Method

Open Access

A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes

Paul M Harrison and Mark Gerstein

Address: Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114, USA.

Correspondence: Paul M Harrison. E-mail: harrison@csb.yale.edu.

Published: 30 May 2003

Genome ***Biology*** 2003, **4:**R40

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/6/R40

Received: 6 February 2003
Revised: 8 April 2003
Accepted: 29 April 2003

Abstract

We have derived a novel method to assess compositional biases in biological sequences, which is based on finding the lowest-probability subsequences for a given residue-type set. As a case study, the distribution of prion-like glutamine/asparagine-rich ((Q+N)-rich) domains (which are linked to amyloidogenesis) was assessed for budding and fission yeasts and four other eukaryotes. We find more than 170 prion-like (Q+N)-rich regions in budding yeast, and, strikingly, many fewer in fission yeast. Also, some residues, such as tryptophan or isoleucine, are unlikely to form biased regions in any eukaryotic proteome.

Background

Amyloidogenesis involving domains that contain glutamine and/or asparagine ((Q+N)-rich domains) is linked to prion phenomena in budding yeast, as well as a number of neurological disorders in humans, including Huntington's disease.

A prion is an alternative conformation for a protein that can direct its own propagation [1,2]. In budding yeast, there are currently four identified prions: [PSI+], [URE3], [RNQ+] and [NU+] [3–7]. [PSI+] arises from the propagation of an alternatively folded amyloid-like form of Sup35p [3]. Sup35p is part of the complex in budding yeast that controls translation termination and nonsense-codon readthrough [8,9]. [URE3] is caused by an alternatively folded form of Ure2p, a protein involved in nitrogen metabolism [4,10]. The determinant sequences of [PSI+] and [URE3] are characterized chiefly by bias for glutamine (Q) and asparagine (N) residues (Table 1). [RNQ+] and [NU+] arise from alternative propagatable forms of parts of the Rnq1p and New1p sequences, and were found by searches for further sequences with Q/N compositional bias (Table 1) [7,11]. Prokaryotic and eukaryotic proteomes were assessed for yeast-prion-like domains that comprised a total of 30 or more glutamines and asparagines in an 80-amino-acid stretch [12]. [PIN+] is a non-Mendelian inherited trait that is required for the *de novo* appearance of [PSI+] [7,13]. Eight candidate sequences for [PIN+], which tend to have a Q- and/or N-rich segment, were identified using a genetic screen and remain to be verified [13].

Expanded polyglutamine repeats underlie the pathology of neurodegenerative disorders in humans, the most common of which is Huntington's disease [14]. This disorder is caused by inherited expansions of length equal or greater than 39 amino-acid residues in the polyglutamine region of the protein huntingtin [15]. (Q+N)-rich regions, polyglutamine and polyasparagine are thought to oligomerize or polymerize through a 'polar zipper' of hydrogen bonds between the side chains [15,16].

Here, we have derived a method for identifying biased regions that relies on defining the lowest-probability subsequences (LPSs) for a given amino-acid composition. For six eukaryotic proteomes (budding yeast, fission yeast, nematode worm, fruit fly, human and *Arabidopsis*), we have used this formalism to analyze the prevalence of Q- and N-rich regions in the context of other biases. In general, N-rich regions are rarer than Q-rich regions in the eukaryotic proteomes, most notably so in the human proteome. We use the biases for the four known prions of budding yeast to survey comprehensively for (Q+N)-rich domains, and examine the diversity of their subsidiary amino-acid compositions, their functions and

Table 1

The four prion sequences*

	Notable single-residue bias counts for prion determinant domain (P_{bias} in brackets)†	Notable LPSs (for whole sequence)‡
Prion sequence: Ure2p (YNL229C) Prion determinant is residues 1-65		
MMNNNGNQVSNLSNALRQVNIGNRNSNTTTDQSNINFEFSTG VNNNNNNNSSSNNNNVQNNNSGRNGSQNNDNENNIKNTLEQH RQQQQAFSDMSHVEYSRITKFFQEQPLEGYTLFSHRSAPNGFKV AIVLSELGFHYNTIFLDFNLGEHRAPEFVSVNPNARVPALIDHGM DNLSIWESGAILLHLVNKYYKETGNPLLWSDDLADQSQINAWLF FQTSGHAPMIGQALHFRYFHSQKIASAVERYTDEVRRVYGVVE MALAERREALVMELDTENAAAYSAGTTPMSQSRFFDYPVWLV GDKLTIADLAFVPWNNVVDRIGINIKIEFPEVYKWTKHMMRRPA VIKALRGE	N 27/65 (1.9×10^{-16})	N, 33 in 2 to 78 (1.2×10^{-16}) Q, 7 in 82 to 108 (5.5×10^{-5}) {QN}, 43 in 2 to 88 (1.2×10^{-20}) {DERK}, 10 in 2 to 88 (1.6×10^{-3}) {VILM}, 4 in 2 to 88 (1.8×10^{-3})
Prion sequence: Sup35p (YDR172W) Prion determinant is residues 1-123		
MSDSN**QGNNQQNYQQYSQNGNQQQGNNRYQGYQAYNAQA QPAGGYYQNYQGYSGYQQGGYQQYNPDAGYQQQYNPQGG YQQYNPQGGYQQQFNPQGGRGNYKNFNYNNNLQGYQAGF QPQSQG**MSLNDFQKQQKQAAPKPKKTLKLVSSSGIKLANATKK VGTKPAESDKKEEEKSAETKEPTKEPTKVEEPVKKEEKPVQTEE KTEEKSELPKVEDLKISESTHNTNNANVTSADALIKEQEEEVDDE VVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLTGSVDKRTI EKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYET GFERGGQTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKER YDQCVSNVSNFLRAIGYNIKTDVVFMPVSGYSGANLKDHVDPK ECPWYTGPTLLEYLDTMNHVDRHINAPFMLPIAAKMKDLGTIVE GKIESGHIKKGQSTLLMPNKTAVEIQNIYNETENEVDMAMCGEQ VKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVAQIAIVELKSIIAA GFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPAFAKKGM KVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIAE	Q 35/123 (9.6×10^{-21}) Y 20/123 (4.9×10^{-9}) G 21/123 (5.9×10^{-7}) N 20/123 (3.7×10^{-5})	Q, 39 in 5 to 134 (7.3×10^{-24}) Y, 20 in 12 to 112 (1.4×10^{-10}) E, 23 in 166 to 248 (1.7×10^{-9}) K, 23 in 130 to 218 (5.5×10^{-8}) G, 20 in 19 to 122 (1.5×10^{-7}) N, 20 in 4 to 108 (3.6×10^{-6}) V, 47 in 188 to 653 (4.5×10^{-5}) {QN}, 60 in 4 to 134 (6.1×10^{-26}) {DERK}, 6 in 4 to 134 (8.0×10^{-9}) {VILM}, 2 in 4 to 134 (3.5×10^{-12})
Prion sequence: Rnq1p (YCL028W) Prion determinant is residues 153-405		
MDTDKLISEAESHFSQGNHAEAVAKLTSAAQSNPNDEQMSTIES LIQKIAGYVMDNRSGGSDASQDRAAGGGSSFMNTLMADSKGSS QTQLGKLALLATVMTHSSNKGSSNRGFDVGTVMSMLSGSGGGS QSMGASGLAALASQFFKSGNNSQ**GQGQGQGQGQGQGQGQG QGSFTALASLASSFMNSNNNNQQGQNQSSGGSSFGALASMAS SFMHSNNNQNSNNSQQGYNQSYQNGNQNSQGYNNQQYQGG NGGYQQQQGQSGGAFSSLASMAQSYLGGGQTQSNQQQYNQ QGQNNQQQYQQQGQNYQHQQQGQQQQQGHSSSFSALASM ASSYLGNNSNSNSSYGGQQQANEYGRPQHNGQQQSNEYGRP QYGGNQNSNGQHESFNFSGNFSQQNNNGNQNRY**	Q 66/253 (4.2×10^{-35}) G 42/253 (6.6×10^{-12}) N 41/253 (7.4×10^{-9})	Q, 67 in 152 to 401 (2.1×10^{-36}) G, 60 in 50 to 399 (3.1×10^{-17}) N, 41 in 185 to 402 (8.3×10^{-11}) {QN}, 110 in 149 to 402 (3.2×10^{-43}) {DERK}, 5 in 149 to 402 (1.2×10^{-23}) {VILM}, 12 in 149 to 402 (8.9×10^{-17})

Table 1 *(Continued)*

The four prion sequences*		
Prion sequence: New1p (YPL226W) Prion determinant is residues 1-153		
MPPKKFKDLNSFLDDQPKDPNLVASPFGGYFKNPAADAGSN **NASKKSSYQQQRNWKQGGNYQQGGYQSY<u>NSNYNNYNNYNN</u>** **<u>YNNYNNYNNYNKYN</u>GQGYQKSTYKQSAVTPNQSGTPTPSAS** **TTSLTSLNEKLSNLELTPISQFLSKIPECQS**ITDCKNQIKLIIEEF GKEGNSTGEKIEEWKIVDVLSKFIKPKNPSLVRESAMLIISNIAQF FSGKPPQEAYLLPFFNVALDCISDKENTVKRAAQHAIDSLLNCFP MEALTCFVLPTILDYLSSGAKWQAKMAALSVVDRIREDSANDL LELTFKDAVPVLTDVATDFKPELAKQGYKTLLDYVSILDNLDLS PRYKLIVDTLQDPSKVPESVKSLSSVTFVAEVTEPSLSLLVPILNR SLNLSSSSQEQLRQTVIVVENLTRLVNNRNEIESFIPLLLPGIQKV VDTASLPEVRELAEKALNVLKEDDEADKENKFSGRLTLEEGRDF LLDHLKDIKADDSCFVKPYMNDETVIKYMSKILTVDSNVNDWK RLEDFLTAVFGGSDSQREFVKQDFIHNLRALFYQEKERADEDEGI EIVNTDFSLAYGSRMLLNKTNLRLLKGHRYGLCGRNGAGKSTL MRAIANGQLDGFPDKDTLRTCFVEHKLQGEEGDLDLVSFIALDE ELQSTSREEIAAALESVGFDEERRAQTVGSLSGGWKMKLELARA MLQKADILLLDEPTNHLDVSNVKWLEEYLLEHTDITSLIVSHDSG FLDTVCTDIIHYENKKLAYYKGNLAAFVEQKPEAKSYYTLTDSN AQMRFPPPGILTGVKSNTRAVAKMTDVTFSYPGAQKPSLSHVSC SLSLSSRVACLGPNGAGKSTLIKLLTGELVPNEGKVEKHPNLRIG YIAQHALQHVNEHKEKTANQYLQWRYQFGDDREVLLKESRKIS EDEKEMMTKEIDIDDGRGKRAIEAIVGRQKLKKSFQYEVKWKY WKPKYNSWVPKDVLVEHGFEKLVQKFDDHEASREGLGYRELIP SVITKHFEDVGLDSEIANHTPLGSLSGGQLVKVVIAGAMWNNPH LLVLDEPTNYLDRDSLGALAVAIRDWSGGVVMISHNNEFVGAL CPEQWIVENGKMVQKGSAQVDQSKFEDGGNADAVGLKASNLA KPSVDDDDSPANIKVKQRKKRLTRNEKKLQAERRRLRYIEWLSS PKGTPKPVDTDDEED	N 26/153 (1.4×10^{-6})	N, 16 in 69 to 94 (9.8×10^{-14}) Y, 13 in 60 to 103 (1.2×10^{-9}) L, 74 in 253 to 738 (2.4×10^{-5}) Q, 11 in 49 to 112 (2.9×10^{-5}) R, 7 in 1149 to 1173 (7.5×10^{-5}) {QN}, 27 in 49 to 99 (1.8×10^{-14}) {DERK}, 3 in 49 to 99 (5.3×10^{-4}) {VILM}, 0 in 49 to 99 (8.1×10^{-7})

*The prion determinant regions (found from experiment) are in bold, the LPSs for the whole protein sequence for the most pronounced single-amino-acid bias, are underlined. †All biases with a $P_{bias} = 1 \times 10^{-4}$ are listed for each prion sequence. ‡As examples, the counts for the sets of residues {DERK} and {VILM} that correspond to the {QN} lowest probability sequence are listed for each prion.

their cellular compartments. We find up to around 170 (Q+N)-rich regions in budding yeast, and a relative dearth of such regions in fission yeast. In addition, to provide more context, we discuss some overarching observations on biased regions of any sort.

Results and discussion

Our analysis can be broken up as follows. First we analyze the Q, N, and (Q+N) biases in the four known prion sequences of budding yeast, as well as other subsidiary biases for and against certain residue types. We discuss how this relates to prion-determinant domains (that is, regions of the prion sequences that are necessary for the prion phenomenon). Our analysis is performed using a simple algorithm to find the lowest probability subsequences (LPSs) for a given residue bias (see Materials and methods).

Second, we ask how prevalent are Q- and N-biases in eukaryotes? Motivated by the fact that prion-determinant sequences correspond to LPSs, we examine LPSs for Q and N biases in the context of single-residue biases for all residue types, in all six proteomes. Focusing on the budding-yeast proteome, we also compare single-residue biases observed for known and hypothetical proteins, and for conceptually translated intergenic DNA (igDNA) in all six potential reading frames.

Then, on the basis of biases for Q and N in combination, we examine the abundance and diversity of (Q+N)-rich regions in the six eukaryotic proteomes. We survey their subsidiary biases for and against certain residues and groups of residues, and their sizes, functional classes and cellular compartments. Finally, to provide more general context, we discuss some overarching perspectives on compositional bias in the eukaryotic proteomes.

Analysis of the identified yeast prion sequences for their biases

Four identified prion protein sequences of budding yeast are Sup35p, Ure2p, Rnq1p and New1p, which form the prions [PSI+], [URE3], [RNQ+] and [NU+] respectively. We extracted the domains that are determinant for prion formation from each of these protein sequences, which have been found previously by experimental study (Table 1) [3,4,6,7]. Using

the formalism described in Materials and methods, we determined the main biases for each prion-determinant domain and an associated probability for each bias. Results are shown for single-residue biases, with examples for the sets of residues {QN}, {DERK} and {VILM} (single-letter amino-acid code; Table 1). The groupings {DERK} and {VILM} are 'charged residues' and 'major hydrophobics' respectively [17]. Charged residues and the major hydrophobics appear to be disfavored for the yeast prions (Table 1) [18]; mutation of Q or N to charged residues can lead to loss of prion-forming capability [19]. We also derived the LPSs for a given bias for the whole protein sequences (not just the prion-determinant domains). In addition to the well-documented Q and N biases, we also note that three of the four budding-yeast prions have subsidiary biases for tyrosine, glycine and/or serine (Table 1). The mild bias for tyrosine is conserved for homologs of the Sup35 prion determinant in other fungi, although it is not clear how this is related to the prion phenomenon [18]. It has been suggested that π-stacking of aromatic groups, as in tyrosine and phenylalanine, may play a part in stabilizing amyloid conformations [20].

Interestingly, the prion-determinant domains for three of the prion sequences (Sup35p, Ure2p, Rnq1p) are congruent with the top-ranking single-residue LPSs for the whole sequences (these are the underlined sequence regions in Table 1). That is, the most biased regions coincide with the experimentally derived prion-determinant domains. These are either for Q or N biases (Table 1). However, for the fourth prion sequence (New1p), the prion-determinant domain is comparatively poorly biased for N or Q or {QN} (for example, for N, $P_{bias} = 1.4 \times 10^{-6}$, where P_{bias} is the probability of bias). Its LPS for N bias does, however, coincide with a region derived from a repeat of the amino-acid triplet NYN that has been shown to be necessary for [NU+] prion propagation [7].

In the next section, we show how single-residue biases for Q and N rank in terms of their relative abundance in eukaryotic proteomes. After that, we use the Q- and N-bias levels of the LPSs of the four prions in combination to derive a refined set of (Q+N)-rich domains in the six eukaryotic proteomes (see below).

Abundance of Q and N biases in a proteomic context, for budding yeast and five other eukaryotic proteomes

How abundant are the biases for Q and N observed for the yeast prion domains compared to biases for all the other residue types? Are they noticeably more or less prevalent in budding yeast compared to other eukaryotic proteomes? We examined the most prevalent single-residue biases for the six eukaryotic proteomes at P_{bias} values corresponding to the LPSs observed in the prion-determinant domains (Table 1). This data gives us a perspective on the relative abundance of such biases (arrayed in Table 2; the exact threshold used to make this table is $P_{bias} < 1 \times 10^{-13}$).

It is clear that biases for Q and N are relatively more prevalent in the budding yeast proteome than in the other eukaryotic proteomes. Both Q and N are among the top six biases for this organism at this bias level (Table 2). This observation is the same regardless of whether the biases are ranked in terms of the total number of bias residues, or the total number of biased regions (Table 3), or as a weighted count in which the number of bias residues is multiplied by a factor derived from the amino-acid composition of the proteome (see Additional data file 1 or Supplementary Table A at [21]). For all the proteomes, N biases are always less prevalent than Q biases, being most disfavored in the human proteome, where they are up to 12 times rarer than Q-rich regions (Table 2c). The small number of N-rich regions in human sequences is intriguing, and may be due to a cellular toxicity of such regions in higher eukaryotes.

Interestingly, as noted in Table 2, there are eight examples of predicted coiled-coil domains [22] that are in our list of (Q+N)-rich domains. Coiled coils are alpha-helical, whereas the prions form beta-sheet-rich aggregates; this may be an artifact of the coiled-coil prediction program [22], although there are some known viral coiled coils that have short runs of up to five Q residues, and a mild overall Q bias over their whole sequence [23].

The prevalent biases in the budding-yeast proteome were broken down into those for hypothetical and known proteins, and compared at three bias levels (Table 3). Known proteins are those in open reading frame (ORF) classes 1 through 3 in the MIPS database [24] (these are either characterized proteins or sequences that have homology to a characterized sequence). 'Hypothetical' proteins are the remaining annotations (ORF classes 4 through 6 in the MIPS database). There is little difference in the rankings for biases for the whole proteome, the set of known proteins and the set of hypothetical proteins (Table 3a,c,d). Surprisingly, however, total amounts of biased regions are substantially higher for known proteins (Table 3); for example at $P_{bias} < 1 \times 10^{-9}$, eight times as prevalent. Q and N biases both remain high-ranking in the 'known' and 'hypothetical' protein lists, and are lowly ranked for conceptually translated igDNA (Table 3a,b,e). In general, the prevalent biases observed for conceptually translated igDNA are very different from those for the annotated proteome (Table 3a,b,e). Notably, there is also very little implied bias for negatively charged residues (aspartic acid (D) and glutamic acid (E) combined), relative to positively charged residues (lysine (K) and arginine (R) combined) in the translated igDNA biases. This suggests that negatively charged bias regions in protein-coding sequences would take longer to evolve or need much greater selective pressure than those for positively charged biased regions, and that underlying replication 'slippage' tendencies [25] and mutation biases for the formation of cryptically simple sequences [26] may disfavor such regions.

Table 2

Abundance of biased regions that have biases at the same level as the Q and N biases in the four budding-yeast prions

(a) Biases in terms of total number of regions*

Rank	Budding yeast		Fission yeast		Fruit fly		Nematode		*Arabidopsis*		Human	
1	S	108 [1]†	S	74	Q	725 (20.7)	P	494	P	345	P	549
2	Q	104 [8] (17.8)	P	40	G	400	G	448	E	292	E	322
3	N	73 [1] (12.5)	E	37	P	359	E	286	G	242	C	302
4	E	68 [18]	T	32	S	327	Q	270 (8.9)	Q	153 (6.0)	G	294
5	T	58	Q	17 (5.6)	A	264	C	220	C	150	S	233
6	P	37 [1]	G	17	H	231	T	199	S	134	K	188
7	D	35	K	16	E	212	K	184	D	90	Q	176 (5.9)
8	K	24 [9]	A	15	T	188	S	146	K	86	A	136
9	G	20	C	13	K	170	R	132	R	81	R	83
10	A	19 [1]	R	11	N	144 (4.1)	A	102	L	56	H	80
11	H	8	V	6	C	144	D	55	A	56	T	59
12	C	8	H	6	R	118	H	46	H	47	D	49
13	R	6	M	5	D	74	N	40 (1.3)	Y	28	L	31
14	M	5	N	4 (1.3)	L	47	F	17	N	28 (1.1)	M	23
15	L	5	F	4	Y	28	Y	16	T	17	F	21
16	V	3	L	3	M	24	M	15	M	11	V	18
17	Y	2	D	2	V	22	L	11	V	8	N	15 (0.6)
18	F	2	Y	1	F	8	V	6	W	5	Y	13
19	W	0	W	0	W	5	I	5	F	3	W	11
20	I	0	I	0	I	5	W	1	I	1	I	10
	Total	585	Total	303	Total	3,495	Total	2,693	Total	1,833	Total	2,613

(b) Biases in terms of total number of residues‡

Rank	Budding yeast		Fission yeast		Fruit fly		Nematode		*Arabidopsis*		Human	
1	S	10,630	S	9,035	Q	39,186 (16.3)	P	31,917	E	23,229	P	44,427
2	T	5,900	T	5,805	S	31,936	E	31,216	P	21,124	E	27,352
3	E	4,704	P	2,887	P	29,345	G	28,192	G	13,462	S	26,363
4	Q	3,924 (10.4)	E	2,657	G	24,320	Q	18,126 (8.9)	S	10,313	G	22,131
5	N	3,745 (10.0)	A	1,854	E	23,384	T	15,994	L	9,459	C	16,681
6	P	2,049	G	1,669	A	14,730	S	15,262	C	6,852	K	15,459
7	K	1,910	C	1,185	K	14,448	C	15,224	Q	6,835 (6.0)	Q	12,156 (5.9)
8	D	1,292	Q	1,107 (3.6)	T	12,560	K	14,518	K	6,122	A	9,587
9	G	961	L	1,087	C	10,067	A	9,124	R	4,061	T	5,667
10	A	916	V	851	L	9,331	R	7,501	A	3,244	L	5,646
11	L	554	K	680	R	6,847	D	6,950	D	3,176	R	5,165
12	C	256	N	486 (1.6)	H	6,302	N	2,606 (1.3)	Y	2,315	H	3,189
13	R	204	R	425	D	5,695	H	2,361	N	1,259 (1.1)	V	2,964
14	H	195	F	257	N	5,690 (2.4)	L	1,352	H	1,044	D	2,085
15	M	163	H	238	V	2,651	F	827	T	697	N	1,714 (0.8)
16	F	94	M	217	Y	1,179	M	746	V	549	F	1,433
17	V	90	D	127	M	915	Y	692	M	287	M	1,081
18	Y	33	Y	60	I	798	V	608	F	221	I	924
19	W	0	I	0	F	667	I	404	W	162	Y	617
20	I	0	W	0	W	147	W	42	I	16	W	541
	Total	37,620	Total	30,627	Total	240,198	Total	203,662	Total	114,427	Total	205,182

Table 2 *(Continued)*

Abundance of biased regions that have biases at the same level as the Q and N biases in the four budding-yeast prions

(c) Ratios of numbers for Q-rich and N-rich biased regions compared to the ratios of their overall abundances as residues§

	Budding yeast	Fission yeast	Fruit fly	Nematode	*Arabidopsis*	Human
$R_{Q/N}$ (total residues)	1.05	2.28	6.89	6.96	5.43	7.09
$R_{Q/N}$ (total regions)	1.42	4.25	5.03	6.75	5.46	11.73
Q/N (composition)	(0.039/0.061) = 0.64	(0.038/0. 052) = 0.73	(0.052/0 .047) = 1.12	(0.041/ 0.049 = 0.83	(0.035/ 0.044 = 0.79	(0.047/ 0.037 = 1.28

*The total numbers of regions that have a compositional biased LPS with P_{bias} < 1 × 10-13. †The number of LPSs for a particular compositional bias in the budding yeast proteome that overlap a region assigned as coiled coil by the MULTICOIL program [22]. ‡The total numbers of bias residues (for example, total number of serines for a serine bias) for all of the regions tallied for part (a) of the table. § $R_{Q/N}$ is the ratio of the number of Q-rich regions to N-rich regions as listed in parts (a) and (b) of the table. The overall abundance of the residues is simply the fraction of the total proteome that is either Q or N.

(Q+N)-rich domains

We derived a list of (Q+N)-rich domains using Q, N, and {Q+N} compositional bias in combination (Table 4, see footnotes for details of P_{bias} thresholds used). The longest LPS was chosen to define the domain where any of the three LPSs overlap substantially (a threshold of 15 residues was found to be suitable). There are up to approximately 170 such (Q+N)-rich domains in budding yeast. Most strikingly, we note that (Q+N)-rich domains are relatively rarer in fission yeast, with a comparatively large number in fruit fly (Table 4). The four known budding-yeast prions have biases against the major hydrophobics {VILM} and charged residues {DERK} (Table 1). When these negative biases are accounted for, the number of (Q+N)-rich domains in budding yeast reduces by half to around 100 (Table 4). This may be due to selection against amyloidogenesis mechanisms, where such bias is used for a different reason (perhaps in some cases as part of a coiled coil, see above). Subsidiary biases for glycine (G), tyrosine (Y) and serine (S) occur for three of the four yeast prions (Table 1). When these are accounted for, a substantial number (30) still remains (Table 4). The thresholds used in Table 4 are derived from the highest P_{bias} values for the LPSs of any yeast prion sequence (rounded up to two significant figures) (Table 4). These observations on subsidiary biases demonstrate the diversity of (Q+N)-rich domains in eukaryotes, showing that about half of them have other biases that are predicted to be incompatible with prion-like amyloidogenesis mechanisms (Table 4).

[PIN+] is a non-Mendelian inherited trait required for the *de novo* appearance of the [PSI+] prion in budding yeast [13]. A recent study derived a list of nine candidate genes responsible for the [PIN+] phenomenon [13]. Seven of these nine are in found in the (Q+N)-rich domain list here. With regard to the other two, one (*PIN2*, YOR104W) has a notable subsidiary bias for Y (11 in 51 residues, $P_{bias} = 9.0 \times 10^{-7}$), and the other (*STE18*, YJR086W) has a very short Q-rich region (12 in 25 residues, $P_{bias} = 3.9 \times 10^{-11}$).

To characterize the (Q+N)-rich domains further, we examined their lengths (Figure 1), and also their prevalent gene Ontology (GO) annotations [27] for the proteins that contain them (Table 5), focusing on budding yeast, fruit fly and human. The GO annotations can be considered as 'keywords' that give an indication of the biological role of the (Q+N)-rich domains (Table 5). The distribution of lengths for the regions with (Q+N) bias varies markedly from organism to organism, with humans having the largest proportion of very long regions with (Q+N) bias (44% > 275 residues; see Figure 1 legend). The fly (Q+N)-rich regions tend to be short, like those in budding yeast (see Figure 1 legend). They have a large proportion (around 18%) that localize to the nucleus, with some of these appearing to be related to transcription (Table 5). In budding yeast, the distribution of GO compartment annotations for proteins with (Q+N)-rich domains shows that these sequences occur most often in the nucleus (23 annotations), in preference to the cytoplasm (16), and the plasma membrane (9). Those that are placed in the nucleus tend to be transcription factors (see function categories in Table 5). Along with transcription, the preferred processes for proteins with (Q+N)-rich domains are 'endocytosis', 'pseudohyphal growth' and 'nuclear pore organization'.

Some overarching perspectives on biased regions

To put our case study of Q- or N-rich regions in a general context, we will discuss some overarching perspectives on compositional bias in the eukaryotic proteomes. The behavior of all 20 single-residue biases as a function of decreasing P_{bias} in the proteomes of budding yeast, fission yeast, nematode worm, fruit fly, *Arabidopsis* and human was examined. The

Table 3

Comparison of prevalent compositionally biased regions for the whole proteome, translated intergenic DNA, known proteins, hypothetical proteins and dORFs in budding yeast

(a) Proteome

$P_{bias} < 1 \times 10^{-5}$		$P_{bias} < 1 \times 10^{-9}$		$P_{bias} < 1 \times 10^{-13}$	
S	37,006	S	18,502	S	10,630
E	21,163	E	9,147	T	5,900
L	18,064	T	6,836	E	4,704
K	17,067	N	6462 (9.3)	Q	3,924 (10.4)
N	15,577 (7.4)	Q	5,212 (7.5)	N	3,745 (10.0)
A	13,974	K	4,280	P	2,049
G	12,927	P	3,831	K	1,910
D	10,004	L	3,512	D	1,292
P	9,892	D	3,176	G	961
T	9,866	A	2,473	A	916
F	8,934	G	2,115	L	554
Q	8,689 (4.1)	C	810	C	256
I	6,939	F	764	R	204
R	5,333	H	662	H	195
V	4,121	R	509	M	163
C	3,293	I	264	F	94
Y	2,960	Y	262	V	90
H	2,645	M	245	Y	33
W	2,009	V	150	W	0
M	850	W	0	I	0
Total	211,313	Total	69,212	Total	37,620

(b) Translated igDNA*

$P_{bias} < 1 \times 10^{-5}$		$P_{bias} < 1 \times 10^{-9}$		$P_{bias} < 1 \times 10^{-13}$	
F	28,949	F	5,692	F	1,211
C	10,074	C	1,280	H	602
K	7,800	H	908	V	490
R	7,551	V	814	T	448
Y	6,450	K	753	C	377
L	6,283	Y	690	L	366
I	3,789	T	681	Y	282
H	3,157	P	675	P	243
P	1,650	R	594	S	222
S	1,613	L	576	K	186
V	1,566	S	380	I	185
T	1,299	G	380	R	178
G	1,136	I	353	N	173 (3.2)
N	798 (0.9)	W	299	G	166
W	746	N	242 (1.7)	W	98
Q	498 (0.6)	Q	125 (0.9)	Q	51 (1.0)
M	282	E	85	E	39
A	268	M	26	D	16
E	241	D	16	M	15
D	16	A	0	A	0
Total	84,166	Total	14,569	Total	5,348

Table 3 *(Continued)*

Comparison of prevalent compositionally biased regions for the whole proteome, translated intergenic DNA, known proteins, hypothetical proteins and dORFs in budding yeast

(c) Known yeast proteins†

$P_{bias} < 1 \times 10^{-5}$		$P_{bias} < 1 \times 10^{-9}$		$P_{bias} < 1 \times 10^{-13}$	
S	27,539	S	15,328	S	9,819
E	17,519	E	8,074	T	5,900
L	13,928	N	5,716 (9.9)	E	4,289
K	13,785	T	5,413	N	3,551 (11.9)
N	12,854 (7.7)	Q	4,520 (7.8)	Q	3,348 (11.3)
A	12,482	K	3,653	K	1,723
G	11,783	L	2,864	P	1,669
D	1,934	L	595	P	170
P	1,883	P	453	G	62
Q	7,299 (4.4)	A	2,434	G	899
P	7,045	G	1,969	L	451
F	6,154	C	608	C	207
I	5,495	H	530	H	162
R	3,973	R	447	R	155
V	3,415	F	443	M	113
C	2,400	I	264	F	78
Y	2,158	Y	218	V	0
H	1,536	M	195	Y	0
W	1,484	V	60	W	0
M	656	W	0	I	0
Total	166,920 (13)	Total	57,938 (38)	Total	33,070 (19)

(d) Hypothetical yeast proteins†

$P_{bias} < 1 \times 10^{-5}$		$P_{bias} < 1 \times 10^{-9}$		$P_{bias} < 1 \times 10^{-13}$	
S	8,621	S	2,958	T	1,240
L	3,905	T	1,423	S	772
E	3,630	E	1,073	Q	576 (13.7)
K	3,043	Q	680 (6.8)	E	415
F	2,747	N	664 (6.6)	D	262
N	2,506 (6.4)	K	602	N	194 (4.6)
T	2,050	D	600	K	187
D	1,934	L	595	P	170
P	1,883	P	453	G	62
A	1,386	F	321	V	55
I	1,267	C	202	M	50
R	1,264	G	146	L	50
Q	1,171 (3.0)	H	106	R	49
G	882	R	62	C	49
C	863	V	55	Y	33
H	528	M	50	H	33
W	514	Y	44	F	16
Y	512	A	14	A	0
V	389	V	0	W	0
M	179	W	0	I	0
Total	39,274 (16)	Total	10,048 (221)	Total	4,213 (150)

Table 3 *(Continued)*

Comparison of prevalent compositionally biased regions for the whole proteome, translated intergenic DNA, known proteins, hypothetical proteins and dORFs in budding yeast

(e) dORFs

$P_{bias} < 1 \times 10^{-5}$		$P_{bias} < 1 \times 10^{-9}$		$P_{bias} < 1 \times 10^{-13}$	
R	459	R	254	R	254
H	307	L	204	L	204
S	288	T	138	H	122
G	271	Q	129 (11.0)	T	120
L	248	H	122	C	99
Q	225 (6.8)	C	99	Q	74 (8.3)
T	208	S	82	N	23 (2.6)
N	172 (5.2)	P	72	A	0
F	168	Y	50	D	0
C	163	N	23 (2.0)	E	0
V	151	A	0	F	0
A	149	D	0	G	0
D	111	E	0	I	0
I	98	F	0	K	0
P	84	G	0	P	0
Y	67	I	0	S	0
E	45	K	0	V	0
K	37	M	0	Y	0
W	23	V	0	W	0
M	14	W	0	M	0
Total	3,288	Total	1,173	Total	896

*Translated igDNA ('intergenic DNA') is conceptually translated in six frames. For analysis of intergenic DNA in budding yeast, we used the 'Not Feature' file of sequences in FASTA format distributed by SGD (this contains all genomic DNA that does not overlap an annotated feature [32]). This set of nucleotide sequences was conceptually translated in all six reading frames, and the amino-acid compositional biases were tallied up as for the annotated budding-yeast proteome. A dORF is an open reading frame that is disrupted by one or more frameshifts or premature stop codons, and which is likely to be a pseudogene. A data set of dORFs has been derived previously for the budding-yeast genome [9]. †In the totals for known and hypothetical proteins, the number of bias residues per residue of protein is given in parentheses.

curves for seven selected residues are shown for the budding yeast, fission yeast, fruit fly and human proteomes (Figure 2). Each eukaryotic proteome has a characteristic profile of bias proportions (Figure 2). For budding yeast, serine (S) is an abundant bias regardless of the P_{bias} threshold (Figure 2). For lower P_{bias} values, less than 1×10^{-15}, these biases arise mainly from serine-rich mannoproteins that are involved in the cell wall (for example, FLO8 [28]). N and Q, however, are prevalent biases only for the lowest P_{bias} levels ($P_{bias} \leq 1 \times 10^{-13}$). In all the proteomes, biases for individual hydrophobic residues (for example, isoleucine (I) and leucine (L)) fall off at much milder levels of probability, although less so for leucine because of its involvement in coiled-coil regions (Figure 2). There are no I or tryptophan (W) biases at $P_{bias} = 1 \times 10^{-10}$ or lower for each of the eukaryotes. It is noticeable that cysteine (C) bias is maintained at relative abundance in the human proteome (Figure 2) to much lower P_{bias} levels than in the other eukaryotes studied; this arises from the occurrence of large tandem arrays of cysteine-rich domains that are disulfide-bridged (for example, epidermal growth factor-like domains [29] and/or metal-binding proteins (such as the zinc finger)).

Conclusions

We have carried out an analysis of (Q+N)-rich domains in the complete proteomes of six eukaryotes, using a simple formalism based on finding the LPSs for a given set of amino acids within a protein sequence. We were motivated to use LPSs by the fact that the four known (Q+N)-rich prion-determinant sequences in budding yeast (found previously by experiment) each correspond to an LPS.

Analysis of budding-yeast prion sequences

We have examined the characteristic biases of the four known budding-yeast prion sequences. Supplementary to the well-

Table 4

Numbers of (Q+N)-rich domains for the six proteomes

Category		Budding yeast	Fission yeast	Fruit fly	Nematode	*Arabidopsis*	Human
1	(Q+N)-rich domains according to a Q, N or Q+N bias*	172	22	853	315	213	194
2	(1) plus filter for charged and hydrophobic residues†	96	14	473	216	125	69
3	(2) plus requirement for a subsidiary bias for G, Y or S‡	31	7	86	80	35	21

*The total number of (Q+N)-rich domains. These are all LPSs that have a Q or N bias with $P_{bias} < 1 \times 10^{-13}$ or a {QN} bias with $P_{bias} < 1.8 \times 10^{-14}$. †A filter is used so that only LPSs that have a subsidiary bias against {DERK} with a $P_{bias} < 6.5 \times 10^{-3}$ and against {VILM} with a $P_{bias} < 2 \times 10^{-2}$ are considered. ‡A filter is used so that only LPSs that have a subsidiary bias for one of the residues G, Y or S with a $P_{bias} < 5 \times 10^{-4}$ are considered.

documented Q and N biases, there are also mild biases for Y, G and/or S in three of the prion-determinant sequences. A substantial fraction (30/172) of the (Q+N)-rich domains found in this survey for budding yeast have such a subsidiary bias. In particular, for Sup35p, the bias for Y is conserved in homologs in other fungi [18], and could potentially contribute to amyloid formation through π-stacking of aromatic groups [20]. It is likely that some Q- or N-rich regions may have subsidiary compositions that are there to decrease the likelihood of prion-like amyloidogenesis in higher eukaryotes; this would explain the large number of (Q+N)-rich domains that are deleted when a mild bias against charged and major hydrophobic residues is considered. Interestingly, the prion-determinant domains of three of the four prions correspond closely with the LPSs for the single most abundant residue types. For the fourth, New1p, the LPS corresponds to a triplet repeat $(NYN)_n$ that appears necessary for prion propagation [13].

Relative abundance of Q and N bias in eukaryotic proteomes

We examined the relative abundance of biases for all 20 residue types for the six different eukaryotic proteomes. When biases that are at least at the level of the Q and N biases observed for the yeast prion sequences are considered, regions with N bias are always less common than those with Q bias, and become substantially less favored for the human proteome (being 12 times rarer than Q-biased regions). Disfavoring of N-rich regions in the human proteome (and other mammalian proteomes) has also been observed for homopolymeric runs of sequence [30,31].

Occurrence of (Q+N)-rich regions

As a suitable standard, we determined a refined list of domains that are at least as biased as the budding-yeast prion domains (either in terms of Q, N, or {Q+N} compositional bias). Fission yeast appears to have rather fewer (Q+N)-rich domains than budding yeast, which may indicate a relative intolerance to Q/N-based 'polar zipper' oligomerization/polymerization [16]. In the fruit fly, the large number of apparent (Q+N)-rich domains tend to be as short as those in budding yeast, with about a fifth (around 18%) of them localizing to the nucleus, some of which are annotated as involved in transcription (by GO classification [27]).

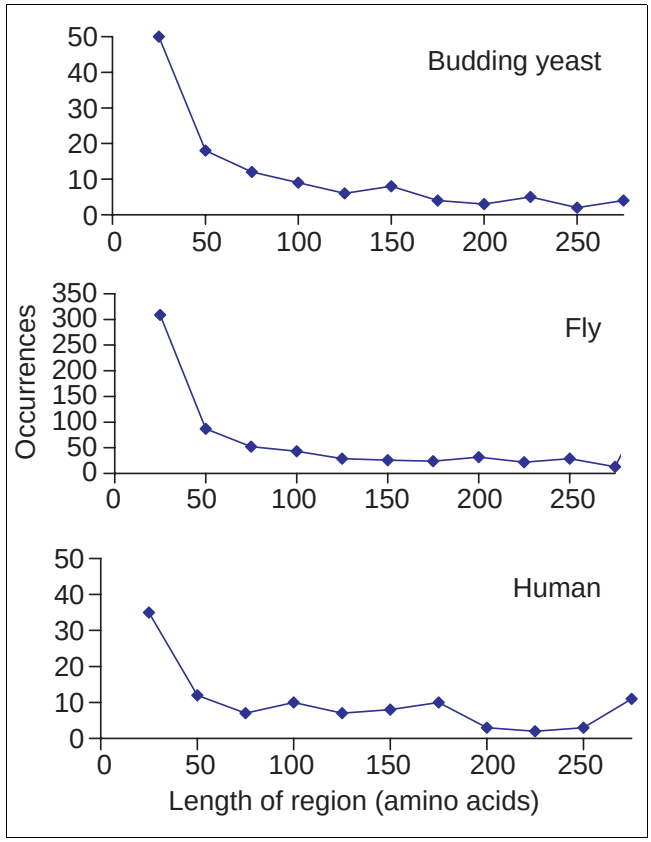

Figure 1
Histogram of the lengths of the (Q+N)-rich domains for budding yeast, fruit fly and human. The distribution of sequence lengths for the (Q+N)-rich domains are shown for budding yeast (top panel), fruit fly (middle panel) and human (bottom panel). The *y*-axis is the number of regions per bin, and the *x*-axis is for bins with labels *x* such that each bin contains all sequences with length *x* to *x* + 24 inclusive. The mean and median lengths for each of these distributions are as follows (organism, mean (± SD), median): budding yeast, 209 ± 209, 116; fruit fly, 236 ± 389, 89; human, 553 ± 730, 268. Only the distributions up to bin *x* = 275 are shown; a sizeable proportion of each distribution is longer than 275 residues (budding yeast 30% of sequences, fruit fly 22% and human 44%).

Table 5

Functional categories for the (Q+N)-rich domains for budding yeast fruit, fly and human

Organism	GO ontology	Five most frequent category annotations*
Budding yeast	Component	Nucleus GO:0005634 (23), Cytoplasm GO:0005737 (16), cellular_component_unknown GO:0008372 (14), Plasma_membrane GO:0005886 (9), actin_cortical_path GO:0005857 (8), nuclear pore GO:0005643 (6)
	Function	Molecular_function_unknown GO:0005554 (59), transcription_factor GO:0003700 (19), cytoskeletal_adaptor GO:0008093 (7), general transcriptional repressor GO:0016565 (6), general_RNA_polymerase_II_transcription_factor GO:0016251 (6), structural molecule GO:0005198 (6)
	Process	Biological_process_unknown GO:0000004 (52), endocytosis GO:0006897 (10), pseudohyphal_growth GO:0007124 (9), transcription GO:0006350 (9), nuclear pore organization GO:0006999 (8), protein amino acid phosphorylation GO:0006468 (7), regulation of cell cycle GO:0000074 (7)
Fly	Component	Nucleus GO:0005634 (157), TFIID_complex GO:0005669 (13), plasma_membrane GO:0005886 (19), cytoplasm GO:0005737 (23), microtubule_associated_protein GO:0005875 (9)
	Function	RNA_polymerase_II_transcription_factor GO:0003702 (52), transcription_factor GO:0003700 (39), specific_RNA_polymerase_II_transcription_factor GO:0003704 (36), RNA binding GO:0003723 (30), general RNA polymerase II transcription factor GO:0016251 (17)
	Process	Notch receptor signaling pathway GO:0007219 (18), protein amino acid phosphorylation GO:0006468 (18), transcription initiation GO:0006367 (13), gene silencing GO:0016458 (9), neuroblast determination GO:0004725 (9)
Human	Component	Nucleus GO:0005634 (52), integral membrane protein GO:0016021 (9), extracellular space GO:0005615 (9), plasma membrane GO:0005887 (7), cytoskeleton GO:0005856 (7).
	Function	Transcription_factor GO:0003700 (22), GO:0003677 DNA binding (20), calcium ion binding GO:0005509 (11), ATP binding GO:0005524 (10), transcription coactivator GO:0003713 (10)
	Process	Regulation of transcription GO:0006355 (34), signal transduction GO:0007165 (15), protein amino acid phosphorylation GO:0006468 (7), transcription from PolII promoter GO:0006366 (7), oncogenesis GO:0005198 (7)

*A description of each GO category is followed by the number in the ontology and the total number of such designations found, in brackets.

The analysis of Q/N can be of use to those studying the prion phenomenon in budding yeast and aggregation/amyloidogenesis in the eukaryotic cell. The data for (Q+N)-rich domains is available [21]. We are not suggesting that these regions are indeed prion-like; on the contrary, we have shown the diversity and abundance of these domains in a genomic context, and that they can have a variety of functions, compartments, and importantly, that they very often have subsidiary biases which would be disruptive to prion-like amyloidogenesis. The main results here on Q/N bias are robust to the underlying probability model, as a uniform frequency expectation for amino acids (all $f_x = 0.05$ in Equation 1, see Materials and methods) produces essentially the same trends (see, for example, Additional data file 2 or Supplementary Table B at [21]).

Some general perspectives on compositional bias

To put this 'case study' of Q/N bias in context, we have also presented some results that offer a more general perspective on the phenomenon of compositional bias. We found that the prevalence of different biases as a function of P_{bias} is characteristic for each proteome (Figure 2); however, there are some common trends, such as a disfavorment of regions with pronounced biases for I or W. For budding yeast, compositional biases extracted from conceptually translated igDNA, and for disabled ORFs (so-called dORFs) are very different from those for the annotated proteome. Also, there is surprisingly little difference between the prevalent biases observed for the approximately 2,000 annotated 'hypothetical' proteins and the approximately 4,000 known proteins in the budding-yeast proteome. However, biased regions are substantially more common (at some bias levels more than eight times as common) for known proteins than for hypothetical proteins. These observations may be applicable to gene prediction and verification.

The algorithm presented here can be developed for other investigations of compositional bias, for structural genomics, and for topics in protein folding and design, and is also readily applicable to nucleic acid sequences.

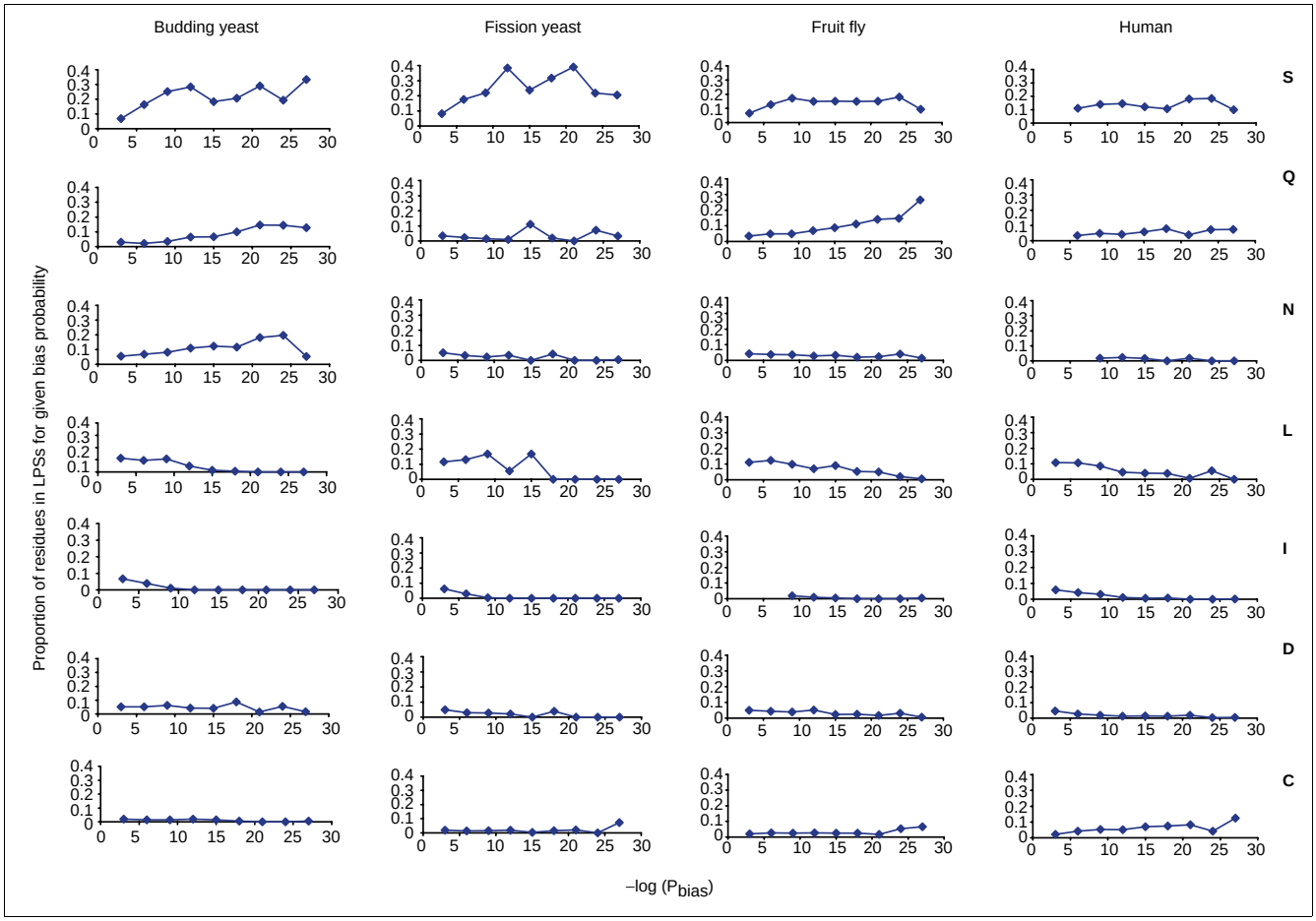

Figure 2
Each proteome has a characteristic distribution of biases. The proportion of bias residues (*y*-axis) counted up for each of the following seven residues (S, Q, N, L, I, D, C) are shown as a function of the bias probability (*x*-axis). The *x*-axis comprises bins labeled with -log(P) such that all regions with probabilities from -log(P) to 3.0 -log(P) are included. The end (right-most) bin includes all regions with log probability greater than -log(P). From left to right, the first set of panels is for budding yeast, the second set for fission yeast, the third set for fruit fly and the fourth for human. The rows of panels are labeled at the far right with the appropriate one-letter amino-acid symbol (S, Q, N, L, I, D and C).

Materials and methods

Calculating regions that are Q+N-rich or have other biases

Six complete eukaryotic proteomes were downloaded from the web: budding yeast (*Saccharomyces cerevisiae* from the SGD [32]), nematode worm (*Caenorhabditis elegans* Wormpep25 [33]), fruit fly (*Drosophila melanogaster* [34]), mustard weed (*Arabidopsis thaliana* [35]) and the Ensembl data set for human [36]). In each protein sequence of these proteomes, we searched for biased regions for each of the 20 amino-acid types as follows. For each individual amino-acid type x, and for the range of window sizes (w) from 25 residues to 2,500 residues, we searched each protein sequence for segments that have compositional bias of the lowest probability ($P_{bias,min}$):

$$P_{bias,min} = \min P(i,w) \text{ for all } i \text{ and } x \quad (1)$$

where i is each possible start position for a window w in the sequence. The probability $P(i,w)$ is given by a binomial distribution:

$$P(i,w) = [\{w!/[n!(w - n)!]\}.(f_x)^n.(1-f_x)^{w-n}] \quad (2)$$

where f_x is the proportion of amino-acid type x in all of the sequences of the proteome taken together (or a uniform expected proportion for each amino acid = 0.05). The count for x is denoted n in the window w starting at position i. Such segments with $P_{bias,min}$ are termed LPSs. Once an initial LPS is found in a protein sequence, the remainder of the sequence is resubmitted to the procedure until no further LPSs can be found. This is somewhat similar to the procedure in the program SEG for assignment of low-complexity or compositionally biased regions (which is based on the calculation of sequence information entropy), and which also determines

an LPS [37]. To save on computational time, an initial filtering is applied (before using the procedure described above) using pre-computed threshold tables for each window length for all residue types for a fixed relatively high probability value (P_{bias} = 0.001 was found to be suitable).

This procedure differs from those previously reported as it allows for calculation of biases both for and against amino-acid types, while allowing calculation of subsidiary biases for any predefined sequence or subsequence [37–39]. We applied this formalism, because we noted that prion-determinant domains for the budding-yeast prions correspond closely to LPSs. The results and trends for Q/N biases reported in this paper use f_x values derived from the eukaryotic proteomes, but do not differ substantially if a uniform probability model for residue bias is used (all residues having f_x = 0.05; see [21]).

Calculating biases for any set of amino acids

Equation (2) can be generalized to calculate a bias for any set of amino acids {*xyz*...}, by summing up the number of residues over the whole set. This is studied in particular for the sets {QN}, {DERK} (charged residues) and {VILM} (major hydrophobics). As for single-residue biases, the LPSs for a sequence are identified.

GO annotations

Annotations for GO categories [27] for the eukaryotic proteomes were downloaded from the Gene Ontology website [40] and counted up as lists of keywords indicative of biological role.

Additional data files

The abundance of biases counted up in different ways for different bias probability thresholds is available in Additional data file 1. A table showing the number of biased regions for all the eukaryotes (for a uniform probability model) is available in Additional data file 2. The coordinates of (gln+asn)-rich domains are available for the following organisms: *S. cerevisiae* (Additional data file 3) *S. pombe* (Additional data file 4), *C. elegans* (Additional data file 5), *Arabidopsis* (Additional data file 6), *Drosophila* (Additional data file 7) and human (Additional data file 8). The format for each of these files is as follows: field #1 = name, field #2 = sequence length, field #3 = bias (Q or N or {QN}), field #4 = number of bias residues, field #5 = start of QN-rich region, field #6 = end of QN-rich region, field #7 = probability of bias (see manuscript for details).

The sequences of the proteomes can be found in *S. cerevisiae* (Additional data file 9), *S. pombe* (Additional data file 10), *C. elegans* (Additional data file 11), *Arabidopsis* (Additional data file 12), *Drosophila* (Additional data file 13) and human (Additional data file 14). All Additional data files are also available at [21].

Acknowledgements

We thank Tricia Serio (Brown University) for discussions about budding-yeast prions and comments on the manuscript.

References

1. Harrison PM, Bamborough P, Daggett V, Prusiner SB, Cohen FE: **The prion folding problem.** *Curr Opin Struct Biol* 1997, **7:**53-59.
2. Harrison PM, Chan HS, Prusiner SB, Cohen FE: **Thermodynamics of model prions and its implications for the problem of prion protein folding.** *J Mol Biol* 1999, **286:**593-606.
3. Serio TR, Lindquist SL: **Protein-only inheritance in yeast: something to get [PSI+]-ched about.** *Trends Cell Biol* 2000, **10:**98-105.
4. Wickner RB, Taylor KL, Edskes HK, Maddelein ML: **Prions: portable prion domains.** *Curr Biol* 2000, **10:**R335-R337.
5. Wickner RB, Taylor KL, Edskes HK, Maddelein ML, Moriyama H, Roberts BT: **Prions of yeast as heritable amyloidoses.** *J Struct Biol* 2000, **130:**310-322.
6. Sondheimer N, Lopez N, Craig EA, Lindquist S: **The role of Sis1 in the maintenance of the [RNQ+] prion.** *EMBO J* 2001, **20:**2435-2442.
7. Osherovich LZ, Weissman JS: **Multiple Gln/Asn-rich prion domains confer susceptibility to induction of the yeast [PSI(+)] prion.** *Cell* 2001, **106:**183-194.
8. Serio TR, Lindquist SL: **[PSI+]: an epigenetic modulator of translation termination efficiency.** *Annu Rev Cell Dev Biol* 1999, **15:**661-703.
9. Harrison P, Kumar A, Lan N, Echols N, Snyder M, Gerstein M: **A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution.** *J Mol Biol* 2002, **316:**409-419.
10. Speransky VV, Taylor KL, Edskes HK, Wickner RB, Steven AC: **Prion filament networks in [URE3] cells of *Saccharomyces cerevisiae*.** *J Cell Biol* 2001, **153:**1327-1336.
11. Sondheimer N, Lindquist S: **Rnq1: an epigenetic modifier of protein function in yeast.** *Mol Cell* 2000, **5:**163-172.
12. Michelitsch MD, Weissman JS: **A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions.** *Proc Natl Acad Sci USA* 2000, **97:**11910-11915.
13. Derkatch IL, Bradley ME, Hong JY, Liebman SW: **Prions affect the appearance of other prions: the story of [PIN(+)].** *Cell* 2001, **106:**171-182.
14. Perutz MF, Finch JT, Berriman J, Lesk A: **Amyloid fibers are water-filled nanotubes.** *Proc Natl Acad Sci USA* 2002, **99:**5591-5595.
15. Perutz MF, Windle AH: **Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats.** *Nature* 2001, **412:**143-144.
16. Perutz M: **Polar zippers: their role in human disease.** *Protein Sci* 1994, **3:**1629-1637.
17. Taylor WR: **The classification of amino acid conservation.** *J Theor Biol* 1986, **119:**205-218.
18. Santoso A, Chien P, Osherovich LZ, Weissman JS: **Molecular basis of a yeast prion species barrier.** *Cell* 2000, **100:**277-288.
19. DePace AH, Santoso A, Hillner P, Weissman JS: **A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion.** *Cell* 1998, **93:**1241-1252.
20. Gazit E: **A possible role for pi-stacking in the self-assembly of amyloid fibrils.** *FASEB J* 2002, **16:**77-83.
21. **Additional data** [http://bioinfo.mbb.yale.edu/~harrison/qn]
22. Wolf E, Kim PS, Berger B: **MultiCoil: a program for predicting two- and three-stranded coiled coils.** *Protein Sci* 1997, **6:**1179-1189.
23. Singh M, Berger B, Kim PS: **LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins.** *J Mol Biol* 1999, **290:**1031-1041.
24. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, *et al.*: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2000, **28:**37-40.
25. Mar Alba M, Santibanez-Koref MF, Hancock JM: **Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process.** *J Mol Evol* 1999, **49:**789-797.
26. Tautz D, Trick M, Dover GA: **Cryptic simplicity in DNA is a**

major source of genetic variation. *Nature* 1986, **322:**652-656.
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
28. Liu H, Styles CA, Fink GR: ***Saccharomyces cerevisiae* S288C has a mutation in *FLO8*, a gene required for filamentous growth.** *Genetics* 1996, **144:**967-978.
29. Harrison PM, Sternberg MJ: **The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds.** *J Mol Biol* 1996, **264:**603-623.
30. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes and disease associations.** *Proc Natl Acad Sci USA* 2002, **99:**333-338.
31. Kreil DP, Kreil G: **Asparagine repeats are rare in mammalian proteins.** *Trends Biochem Sci* 2000, **25:**270-271.
32. Chervitz SA, Hester ET, Ball CA, Dolinski K, Dwight SS, Harris MA, Juvik G, Malekian A, Roberts S, Roe T, *et al.*: **Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure.** *Nucleic Acids Res* 1999, **27:**74-78.
33. *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: A platform for investigating biology.** *Science* 1998, **282:**2012-2018.
34. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, Venter JC: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
35. *Arabidopsis* Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
36. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
37. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266:**554-571.
38. Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S: **Methods and algorithms for statistical analysis of protein sequences.** *Proc Natl Acad Sci USA* 1992, **89:**2002-2006.
39. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA: **CAST: an iterative algorithm for the complexity analysis of sequence tracts.** *Bioinformatics* 2000, **16:**915-922.
40. **Gene Ontology Consortium** [http://www.geneontology.org]