Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 1560–1565, February 1996
Genetics

Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development

SAMUEL KARLIN AND CHRIS BURGE

Department of Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, October 27, 1995

ABSTRACT **Several human neurological disorders are associated with proteins containing abnormally long runs of glutamine residues. Strikingly, most of these proteins contain two or more additional long runs of amino acids other than glutamine. We screened the current human, mouse, *Drosophila*, yeast, and *Escherichia coli* protein sequence data bases and identified all proteins containing multiple long homopeptides. This search found multiple long homopeptides in about 12% of *Drosophila* proteins but in only about 1.7% of human, mouse, and yeast proteins and none among *E. coli* proteins. Most of these sequences show other unusual sequence features, including multiple charge clusters and excessive counts of homopeptides of length ≥ two amino acid residues. Intriguingly, a large majority of the identified *Drosophila* proteins are essential developmental proteins and, in particular, most play a role in central nervous system development. Almost half of the human and mouse proteins identified are homeotic homologs. The role of long homopeptides in fine-tuning protein conformation for multiple functional activities is discussed. The relative contributions of strand slippage and of dynamic mutation are also addressed. Several new experiments are proposed.**

Several human dominantly inherited neurodegenerative diseases are associated with abnormally expanded tracts of glutamine codons (for recent reviews, see refs. 1–3). Principal examples include Huntington disease (HD) protein (see Table 1), containing near the N terminus Q_{23} (Q = glutamine residue in the one-letter code); ataxin I (spinocerebellar ataxia type 1, SCA1), containing the glutamine cluster $Q_{12}HQHQ_{15}$; atrophin 1 (denatatorubral–pallidoluysian atrophy 1, DRPLA), containing the homopeptide Q_{20}; and androgen receptor (AR; spinobulbar muscular atrophy, Kennedy disease), containing the glutamine homopeptides Q_{21}, Q_6, Q_5. In these proteins the length of the polyglutamine run is polymorphic but of limited size in healthy individuals (1). Disease severity appears to be correlated with the extent of tandem reiteration of the CAG (glutamine) codon above a certain threshold number (1, 2).

Proposed consequences of hyperexpansion of DNA triplet repeats (especially of CTG and CAG) include loss of or altered rate of transcription or translation, mRNA instability, aberrant DNA hairpin structures, etc. (1–3). Hypermethylation has been shown as a consequence of extended reiterations of CGG (1, 2). On the protein structural level, difficulties may arise from protein aggregation due to attachment of glutamine-rich proteins to unrelated molecules (6), to inappropriate multimerization, or to formation of "polar zippers," in which a long stretch of glutamine residues link β-strands by hydrogen bonds (7–9).

It is striking that three of the trinucleotide disease proteins contain multiple long homopeptides, not only of glutamine (Table 1). Specifically, listing all homopeptides of lengths ≥ five residues, HD also contains P_{11}, P_{10}, E_5, and E_6; DRPLA also contains S_7, S_{10}, P_6, and H_5; and AR also contains P_8, A_5, and G_{24}. What kinds of proteins contain multiple long homopeptides? What specific functions or structures are associated with particular long homopeptides? Under what conditions will hyperexpansion of a DNA triplet occur, and when will such expansions cause disease? Why are the triplet-repeat disease genes conspicuous in neuronal tissues?

We identified all human, mouse *Drosophila melanogaster*, yeast (*Saccharomyces cerevisiae*), and *Escherichia coli* protein sequences of length ≥ 200 residues that contain multiple long homopeptides. (For precise criteria and rationale, see the legend to Table 1.) A search of the SWISS-PROT data base release 31 (February 1995) for sequences containing multiple long homopeptides identified 36 of 2234 (1.6%) human sequences (Table 1), 68 of 551 (12%) *D. melanogaster* sequences (Table 2), 22 of 1242 (1.8%) mouse sequences (not shown), 42 of 2454 (1.7%) yeast sequences (not shown), and 0 of 1790 *E. coli* sequences. These examples feature homopeptides predominantly of the uncharged polar amino acids Q, N, S, T, P, and H; the small residues G and A; or acidic types D and E. Apart from three cases of leucine runs among the human examples, no other aliphatic or aromatic runs appeared.

The examples of Tables 1 and 2 are intriguing. (*i*) With few exceptions, the *Drosophila* proteins contribute prominently in developmental regulatory capacities, and the majority of these proteins are active in the central or peripheral nervous system (CNS or PNS). (*ii*) Most of these proteins contain many unusual sequence features, including significantly high multiplet counts, multiple charge clusters, and multiple significantly long uncharged segments (see below for definitions and further discussion). (*iii*) Several contrasts between the *Drosophila* examples (Table 2) and the human examples (Table 1) stand out. (*a*) The proportion of *Drosophila* proteins with multiple long homopeptides (12%) is dramatically higher than in human proteins (1.6%). (*b*) In the *Drosophila* sequences, the predominant homopeptides are glutamine runs, found in 51 of 68 cases, whereas polyglutamine (Q_n, $n \geq 5$) occurs in only six of the 34 human proteins. (*c*) The *Drosophila* proteins tend to be longer than the human proteins (median amino acid length of sequences: human, 527: *Drosophila*, 748; and mouse, 460).

Before discussing in detail the sequences of Tables 1 and 2, we present some background discussion on charge clusters, uncharged runs, and multiplet counts in protein sequences.

BACKGROUND AND METHODS

Charge Clusters and Uncharged Runs. A charge cluster is a protein segment (typically 20–75 residues) with significantly high specific charge content relative to the charge composition of the whole protein; see refs. 5, 10, 11 for precise characterizations. The percentage of proteins with at least one significant charge cluster is about 20–25% in most eukaryotic species (e.g., human, mouse, and yeast), about 35% in *Dros-*

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CNS and PNS, central and peripheral nervous systems; DRPLA, denatatorubral–pallidoluysian atrophy 1; AR, androgen receptor; HD, Huntington disease; SCA1, spinocerebellar ataxia type 1.

Table 1. Sequence features of human triplet repeat disease proteins and other human proteins containing multiple long homo-amino acid runs

Protein name (accession no.)	Length, aa	Homo-amino-acid runs of length ≥ 5 aa	Charge clusters +, −, ±	Uncharged runs, no.	Multiplets (aa, codon)
Triplet repeat disease proteins					
Androgen Receptor (Kennedy's Disease) (P10275)	919	Q_{21}, Q_6, Q_5, P_8, A_5, G_{24}	1, 0, 0	0	0, +
Atrophin 1 (DRPLA) (N/A)	1191	S_7, S_{10}, P_6, H_5, Q_{20}	0, 0, 3	5	0, n/a
Ataxin 1 (SCA1) (S46268)	816	Q_{12}, Q_{15}	0, 0, 2	2	0, 0
Huntingtin (Huntington Disease) (P42858)	3144	Q_{23}, P_{11}, P_{10}, E_5, E_6	0, 1, 0	1	++, ++
Other human proteins with multiple long homopeptides					
RDC-1 (Q01851)	331	H_9, G_5, G_6, G_6, A_5, A_{10}	0, 0, 1	1	0, 0
Nervous system-specific Octamer Binding Factor N-OCT 3 (P20265)	443	G_{21}, Q_{21}, A_5, P_7	0, 0, 2	1	+, ++
Calcium-dependent Protease, Small Subunit (P04632)	268	G_{11}, G_{20}	0, 0, 0	2	0, 0
SOX-3 Protein (P41225)	443	G_5, A_{15}, P_5, A_7, A_6, A_{10}	1, 0, 1	1	0, 0
Sarcoplasmic Reticulum Histidine-rich Calcium-binding Protein (P23327)	699	E_{12}, D_{16}, E_7, H_5, E_6, E_7, E_8	0, 2, 0	1	0, +
Thyroliberin (P20396)	242	E_6, E_7, E_8	0, 0, 0	1	0, n/a
Hox-B4 (P17483)	251	P_{15}, P_5	0, 0, 0	0	0, n/a
Octamer-binding Transcription Factor 6 (Q03052)	398	G_6, G_6, G_6, P_6, H_6	0, 0, 2	1	0, 0
OTX1 (P32242)	354	S_5, A_6, H_5, H_{10}	1, 0, 1	3	+, n/a
SOX-4 (Q06945)	474	G_7, G_6, G_6, A_7, S_7	0, 1, 1	0	+, ++
Engrailed-1 (Q05925)	391	P_6, A_{10}, A_9	0, 0, 1	1	0, ++
Factor VIII Intron 22 Protein (P23610)	364	A_5, A_5, P_6, L_6	0, 0, 0	0	0, 0
Major Centromere Autoantigen B (P07199)	599	E_{14}, E_6, E_6, E_5, D_5	0, 1, 0	0	++, ++
Translational Initiation Factor 2 β (P20042)	333	K_8, K_6, K_6	0, 0, 0	0	0, 0
Heterogeneous Nuclear Ribo-NP L (P14866)	558	G_7, G_9, G_6, P_8	0, 0, 0	0	0, 0
Acrosin (P10323)	421	P_5, P_8, P_9	0, 0, 0	0	0, 0
Transcriptional Repressor Prot. YY1 (P25490)	414	E_5, G_5, H_{11}	0, 1, 0	1	++, ++
Hox-B3 Protein (P14651)	431	G_{11}, G_{10}	1, 0, 1	0	0, 0
Chromogranin A (P10645)	457	E_8, E_9, E_5	0, 0, 0	1	0, 0
Protein Kinase C Substrate, 80 KD (P14314)	527	L_5, E_{12}, E_7	0, 1, 0	0	0, 0
Ring3 (P25440)	754	E_6, E_5, E_5, S_5, S_{11}	1, 1, 1	2	++, +
Early Growth Response 2 (P11161)	406	P_7, A_{10}	0, 0, 1	1	0, 0
Early Growth Response 1 (P18146)	543	S_6, G_9, S_6	1, 0, 0	1	0, 0
Nucleolin (P19338)	706	A_5, D_5, D_5, E_{10}	0, 2, 1	0	++, ++
Amyloid-like Protein 2 (Q06481)	763	A_5, L_6, E_5, E_7	0, 1, 0	2	0, 0
MHC Class II Reg. Factor RFX-1 (P22670)	979	P_7, G_{14}, E_8	1, 1, 0	2	0, 0
Insulin Receptor Substrate-1 (P35568)	1242	G_5, S_7, Q_6, G_5, S_5, P_7	0, 0, 0	0	++, +
SP4 Transcription Factor (Q02446)	784	E_5, A_6, S_9	0, 1, 1	2	0, n/a
Heterogeneous Nuclear Ribo-NP U (Q00839)	806	E_{11}, G_6	0, 1, 0	0	++, ++
Macrophage CSF I Receptor (P07333)	972	L_5, L_{10}	0, 0, 0	0	0, 0
HRX (Q03164)	3969	G_7, A_7, S_9, G_5, E_5	0, 0, 1	6	++, ++
Adenomatous Polyposis Coli (P25054)	2843	S_5, P_5, A_5, S_6	0, 2, 0	0	0, 0

SWISSPROT release 31 (Feb 1995) was screened for human proteins satisfying either or both of the following: (*i*) three or more homopeptides of length ≥ 5 amino acids (aa), whose combined lengths total at least 20 aa; or (*ii*) at least one homopeptide of length ≥ 10 aa and at least one other of length ≥ 5 aa. Proteins with extremely biased aa composition (1 or more aa with frequency > 20%) were excluded, since long homopeptides are expected to occur quite frequently by chance in sequences of very biased composition. For each group of two or more substantially identical proteins, only the longest was chosen for analysis to avoid redundancy. This resulted in 36 out of 2234 (1.6%) human proteins being chosen, including the three triplet-repeat disease gene products androgen receptor, ataxin 1, and huntingtin. The atrophin 1 (DRPLA) sequence, not in SWISSPROT but included in the counts above for simplicity, was provided to us by R. Margolis (Johns Hopkins Medical School). The human TATA box-binding protein TFIID, which contains Q_{36} but no other long homopeptide, was excluded by these criteria. The proteins in each of the groups (disease-related and other proteins) were ordered on the basis of their multiplicity index, M = (no. of aa in runs of length ≥ 5 aa)/(protein length), from highest to lowest. Column 2 lists all homopeptide runs of length ≥ 5 aa in each protein in the order in which they occur along the protein sequence. Column 3 lists the number of significant positive charge clusters (clusters of R and K, one letter code), negative charge clusters (D and E), and mixed charge clusters (R, K, D, and E). Charge clusters are defined statistically as regions unusually rich in a particular charged residue type (4). Column 4 lists the number of significantly long uncharged runs—unusually long stretches devoid of charge or with only one or two charged residues (4, 5). Column 5 indicates the significance of the number of amino acid "multiplets" and codon "multiplets"—patterns of the type X_n, where X is any of the 20 amino acid types or one of the 61 codon types, respectively, and $n \geq 2$ (runs longer than 2 are counted only once). The number of multiplets in a protein (or its corresponding gene sequence) is evaluated in terms of the number of standard deviation (SD) units above or below the mean, given the length and amino acid/codon composition of the protein/gene (4): counts > 2.3 SD above the mean are indicated by "+"; those > 3.0 SD above the mean, by "++"; those within 2.3 SD of the mean, by "0." (No negative extremes were observed.) n/a, gene not available; Dis., disease.

ophila, and only 7% in *E. coli* (and most prokaryotes) (10). In eukaryotes, charge clusters are associated with transcriptional activation, membrane receptor activity, and developmental regulation, the last category apparently overrepresented among available *Drosophila* sequences (see refs. 10 and 11 for specific examples and classifications). In contrast, hardly any distinctive charge configurations occur among the bulk of housekeeping proteins, cytoplasmic enzymes, or among prokaryotic proteins (10). Proteins with multiple charge clusters are uncommon, about 3.5% in human, mouse, and yeast, but are relatively more frequent (13.5%) in *Drosophila* (11). Primary families of proteins with multiple charge clusters are

Table 2. Sequence features of *Drosophila* proteins containing multiple long homo-amino acid runs

Protein Name (accession no.)	Length, aa	Homo-amino acid runs of length ≥ 5 aa	Charge clusters +, −, ±	Uncharged runs, no.	Multiplets (aa, codon)	Devel./ Neuro./ TF	Description/function
E74-B (P11536)	883	S_{14}, S_8, A_5, Q_7, Q_8, A_{13}, Q_8, Q_5, Q_{11}, A_9, S_5, A_5	2, 2, 3	3	++, ++	D, T	Control of molting, ecdysome inducible
Abdominal-B (P09087)	491	Q_{10}, Q_8, Q_6, Q_5, Q_{10}, Q_7, A_6, N_5	1, −, 1	1	+, ++	D, N, T	Segment/PNS devel.
Odd-skipped (P23803)	392	Q_{12}, Q_6, Q_8, H_6, S_7	0, 1, 1	1	0, ++	D, N, T	Pair-rule segmentation
Big brain (P23645)	700	Q_{13}, Q_{10}, P_5, Q_{25}, Q_5	0, 1, 1	1	0, n/a	D, N	Neural/epidermal devel.
Hairy (P14003)	337	Q_5, Q_6, Q_6, A_{10}	0, 0, 1	2	0, 0	D, N, T	Segment/sense organ/PNS development
Dorsal (P15330)	678	Q_5, Q_8, Q_6, Q_{14}, A_5, Q_7	0, 0, 1	2	0, ++	D	Controls dorsal polarity
AEF1 (P39413)	308	N_5, Q_{13}, Q_6	0, 0, 0	2	+, n/a	T	Regulates *Adh*
E75-C (P13055)	1443	Q_5, Q_7, P_{15}, Q_6, Q_6, S_7, S_6, Q_8, P_8, S_5, S_5, Q_9, Q_5, S_{17}	1, 0, 0	2	++, ++	D, T	Control of molting, ecdysome inducible
Zeste (P09956)	574	Q_6, Q_7, A_6, Q_5, Q_7, A_6, G_5	1, 1, 2	1	0, ++	D, T	Activates *Ubx*
Ubx (P02834)	389	A_5, G_{13}, A_{10}	1, 0, 1	2	0, n/a	D, T	Anter./post. devel.
Giant (P39572)	448	Q_{11}, Q_{12}, A_9	1, 0, 0	1	++, ++	D, T	Represses Kruppel, Knirps
Caudal (P09085)	472	H_5, N_{10}, N_7, R_{11}	1, 0, 1	2	0, n/a	D, T	Segment devel. (early)
Abdominal-A (P29555)	330	Q_{17}, Q_6	0, 0, 1	1	0, +	D, N, T	Segment/PNS devel.
Engrailed (P02836)	552	Q_{11}, A_{14}, A_6, S_7	0, 1, 1	3	++, ++	D, N, T	Intrasegment devel.
P62/RM62 (P19109)	575	G_6, G_7, G_5, G_6, G_7, G_8	0, 0, 0	1	0, 0	—	RNA helicase
C/EBP (Q02637)	444	Q_{20}, G_5, Q_5	1, 1, 0	1	++, n/a	D, T	Border cell migration
Prospero (P29617)	1407	A_6, A_5, Q_6, N_7, N_5, D_5, Q_{18}, Q_6, Q_5, Q_{24}, P_7	0, 2, 1	4	++, ++	D, N, T	Embryonic CNS devel.
Female sterile homeotic (P13709)	2038	A_{14}, S_7, S_7, Q_{10}, Q_{17}, Q_6, Q_6, Q_5, Q_6, Q_{11}, A_6, S_7, S_5, A_6, G_5, G_8, G_6	0, 1, 5	6	++, ++	D	Oogenesis, regulates homeotic genes
RBP-J kappa (P28159)	594	Q_6, Q_6, Q_{14}, N_6, Q_6	1, 0, 1	2	0, ++	D, N, T	Suppressor of hairless
Maternal Pumilio (P25822)	1533	A_6, G_7, A_5, Q_8, Q_{13}, A_{11}, Q_{15}, A_{11}, Q_6, A_{10}, A_6	1, 0, 1	4	++, ++	D	Anter./post. devel.
Ftz-F1 (P33244)	1043	Q_9, A_6, T_5, N_7, N_7, T_5, Q_7, G_9, G_5, G_6	1, 1, 0	3	++, ++	D, T	Activation of *Ftz*
DSX, male (P23023)	549	A_8, A_6, H_5, G_9, P_6	1, 1, 2	2	++, n/a	D, T	Controls sexual diff.
ELAV (P16914)	483	A_6, A_6, Q_5, Q_6, A_6	0, 0, 1	1	0, 0	D, N	Nerve development
Deformed (P07548)	590	G_6, Q_5, N_5, N_{11}, N_7	1, 1, 1	3	++, ++	D, N, T	Regulates segmentation
DGK (Q01583)	517	S_5, A_5, Q_{13}, Q_6	0, 0, 0	0	0, 0	—	Signal transduction
Knirps (P10734)	429	A_5, Q_7, A_6, S_6	1, 0, 0	0	++, ++	D, T	Abdominal segmentation
Orthodenticle (P22810)	671	Q_5, A_5, A_5, A_5, A_7, G_{10}	1, −, 1	1	++, ++	D, N, T	Segmentation (head)
Topoisomerase I (P30189)	972	H_6, S_7, S_5, S_6, S_5, S_5, S_8, S_6, D_5	0, 1, 1	2	0, ++	—	Alters DNA supercoiling
Eyes absent (Q05201)	766	Q_7, Q_9, G_5, A_6, A_7, A_7	2, 1, 1	1	++, ++	D, T	Eye development
Antennapedia (P02833)	378	Q_9, Q_5, Q_6	1, 0, 1	2	0, 0	D, T	Anter./post. devel.
Cut (P10180)	2175	A_5, A_9, Q_7, A_9, N_7, N_6, N_5, A_{11}, Q_5, A_5, A_7, A_5, A_7, A_5, A_8, P_6, S_6	0, 2, 2	5	++, ++	D, N, T	Embryonic CNS devel.
Hairless (Q02308)	1077	N_5, T_5, S_5, S_5, S_7, A_9, A_{10}, A_8	1, 1, 1	2	++, ++	D, N, T	Sense organ/PNS devel.
Polycomb (P26017)	390	H_{10}, H_8	0, 0, 0	0	0, ++	D, N	Regulates homeotic genes
Polyhomeotic-proximal chromatin (P39769)	1589	Q_7, Q_5, Q_6, Q_5, Q_6, Q_7, Q_6, Q_5, Q_7, Q_7, Q_6, T_6	1, 0, 2	3	++, n/a	D, T	Regulates segment development
Disconnected (P23792)	568	A_6, P_5, A_8, H_7	1, 2, 1	3	++, ++	D, N, T	Embryonic nerve devel.
5HT-receptor (P28285)	834	N_5, T_5, A_5, A_5, G_5, A_5, Q_8	1, 2, 0	1	0, +	N	Serotonin receptor
Shaker-ε (P08513)	656	Q_6, Q_7, Q_{11}, Q_5	1, 0, 2	1	++, ++	D, N	Voltage-dep. K^+ channel
Labial (P10105)	635	Q_5, Q_5, Q_6, Q_6, G_6	−, −, 1	0	++, ++	D, T	Anter./post. devel.
GRH/ELF1 (P13002)	1063	Q_9, A_8, Q_{17}, Q_5, Q_7	0, 0, 1	3	+, +	D, T	Nerve devel./cuticle
Suppressor 2 of zeste (P25172)	1365	Q_{11}, Q_{14}, Q_5, N_{20}, S_9	0, 1, 1	0	0, 0	D, N, T	Segmentation, regulates *Ubx*
Yan (Q01842)	732	Q_{10}, Q_8, Q_5, A_7	0, 1, 1	2	0, ++	D, N, T	Neural differentiation
Daughterless (P11420)	710	G_5, Q_5, G_7, A_5, Q_7	0, 0, 1	2	0, ++	D, N, T	Neurogenesis, sex determ.
Stubble (Q05319)	786	S_{10}, Q_6, T_8, S_8	0, 1, 0	0	+, +	D	Epithelial devel.
Empty spiracles (P18488)	497	Q_6, Q_6, A_8	1, 1, 1	1	+, ++	D, T	Antenna/mandible devel.
KNRL (P13054)	647	A_6, G_9, N_{11}	1, 2, 0	2	++, ++	D, T	Embryonic segmentation
PSC (P35820)	1603	T_5, T_6, T_8, T_6, S_{17}, S_{10}, P_5	0, 1, 1	1	++, ++	D, N, T	Regulates homeotic genes
Runt (P22814)	509	A_{12}, A_6	0, 1, 0	2	0, +	D, N, T	Pair-rule segmentation
Fork head (P14734)	510	G_{13}, H_5	0, 0, 1	1	+, ++	D, T	Embryonic CNS devel.
CF2, III (Q01522)	514	P_6, Q_{11}	0, 0, 0	0	0, 0	D, T	Chorion development
SHAB11 (P17970)	924	Q_7, Q_9, Q_5, Q_9	0, 0, 2	3	+, ++	N	Voltage-dep. K^+ channel
TKR (P14083)	753	E_5, A_6, A_6, A_8	0, 1, 0	1	++, n/a	D	Protein-tyrosine kinase
SOL (P27398)	1597	Q_6, H_5, H_6, S_5, Q_5, Q_9, H_8, G_5	0, 0, 0	0	0, ++	N	Neuronal devel.

Table 2. (*Continued*)

Protein Name (accession no.)	Length, aa	Homo-amino acid runs of length ≥ 5 aa	Charge clusters +, −, ±	Uncharged runs, no.	Multiplets (aa, codon)	Devel./ Neuro./ TF	Description/function
Rutabaga (P32870)	2248	G_{10}, Q_5, Q_8, Q_5, Q_5, Q_5, H_6, Q_5, D_5, Q_{13}	1,2,1	4	++, ++	N	Mutant has learning defect
Single-minded (P05709)	673	S_5, Q_5, S_5, Q_5	0,0,1	1	+, +	D, N, T	CNS development
Staufen (P25159)	1026	Q_5, P_5, Q_7, P_5	0,0,1	2	++, ++	D	Controls oocyte polarity
Inflated (P12080)	1394	S_7, S_5, S_6, S_{10}	0,0,0	2	++, 0	D	Intercellular comm.
PTP99A (P35832)	1301	S_5, Q_{16}, S_5	0,0,0	2	+, ++	D, N	Intercellular comm.
Zinc-finger 1 (P28166)	1060	Q_6, Q_5, Q_5, Q_5	0,2,1	2	+, +	D, N, T	Embryonic CNS devel.
DWNT-3 (P28466)	1010	G_6, S_{14}	0,0,0	0	0, +	D, N	CNS/wing devel.
Spalt-major (P39770)	1356	S_5, A_6, A_{14}	0,2,2	4	++, n/a	D, N, T	Segment development
Periodic clock (P07663)	1155	G_5, A_5, A_6, A_5	0,1,1	3	+, +	N	Biological rhythms
SU(S) (P22293)	1322	G_5, G_5, Q_6, G_8	1,2,1	2	++, ++	D	RNA-binding suppressor
Notch (P07207)	2703	A_8, G_9, Q_{13}, Q_{17}	0,0,0	2	0, 0	D, N	Neurogenic, oogenic
Brahma (P25439)	1638	Q_5, Q_6, E_5, E_5, D_6	0,3,2	4	++, ++	D, T	Regulates homeotic genes
SOS (P26675)	1595	G_5, G_5, A_6, Q_6	0,0,0	1	0, 0	D, N	Neuronal development
Trithorax (P20659)	3759	S_5, D_6, Q_7, Q_5, Q_7, S_7	0,2,1	6	++, ++	D, T	Segmentation
Zinc-finger 2 (P28167)	3005	S_5, A_7, Q_5, S_7, Q_5	0,0,0	4	++, 0	D, N, T	Embryonic CNS devel.
FAT (P33450)	5147	S_5, D_5, S_7, Q_5, Q_6, P_5, Q_6	0,0,1	0	0, 0	D, N	Cell adhesion protein

SWISSPROT release 31 (February 1995) was screened for *D. melanogaster* proteins satisfying the criteria described in the legend to Table 1. This resulted in 68 out of 551 *Drosophila* proteins (12%) being chosen. The layout of Table 2 is the same as that of Table 1, except that two new columns have been added. Column 7 (Devel./Neuro./TF) describes whether or not the protein is a developmental protein (D), a nervous system-related protein (N), and/or a transcription factor (T). Column 8 provides a very brief indication of the major developmental role(s) or other known functions of each protein. A "−" in column 4 indicates that the frequency of the corresponding charged residue type was too low in the given protein for evaluation of charge clusters. The major *Drosophila* developmental protein mastermind, which exhibits (in order of occurrence) the plethora of runs Q_{10}, Q_5, Q_{16}, Q_5, N_5, G_{10}, N_5, G_7, N_5, Q_5, Q_7, Q_{14}, Q_6, Q_7, Q_6, Q_{16}, Q_5, Q_5, Q_7, Q_8, Q_6, Q_5, Q_6, A_{10}, Q_5, Q_{12}, Q_5, T_5, G_8, and G_6, is excluded because of its biased composition (Q = 21.7%). diff., Differentiation; anter./post., anterior/posterior; dep., dependent; determ., determination; comm., communication; 5HT, seritonin; CNS and PNS, central and peripheral nervous systems.

(*i*) developmental regulatory proteins, (*ii*) voltage-gated ion-channel proteins, and (*iii*) regulatory proteins of large eukaryotic DNA viruses (10, 11).

Sequence Features of Transcription Factors. Transcription factors tend to have a modular organization, generally involving a DNA-binding domain, activating domain(s), and other structural/functional domains separated by linker segments (12, 13). Generally, transactivating domains in transcription factors have been associated with acidic clusters (e.g., GAL-4, GCN-4), with proline or glutamine clusters (e.g., SP1, CREB), or with serine/threonine-rich regions (e.g., POU proteins) (for reviews, see refs. 12 and 13). For example, the charge cluster overlapping the POU domain of human Oct-1 is flanked by a Q-rich region on the N-terminal side and by an abundance of S and T residues on the C-terminal side. The human transcription factor Sp1 has a charge cluster centering on three tandem zinc-finger motifs that are responsible for DNA binding and a long uncharged region near the N terminus rich in S, T, and Q residues. The Q-rich portions are considered important for transcriptional activation (14, 15). Recently, based on a series of GAL-4 chimeric constructs manipulating different sizes of Q_n-runs, the highest levels of transcription were achieved with polyglutamine tracts in the length range of 6 to 34 residues (16). Longer tracts showed a reduced ability to mediate transcriptional activation. Similar results were achieved for P_n runs, but with a steeper dropoff in activation activity for longer runs (16). Multiple significant charge clusters, in conjunction with one or more significantly long uncharged stretches of mostly polar residues, occur in many *Drosophila* developmental regulatory proteins (e.g., engrailed, paired, cut, deformed, and doublesex; see Table 2). Among the trinucleotide disease genes, only the androgen receptor is an established transcription factor.

Multiplet Counts. A measure of the homopeptide density of a protein sequence is the multiplet count: the number of distinct homooligopeptide runs of 2 or more residues (for details, see the legend to Table 1 and ref. 4). A complete absence of proteins with significantly high multiplet counts in *E. coli* and a pronounced overrepresentation of proteins with significant multiplet counts in *Drosophila* (about 10%) have been observed (4). The percentage of human proteins with significant multiplet counts is about 1.6% with similar percentages observed in mouse and yeast. We also assessed for the genes of Tables 1 and 2 the multiplet count on the codon level, indicating the number of distinct homocodon runs of length two or more. Significance is evaluated similarly to the amino acid case.

RESULTS

Distinctive Sequence Features of Proteins with Multiple Long Homopeptides. The presence of multiple long homopeptides correlates strongly with the presence of unusual sequence features including multiple charge clusters, significantly long uncharged segments, and high multiplet counts (Tables 1 and 2). It is useful to highlight and contrast the distinctive sequence features of the proteins with multiple long homopeptides versus general proteins.

D. melanogaster (68 with multiple long homopeptides). (*i*) Fifty-five (80%) are involved in developmental control, and 31 (45%) are involved in CNS development; (*ii*) 47 (69%) have multiple distinct charge clusters (overall *Drosophila*, 13.5%); (*iii*) only 12 (18%) are absent of any charge cluster (overall, 65%); (*iv*) 38 (56%) have significant amino acid multiplet counts (overall, 12%); 40 of 63 (five gene sequences were unavailable) show significant codon multiplet counts, including most of those with high counts of amino acid multiplets.

Human (36 with multiple long homopeptides). (*i*) Eight (22%) have at least two charge clusters (overall, 3–4%); (*ii*) 25 (69%) have at least one charge cluster (overall, 20–25%); (*iii*) 8 (22%) have significant amino acid multiplet counts (overall, <1.5%); and HOX-B3, HOX-B4, RDC-1, engrailed-1, N-OCT3, OTX-1, and OCT6 are likely developmental proteins (N-OCT3 participates in neural development).

Mouse (22 with multiple long homopeptides). (*i*) Eight (36%) have multiple charge clusters; (*ii*) 18 (82%) have at least one

charge cluster; (*iii*) 5 (23%) have significant multiplet count (including the mouse homolog of AR); apparently active in development are BRN1 (brain specific homeobox), HME1 (engrailed-1), HOX-AP, HOX-B4, HOX-D8, HOX-D9, HOX-D11, OTX1, FORMIN-4, N-OCT3, and OCT6.

Long acidic runs. There are several examples of Table 1 that possess hyperacidic runs characterized in ref. 11: CENP-B, nucleolin, sarcoplasmic reticulum, histidine-rich calcium-binding protein (HRC), chromogranin A, thyroliberin, and amyloid-like protein 2 (APLP2). Activity of several of these proteins has been detected in association with chronic systemic autoimmune diseases (17). Actually, more than half of all proteins eliciting autoimmune diseases show unusual charge configurations and especially long acidic runs (11, 17), suggesting that hyper-expansion of acidic homopeptides might lead to illness. The listing of Table 1 also includes the oncogenes adenomatous polyposis coli (APC) protein and HRX (chromosomal translocation associated with leukemia).

Codon content of long homopeptides. The glutamine runs of the triplet-repeat disease proteins are almost entirely encoded from the codon CAG. However, most of the other long homopeptides show considerable variation in codon usage. For example, P_8 in the androgen receptor is coded by $(CCG)_2(CCC)(CCT)(CCG)_2(CCT)(CCC)$. Interestingly, G_{24} is encoded from $(GGT)_3(GGG)(GGT)(GGC)_{18}$. The P_{11} peptide of the HD protein immediately downstream of Q_{23} is encoded from $(CCG)(CCA)(CCG)_7(CCT)_2$. The nervous system-specific octomer binding factor N-OCT3 has Q_{21} encoded by a genuine mixture of CAG and CAA codons, and the same is true for most examples of Tables 1 and 2. Apart from a few serine runs encoded exclusively from AGC—e.g., early growth response protein-1 (EGR1) and insulin receptor substrate-1 (IRS1)—most other serine runs of Table 2 use both serine codon types AGY and TCN. In general, the codon content of long homopeptides does not present a consistent pattern. Homopeptides encoded by iterations of a single codon are likely the result of strand slippage; substantial use of multiple codons suggests possible importance of the homopeptide for protein structure or function.

In addition to the multiple long homopeptides, some of the human triplet repeat disease genes contain unusual alternating patterns.

Histidine patterns. The alternating histidine pattern $(HX)_8$ = $H_2HQHSHIHSHLHLHQ$ in the DRPLA protein (Table 1) is notable in view of the plethora of histidine periodic patterns in *Drosophila* developmental proteins. These include the *Drosophila* female sterile homeotic protein (FSH; see Table 2), featuring $(HG)_{14}^{(1)}$ = $HH(HG)_{13}$ (the superscript indicates the number of errors—i.e., departures from the pattern); homeotic bicoid, $(HX)_{12}^{(1)}$ = $(HT)_2(HP)_2HS(HP)_2HS(HP)_2HHQHP$; segmentation protein paired, $(HX)_9^{(2)}$; homeotic spalt, $[(E/G)HH]_7$; homeobox protein NK-1, $(HX)_{13}^{(2)}$; male specific DsX showing three separated patterns, $(HHX)_9^{(2)}$, $(HX)_6^{(1)}$, and $(HX)_8^{(2)}$; ecdysone-inducible proteins E74A, $(HX)_{11}^{(1)}$, and E75B, $(HQ)_5$; segmental pattern protein deformed, $(HX)_{10}^{(2)}$ and $(HX)_6$; protein grainy-head (ELF1), $(HX)_8^{(1)}$; homeobox protein HM(10), $(HX)_{13}^{(2)}$. Drosophila extracellular structural proteins chorion 36 and chorion 38 show $(GH)_5^{(1)}$ and $(GH)_{11}^{(2)}$, respectively. There are also many developmental proteins with long histidine runs—e.g., polycomb H_{10} and H_8 and hunchback H_6. Histidine is a versatile amino acid that can adopt flexible roles in conformation, in catalytic actions, in metal coordination, and various enzymatic activities. Histidine patterns and runs provide opportunities for differential charge gradients, hydrogen bonding, and metal coordination. The only mouse proteins with long histidine runs found in the current SWISSPROT data base are the homeotic protein HXA1, featuring H_{11}; the homeobox proteins MOX-2A, containing H_{12}; and the regulatory nervous-system protein OTX1, containing H_{11}, H_5, and an alternating pattern, $(HX)_4$. The mouse homeobox protein HOX43 features $HA(HP)_4$. In the human N-OCT3 (nervous-system specific octomer binding) we observe the pattern $HHADH(HP)_2HSHPHQ$.

Extensive alternating charge patterns. The two alternating charge runs $(+,-)_{10}$ = $EK(ER)_4EK(ER)_2EKER$ and $(RE)_3ARERD$ of the disease gene atrophin 1 (DRPLA) are very unusual. A few other examples of similar alternating charge patterns (see refs. 7 and 11) include the large chain of the *Drosophila* developmental transcription factor 5, with $(KD)_2(KE)_4$; the *Drosophila* female sterile homeotic protein (FSH, see Table 2), with $(DR)_4(ER)_3$; human HRX protein (see Table 1), with $RERD(RE)_2KE$; the 101-kDa malaria antigen P-101 (*Plasmodium falciparum*), featuring $KE_4(KE)_5(EK)_2E(EK)_3E(EK)_2E(EK)_2(E(EK))_2E_3K_2$; the immune system–related RD-protein possessing the precise alternating $(+,-)_{24}$ sequence [+ is K or R, − is E or D]; and MHC-H2 42 kDa (mouse), which contains an unprecedented $(+,-)_{26}$ sequence. For possible function and structure implications of hyper-mixed charge runs, see ref. 11.

DISCUSSION

The human triplet repeat disease genes contain multiple long homopeptides, not only of glutamine. Examination of other human, mouse, and particularly *Drosophila* proteins carrying multiple long homopeptides revealed several associated functional and sequence properties. All but a few of the *Drosophila* examples are essential developmental proteins, and a majority of these regulate CNS and/or PNS development. Many of these proteins exhibit a highly anomalous charge distribution with multiple charge clusters interspersed with significantly long uncharged runs and are rich in multiplets and complex repetitive structures. Almost half of the human proteins identified are considered homeotic proteins. Some possible interpretations and implications of these observations are discussed below.

Structural Roles of Long Homopeptides. Why do many developmental regulatory proteins carry multiple long homopeptides, particularly of the small and/or polar amino acids Q, N, H, P, A, S, T, and G? A common proposal suggests that long homopeptides are tolerated nonessential insertions. In some cases, homopeptides of Q, P, S, and T may play a role in transcriptional activation (13–16). In another view, homopeptides may serve as spacer elements between functional domains. As previously indicated, many of these proteins contain two or more charge clusters putatively providing distinct functional domains that interact with DNA, RNA, or proteins and that contribute to intramolecular conformation. The segments linking these domains might often be uncharged polar regions centered on moderate-length homopeptides. In general, if the concentrated charge regions are the essential functional parts of these proteins, the uncharged stretches may play scaffold or hinge roles and thus provide flexibility to the three-dimensional conformation, fine-tuning domain orientation with respect to protein–protein interactions. In this context, excessive expansion of a homopeptide might lead to inappropriate domain organization, resulting in aberrant intramolecular conformation and protein–protein interaction. Moreover, following Perutz *et al.* (8), a drastically extended glutamine (or polar) run in one protein could form a hydrogen bond to a moderate polyglutamine (or polar) peptide and incapacitate the normal counterpart protein in heterozygotes. No complete NMR or crystal structures are yet available for any of the proteins of Tables 1 and 2.

Generation of Multiplets. What is the molecular mechanism that generates high multiplet counts? A high multiplet number indicates an unusual (compared to random) amount of local amino acid iterations. Strand slippage during DNA replication might account for creation of short multiplets and sister–chromatid exchange could be a factor contributing to expan-

sion of long homopeptides. The fact that most proteins of Tables 1 and 2 with significantly large numbers of amino acid multiplets also have significantly high codon multiplet counts argues in favor of strand slippage as a primary mechanism for creating multiplets. A possible explanation for the large proportion of proteins in Tables 1 and 2 that concomitantly have significantly high amino acid and codon multiplet counts is as follows. If long homopeptides usually arise from preexisting short homopeptide/homocodon runs, then those proteins with large numbers of short homopeptides/homocodons would be more likely to acquire one or more long homopeptides. The question still remains as to why the occurrence of large numbers of short and long homopeptides should be so much more frequent in developmental proteins than in other classes of proteins. Is there a difference between essential proteins (probably including most developmental proteins) and nonessential proteins? Are genes expressed at early developmental stages subject to higher levels of mutation? Since most of the long homopeptides are not derived from a single codon it appears that either strand slippage is not the exclusive mechanism for generating amino acid repeats or that many amino acid repeats have been conserved after their creation as homocodon repeats.

Mouse models of neurological disease. A number of mouse models of the HD syndrome have been engineered (18, 19). In all cases where the HD gene is disrupted or turned off, embryonic lethality results in homozygotes, but even in heterozygotes there are behavioral and morphological changes (18), suggesting that the normal protein is critical to development of the animal. In a parallel mouse model on the progression of SCA1 (ataxin 1), polyglutamine expansion leads to rapid neuronal degeneration and death (1). There are also clear asymmetries with respect to HD, SCA-1, and DRPLA diseases in maternal vs. paternal transmission apparently reflected in differences in spermatogenesis, oogenesis, and zygotic development (e.g., refs. 1 and 2).

Proteins with multiple long homopeptides are considerably more abundant in *Drosophila* (about 12% of all available proteins) compared with human and mouse (about 1.6%). Why? On the one hand, there is a well-known positive bias in present collections of *Drosophila* sequences toward developmental proteins. On the other hand, in human and mouse there may be a negative bias in that relatively few nervous system developmental/regulatory genes have been sequenced.

Why do we not observe excessively long homopeptide runs in *Drosophila* or mouse proteins (longest runs are Q_{25} and Q_{23}, respectively)? Expansions of triplet repeats presumably do occur at some rate in both organisms, but perhaps such mutations are selectively deleterious.

Proposed experimental studies. (*i*) Although long homopeptides have been deleted from some *Drosophila* proteins [e.g., elimination of the glutamine cluster in Sp1 (14) and in the shaker K^+ channel protein (20)], the consequences of significantly expanding these repeats have not been experimentally investigated. (*ii*) Are there diseases (in any tissue) associated with the examples of Table 1? Could expansions of the repeats in the developmental proteins—e.g., engrailed-1, RDC-1, HOX-B3, HOX-B4, OTX-1, and especially the nervous system-specific N-OCT3 (which features Q_{21})—cause disease? A review of the literature on the human proteins of Table 1 reveals at least two other cases of associated disease. The protein insulin receptor substrate 1 (IRS1) is considered a primary candidate for the site of the defect in insulin action seen in patients with non-insulin-dependent diabetes mellitus (21). Acrosin is the major proteinase present in the acrosome of mature spermatozoa: some abnormalities in acrosin produce infertility (22). (*iii*) In the trinucleotide repeat diseases, attention has been focused on expanded polyglutamine tracts. But these proteins contain other long homopeptides, whose expansion might be deleterious. For example, it seems relevant to investigate expansions of the other long homopeptides—e.g., G_{24} in AR, P_{11} in the HD gene, S_{10} in SCA1, etc.—in rodent models and to evaluate phenotypic consequences. Gene sequences from patients suffering infertility or other diseases related to AR could be examined for expansion of the homopeptides G_{24}, P_8, and A_5 as well as Q_{21}. (*iv*) Most nervous system developmental proteins function widely in many tissues; mutations of these genes apparently induce more severe disorders in the nervous system than in other tissues. Epidemiological surveys and experiments might be designed to search for repeat diseases of other tissues. (*v*) One or more of the charge clusters could be deleted from the murine homologs of SCA1, DRPLA, HD, or AR proteins and the resulting phenotype characterized.

We are happy to acknowledge valuable discussions concerning triplet repeat diseases with Dr. C. Laird of Seattle and on *Drosophila* development with Drs. M. Krasnow, R. Nusse, and M. Scott of Stanford. Dr. U. Franke was very helpful in reviewing the disease status of the human examples. We also thank Drs. V. Brendel and J. Mrazek for comments on the manuscript. This work was supported in part by National Institutes of Health Grants 2R01GM10452-31 and 5R01HG00335-07 and by National Science Foundation Grant DMS 9403553.

1. La Spada, A. R., Paulson, H. L. & Fischbeck, K. H. (1994) *Ann. Neurol.* **36,** 814–822.
2. Sutherland, G. R. & Richards, R. I. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3636–3641.
3. Hopkin, K. (1995) *J. NIH Res.* **1,** 45–48.
4. Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 2002–2006.
5. Karlin, S., Ost, F. & Blaisdell, B. E. (1989) in *Mathematical Methods for DNA Sequences*, ed. Waterman, M. (CRC, Boca Raton, FL), pp. 133–157.
6. Green, H. (1993) *Cell* **74,** 955–956.
7. Perutz, M. (1994) *Prot. Sci.* **3,** 1629–1637.
8. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5355–5358.
9. Scott, K., Blackburn, J. M., Butler, P. J. G. & Perutz, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6509–6513.
10. Karlin, S. (1990) in *Structure & Methods: DNA Protein Complexes & Proteins*, eds. Sarma, R. H. & Sarma, M. H. (Adenine Press, Guilderland, NY), Vol. 2, pp. 171–180.
11. Karlin, S. (1995) *Curr. Opin. Struct. Biol.* **20,** 360–371.
12. Mitchell, P. J. & Tjian, R. (1989) *Science* **245,** 371–378.
13. Brendel, V. & Karlin, S. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5698–5702.
14. Courey, A. J. & Tjian, R. (1988) *Cell* **55,** 887–898.
15. Su, W., Jackson, S., Tjian, R. & Echols, H. (1991) *Genes Dev.* **5,** 820–826.
16. Gerber, H. P., Seipel, K., Georgiev, O., Höfferer, M., Hug, M., Rusconi, S. & Schaffner, W. (1994) *Science* **263,** 808–811.
17. Brendel, V., Dohlman, J., Blaisdell, B. E. & Karlin, S. (1990) *Proc. Natl. Acad. Sci. USA* **88,** 1536–1540.
18. Nasir, J., Floresco, S. B., O'Kusky, J. R., Diewert, V. M., Richman, J. M., Zeisler, J., Borowski, A., Marth, J. D., Phillips, A. G. & Hayden, M. R. (1995) *Cell* **81,** 811–823.
19. Duyao, M. P., Auerbach, A. B., Ryan, A., Persichetti, F., Barnes, G. T., McNeil, S. M., Ge, P., Vonsattel, J.-P., Gusella, J. F., Joyner, A. L. & MacDonald, M. E. (1995) *Science* **269,** 407–410.
20. Hoshi, T. & Aldrich, R. (1991) *Neuron* **7,** 547–556.
21. Stoffel, M., Espinosa, R., III, Keller, S. R., Lienhard, G. E., Le Beau, M. M. & Bell, G. I. (1993) *Diabetologia* **36,** 335–337.
22. Florke-Gerloff, S., Topfer-Pertersen, E., Muller-Esteri, W., Schill, W. B. & Engel, W. (1983) *Hum. Genet.* **65,** 61–67.