Amino acid runs in eukaryotic proteomes and disease associations

Samuel Karlin*[†], Luciano Brocchieri*, Aviv Bergman[‡], Jan Mrázek*, and Andrew J. Gentles*

*Department of Mathematics, and [‡]Center for Computational Genetics and Biological Modeling, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, November 14, 2001

We present a comparative proteome analysis of the five complete eukaryotic genomes (human, *Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Arabidopsis thaliana*), focusing on individual and multiple amino acid runs, charge and hydrophobic runs. We found that human proteins with multiple long runs are often associated with diseases; these include long glutamine runs that induce neurological disorders, various cancers, categories of leukemias (mostly involving chromosomal translocations), and an abundance of Ca^{2+} and K^+ channel proteins. Many human proteins with multiple runs function in development and/or transcription regulation and are *Drosophila* homeotic homologs. A large number of these proteins are expressed in the nervous system. More than 80% of *Drosophila* proteins with multiple runs seem to function in transcription regulation. The most frequent amino acid runs in *Drosophila* sequences occur for glutamine, alanine, and serine, whereas human sequences highlight glutamate, proline, and leucine. The most frequent runs in yeast are of serine, glutamine, and acidic residues. Compared with the other eukaryotic proteomes, amino acid runs are significantly more abundant in the fly. This finding might be interpreted in terms of innate differences in DNA-replication processes, repair mechanisms, DNA-modification systems, and mutational biases. There are striking differences in amino acid runs for glutamine, asparagine, and leucine among the five proteomes.

Several human inherited neurodegenerative diseases are triplet-repeat diseases associated with proteins containing long runs of glutamine (long CAG codon iterations; for reviews, see refs. 1 and 2). Disease severity seems to be correlated with the extent of iterations of the CAG codon above a threshold (3). Strikingly, many of the triplet-repeat disease proteins contain multiple long runs of amino acids other than glutamine. Listing all runs of lengths of at least five residues (and using the standard one-letter amino acid code), the huntingtin protein contains Q_{23}, P_{11}, P_{10}, E_5, E_6; atrophin-1 (dentatorubral pallidoluysian atrophy, DRPLA) contains Q_{20}, S_7, S_{10}, P_6, H_5; the androgen-receptor protein (Kennedy's disease) contains Q_{26}, Q_6, Q_5, P_8, A_5, G_{24}; and the brain-voltage-dependent calcium channel protein CCAA (spinocerebellar ataxia 6) contains H_{10} and Q_{11}.

Consequences of hyperexpansion of DNA-triplet repeats might include altered rates of transcription or translation, mRNA instability, and aberrant DNA-hairpin structures (4, 5). Protein aggregation attributed to attachment of glutamine-rich proteins to unrelated molecules may lead to inappropriate multimerization or to formation of "polar zippers," in which a long stretch of glutamine residues link strands by hydrogen bonds (6–8).

The foregoing examples motivate our comparative analysis of eukaryotic proteomes focusing on proteins containing multiple amino acid runs. The complete genomes investigated are those of the Human Genome Project tentative draft,[§] *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm), *Saccharomyces cerevisiae* (yeast), and *Arabidopsis thaliana* (weed). Many eukaryotic proteins with multiple amino acid runs show other unusual protein sequence properties, including anomalous charge distributions, high counts of amino acid multiplets, extended alternating basic and acidic charge residues, and periodic histidine patterns. In each genome, the number of acidic runs exceeds the number of basic runs by a factor of three or four, and acidic runs tend to be longer than basic runs. In human sequences, proteins with multiple long amino acid runs are often associated with diseases (see below). Fly proteins with multiple runs have predominantly developmental functions and/or transcriptional regulatory capacities, with the majority of them active in central or peripheral nervous system function and development.

Runs of Individual Amino Acids in the Five Eukaryotic Genomes. For a "typical" protein of 400 residues and average composition, a run of an individual amino acid is statistically significant (at the 0.1% significance level) if it is five or more residues long (9). Table 1 displays the percentage of all proteins ≥200 residues long in the five eukaryotic genomes that have at least one amino acid run, along with the percentage of runs accounted for by each amino acid type. The percentage of proteins with at least one run ranges from 13% in worm and 15% in yeast to around 20% in human and weed and 27% in the fly. The residues A, S, and Q account for a significant proportion of the runs in each eukaryote; S runs range from 13.7% (human) to 33.4% (weed); A runs range from 4.7% (yeast) to 26.3% (fly), and Q runs range from 5.8% (weed) to 33.9% (fly). Amino acid runs emphasize small polar residues and the acidic residues E and D but avoid aliphatic, aromatic, arginine, and cysteine residues. Runs of the hydrophobic residues I, V, M and runs of the aromatics Y, F, and W are sparse in all five genomes; no human or fly protein has more than one run of each. However, leucine (L) runs occur in 19% of human sequences with a run, whereas in the other genomes, only 3.4–5.2% of proteins with runs have a run of L. In human sequences, ≈90% of L runs occur within 40 amino acids of the amino terminus of the protein, recognizable as part of a signal-peptide sequence. This proportion is much lower in the other eukaryotes, with only 0.4% of fly and weed proteins, 0.2% of worm, and 0.04% of yeast proteins having a similarly located L run. Enigmatically, runs of asparagine (N) are very infrequent in human proteins but are substantial in yeast and fly sequences. Specifically, only 0.06% of human proteins have an N run, compared with 1.0% among weed proteins and 2.7% among fly proteins. The dearth of N runs among human proteins applies to all mammalian species and contrasts sharply with N runs in invertebrate protein sequences (see ref. 10 and *Discussion*).

The data indicate that human proteins are more abundant in specific charged amino acid runs than fly proteins. R runs account for 3.2% of runs in human, 2.1% in fly, and 1.6–2.2% in the other genomes. The corresponding figures for K runs are 6.2% in human, 3.0% in fly and 3–7.2% in the others, with K runs being more frequent than R runs in each proteome. This finding may reflect the relative A + T nucleotide richness of the yeast, weed, and worm genomes. Genes in human and *Drosophila* favor

GENETICS

[†] To whom reprint requests should be addressed. E-mail: karlin@math.stanford.edu.

[§] Human proteins from SwissProt have essentially identical statistics (percentages) on runs as those presented for the genome.

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.

www.pnas.org/cgi/doi/10.1073/pnas.012608599

Table 1. Frequency of runs of amino acids among eukaryotic proteins

	Human	Fly	Worm	Yeast	Weed
≥200 aa	10,651	10,740	13,668	4,685	19,784
One run (at least)[*]					
Number	2,091	2,887	1,779	702	4,005
Percentage	19.6	26.9	13.0	15.0	20.2
Multiple runs					
Number	192	773	147	79	203
Percentage	1.9	7.2	1.1	1.7	1.0
Amino acid, %					
Asp	4.0[†]	6.3	8.1	14.1	8.4
Glu	19.8	7.3	10.7	14.5	15.8
Lys	6.2	3.0	6.9	7.1	7.2
Arg	3.2	2.1	2.2	1.6	2.1
His	2.3	6.1	2.6	2.3	2.7
Ser	13.7	23.7	21.5	28.5	33.4
Thr	2.0	10.8	13.5	4.4	3.5
Asn	0.3	9.9	3.5	13.4	5.0
Gln	6.0	33.9	11.9	19.5	5.8
Trp	—	—	—	—	—
Tyr	—	0.1	0.3	0.4	0.0
Phe	—	0.1	1.1	0.6	1.2
Leu	19.0	4.2	5.2	3.4	5.2
Met	0.0	—	—	—	0.4
Ile	0.2	0.2	0.7	0.6	0.1
Val	0.8	0.4	1.3	0.3	1.1
Ala	16.9	26.3	10.1	4.7	6.3
Gly	11.7	19.5	10.1	1.0	11.0
Cys	0.4	0.1	0.1	—	0.1
Pro	18.2	13.9	13.3	5.6	10.4

[*] Number and percentage of proteins (≥200 aa long) which have at least one run of any amino acid.

[†] Percentage of proteins with a run which have at least one Asp run.

strong amino acid types (G + C-rich codons: alanine, proline, and glycine), whereas yeast, weed, and worm genomes favor A + T rich codons. Glycine runs occur in only 1.0% of yeast proteins with runs, in 11.0% of weed, and in 19.5% of fly. Proline runs account for 18.2% of runs in human proteins and 14% in fly and are least common in the yeast genome (about 6%). The proportion of histidine runs among fly proteins is about 6.1%, exceeding by a factor of two the percent of H runs in the other genomes. Strikingly, the proportion of D runs in yeast proteins (14%) is higher than in the other genomes (4–8%); E runs compose about 20% of runs in human (the highest percentage of all amino acid types), 11% in worm proteins, 14% in yeast, 16% in weed, but only 7% in fly.

Multiple Amino Acid Runs. A protein has multiple amino acid runs if it has one or more runs, each ≥5 residues, with aggregate length ≥20 amino acids. With this definition, a random sequence of 1,000 residues has only a 0.1% chance of containing multiple long runs (9). As shown in Table 1, the proportion of fly proteins with multiple runs (7.2%) is dramatically higher than for human (1.9%), yeast (1.7%), worm (1.1%), or weed (1.0%). Multiple runs differ most for A, G, H, S, and Q. At least one A run is found in 47% of fly sequences with multiple runs, and S runs occur among 37.4% of these proteins. Of human proteins, 43.9% have multiple runs of P, 35.4% have a run of A, 33.3% of G, 28.8% of S, 26.8% of E, 22.2% of Q, and 10.6% of H (data not shown). All are lower than the corresponding assessments in the fly genome.

Except for A and L, hydrophobic runs almost never occur in proteins with multiple runs. Human and fly sequences containing multiple runs are comparable for A (about 40%), G (30–33%), D (6–8%), H (9–14%), S (28–38%), and R (3–4%) but differ significantly with respect to E (human 27%, fly 8%), K (human 7%, fly 3%), L (human 7%, fly 1%), P (human 44%, fly 23%), T (human 3%, fly 16%), N (human 1.0%, fly 22%), and Q (human 22%, fly 70%). Intriguingly, of the five proteomes, human has the lowest percentage of Q runs in proteins with multiple runs.

Table 2. Frequency and length distribution of charge, noncharged polar, and hydrophobic runs (≥6 long) in eukaryotic proteins

		Human	Fly	Worm	Yeast	Weed
Composition, %						
	−	11.9	11.5	11.8	12.4	12.3
	+	11.2	10.9	11.5	11.5	11.8
	p	40.2	41.5	38.7	40.2	38.3
	h	36.7	36.1	38.0	35.8	37.7
Proteins with runs, %						
	−	7.0	6.3	4.4	7.2	6.4
	+	2.4	1.9	1.2	1.5	2.1
	p	81.9	72.6	64.4	69.3	68.7
	h	64.7	57.6	60.6	43.2	54.4
Longest runs						
	−	38	33	31	56	41
	+	14	12	10	14	23
	p	206	192	65	104	123
	h	23	23	27	19	27

Runs of amino acids in eukaryotic proteins from complete genomes classified according to their charge properties by the four-letter Lehninger alphabet: negative (−), DE; positive (+), KR; polar noncharged (p), GHNPQSTY; hydrophobic (h), IVLMFACW.

Hydrophobic and Charge Runs. A useful concept applicable to all sequence statistics is the grouping of letters in one alphabet to form natural new alphabets. In this context, amino acids can be classified according to structural, chemical, charge, hydrophobicity, physical and/or kinetic properties, and associations with secondary structure. For example, the Lehninger functional alphabet is based on four amino acid categories: acidic (−), represented by the amino acids D or E; basic (+), represented by K or R; polar uncharged, (p) = (GHNPQSTY); and hydrophobic (h) = (IVLMFACW). This reduced alphabet (+, −, p, and h) requires a longer run ($n \geq 6$ instead of $n \geq 5$) to achieve the same significance level. In terms of this alphabet, Table 2 reports numbers, percentages, and lengths of charged and hydrophobic runs in the five complete eukaryotic genomes; the distributions are rather similar.

The percentage of acidic runs among the five complete genomes is of the same order, ranging from 5.9 to 7.2% (except for worm, 4.4%). The percentages of basic runs range from 1.2 to 2.2%. The worm has no basic runs exceeding 10 residues in length. The disparity between basic and acidic runs in protein sequences is pronounced, the latter being far more numerous and longer. The longest uninterrupted acidic run is 38 residues in human, 33 residues in fly, 31 residues in worm, 56 residues in yeast and 41 in weed. In contrast, the longest uninterrupted basic runs are of length 14 in human, 12 in fly, 14 in yeast, and 10 in worm. For each eukaryotic genome, the percent of acidic runs exceeds the percent of basic runs by a factor of 3 or 4. By contrast, bacterial genomes have roughly equal proportions of acidic and basic runs (data not shown).

For most extended charge runs, there is considerable variation in codon usage, which argues for an essential function for these charge runs. For example, in *Drosophila*, the codon counts for the positive run R_9 in *sevenless* are $(CGC)_4$, $(CGG)_1$, $(AGA)_2$,

Table 3. Multiple runs in human triplet-repeat (polyglutamine) disease proteins, transport channel proteins, and cancer-related proteins

Protein (Genbank accession)	Size, aa	Chromosome	Amino acid runs
Triplet repeat proteins			
Androgen receptor (NP_000035)	919	X	Q21 Q6 Q5 P8 A5 G24
Achaete-scute complex homolog-like 1 (NP_004307)	238	12	A13 Q14
BAI1-associated protein 1 (NP_004733)	1256	3	Q20 G5
Atrophin-1 (NP_001931)	1184	12	S7 S10 P6 H5 Q14
Huntingtin (NP_002102)	3144	4	Q23 P11 P10 E5 E6
Meningioma 1 (NP_002421)	1319	22	Q5 Q5 P5 Q28 P5 G5 G7 G5
numb-like protein (NP_004747)	609	19	Q20
Ataxin-1 (NP_000323)	816	6	Q12 Q15
Ataxin-2 (NP_002964)	1312	12	Q22
Ataxin-3 (P54252*)	360	14	Q22
Ataxin-6 (NP_075461)	2505	19	H10 Q11
Ataxin-7 (NP_000324)	892	3	A5 A6 Q10 P5 S5 S6 S12
Trinucleotide repeat containing 11 (NP_005111)	2212	X	Q26 Q6 Q26 Q7 Q5
Trinucleotide repeat containing 4 (NP_009116)	358	1	P5 Q15
Channel proteins			
ATP-binding cassette, subfamily F, member 1 (NP_001081)	807	6	Q10 E7 L5
CACNAID, voltage-dependent Ca channel α-1D subunit (NP_000711)	2161	3	M7 L5 E8
CACNAIF, voltage-dependent Ca channel, α-1F subunit (NP_005174)	1966	X	L5 E17 E6
HCN2 potassium channel (NP_001185)	889	19	P7 P7 P7
HRC Histidine-rich calcium-binding protein precursor (NP_002143)	699	19	E12 D16 E7 H5 E6 E7 E8
K^+ voltage-gated channel, shaker-related subfamily, member 4 (NP_002224)	653	11	A6 R5 E10
KCNMA1, large conductance Ca-activated K^+ channel (NP_002238)	1154	10	S22 D5
KCNN3, intermediate/small conductance K^+ channel (NP_002240)	731	1	Q12 Q14 Q5
Nucleoporin 153kD (NP_005115)	1475	6	G6 G5 S5 S5 S6
Ryanodine receptor 3 (NP_001027)	4870	15	E8 E5 E7 E5 E5
Solute carrier family 12 (NP_001037)	1212	5	A15 G5
Solute carrier family 24 (NP_004718)	1099	15	L5 E9 E8 E12
Cancer-related			
Adenomatosis polyposis coli (NP_000029)	2843	5	S5 P5 A5 S6
Achaete-scute complex homolog-like 1 (NP_004307)	238	12	A13 Q14
Brain-specific angiogenesis inhibitor 1 precursor (NP_001693)	1584	8	L8 P12
Breast carcinoma-associated antigen BCAA, isoform 2 (NP_057458)	1225	1	S6 E7 E5 S5 S6
BIRC6 baculoviral IAP repeat-containing 6 (NP_057336)	4829	2	A7 A7 L5 E5 S5
CDC2L1 cell division cycle 2-like protein 1 (NP_001778)	795	1	E13 E13
CHGA chromogranin A (NP_001266)	457	14	E8 E9 E5
D10S170 DNA segment, single copy, probe pH4 (NP_005427)	585	10	G13 P9
Glioma tumor suppressor candidate region gene 1 (NP_056526)	1509	19	G6 P6 S8 P5 P7
MAF v-maf musculoaponeurotic fibrosarcoma oncogene homolog (NP_005351)	403	16	A5 H6 G14
MAZ myc-associated zinc finger protein (NP_002374)	497	16	A13 P7 A5 G5 A9
Matrix metalloproteinase 24 (NP_006681)	645	20	P8 L6 A6
Nuclear receptor coactivator 3 (NP_006525)	1412	20	Q5 Q26
Cutaneous T-cell lymphoma-associated tumor antigen SE20-4 (NP_071400)	693		P9 P5 R9

*SwissProt accession number.

and $(AGG)_2$. Large variation at codon site three also is observed for long acidic runs in the *cut* protein (fly), Rad6p (yeast ubiquitin-protein ligase), Cenp-B (human major centromere autoantigen B), and others. Variable codon usages suggest that the longer runs are likely not generated entirely by strand slippage.

We found acidic runs exceeding 10 residues in 134 sequences of human, 86 of fly, 75 of worm, 63 of yeast, and in 149 of the mustard weed plant (data not shown). The number of proteins with basic runs exceeding 10 residues are far fewer: 2 in human, 2 in fly, 0 in worm, 1 in yeast, and 8 in weed. Paradoxically, on average, proteins show anionic frequencies in $\approx$11.5–12.0% and cationic frequencies in $\approx$11.0–11.5%, yet the numbers of proteins with long (at least six residues in length) acidic runs well exceed the numbers of long basic runs. The longest hydrophobic residue run in the eukaryotic genomes under study is in the range 20–27 amino acids, whereas noncharged polar runs can extend beyond 60 amino acids in length (Table 2). Hydrophobic long runs appear frequently as helical transmembrane segments, but these are generally confined to 17–25 amino acids in length.

Multiple Amino Acid Runs and Disease Associations. There are 192 human protein sequences (of the 10,651 $\geq$ 200 amino acids long) that have multiple amino acid runs (see Table 4, which is published as supporting information on the PNAS website, www.pnas.org). More than 40% of these proteins are associated with diseases [as identified in Online Mendelian Inheritance in Man (OMIM), which can be found at http://www.ncbi.nlm.nih.gov/Omim/], including: triplet-repeat proteins with long glutamine runs that underlie certain neurodegenerative disorders (Table 3); 14 cancer-related proteins [e.g., adenomatosis polyposis coli, breast carcinoma-associated antigen (BCAA), and matrix metalloproteinase 24 (MMP24)]; 10 leukemia-related proteins often resulting from chromosomal translocations (e.g., anaplastic lymphoma kinase K_i-1, myeloid/lymphoid mixed-lineage leukemia 2, meningioma 1); 14 channel proteins, mainly voltage-gated Ca^{2+} and K^+ channel proteins (Table 3; see also ref. 11); 6 proteases including sperm trypsin-like acrosin, calpain 4, and some metalloproteinases (see also ref. 12); 7 kinases; and a variety of disease syndrome-related proteins (e.g., Wiskott-Aldrich syndrome, cat-eye syndrome, and cleidocranial dysplasia). A key aspect of 82 of the 192 human protein sequences is their role in transcription, translation, and development regulation. Many of these proteins are homeotic homologs of *Drosophila* developmental sequences and transcription factors, including *forkhead, frizzled, engrailed, distal-less, timeless, diaphanous 1–3, pumilio, trithorax, runt*-related and *caudal*. Other examples include isoforms of E2F transcription factor, neuregulin, several translation–initiation factors, numerous homeobox genes, global transcription factors such as GATA-binding proteins 4 and 6, POU domain class proteins 3 and 4, and various nuclear-receptor coactivators and corepressors. There are two major immune-system proteins among the 192 with multiple runs: immunoglobin superfamily member 4 (IGSF4) and HLA-B associated transcript 2 (D6S51E).

In marked contrast, no metabolic enzymes (e.g., glycolysis, tricarboxylic acid cycle, pentose phosphate pathway), structural proteins (e.g., actin, myosin, and troponin 1), or housekeeping proteins contain multiple runs. However, several structural–regulatory proteins do have multiple runs, including ankyrin 3, nucleolin, SMARCA2 (actin dependent regulator of chromatin), and synapsin II, which coats synaptic vesicles and may function in the regulation of neurotransmitter release. Major chaperone and degradation proteins, including heat shock protein 70 (Hsp70), Tcp1, and subunits of the proteasome also lack multiple runs. Hsp70, which modulates protein folding and some transport and secretion activities, can counteract the toxic effects of aggregations caused by extended glutamine iterations (13–15). The DNA-repair protein repertoire (e.g., Rad51 and -54, Dmc1, uracil glycosylase, ERCC) does not carry multiple runs. Calcium and potassium channel proteins stand out with multiple runs, but transporters of Cu^{2+}, Fe^{2+}, Mn^{2+}, and Zn^{2+} do not have multiple runs. The hyperpolarization-activated cyclic nucleotide-gated potassium channel 2 (HCN2) is expressed in the heart ventricle and atrium and functions in cardiac pacemaking (16); KCNA4 (CIK4) shaker-related channel protein mediates the voltage-dependent potassium ion permeability of excitable membranes; KCNMA1 (*slowpoke Drosophila* homolog) is a calcium-activated potassium-channel gene exhibiting many alternative splicings; the small conductance calcium-activated potassium channel KCNN3 is voltage-independent and lacks the transmembrane S4 motif $(+,\varphi,\varphi)_{4-6}$ of positive-charge residues separated by two hydrophobic residues. Among the Ca^{2+} channel proteins with multiple runs, CACNA1F is involved with X-linked congenital stationary night blindness and, in addition, is a target for drugs alleviating hypertension. The ataxin-6 calcium channel (SCA6), which also contains extended CAG (polyglutamine) repeats, has been linked to familial hemiplegic migraine.

Strikingly, prokaryote protein analogs/homologs in the human genome do not have multiple amino acid runs. On this basis, multiple runs in human proteins may be a recent evolutionary outcome, concomitant with complex brain development. More than 80% of *Drosophila* proteins with multiple runs seem to function in developmental and transcription regulation. It is plausible that the corresponding human proteins are developmental proteins that function in embryogenesis and/or neurogenesis and become relatively quiescent during normal life. In a few anomalous cases, some maladies could become exacerbated at adult life stages, as with the late-onset triplet-repeat diseases. Screening mouse for proteins with multiple runs reveals substantial conservation with the human proteins. Specifically, we identified 56 SwissProt mouse entries with multiple runs, of which 52 have a known human homolog. In 43 cases (83%), the human homolog also has multiple runs; 5 (10%) of the mouse proteins have a homolog that has amino acid runs but does not meet the criterion for multiple runs; and 4 (7%) have human homologs that have one or no runs (these are DDX9 ATP-dependent RNA helicase A, DUS8 neuronal tyrosine threonine phosphatase 1, HOXD9 homeobox protein D-9, and UBF1 nucleolar transcription factor 1). Prominent examples of mouse/human homologs that share multiple runs include the CREB-binding protein, *diaphanous* 1 homolog, *even-skipped* homolog, GATA-binding proteins 4 and 6, anaplastic lymphoma kinase, MAZ myc-associated zinc finger, and the ZIC2 and ZIC3 proteins.

It is useful to highlight unusual protein sequence features accompanying many proteins with multiple runs. (*i*) *Charge clusters*. A charge cluster refers to a protein segment (typically 20–80 residues) with high specific-charge content relative to the charge composition of the whole protein (see ref. 9 for elaborations). The percentage of proteins with at least one significant charge cluster is about 19–23% in most eukaryotic species. In all current complete prokaryotic genomes, the percent of proteins with one or more charge clusters ranges from 6–10%. Proteins with multiple-charge clusters in eukaryotes are uncommon, about 2–4% and <1% in prokaryotes. In eukaryotes, charge clusters are associated with transcriptional activation, membrane receptor activity, and developmental regulation. By contrast, charge clusters are rare among the bulk of housekeeping and metabolic proteins, cytoplasmic enzymes, and among prokaryotic proteins. Primary families of proteins with multiple charge clusters include essential developmental proteins, voltage-gated Ca^{2+} and K^+ ion channel complexes and transporters, and transactivator proteins of large eukaryotic DNA viruses (17). (*ii*) *Alternating charge runs*. A typical example is the alternating charge run $(-,+)_{10}$ = $EK(ER)_4EK(ER)_2EKER$ observed in the human triplet-repeat disease gene atrophin 1 (DRPLA). The human immune system-related RD-protein possesses an alternating $(+,-)_{24}$ sequence, and the 42-kDa mouse histocompatibility complex MHC-H2 contains an unparalleled $(+,-)_{26}$ sequence. The fly female sterile homeotic protein FSH contains $(DR)_4(ER)_3$. (*iii*) *Histidine patterns*. The period two-histidine pattern $(HX)_8$ = $H_2HQHSHIHSHLHLHQ$ in the DRPLA protein is distinctive. In the human N-OCT3 (nervous-system specific octomer binding) protein, we observe the pattern $HHADH(HP)_2HSHPHQ$. Many histidine periodic patterns occur in *Drosophila* developmental proteins. Histidine is a versatile amino acid that can adopt flexible roles in conformation, in catalytic actions, and in various enzymatic activities. Histidine patterns and runs also provide opportunities for differential charge gradients, hydrogen-bonding networks, and metal coordination. (*iv*) *Multiplets*. There are several levels and forms of repetitive structures (18). Multiplets comprise all homodipeptides XX, homotripeptides XXX, etc., where X denotes any specific amino acid. The count of multiplets provides a measure of the homopeptide density of the protein sequence.

Discussion

Our proteomic analysis comparing the five complete eukaryotic (human, fly, worm, yeast, weed) genomes focuses on proteins containing specific individual amino acid, charge, or hydrophobic runs. Multiple long amino acid runs in human proteins often are associated with diseases; e.g., triplet-repeat diseases, diseases induced by long acidic charge runs (as with lupus antigenic afflictions), CENP-B or nucleolin, and chromosomal translocation proteins, several of which cause leukemia. The fly proteome collection with multiple runs emphasizes proteins involved in developmental activities where glutamine runs are especially profuse. Serine runs are frequent in all genomes at a high level in the range 14–33% of the proteome.

Amino acid runs are significantly more abundant in many respects in the fly proteome compared with the other complete proteomes. (*i*) About 27% of fly proteins contain at least one amino acid run, whereas at most, 20% of protein sequences in the other genomes have long runs (Table 1). (*ii*) For proteins with multiple runs (runs of aggregate length ≥20 residues), fly sequences again stand out, cumulating about 7% of the proteome compared with ≤2.2% for the other genomes. (*iii*) The fly has 81 protein sequences, each with at least 10 runs, whereas worm has only 9 such proteins, human has 7, weed has 8, and yeast has 2. (*iv*) The most common amino acid run among fly sequences consists of Q residues (33.9%), but only 6% of runs in human proteins involve Q (the lowest proportion of the five genomes). Yet the human coding triplet-repeat diseases feature excessively long Q runs (Table 3). The percentage of proteins with runs in fly and human genomes differs significantly for the amino acids Q (fly 33.9%, human 6.0%), N (9.9%, 0.3%), and S (23.7%, 13.7%). What could account for the proliferation of runs in fly sequences compared with human sequences? The fly genome contains (percentage-wise) more protein runs than the other genomes (Table 1). This fact cannot be attributed to a protein sampling bias, because we are dealing with complete genomes. Is this abundance of runs true for all *Drosophila* species (e.g., *D. virilis, pseudoobscura*) and perhaps other insect populations? Is it possible that the current *Drosophila melanogaster* laboratory and/or domesticated strain sequences are significantly inbred? Early protein studies suggested that *Drosophila* exhibits high polymorphism (19). Is there a tie-in between polymorphism and run counts?

Another contingency is that there are innate differences in replication, information processing mechanisms, repair systems, DNA modification operations, and mutational biases between human (mammals in general) and fly, as shown in the following examples. (*i*) There is a lack of methylation activity in the fly and most invertebrates. (*ii*) *Drosophila* (and apparently all protostomes), unlike mouse, lacks embryonic transcription-coupled repair capacity (20). *Drosophila* also lacks mammalian type uracil DNA glycosylase (21). Does this mean that *Drosophila* DNA-replication processes are less accurate than those in mammalian eukaryotes? (*iii*) *Drosophila* is very different from mouse (and apparently also human) in replication processes. First, *Drosophila* DNA replicates frenetically in the first hours after fertilization, with replication bubbles distributed about every 10 kb (22). By 12 h, effective origins are spread to around 40 kb. In mice, the rate of replication seems to be uniform throughout developmental and adult stages. Moreover, cell divisions involve DNA stacking on itself and loopouts that need to be decondensed to undergo segregation. The observed narrow limits to intragenomic heterogeneity putatively correlate with conserved features of DNA structure. Second, *Drosophila* zygotic nuclei divide into 128 copies before the initial cell division (syncitium). It is possible there is DNA exchange (recombination) among these nuclei that generates extra amino acid runs. (*iv*) A difference in mutational patterns is manifest between human and fly genomes. In fact, complex sequence deletions in the fly are more frequent and extensive, especially evidenced by microsatellite changes (23, 24).

There seems to be some influence of the genome G + C content and dinucleotide relative abundances on occurrence of runs. For example, the yeast genome with only 38% G + C content is very low in the strong amino acids A, G, and P. The worm, yeast, and weed genomes are G + C poor (<40%), even in regions rich with genes, whereas human and fly genes favor enriched G + C content around gene-rich regions. The strong-codon amino acid group (A, G, P) is translated from codon types SSN (S is the strong nucleotide C or G, N is any nucleotide) and the weak-codon amino acid group, WWN (W is A or T) emphasize the amino acids (F, I, M, K, N, Y). The G + C-rich human and fly proteins favor use of strong amino acids, compared with the A + T-rich yeast, worm, and weed sequences.

There is obviously strong selection against asparagine runs among mammalian sequences. Structurally, N runs avoid the secondary structures of α-helices and β-strands and tend to establish disordered loops (25). We further speculate that runs of N may be prone to excessive glycosylation in mammals and seem to be selected against among mammalian protein sequences. For unknown reasons, the very A + T-rich malaria parasite *Plasmodium falciparum* is replete with N runs (data not shown). We conjecture that this fact may in some way assist *Plasmodium* in evading the host immune system response. The dearth of N runs in human protein sequences cannot be attributed to differences in amino acid usage. In fact, the median asparagine usage frequency is quite similar across the five genomes: human, 4.3%; fly, 4.5%; worm, 3.7%; yeast, 3.7%; weed, 3.2%. Also, the full quantile usage distributions for asparagine are rather similar across eukaryotes.

Nonspecific hydrophobic runs commonly identify transmembrane segments of receptor or extracellular proteins, and L runs (4–7 residues) stand out in signal peptide sequences near the amino terminus of membrane and extracellular proteins. Unlike other aliphatic and aromatic residues in the human genome, L runs are strikingly high (19.0%). The prominence of L among protein sequences certainly reflects its important role in hydrophobic cores, in transmembrane segments, and in signal peptides, and its prevalence and stability in secondary and tertiary structures. The relatively high alanine frequency in proteins also may reflect on α-helix stability and flexible hydrophobic properties. Interestingly, in human nuclear proteins, serine runs predominate.

Charge Compositional Biases. For all eukaryotes, the median net charge of proteins is slightly negative (around −0.5%). The aggregate positive charge (K + R) per protein is generally constant over species, at 11.5–12.0%. However, the median K and R frequencies per protein vary individually across the different species. For example, in human, R is under-represented, presumably because of CpG suppression, whereas in *E. coli*, K is under-represented. Why are E runs more frequent than D runs? From a structural viewpoint, D is recognized as an α-helix breaker, whereas E is favorable to α-helix formation. Moreover, the side chain of E involves two methylene groups as against a single methylene group in D, thus providing greater conformational flexibility. D and E are encoded by similar codon forms (GAR and GAY, respectively), but the juxtaposition of purine-pyrimidine at codon sites 2 and 3 may be sterically unfavorable compared with a purine-purine arrangement (26).

Residues on the surface of proteins presumably need to be highly selective to be able to interact with appropriate structures or to avoid interacting with other structures. From this viewpoint, a general net negative charge or a negative charge run may more easily avoid (for example, mediated by electrostatic repulsion) undesirable interactions with DNA, RNA, membrane surfaces, and other proteins. The extracellular environment for metazoans is mildly alkaline, with pH ~ 7.2–7.4 (27), whereas the intracellular pH is variable, ranging from 5.0 to 7.2, depending on tissue type and subcellular localizations (28, 29). One might speculate that enzyme activity is "optimal" at a pH similar to the pH of the host cells, which in mammalian organisms tend to be slightly acidic. Moreover, protein negative charge runs can contribute in modulating secretion and intracellular transport, in inducing transcriptional activation, and generally, in mediating rapid and potent interactions of protein assemblages. Mixed charge runs often contribute to protein–protein interaction at the interface of quaternary formations (30).

There is strong correlation between protein sequences with multiple runs and highly anomalous charge distribution. In particular, many of these proteins contain two or more charge clusters that putatively function through domain interactions with DNA, RNA, or other proteins and facilitate intramolecular conformation. Segments linking the domains are often uncharged polar regions involving moderate length polar homopeptides. The charge regions might contribute functional properties, whereas uncharged stretches have scaffold or hinge roles, providing flexibility to the three dimensional conformation, or help in fine-tuning domain organization. However, excessively lengthened homopeptides can induce incorrect domain interactions, producing aberrant conformation and inappropriate protein–protein interactions. Extended polyQ tracts may corrupt protein conformation, causing mis-folding of the protein. Also, long glutamine runs or glutamine-rich domains can recruit proteins into polyQ aggregates with concomitant instabilities (4, 31). Long coding CAG triplets (polyglutamine) are unstable and produce insoluble aggregates that seem to be toxic (1–4). There are dynamic mutations leading to disease based on noncoding triplet nucleotide repeats; e.g., fragile X, myotonic dystrophy, and Friedrich ataxia. It may be the repetitive nature of the nucleotides rather than the ability to code multiple amino acid runs that is critical to the disease mechanism (however, see refs. 6–8 and 32).

What are the potential benefits and the problems of multiple runs in proteins? Extended runs can provide substrates for caspase cleavage, yielding tangles, plaques, dead neurons, and a signal for apoptosis. Runs may provide binding sites for protein–protein interactions. Also, extended runs may trigger inflammatory brain responses, oxidative damage, and protein aggregations that clog the proteasome (15).

Why do the polyglutamine disease genes all encode multiple amino acid runs in addition to the pathogenic repeat? The reasons are not known. The contemporaneous presence of other unusual protein features such as charge clusters, alternating charge runs, periodic histidine patterns, and high numbers of multiplets is fascinating. Multiple runs likely fulfil a role in protein structure, protein–protein interactions, and transcription regulation. Extensive runs prominently feature glutamine which can produce aggregation with consequent toxicity. Apart from long Q runs, long E runs in human sequences do occur and could engender structural distortions, but perhaps contribute positively to function. Two questions spring to mind. First, are multiple runs highly polymorphic, as is the case with the polyglutamine repeat in many triplet repeat diseases? Second, are multiple runs predictive of disease associations? Both questions may be addressed experimentally, by surveying the population for polymorphism at repeat loci, and by testing whether multiple repeats are expanded in disease phenotypes. Further, novel mouse disease models can be made by expanding the repeats in candidate proteins.

We are grateful to Allan Campbell, Dmitri Petrov, and Lubert Stryer (Stanford) and Dan Geschwind (Univ. of California, Los Angeles) for helpful discussions and suggestions regarding this manuscript. This research was supported in part by National Institutes of Health Grants 5R01GM10452-36 and 5R01HG00335-14.

1. Zoghbi, H. Y. & Orr, H. T. (1997) *FASEB J.* **11,** 864–864.
2. Cummings, C. J. & Zoghbi, H. Y. (2000) *Annu. Rev. Genomics Hum. Genet.* **1,** 281–328.
3. Sutherland, G. R. & Richards, R. I. (1995) *Curr. Opin. Genet. Dev.* **5,** 323–327.
4. Sinden, R. R. (2001) *Nature (London)* **411,** 757–758.
5. Kovtun, I. V., Goellner, G. & McMurray, C. T. (2001) *Biochem. Cell. Biol.* **79,** 325–336.
6. Green, H. (1993) *Cell* **74,** 955–956.
7. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5355–5358.
8. Perutz, M. F. (1999) *Trends Biochem. Sci.* **24,** 58–63.
9. Karlin, S. (1995) *Curr. Opin. Struct. Biol.* **5,** 360–371.
10. Kreil, D. P. & Kreil, G. (2000) *Trends Biochem. Sci.* **25,** 270–271.
11. Rolfs, A. & Hediger, M. A. (1999) *J. Physiol.* **518,** 1–12.
12. Yong, V. W., Power, C., Forsyth, P. & Edwards, D. R. (2001) *Nat. Rev. Neurosci.* **2,** 502–511.
13. Warrick, J. M., Chan, H. Y. E., Gray-Board, G. L., Chai, Y., Paulson, H. & Bonini, N. M. (1999) *Nat. Genet.* **23,** 425–428.
14. Chai, Y., Koppenhafer, S. L., Bonini, N. M. & Paulson, H. L. (1999) *J. Neurosci.* **19,** 10338–10347.
15. Bence, N. F., Sampat, R. M. & Kopito, R. R. (2001) *Science* **292,** 1552–1555.
16. Ludwig, A., Zong, X., Stieber, K., Hullin, R., Hofmann, F. & Biel, M. (1999) *EMBO J.* **18,** 2323–2329.
17. Karlin, S., Blaisdell, B. E. & Brendel, V. (1990) *Methods Enzymol.* **183,** 382–402.
18. Karlin, S., Blaisdell, B. E. & Bucher, P. (1992) *Protein Eng.* **5,** 729–738.
19. Nevo, E., Beiles, A. & Ben-Shlomo, R. (1984) *Lect. Notes Biomath.* **53,** 13–213.
20. deCock, J. G., Klink, E. C., Ferro, W., Lohman, P. H. & Eeken, J. C. (1992) *Mutat. Res.* **293,** 11–20.
21. Aravind, L. & Koonin, E. V. (2000) *Genome Biol.* **1,** RESEARCH0007.
22. Blumenthal, A. B., Kriegstein, H. J. & Hogness, D. S. (1974) *Cold Spring Harbor Symp. Quant. Biol.* **38,** 205–223.
23. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. (1996) *Nature (London)* **384,** 364–369.
24. Petrov, D. A. & Hartl, D. L. (1998) *Mol. Biol. Evol.* **15,** 293–302.
25. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648–1652.
26. Hunter, C. A. (1993) *J. Mol. Biol.* **230,** 1025–1054.
27. Roos, A. & Boron, W. F. (1981) *Physiol. Rev.* **62,** 296–434.
28. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994) in *Molecular Biology of the Cell* (Garland, New York), 3rd Ed.
29. Stryer, L. (1995) *Biochemistry* (Freeman, New York), 4th Ed.
30. Zhu, Z.-Y. & Karlin, S. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 8350–8355.
31. Michelitsch, M. D. & Weissman, J. S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11910–11911.
32. Karlin, S. & Burge, C. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1560–1565.