Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

WWWWhy does nature stutter? A survey of strands of repeated amino acids

Edgar F. Meyer* and W. John Tollett Jr†

Biographics Laboratory, Department of Biochemistry and Biophysics, A&M University, College Station, TX 77843-2128, USA

† Current address: A&M Consolidated High School, College Station, TX 77840, USA.

Correspondence e-mail: e-meyer@tamu.edu

Received 29 August 2000
Accepted 20 November 2000

Human stuttering is a simple example of the repetition of sounds or symbols, sometimes associated with single letters, and may be used to illustrate the amazing repetition of amino acids (symbolized by a letter, *e.g.* W) in proteins. A survey of available databases with highly improbable strings of single amino acids is tabulated. This paper concludes with a challenge to the crystallographic community to probe the structural origins of the structure–function relationship in this neglected area. When nature stutters, we should pay attention.

1. Introduction

As macromolecular sequence information continues to become available, a striking relationship emerges from the exploding mass of data. While proteins are composed of varying combinations of the 20 natural (L-) amino acids, an improbable condition is reported whereby homo-repeats of poly amino acids are found in natural sequences, ranging from four (tryptophan) to 67 (threonine). The challenge is to relate sequence to structure to function using genetic, protein and structural databases.

For a homework problem, 410 undergraduate biochemistry students searched for repeated sequences of the amino acid related to the first letter of their family names (corresponding to the 20 amino acids) using the *Biology Workbench* 3.2 (Subramaniam, 1998; http://workbench.sdsc.edu/). With the *PATTERNMATCHDB* program, a repetitive string, *e.g.* WWWW, received 25 'hits' using the SDSC (composite gene + protein) database (935 147 entries) and 24 'hits' in the comprehensive protein database (SwissProt + EMBL + PIR, 634 784 entries). WWWW received 34 'hits' (all virus protein structures) in the Protein Data Bank (12 775 entries; no tetraWs were found as of August 2000).

If we assume that the occurrence probability of any specific sequence and locus can be estimated by the simple function: $(\text{letters})^{\text{places}}$ and that there are no correlations among these sequences, then the probability of occurrence of any homo-tetrapeptide at a specific site in an extended chain is approximately 1 in 20^4 = 1/160 000 *versus* a random sample of the 20 amino acids; roughly six in 10^6 events.[1]

[1] That 34 virus structures were detected suggests that this model may be overly simplistic and that cross-correlations may occur, but our purpose here is to report a finding and encourage others to explore its implications, be they probabilistic, statistical, genetic, functional or structural.

As gene, protein and structural databases were searched, who would have guessed that 67 consecutive threonines would be found in *Cryptosporidium parvum* (Barnes *et al.*, 1998)? The probability of 67 repeats in a random sequence at a specific site is ~1 in 20^{67} = $1/1.5 \times 10^{87}$ events; the difference in probabilities is exponentially significant. Even though this statistical approximation begs for a more rigorous treatment, it is amazing. WWWWhat is nature telling us?

Long consecutive strands of positively or negatively charged amino acids must carry electrostatic penalties, yet these too abound. In a nuclear transport protein (PDB code 1qbk), polyaspartate is augmented by two glutamates to create a startling exposed strand of 14 consecutive negatively charged residues. Intuitively, one could assume that uncharged amino acids would be more likely to occur repetitively, but polymethionine also has a relatively low occurrence (7). Because of pronounced peptide backbone angular constraints, proline was considered to be a 'helix breaker', but polyPro actually forms a left-handed helix (1jvr). In HIV-1 reverse transcriptase (1c9r; residues 315–326), an extended polyAla strand is parallel to an α-helix that is also rich in Ala. Conversely, a 12-Ala repeat forms a cluster of three α-helices at the tip of a tumor necrosis factor receptor (1czz). At this stage, it appears that while polyPro may be structurally conserved, polyAla is not. PolyCys is one of the few repeat sequences which is generally buried, forming a tight trimer knot in a spider toxin (1qdp), a triple S–S knot (1ag8), and a tight buried loop central to an amazing chain of seven S–S linkages in the ferric hydroxamate uptake receptor (1cw3, 1a4z).

These searches reveal a wide range of structures, populations and probabilities, summarized by abbreviated tables [tables also

have been posted on the web site (http://www.tamu.edu/struct/research/WWWW/searches.html; for help with *Chime* and *Isis* plug-ins, *cf*. http://www.mdli.com/cgi/dynamic/downloadsect.html?uid=$uid&key=$key&id=1); the related Chime links will make the structural results more readily accessible to a broader audience]. While some entries of gene sequences are deposited without comment and/or literature citation (Table 1), many protein sequence entries (*e.g.* PIR, SwissProt, EMBL) are cited (Table 2) and infer functional roles. Although smallest in size, the Protein Data Bank (Bernstein *et al.*, 1977; Meyer, 1997;

Table 1
GenBank results, 23 June 2000.

Amino acid	No. of repeats	Protein name	Residues	GenBank ID#
Alanine	20	*Drosophila melanogaster* additional sex combs cDNA sequence	129–148	GBINV:DMJ001164
	20	*D. melanogaster* genomic scaffold 142000013386047 section 24 of 52, complete sequence	129–148	GBINV:AE003814
	21	*D. melanogaster* Ovo-1028aa (ovo) mRNA, complete cds	497–517	GBINV:DMU11383
	21	*D. melanogaster* ovo gene required for female germ-line development (zinc-finger protein)	497–517	GBINV:DMOVO
	20	*Homo sapiens* homeodomain transcription factor NBPHOX (PMX2B) gene	241–260	GBPRI:AF117979
	20	*H. sapiens* mRNA for NBPhox, complete cds	241–260	GBPRI:D82344
	20	*Mus musculus* mRNA for PHOX2b protein	241–260	GBROD:MMPHOX2B
	20	*M. musculus* NBPhox gene, complete cds	241–260	GBROD:AB015672
	20	NBPhox	241–260	GBPRI:AB015671
Arginine	30	Human Fragile X mental retardation 1FMR-1 gene, 3′ end, clones BC72 and BC22	13–42	GBPRI:HUMFMR1
Asparagine	50	*Dictyostelium discoideum* protein tyrosine phosphatase (PTP3) mRNA, complete cds	138–187	GBINV:DDU38197
	46	*D. discoideum* HelE (*helE*) gene, partial cds	24–69	GBINV:AF019981
	49	*D. discoideum* mRNA for guanylyl cyclase	720–768	GBINV:DDI238883
	46	*D. discoideum* prespore-specific protein (*pspC*) gene, partial cds and unknown gene	266–311	GBINV:AF104350
	46	*Plasmodium falciparum* chromosome 2, section 53 of 73 of the complete sequence	50–95	GBINV:AE001416
	46	*P. falciparum* chromosome 2, section 55 of 73 of the complete sequence	777–822	GBINV:AE001418
Aspartic acid	41	*Arabidopsis thaliana* chromosome 1 BAC F11A17 sequence, complete sequence	285–325	GBPLN:F11A17
Cystine	23	Human super-cysteine-rich protein mRNA, partial cds	11–33	GBPRI:HSU63332
Glutamine	45	*D. iscoideum* adenylyl cyclase (*acrA*) gene, complete cds	1856–1900	GBINV:AF153362
	41	*Coturnix coturnix japonica qMEF2D* gene	362–402	GBVRT:CCJ002238
	40	*H. sapiens* CAGH44 mRNA, partial cds	152–191	GBPRI:HSU80741
	38	*H. sapiens* transcription factor IID mRNA, complete cds	58–95	GBPRI:HUMTFIIDA
	38	Human DNA sequence from clone RP1-191N21 on chromosome 6q27. Contains a 7 transmembrane receptor (rhodopsin family) (olfactory receptor like) pseudogene, the *PDCD2* gene for programmed cell death 2 (RP8 homolog), the *TBP* gene for TATA box binding protein, the gene for proteasome subunit HC5, ESTs, STSs and GSSs, complete sequence	58–95	GBPRI:HS191N21
	38	Human TATA-binding protein mRNA, complete cds	58–95	GBPRI:HUMTFIID
	37	*D. discoideum* hybrid histidine kinase DHKB (*dhkB*) gene, complete cds	1719–1755	GBINV:AF024654
	37	*D. melanogaster* genomic scaffold 142000013386054 section 30 of 35, complete sequence	573–609	GBINV:AE003446
	37	*M. musculus* mRNA for hyperpolarization-activated cation channel HAC2	739–755	GBROD:MMJ225123
	37	*M. musculus* brain cyclic nucleotide gated 1 (Bcng-1) mRNA, complete cds	739–775	GBROD:AF028737
	37	*Saccharomyces cerevisiae* chromosome II reading frame ORF YBR289w	232–268	GBPLN:SCYBR289W
	37	*S. cerevisiae* (s288c) *RIF1*, *DPB3*, *YmL27* and *SNF5* genes	232–268	GBPLN:SCDPB3
	37	*S. cerevisiae* SNF5 protein gene, complete cds	232–268	GBPLN:YSCSNF5
	35	*D. melanogaster* genomic scaffold 142000013386050 section 23 of 54, complete sequence	222–256	GBINV:AE003536
Glutamic acid	39	*A. thaliana* DNA chromosome 5, BAC clone F17C15	6–44	GBPLN:ATF17C15
Glycine	29	*A. thaliana* DNA chromosome 4, BAC clone F23E13	153–181	GBPLN:ATF23E13
	29	*A. thaliana* DNA chromosome 4, contiguous fragment No. 85	153–181	GBPLN:ATCHRIV85
	27	Human androgen receptor mRNA, complete cds	445–471	GBPRI:HUMARB
	27	Human androgen receptor mRNA, complete cds	261–287	GBPRI:L29496
	25	Human androgen receptor gene, partial cds	4–28	GBPRI:HSU16371
Histidine	17	*A. thaliana* chromosome III BAC T7M13 genomic sequence, complete sequence	519–535	GBPLN:ATAC011708
	16	*D. melanogaster* genomic scaffold 142000013386054 section 35 of 35, complete sequence	183–198	GBINV:AE003451
	15	*D. melanogaster* genomic scaffold 142000013386054 section 14 of 35, complete sequence	1073–1087	GBINV:AE003430
	15	*D. melanogaster* sequence EG0007: concatenation of cosmids 30D3:40623-16930, 65G3:42474-1	73–87	GBINV:DMSEG0007
Isoleucine	10	*Spodoptera exigua* nucleopolyhedrovirus complete genome	6–15	GBVRL:AF169823
	8	*Caenorhabditis elegans* cosmid C15C7	149–156	GBINV:CELC15C7
	8	Synthetic construct, reconstruction of a *Schistosoma mansoni* SR2 subfamily A non-LTR retrotransposon, complete sequence	27–34	GBSYN:AF025672

Table 1 (continued)

Amino acid	No. of repeats	Protein name	Residues	GenBank ID#
Leucine	19	*S. cerevisiae* chromosome X reading frame ORF YJR022w	10–28	GBPLN:SCYJR022W
	18	*S. cerevisiae* chromosome X DNA (cosmid 83)	10–27	GBPLN:SCXCOSM83
	15	*S. cerevisiae* chromosome III complete DNA sequence	14–28	GBPLN:SCCHRIII
Lysine	23	*A. thaliana* DNA chromosome 3, BAC clone T20N10	444–466	GBPLN:ATT20N10
	22	Nucleoporin p62 homolog {inverted repeats} [rat, Sprague-Dawley, testis, mRNA Partial, 1134 nt]	35–56	GBROD:S75997
Methionine	7	*D. melanogaster* genomic scaffold 142000013386050 section 4 of 54, complete sequence	40–46	GBINV:AE003517
	7	*H. sapiens* voltage-gated calcium channel α-1 subunit (CACNA1D) gene, partial cds	1–7	GBPRI:AF055575
	7	Human neuronal DHP-sensitive, voltage-dependent, calcium channel α one-dimensional subunit mRNA, complete cds	1–7	GBPRI:HUMCACNLS
	7	Human neuroendocrine/β-cell-type calcium channel α-1 subunit mRNA, complete cds	1–7	GBPRI:HUMCACH1A
	7	Rat rCACN4A mRNA for L-type voltage-dependent calcium channel α-1 subunit, complete cds	1–7	GBROD:RATRCACN4A
	7	Rat rCACN4B mRNA for L-type voltage-dependent calcium channel α-1 subunit, complete cds	1–7	GBROD:RATRCACN4B
	7	*Schistosoma mansoni* female-specific 800 protein (fs800) mRNA, complete cds	47–53	GBINV:SCMFS800
	7	*X. laevis* α-amidating enzyme (*AE-II*) gene, complete cds.	358–364	GBVRT:XELCAM
Phenylalanine	20	*Oryza sativa* genomic DNA, chromosome 6, clone:P0514G12	39–58	GBPLN:AP000616
	13	Genomic sequence for *A. thaliana* BAC T4O12 from chromosome I, complete sequence	710–722	GBPLN:AC007396
	11	*P. falciparum* chromosome 2, section 31 of 73 of the complete sequence	80–90	GBINV:AE001394
	11	*O. sativa* chloroplast rubisco large subunit (rbcL) mRNA, complete cds	59–69	GBPLN:RICCHRBCLA
Proline	27	*Oryctolagus cuniculus* preproacrosin mRNA, complete cds	352–378	GBMAM:OCU05204
	26	Epstein–Barr virus (EBV) genome, strain B95-8	63–88	GBVRL:EBV
	26	Epstein–Barr virus (B95-8 isolate) U2-IR2 domain encoding nuclear protein EBNA2, complete cds	63–88	GBVRL:HS4U2IR2A
	26	*V. carteri* mRNA for pherophorin-S	277–302	GBPLN:VCPHEROPH
Serine	52	*Schizosaccharomyces pombe* chromosome II cosmid c3D6	90–142	GBPLN:SPBC3D6
	52	*S. pombe* chromosome II cosmid c30B4	126–177	GBPLN:SPBC30B4
	52	*S. pombe* gene for hypothetical protein, partial cds, clone:TA46	20–71	GBNEW:AB027890
	42	*H. sapiens* chromosome 16, cosmid clone 373C8 (LANL), complete sequence	1639–1680	GBPRI:AC004493
	42	*H. sapiens* mRNA for KIAA0324 protein, partial cds	1638–1679	GBPRI:AB002322
	42	*H. sapiens* mRNA for RNA binding protein, complete cds	2607–2648	GBPRI:AB016092
	42	*H. sapiens* mRNA for RNA binding protein, partial cds, clone:R86	1130–1171	GBPRI:AB016091
	42	Human AF-9 mRNA, complete cds	149–190	GBPRI:HUMAF9X
	41	*H. sapiens* splicing coactivator subunit SRm300 (SRM300) mRNA, complete cds	2188–2228	GBPRI:AF201422
Threonine	67	*Cryptosporidium parvum* GP900 gene, complete cds	234–300	GBINV:AF068065
	53	*C. parvum* polythreonine protein gene, partial cds	175–227	GBINV:CPU83169
Tryptophan	4	*C. elegans* cosmid F58D2, complete sequence	504–508	GBINV:CEF58D2
	4	*C. elegans* cosmid T08B6	524–527	GBINV:CELT08B6
	4	*H. sapiens* mRNA for KIAA1399 protein, partial cds	52–55	GBPRI:AB037820
	4	*Mycobacterium leprae* cosmid B1764	21–24	GBBCT:MLU15181
	4	*Mycobacterium tuberculosis* H37Rv complete genome; segment 120/162	21–24	GBBCT:MTCY05A6
	4	*O. sativa* chromosome 10 clone OSJNBa0020E23, complete sequence	75–78	GBNEW:AC025098
	4	*Paramecium bursaria* Chlorella virus 1, complete genome	170–173	GBVRL:PBU42580
	4	*Pseudomonas putida* glucose-binding protein (*gltB*) gene, complete cds	192–195	GBBCT:PPU74323
	4	*Pyrococcus abyssi* complete genome; segment 5/6	273–276	GBBCT:CNSPAX05
	4	*Pyrococcus horikoshii* OT3 genomic DNA, 287001–544000 nt position (2/7)	268–271	GBBCT:AP000002
	4	*Synechocystis* sp. PCC6803 complete genome, 5/27, 524346–630554	55–58	GBBCT:D90903
Tyrosine	13	*Plasmodium falciparum* MAL3P5, complete sequence	214–226	GBINV:PFMAL3P5
	13	*P. falciparum* MAL3P5, complete sequence	214–226	GBNEW:PFMAL3P5
	12	*S. cerevisiae* chromosome XI reading frame ORF YKL030w	22–33	GBPLN:SCYKL030W
Valine	10	Synthetic *Escherichia coli* alkaline phosphatase gene, partial cds	3–12	GBSYN:SYNALKPHX
	9	*D. melanogaster* genomic scaffold 142000013386050 section 33 of 54, complete sequence	16–24	GBINV:AE003546
	9	*O. sativa* genomic DNA, chromosome 1, clone:P0434D08	82–90	GBPLN:AP001278
	9	*Trypanosoma cruzi* insect stage-specific antigen (GP72) mRNA, complete cds	18–26	GBINV:TRBGP72
	8	*A. thaliana* DNA chromosome 4, BAC clone T19F6, partial sequence (ESSA project)	15–22	GBPLN:ATT19F6
	8	*A. thaliana* DNA chromosome 4, contiguous fragment No. 60	15–22	GBPLN:ATCHRIV60
	8	*Neurospora crassa* DNA linkage group II BAC clone B24H17	214–221	GBNEW:NCB24H17
	8	*O. sativa* genomic DNA, chromosome 1, clone:P0693B08	171–178	GBPLN:AP001081
	8	*S. haematobium* eggshell protein gene ORFs	80–87	GBINV:SCMESPAA
	8	*S. hygroscopicus* gene cluster for polyketide immunosuppressant rapamycin	69–76	GBBCT:SHGCPIR

Table 2
Multiple protein database results, 23 June 2000.

Amino acid	No. of repeats	Protein name	Residue	Protein database ID#
Alanine	Many	Fibroin, Chinese oak silkmoth	76 separate repeats	PIR2:T31328
	21	DNA-binding protein ovo, fruit fly (*D. melanogaster*)	497–517	PIR2:A56038
	21	Ovo protein, fruit fly (*D. melanogaster*)	860–880	PIR2:S16356
	21	Ovo protein (shaven baby protein)	497–517	SWISSPROT:OVO_DROME
	20	Paired-type homeobox protein NBP, human	31–50	PIR2:JC5273
	20	Paired mesoderm homeobox protein 2B	241–260	SWISSPROT:PMXB_MOUSE
	20	Sex comb protein, fruit fly (*D. melanogaster*)	129–148	PIR2:T13748
Arginine	14	Hypothetical 95.1 kDa protein in ACT5-YCK1 intergenic region	32–45	SWISSPROT:YHT1_YEAST
	14	Hypothetical protein YHR131c, yeast (*S. cerevisiae*)	32–45	PIR2:S48975
	14	Protamine I, American alligator	49–52	PIR2:B58213
	13	ORF YOR053W, *S. cerevisiae*	50–62	TrEMBL:Q08428
	13	Probable membrane protein YOR053w, yeast (*S. cerevisiae*)	50–62	PIR2:S66936
Asparagine	50	Protein-tyrosine phosphatase 3	137–186	SWISSPROT:PTP3_DICDI
	49	Guanylyl cyclase (EC 4.6.1.2)	720–768	TrEMBL:Q9XZS0
	46	HelE (fragment)	233–278	TrEMBL:O15739
	46	Hypothetical protein HelE, slime mold (*D. discoideum*) (fragment)	234–279	PIR2:T08605
	46	Hypothetical protein PFB0800c, malaria parasite (*P. falciparum*)	707–752	PIR2:D71606
	46	Mtn3/RAG1IP-like protein PFB0760w, malaria parasite (*P. falciparum*)	50–95	PIR2:A71607
	46	Prespore-specific protein (fragment)	266–311	TrEMBL:O97140
Aspartic acid	45	Probable membrane protein YOR054c, yeast (*S. cerevisiae*)	606–650	PIR2:S66937
	44	Calsequestrin precursor, skeletal muscle, edible frog	377–420	PIR1:S22418
	41	F11A17.5 protein	285–325	TrEMBL:Q9SX72
Cystine	23	Cysteine-rich protein (fragment)	11–33	TrEMBL:Q16861
Glutamine	45, 42	Adenylyl cyclase	1856–1900, 2004–2045	TrEMBL:Q9U9S7
	41	QMEF2D protein	362–402	TrEMBL:O42323
	40	CAGH44 (fragment)	152–191	TrEMBL:O15409
Glutamic acid	61	Hypothetical 10.3 kDa protein	6–66	TrEMBLnew:CAB82936
Glycine	29	Glycine-rich cell-wall structural protein homolog F23E13.120, *A. thaliana*	153–181	PIR2:T04592
Histidine	17	Putative ring zinc finger protein	519–535	TrEMBL:Q9SG87
	16	CG9732 protein	183–198	TrEMBL:Q9W2S4
	15	EG:EG0007.4 protein	73–87	TrEMBL:O97423
	15	EG:EG0007.12 protein	1073–1087	TrEMBL:Q9W4M7
Isoleucine	11	T27G7.3	254–264	TrEMBL:Q9SJF7
	10	ORF68	6–15	TrEMBLnew:AAF33598
	8	Hypothetical protein C15C7.1, *C. elegans*	149–156	PIR2:T15511
Leucine	19	Probable membrane protein YJR023c, yeast (*S. cerevisiae*)	9–27	PIR2:S57038
	19	Yeast hypothetical 15.3 kDa protein in MER2-BNA1 intergenic region	10–28	SWISSPROT:YJZ3
	16	F7F22.5	307–322	TrEMBL:Q9SHN5
	16	Hypothetical protein C24A3.3, *C. elegans*	313–328	PIR2:T15587
	16	Similar to *C. elegans* protein ZK892.4	33–48	TrEMBL:Q18123
	15	CAGH1 alternate open reading frame	124–138	TrEMBL:O15420
	15	Hypothetical protein YCR087w, yeast (*S. cerevisiae*)	84–98	PIR2:S19502
	15	Yeast very hypothetical 19.8 kDa protein in ABP1 5′ region	84–98	SWISSPROT:YCX7
Lysine	26	Hypothetical 42.7 kDa protein (fragment)	355–380	TrEMBLnew:CAB70810
	26	Hypothetical protein DKFZp434I1120.1, human	355–380	PIR2:T46395
	23	Hypothetical 59.7 kDa protein	444–466	TrEMBLnew:CAB88307
	22	Nucleoporin p62 homolog, rat (fragment)	35–56	PIR2:I52523
	21	Microvascular endothelial differentiation protein 2	85–105	TrEMBL:O35807
Methionine	7	Calcium channel α-1 chain, pancreatic, human	1–7	PIR2:A38198
	7	Calcium channel α one-dimensional chain, human	1–7	PIR2:JH0564
	7	CG14094 protein	40–46	TrEMBL:Q9VVZ2
	7	Human voltage-dependent L-type calcium channel α one-dimensional subunit	1–7	SWISSPROT:CCAD
	7	Peptidylglycine monooxygenase (E.C. 1.14.17.3) II precursor, African clawed frog	358–364	PIR1:URXLA2
	7	SCHMA female specific 800 protein	47–53	SWISSPROT:F802
	7	Voltage-dependent calcium channel α-1 chain, isoform CACN4A, rat	1–7	PIR2:T42742
	7	Voltage-gated calcium channel α-1 subunit	1–7	TrEMBLnew:AAD08651
	7	XENLA peptidyl-glycine α-amidating monooxygenase II precursor	358–364	SWISSPROT:AMD2
Phenylalanine	22	Similar to ring-H2 finger protein RHA1A	39–60	TrEMBL:Q9SNH1
	13	T4O12.11	710–722	TrEMBLnew:AAF26757
	11	Cytochrome *b* (fragment)	129–139	TrEMBL:O48239
	11	Cytochrome *b* (fragment)	129–139	TrEMBL:O48240
	11	Probable integral membrane protein PFB0415c, malaria parasite	80–90	PIR2:A71615
	11	Predicted integral membrane protein	80–90	TrEMBL:O96177
	11	Ribulose-bisphosphate carboxylase	339–349	PIR2:T02958
	11	Ribulose-bisphosphate carboxylase large chain	339–349	TrEMBL:Q37247
Proline	50	Hypothetical proline-rich protein 1, polychaete (*Owenia fusiformis*)	9–58	PIR2:A34043
	29	U2-IR2 domain encoding nuclear protein EBNA2, complete cds	63–91	TrEMBL:Q69023
	28	Acrosin (E.C. 3.4.21.10) precursor, rabbit	352–378	PIR2:S47538
	27	Rabbit acrosin precursor	352–378	SWISSPROT:ACRO
	26	Nuclear protein EBNA2, human herpesvirus 4	63–88	PIR2:S42442
	26	Pherophorin-S precursor	277–302	TrEMBL:P93797
	26	Pherophorin-S, *Volvox carteri*	277–302	PIR2:T10798

Table 2 (continued)

Amino acid	No. of repeats	Protein name	Residue	Protein database ID#
Serine	52	Hypothetical protein C30B4.01C in chromosome II (fragment)	126–177	TrEMBL:P87179
	52	Hypothetical protein (fragment)	21–72	TrEMBL:Q9USB7
	52	Hypothetical protein SPBC3D6.14c, fission yeast (*S. pombe*) (fragment)	91–142	PIR3:T40374
	52	Hypothetical protein SPBC30B4.01c, fission yeast (*S. pombe*) (fragment)	126–177	PIR2:T40167
Threonine	67	GP900	234–300	TrEMBL:O96503
	67	Mucin-like glycoprotein 900, *C. parvum*	234–300	PIR2:T31113
	53	Polythreonine protein (fragment)	175–227	TrEMBL:O00908
Tryptophan	4	304aa long hypothetical oligopeptide transport system permease protein APPC	268–271	TrEMBL:O58240
	4	Abc transporter	273–276	TrEMBL:Q9UYG8
	4	Abc transporter PAB1345, *Pyrococcus abyssi* (strain Orsay)	273–276	PIR2:G75068
	4	ENTCO cytolytic protein enterolobin	385–388	SWISSPROT:ENT
	4	Enterolobin (cytolysin), *Enterolobium contortisiliquum*	385–388	PIR2:A57982
	4	F58D2.2 protein	505–508	TrEMBL:Q9XVM2
	4	Hypothetical protein F58D2.2, *C. elegans*	505–508	PIR2:T22909
	4	Hypothetical protein T08B6.4, *C. elegans*	524–527	PIR2:T15073
	4	Glucose-binding protein	193–196	TrEMBL:Q9Z422
	4	Hypothetical protein Rv2698, *M. tuberculosis* (strain H37RV)	21–24	PIR2:E70530
	4	Hypothetical protein A333L, *Chlorella* virus PBCV-1	170–173	PIR2:T17832
	4	Hypothetical protein sll1835, *Synechocystis sp*. (strain PCC 6803)	55–58	PIR2:S75223
	4	Hypothetical protein F17M5.70, *A. thaliana*	83–86	PIR2:T05982
	4	Hypothetical 10.4 kDa protein	83–86	TrEMBL:Q9SZA8
	4	Hypothetical 10.4 kDa protein	83–86	TrEMBLnew:CAB80048
	4	Hypothetical 17.6 kDa protein	21–24	TrEMBL:O07200
	4	Hypothetical 21.3 kDa protein	75–78	TrEMBLnew:AAF72073
	4	Hypothetical 28.8 kDa protein	55–58	TrEMBL:P73111
	4	KIAA1399 protein (fragment)	52–55	TrEMBLnew:BAA92637
	4	NONO/P54NRB homolog (fragment)	162–165	TrEMBL:O54725
	4	PLPRNLLL	170–173	TrEMBL:Q84647
	4	Probable oligopeptide transport system permease protein AppC, *Pyrococcus horikoshii*	268–271	PIR2:C71163
	4	T08B6.4 protein	524–527	TrEMBL:O44527
	4	U1764I	21–24	TrEMBL:Q49991
Tyrosine	14	CG11931 protein	39–52	TrEMBL:Q9VR14
	13	PFC0615W protein	214–226	TrEMBL:O97262
	12	Yeast hypothetical 23.0 kDa protein in IXR1-TFA1 intergenic region	22–33	SWISSPROT:YKD0
	12	Hypothetical protein YKL030w, yeast (*S. cerevisiae*)	22–33	PIR2:S37847
Valine	12	CG13535 protein	122–133	TrEMBL:Q9W1Y6
	9	CG12522 protein	16–24	TrEMBL:Q9VTF3
	9	Hypothetical protein	82–90	TrEMBLnew:BAA92199
	9	Hypothetical protein (eggshell protein gene region), fluke (*Schistosoma haematobium*) (clone SH.E 2-1) (fragment)	105–113	PIR2:B44805
	9	Insect stage-specific protein precursor, *Trypanosoma cruzi*	18–26	PIR2:A45551
	9	Stage-specific antigen precursor	18–26	TrEMBL:Q03876
	9	Syntaxin-related protein SSO1, yeast (*S. cerevisiae*)	275–283	PIR2:S39569
	9	Yeast SSO1 protein	275–283	SWISSPROT:SSO1

Table 3
Protein Data Bank Results, 30 July 2000.

Amino acid	No. of repeats	Name of protein	Residue	Repeat structure	PDB ID
Alanine	13	HIV-1 reverse transcriptase	B240–B252	Chain	1c9r
	12	Apoptosis/tumor necrosis factor receptor associated protein 2	315–326	3*Helix	1czz
	10	Hydrolase (O-glycosyl)	40–49	Helix	1l64
Arginine	4/65	Retinoic acid receptor (nuclear receptor)	366–369	Loop	2lbd
Asparagine	4/42	Tetanus neurotoxin	1078–1081	Loop	1af9
Apartic acid	6	Karyopherin 2 nuclear transport protein complex	58–62	N-terminal helix	1qbk
Cystine	3	Human T-cell leukemia virus type II matrix protein (buried knot)	301–303	Buriedknot	1ag8
	3	Robustoxin (knot)	14–16	3*knot/S-S chain	1qdp
Glutamine	5	Siv Gp41 ectodomain (envelope glycoprotein)	38–42	Helix/slab	1qbz
	5	Aldehyde oxidoreductase	451–455	Helix	1alo
Glutamic acid	5	Mucosal addressin cell adhesion molecule-1 (membrane protein)	148–152	Sheet/Strand	1bqs
	5	Adenovirus type 2 hexon	155–161	Coil	1dhx
Glycine	5	Canine parvovirus protein 2	28–34	Strand	4dpv
Histidine	6	Ferric hydroxamate uptake receptor (His tag right helix)	407–412	Helix-tag	1fcp
Isoleucine	3/274	Phosphoribosylaminoimidazole-succinocarboxamide synthase (ATP binding protein)	203–205	Helix	1a48
Leucine	5	D-3-Phosphoglycerate dehydrogenase (crystal dyad)	120–124	Helix-dyad	1psd
	5	NADH oxidase (flavoenzyme)	143–147	Helix	1nox
Lysine	4	Synaptotagmin I (calcium/phospholipid binding protein)	189–192	Loop	1rsy
Methionine	3/55	*Escherichia coli* topoisomerase I	305–307	Helix	1ecl
Phenylalanine	3/136	Aspartate aminotransferase	360–362	Strand	1yoo
Proline	6/6	Human T-cell leukemia virus type II matrix protein	111–117	L-helix	1jvr
Serine	5	Glucoamylase-471 complexed with acarbose	364–368	Loop/bend	1agm
	5	Glucoamylase-471 complexed with 1-deoxynojirimycin	364–368	Bend/loop	1dog

Table 3 (continued)

Amino acid	No. of repeats	Name of protein	Residue	Repeat structure	PDB ID
Threonine	5	Echovirus 1 (viral coat protein)	156–159 (2)	Bend	1ev1
Tryptophan	3/34	Human rhinovirus 16 coat protein	78–80 (2)	Pocket	1aym
Tyrosine	4	Cricket paralysis virus (viral protein)	A93–A96	Helix	1b35
Valine	4/80	Helicase complex (RNA helicase/DNA)	A406–A409	Sheet	1a1v

Berman *et al.*, 2000) (celebrating its 30th anniversary in 2001) provides definitive structural information for visual three-dimensional exploration of homo-polypeptides. Table 3 lists structures with homo-repeat sequences of the 20 amino acids; when multiple events are detected, an example is given together with the number of events (*e.g.* for Met, 55 entries contain the Met-Met-Met sequence; the PDB lists homopolymeric proteins separately).

Literature searches of papers discussing homo-polypeptide sequences (Henderson *et al.*, 1999; Hudgins & Jarrold, 2000; Higashi *et al.*, 1999; Hirakura *et al.*, 2000; Winter *et al.*, 1999) offer some hints for the significance of homo-repeats; a recent paper (Nishizawa *et al.*, 1999) discussed a possible origin of the repetition of amino acids in human sequences and observed that 'modern' neuroproteins and immunoproteins are more likely to exhibit amino-acid repetition. However, our searches do not confirm a temporal relationship. The naturalist would suppose that a functional imperative would encourage improbable repeat domains and certainly would penalize misbegotten molecular stuttering. Except for CCC repeats in this survey of PDB entries, homo-domains occur primarily on molecular surfaces, suggesting functional roles in molecular recognition and signaling, *e.g.* receptors, reproduction and viral invasion. The Cys-Cys-Cys trimer has been sculpted (Meyer *et al.*, 2000; http://www.tamu.edu/struct/research/trinity/trinitatis.html).

Structural genomics is still in its infancy (Atwood, 2000); this report presents our observations and invites others to help discover their significance. It further suggests that these highly unlikely patterns may offer a useful link within sequence–structural databases. As we begin to explore ourselves in the molecular maze of human genomics, additional sequence–structure–function patterns will emerge. Here, we especially challenge the crystallographic community to elucidate the structures of these highly improbable repeats.

We thank the Robert A. Welch Foundation and the Texas Advanced Technology Program for financial support and Drs J. Hu, R. Swanson, S. Swanson and J. Zinn for helpful discussions.

References

Atwood, T. K. (2000). *Science*, **290**, 471–473.

Barnes, D. A., Bonnin, A., Huang, J. X., Gousset, L., Wu, J., Gut, J., Doyle, P., Dubremetz, J. F., Ward, H. & Petersen, C. (1998). *Mol. Biochem. Parasitol.* **96**(1/2), 93–110.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Henderson, S. C., Li, J. W., Counterman, A. E. & Clemmer, D. E. (1999). *J. Phys. Chem. B*, **103**(41), 8780–8785.

Higashi, N., Nishikawa, R., Koga, T. & Niwa, M. (1999). *J. Colloid Interface Sci.* **220**(2), 362–366.

Hirakura, Y., Azimov, R., Azimova, R. & Kagan, B. L. (2000). *J. Neurosci. Res.* **60**(4), 490–494.

Hudgins, R. R. &. Jarrold, M. F. (2000). *J. Phys. Chem. B*, **104**(9), 2154–2158.

Meyer, E. F. (1997). *Protein Sci.* **6**, 1591–1597.

Meyer, E. F., Swanson, S. M. & Williams, J. A. (2000). *Pharmacol. Ther.* **85**, 113–121.

Nishizawa, K., Nishizawa, M. & Kim, K.-S. (1999). *J. Mol. Biol.* **294**, 937–953.

Subramaniam, S. (1998). *Proteins Struct. Funct. Genet.* **32**, 1–2.

Winter, C., tom Dieck, S., Boeckers, T. M., Bockmann, J., Kampf, U., Sanmarti-Vila, L., Langnaese, K., Altrock, W., Stumm, M., Soyke, A., Wieacker, P., Garner, C. C. & Gundelfinger, E. D. (1999). *Genomics*, **57**(3), 389–397.