12	10	7	8
Q	21	21	8	25	20	14
R	4	4	6	4	4	4
S	9	11	11	6	8	7
T	6	7	8	5	8	8
V	4	4	4	4	4	4
W	3	3	3	2	2	2
Y	4	4	3	2	4	3

Italics indicate where the length is at least 15 residues. Underlines indicate where the length is ≥20 residues.

(alanine, glutamic acid, glycine, proline, glutamine, serine, and histidine) are considered *disorder promoting* (Romero *et al.* 2001). An in-depth survey of 90 regions of protein disorder determined that these proteins were typically involved in molecular recognition and suggested that many may function in signaling pathways (Dunker *et al.* 2002). It is argued that due to the conformational flexibility, intrinsic disorder would enable a single binding site to recognize differently shaped partners and have faster rates of association and dissociation (Dunker *et al.* 2002).

Although some repeat regions may arise and be maintained by selection, most appear to have arisen via slipped-strand mispairing, like microsatellite expansion. Our analysis of the serine homopolymers from Tables A1 and A2 shows evidence for slippage, in contrast to the results found in yeast serine homopolymers (Mar Alba *et al.* 1999) and Drosophila serine homopolymers (Karlin and Burge 1996). However, Mar Alba *et al.* (1999) examined specific codons, rather than codon types, while Karlin and Burge (1996) did not provide a statistical analysis of the serine codon type homogeneity. We also know that the repetitive regions have a higher rate of evolution (Huntley and Golding 2000). While one might anticipate a rapid expansion of an amino acid repeat within a protein to be detrimental, the question then remains why these extended repeat regions are so abundant and present in such important proteins.

A study on glutamine, alanine, and glycine repeats being inserted into the loop of a protein showed that the stability and folding rates of the proteins were minimally affected (Ladurner and Fersht 1997). In fact, the largest penalty comes with the addition of the first few residues and not the increased expansion of the repeat. Yet there are numerous deleterious conditions associated with these protein repeats, including the neurodegenerative disorders associated with triplet repeat expansions.

One hypothesis suggests that these repeats allow for protein elongation, followed by functional specialization of the repeat region via mutation (Green and Wang 1994). In support of this hypothesis it has been demonstrated that microsatellite expansion can occur quite rapidly and thus protein repeat expansion via slippage may occur rapidly as well. The importance of protein repeats as a mechanism for creating new protein domains may be increased by the findings of mutational bias in trinucleotide repeat evolution (Cooper *et al.* 1999; Rubinsztein *et al.* 1999). Originally it was assumed that microsatellites had equal probabilities of gaining and losing repeat units. However, these studies indicate that there is a bias toward adding repeat units.

Another argument in support of this hypothesis is that eukaryotes may compensate for longer generation times, using the extra variability afforded by protein repeats to rapidly create novel protein domains (Marcotte *et al.* 1999). Indeed, these protein repeats are a eukaryotic phenomenon, and the predominant amino acid differs from species to species. This would indicate that the particular characteristics of the amino acid in the repeat are not important; only the presence of a

new domain that can be quickly modified and either selected for a new function or deleted is important.

This speculation of the function of protein repeats still does not clearly explain why they are overly abundant in the critically important developmental proteins, but not so in neurological proteins.

We thank two anonymous reviewers for their valuable comments on the manuscript. This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to G.B.G. and an NSERC scholarship to M.A.H.

LITERATURE CITED

Alba, M. M., R. A. Laskowski and J. M. Hancock, 2002 Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics **18:** 672–678.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Banfi, S., A. Servadio, M. Y. Chung, T. J. Kwiatkowski Jr., A. E. McCall *et al.*, 1994 Identification and characterization of the gene causing type 1 spinocerebellar ataxia. Nat. Genet. **7:** 513–520.

Burke, J. R., M. S. Wingfield, K. E. Lewis, A. D. Roses, J. E. Lee *et al.*, 1994 The Haw River syndrome: dentatorubropallidoluysian atrophy (DRPLA) in an African-American family. Nat. Genet. **7:** 521–524.

Cariello, L., T. De Cristofaro, L. Zanetti, T. Cuomo, L. Di Maio *et al.*, 1996 Transglutaminase activity is related to CAG repeat length in patients with Huntington's disease. Hum. Genet. **98:** 633–635.

Cooper, G., N. J. Burroughs, D. A. Rand, D. C. Rubinsztein and W. Amos, 1999 Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Proc. Natl. Acad. Sci. USA **96:** 11916–11921.

David, G., N. Abbas, G. Stevanin, A. Durr, G. Yvert *et al.*, 1997 Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat. Genet. **17:** 65–70.

Dunker, A. K., C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, 2002 Intrinsic disorder and protein function. Biochemistry **41:** 6573–6582.

Duyao, M., C. Ambrose, R. Myers, A. Novelletto, F. Persichetti *et al.*, 1993 Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat. Genet. **4:** 387–392.

Golding, G. B., 1999 Simple sequence is abundant in eukaryotic proteins. Protein Sci. **8:** 1358–1361.

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

Hood, D. W., M. E. Deadman, M. P. Jennings, M. Bisercic, R. D. Fleischmann *et al.*, 1996 DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc. Natl. Acad. Sci. USA **93:** 11121–11125.

Huntley, M., and G. B. Golding, 2000 Evolution of simple sequence in proteins. J. Mol. Evol. **51:** 131–140.

Huntley, M. A., and G. B. Golding, 2002 Simple sequences are rare in the Protein Data Bank. Proteins **48:** 134–140.

Karlin, S., and C. Burge, 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.

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

Kawaguchi, Y., T. Okamoto, M. Taniwaki, M. Aizawa, M. Inoue *et al.*, 1994 CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat. Genet. **8:** 221–228.

Kieburtz, K., M. MacDonald, C. Shih, A. Feigin, K. Steinberg *et al.*, 1994 Trinucleotide repeat length and progression of illness in Huntington's disease. J. Med. Genet. **31:** 872–874.

Koide, R., T. Ikeuchi, O. Onodera, H. Tanaka, S. Igarashi *et al.*, 1994 Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nat. Genet. **6:** 9–13.

Krull, L., J. Wall, H. Zobel and R. Dimler, 1965 Synthetic polypeptides containing sidechain amide groups: water insoluble polymers. Biochemistry **4:** 626–632.

La Spada, A. R., E. M. Wilson, D. B. Lubahn, A. E. Harding and K. H. Fischbeck, 1991 Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature **352:** 77–79.

Ladurner, A. G., and A. R. Fersht, 1997 Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. J. Mol. Biol. **273:** 330–337.

Li, S. H., M. G. McInnis, R. L. Margolis, S. E. Antonarakis and C. A. Ross, 1993 Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms. Genomics **16:** 572–579.

Lindquist, S., S. Krobitsch, L. Li and N. Sondheimer, 2001 Investigating protein conformation-based inheritance and disease in yeast. Philos. Trans. R. Soc. Lond. B **356:** 169–176.

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

Marcotte, E. M., M. Pellegrini, T. O. Yeates and D. Eisenberg, 1999 A census of protein repeats. J. Mol. Biol. **293:** 151–160.

Mitchell, P. J., and R. Tjian, 1989 Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science **245:** 371–378.

Moxon, E. R., P. B. Rainey, M. A. Nowak and R. E. Lenski, 1994 Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4:** 24–33.

Myers, E. W., and W. Miller, 1988 Optimal alignments in linear-space. Comput. Appl. Biosci. **4:** 11–17.

Nagafuchi, S., H. Yanagisawa, K. Sato, T. Shirayama, E. Ohsaki *et al.*, 1994 Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nat. Genet. **6:** 14–18.

Nakamura, K., S. Y. Jeong, T. Uchihara, M. Anno, K. Nagashima *et al.*, 2001 SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum. Mol. Genet. **10:** 1441–1448.

Perutz, M. F., and A. H. Windle, 2001 Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats. Nature **412:** 143–144.

Perutz, M. F., T. Johnson, M. Suzuki and J. T. Finch, 1994 Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. Proc. Natl. Acad. Sci. USA **91:** 5355–5358.

Perutz, M. F., B. J. Pope, D. Owen, E. E. Wanker and E. Scherzinger, 2002 Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques. Proc. Natl. Acad. Sci. USA **99:** 5596–5600.

Pulst, S. M., A. Nechiporuk, T. Nechiporuk, S. Gispert, X. N. Chen *et al.*, 1996 Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat. Genet. **14:** 269–276.

Rohl, C. A., W. Fiori and R. L. Baldwin, 1999 Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. Proc. Natl. Acad. Sci. USA **96:** 3682–3687.

Romero, P., Z. Obradovic and A. K. Dunker, 1999 Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett. **462:** 363–367.

Romero, P., Z. Obradovic, X. Li, E. C. Garner, C. J. Brown *et al.*, 2001 Sequence complexity of disordered protein. Proteins **42:** 38–48.

Rubinsztein, D. C., B. Amos and G. Cooper, 1999 Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos. Trans. R. Soc. Lond. B **354:** 1095–1099.

Saunders, N. J., A. C. Jeffries, J. F. Peden, D. W. Hood, H. Tettelin *et al.*, 2000 Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. Mol. Microbiol. **37:** 207–215.

Silveira, I., C. Miranda, L. Guimaraes, M. C. Moreira, I. Alonso *et al.*, 2002 Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)n allele at the SCA17 locus. Arch. Neurol. **59:** 623–629.

Snell, R. G., J. C. MacMillan, J. P. Cheadle, I. Fenton, L. P. Lazarou *et al.*, 1993 Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat. Genet. **4:** 393–397.

Stern, A., M. Brown, P. Nickel and T. F. Meyer, 1986 Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell **47:** 61–71.

Triezenberg, S. J., 1995 Structure and function of transcriptional activation domains. Curr. Opin. Genet. Dev. **5:** 190–196.

Wharton, K. A., B. Yedvobnick, V. G. Finnerty and S. Artavanis-Tsakonas, 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.

Wootton, J. C., and S. Federhen, 1993 Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. **17:** 149–163.

Yanagisawa, H., M. Bundo, T. Miyashita, Y. Okamura-Oho, K. Tadokoro *et al.*, 2000 Protein binding of a DRPLA family through arginine-glutamic acid dipeptide repeats is enhanced by extended polyglutamine. Hum. Mol. Genet. **9:** 1433–1442.

Zhuchenko, O., J. Bailey, P. Bonnen, T. Ashizawa, D. W. Stockton *et al.*, 1997 Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. Nat. Genet. **15:** 62–69.

Communicating editor: S. W. Schaeffer

APPENDIX A1

Human proteins with multiple homopeptides, where at least one must be ≥15 residues long

Protein	Accession	Length	Homo-amino-acid runs of length ≥5 amino acids
Neurological			
Huntington's disease protein	P42858	3144	E_6, E_5, P_{11}, P_{10}, Q_{23}
Small conductance calcium-activated potassium channel protein 3	Q9UGI6	736	Q_{19}, Q_{12}, Q_5
Dachshund homolog (mushroom body, brain development)	XP_029528	676	A_9, A_6, G_{10}, G_{10}, G_6, G_5, S_{24}
Zinc family member 2 (zinc finger protein of the cerebellum)	NP_009060	533	A_{15}, A_9, A_9, A_5, G_6, G_6, H_9
Developmental			
AT-binding transcription factor 1	XP_051813	3703	A_7, A_5, E_{15}, E_8, G_{14}, G_6, P_{12}, P_5, Q_{19}, Q_8, Q_6, Q_5, Q_5
X-linked nuclear protein (XNP/ATRX)	AAC51657	2375	D_6, E_{15}, E_6, K_5, Q_6, S_5
Similar to mastermind homolog (mesoderm determination)	XP_003460	1138	Q_{25}, Q_{18}, Q_9
Homeobox protein HB9	P50219	401	A_{16}, A_7, G_8, G_6, G_5
Other			
Nuclear receptor corepressor 2	Q9Y618	2517	G_5, P_6, Q_{17}
CREB-binding protein	Q92793	2442	Q_{18}, Q_5
Endocrine regulator	XP_004505	2099	E_5, P_{15}
T-cell transcription factor NFAT5	O94916	1531	Q_{17}, Q_{10}, Q_5
Probable tumor suppressor protein MN1	Q10571	1319	G_7, G_5, G_5, P_5, P_5, Q_{28}, Q_5, Q_5
Serine arginine-rich pre-mRNA splicing factor SR-A1	XP_046313	1312	E_{18}, E_{15}, E_5, K_5, P_8, R_6, S_7, S_6, S_6
KIAA1855 protein (contains protein kinase domains)	XP_052751	1270	E_{23}, E_8, E_6, S_5
Bumetanide-sensitive sodium (potassium) chloride cotransporter 1	P55011	1212	A_{15}, G_5
Similar to gene trap ankyrin repeat	XP_031324	1010	G_{15}, Q_6
Sarcoplasmic reticulum histidine-rich calcium-binding protein precursor	P23327	699	D_{16}, E_{12}, E_8, E_7, E_7, E_6, H_5
Membrane bound transcription factor site 2 protease	O43462	519	S_{23}, S_5, V_5
Basic helix-loop-helix domain containing, class B, 3	NP_110389	482	A_{16}, A_7, A_5, A_5, A_5, G_5
Unknown or hypothetical			
Hypothetical protein KIAA0192	Q93074	2124	Q_{26}, Q_{26}, Q_7, Q_6, Q_5
Nuclear receptor coactivator RAP250	XP_047155	2063	Q_{25}, Q_9
Similar to KIAA0595 protein	XP_016112	1578	E_5, S_{20}, S_{11}
Unknown	AAH10457	1052	E_{21}, E_9, E_7, E_6, E_5, P_9, P_5
Hypothetical protein	T46347	853	E_5, H_8, Q_{27}
Hypothetical protein	XP_084369	646	A_6, A_5, A_5, A_5, G_{16}, G_{11}
Similar to unknown protein	XP_044459	494	P_9, P_8, S_{18}
Unnamed protein product	BAB13872	399	A_{19}, P_5
HRIHFB2206 protein	XP_043054	314	A_7, Q_{29}

APPENDIX A2

Drosophila proteins with multiple homopeptides, where at least one must be ≥15 residues long

Protein	Accession	Length	Homo-amino-acid runs of length ≥5 amino acids
Neurological			
Notch gene product (determination of glial fate)	AAF45848	2703	A_8, G_9, Q_{17}, Q_{13}
prospero gene product (specific RNA polymerase II transcription factor involved in asymmetric cytokinesis, a critical regulator of the transition from mitotically active cells to terminal differentiated neurons during neurogenesis)	AAF54628	1703	A_6, D_5, N_7, N_5, P_7, Q_{20}, Q_{18}, Q_{12}, Q_6, Q_6, Q_5, S_7, T_5
dachshund gene product (RNA polymerase II transcription factor involved in mushroom body, brain development)	AAF53538	1074	A_{15}, A_{12}, A_{10}, A_5, A_5, E_5, G_5, N_{12}, Q_{15}, Q_6, Q_5
sequoia (functions in dendritic development) gene product	AAF58415	518	A_7, G_5, Q_{18}, S_6
Developmental			
fs(1)h [female sterile (1) homeotic] gene product	AAF46312	1937	A_{14}, A_6, A_6, G_8, G_6, G_5, Q_{16}, Q_{11}, Q_{10}, Q_7, Q_6, Q_6, Q_5, S_7, S_5
Additional sex combs gene product [chromatin binding involved in negative regulation of homeotic gene (Polycomb group), which is localized to the polytene chromosome]	AAF58239	1669	A_{20}, H_5, Q_9, Q_9, Q_8, Q_7, Q_6, Q_5, Q_5, Q_5, S_6, S_5, T_5
Posterior sex combs gene product [DNA binding involved in negative regulation of homeotic gene (Polycomb group)]	AAF58434	1601	A_5, P_5, S_{17}, S_{10}, T_7, T_6, T_6, T_5
mastermind gene product (mesoderm determination)	AAF58300	1366	A_{10}, G_{11}, G_8, G_7, G_6, N_5, N_5, N_5, Q_{17}, Q_{14}, Q_{11}, Q_8, Q_8, Q_7, Q_7, Q_6, Q_6, Q_6, Q_6, Q_5, Q_5, Q_5, Q_5, Q_5, Q_5, T_5
odd-paired gene product (specific RNA polymerase II transcription factor involved in periodic partioning; blastoderm segmentation development)	AAF52084	609	H_7, H_5, Q_{16}, Q_8, Q_7, Q_5, S_5
abdominal A gene product (specific RNA polymerase II transcription factor; contains a homeobox domain)	AAF55359	330	Q_{17}, Q_6
mindmelt gene product (photoreceptor differentiation)	AAF57881	297	A_{15}, N_5, Q_5
Other			
Smrter gene product (transcription corepressor)	AAF48196	3502	A_{14}, A_{10}, A_5, A_5, G_7, G_7, G_7, G_6, H_5, N_5, Q_{23}, Q_{23}, Q_{23}, Q_{18}, Q_{16}, Q_{12}, Q_{11}, Q_{10}, Q_{10}, Q_{10}, Q_9, Q_9, Q_9, Q_7, Q_7, Q_7, Q_7, Q_7, Q_7, Q_7, Q_6, Q_6, Q_5, Q_5, Q_5, S_7, S_6, S_6, S_6
Eip75B (Ecdysone-induced protein 75B) gene product (specific RNA polymerase II transcription factor)	AAF49282	2065	A_5, P_{15}, P_8, Q_9, Q_9, Q_8, Q_8, Q_7, Q_6, Q_6, Q_6, Q_6, Q_6, Q_5, Q_5, Q_5, S_{17}, S_7, S_6, S_5, S_5
canoe gene product (actin-binding component of the adherens junction)	AAF52067	1954	N_5, N_5, P_6, P_6, Q_{15}, Q_9, Q_9, Q_7, Q_7, Q_7, Q_7, Q_5
taiman gene product (transcription coactivator involved in border cell migration)	AAF52755	1778	A_8, A_5, G_7, G_5, G_5, N_5, Q_{16}, Q_{14}, Q_9, Q_8, Q_8, Q_7, Q_6, Q_6, Q_6, Q_6, Q_5
dystrophin gene product (structural constituent of muscle that is a component of the dystrophin-associated glycoprotein complex)	AAF55676	1629	A_{15}, A_5, G_5, P_6, P_5, S_6
similar gene product (RNA polymerase II transcription factor)	AAF57008	1507	Q_{15}, Q_{12}, Q_{10}, Q_8, Q_6, Q_6, Q_5, Q_5, Q_5
CG3695 gene product (RNA polymerase II transcription mediator involved in transcription, from Pol II promoter)	AAF46925	1439	Q_{22}, Q_5
CG5960 gene product (RAS GTPase activator)	AAF48761	1436	A_6, Q_{16}, Q_{10}, Q_6, Q_5
Misexpression Suppressor of Ras 1 gene product (transcription factor)	AAF48297	1419	A_{10}, A_7, A_5, N_8, Q_{25}, Q_{19}, Q_8, Q_7

(*continued*)

APPENDIX A2

(Continued)

Protein	Accession	Length	Homo-amino-acid runs of length ≥5 amino acids
CG5797 gene product (putatively actin binding)	AAF50366	1375	P_{15}, P_6
Suppressor of zeste 2 gene product (DNA binding)	AAF58433	1368	N_{22}, Q_{15}, Q_{11}, Q_5, S_9
CG6227 gene product (RNA binding involved in mRNA splicing)	AAF48446	1224	A_{15}, A_8, A_5, G_5
JIL-1 gene product [protein serine/threonine kinase (EC:2.7.1.37) involved in protein amino acid phosphorylation]	AAF50105	1207	A_{16}, N_8
Checkpoint suppressor homologue gene product (transcription factor)	AAF46265	1161	G_9, G_8, N_6, Q_{31}, Q_8, Q_6, S_6, S_5, T_5
Eip93F gene product (involved in autophagy that is expressed in the prepupa and pupa)	AAF55940	1140	A_{15}, G_8, Q_9, Q_7, Q_6, S_6, S_5, S_5
bifocal gene product (microtubule binding involved in female meiosis chromosome segregation)	AAF48076	1108	A_6, Q_{15}
Hormone receptor-like in 38 gene product (ligand-dependent nuclear receptor)	AAF53914	1078	A_5, A_5, A_5, Q_{19}, Q_{16}, Q_8, Q_5, T_5
grainy head gene product (specific RNA polymerase II transcription factor)	AAF57782	1064	A_8, Q_{17}, Q_9, Q_7, Q_5
Clock gene product (RNA polymerase II transcription factor involved in circadian rhythm)	AAF50516	1023	Q_{33}, Q_{11}, Q_5
CG14650 gene product (heat-shock protein)	AAF52123	970	P_5, Q_{17}, Q_9, S_7
oo18 RNA-binding protein gene product (uridine-rich cytoplasmic polyadenylation element binding involved in mRNA polyadenylation)	AAF56120	915	A_9, G_6, Q_{18}, Q_7, Q_7
jim gene product (transcription factor)	AAF51862	820	A_5, G_5, Q_{17}, Q_9, Q_5
multiple ankyrin repeats single KH domain gene product (structural constituent of cytoskeleton involved in cytoskeletal anchoring)	AAF56266	784	Q_{30}, Q_6
GATAe gene product (nonspecific RNA polymerase II transcription factor)	AAF55262	734	A_5, Q_{15}, Q_6, Q_5, Q_5
brinker gene product (RNA polymerase II transcription factor)	AAF46251	704	A_{17}, H_8, H_5, Q_{19}, Q_{10}, Q_7, S_6, S_5
dusky gene product (component of the plasma membrane)	AAF48089	699	P_5, Q_{18}, Q_{14}, Q_9
big brain gene product (connexon channel)	AAF52844	696	P_5, Q_{20}, Q_{13}, Q_{11}, Q_5
CG6952 gene product (inward rectifier potassium channel)	AAF45991	592	A_6, S_{16}
EG:114E2.2 gene product (DNA binding)	AAF45888	582	S_{22}, S_5
CG18610 gene product (heterogeneous nuclear ribonucleoprotein)	AAF57661	468	G_{15}, G_6, Q_5, Q_5
Unknown or hypothetical			
CG10631 gene product	AAF53840	3781	E_5, N_5, Q_{17}, Q_8, Q_6
CG10115 gene product	AAF50647	3080	A_7, A_7, A_6, A_5, D_6, H_8, H_5, P_6, P_5, P_5, Q_{16}, Q_{12}, Q_{12}, Q_{10}, Q_{10}, Q_9, Q_8, Q_6, Q_6, Q_6, Q_6, Q_6, Q_5, Q_5, Q_5, Q_5, S_6, S_5
CG14023 gene product	AAF52239	2439	A_{12}, G_5, N_6, N_5, P_6, P_5, Q_{21}, Q_{16}, Q_{12}, Q_{11}, Q_{11}, Q_9, Q_7, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_5, Q_5, Q_5, Q_5, Q_5, Q_5, Q_5, Q_5
CG3363 gene product	AAF47179	2175	A_{17}, S_5
CG9817 gene product	AAF46608	2090	A_{14}, A_{10}, A_9, A_9, A_5, A_5, A_5, D_5, G_{11}, G_{10}, G_9, G_7, G_6, Q_{15}, Q_{12}

(*continued*)

APPENDIX A2

(Continued)

Protein	Accession	Length	Homo-amino-acid runs of length ≥5 amino acids
CG4857 gene product	AAF45924	1911	A_7, A_5, D_5, H_{15}, N_5, Q_7, Q_5, S_5, T_5
EG:EG0002.3 gene product	AAF45898	1761	P_7, P_6, Q_{34}, Q_{11}, Q_9, Q_5
CG17233 gene product	AAF49048	1543	A_{11}, A_6, A_5, N_5, Q_{17}, Q_9, Q_8, Q_7, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_6, Q_5, Q_5, Q_5
CG4744 gene product	AAF58379	1503	Q_{20}, Q_5
CG12188 gene product	AAF47681	1417	A_7, A_5, P_5, Q_{17}, Q_{12}, Q_9, Q_5, T_5
CG4297 gene product	AAF51491	1403	Q_9, S_{17}, S_9, S_6
CG14351 gene product	AAF51357	1316	A_7, G_5, G_5, L_5, Q_{15}, Q_5, S_9, S_7, S_7, T_8
CG9461 gene product	AAF54459	1182	A_7, A_5, A_5, S_{15}, S_7
CG11203 gene product	AAF47976	1133	A_{11}, A_7, A_7, A_6, A_6, P_6, Q_{21}, Q_{17}, S_5
CG14441 gene product	AAF46192	1125	A_7, A_7, Q_{16}, Q_{10}, Q_7, Q_5, Q_5, S_6
CG18375 gene product	AAF46699	1071	A_{16}, A_9, A_6, Q_{10}, Q_9, Q_9, Q_5, S_5
CG5166 gene product	AAF55196	1069	Q_{18}, Q_{12}, Q_{11}, Q_5
CG14442 gene product	AAF46190	1068	A_{10}, G_7, Q_{18}, Q_8, Q_6, Q_6, Q_5, S_5, T_5
BcDNA:LD21293 gene product [alt 2]	AAF51804	969	N_{16}, N_7, Q_5, S_5
CG10082 gene product	AAF46743	893	A_6, Q_{15}, Q_6, S_5
CG6700 gene product	AAF53017	874	P_7, Q_{15}, Q_{10}, Q_7, Q_7, Q_6, Q_5
CG6619 gene product	AAF50702	861	H_{10}, H_7, Q_{23}, T_7
CG15365 gene product	AAF46433	821	G_5, Q_{37}, Q_{15}, Q_5, Q_5, Q_5, Q_5, Q_5, S_5
CG14982 gene product	AAF47814	731	Q_{15}, S_5
CG13830 gene product	AAF56093	660	Q_{17}, Q_6
CG9369 gene product	AAF48088	638	E_5, G_6, Q_{16}
CG4360 gene product	AAF55756	556	G_{15}, G_6, G_5
CG15784 gene product	AAF46035	554	Q_6, S_{23}, S_{22}, S_{11}
CG12236 gene product	AAF46100	553	A_9, Q_{24}
CG15753 gene product	AAF48279	458	P_5, Q_{15}, Q_6, S_5
EG:171E4.3 gene product	AAF45655	449	A_{13}, G_5, P_6, Q_{16}, S_6
CG17203 gene product	AAF55760	242	P_7, T_{19}, T_8
CG8294 gene product	AAF50512	181	N_{15}, Q_8