A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions

Melissa D. Michelitsch* and Jonathan S. Weissman*†‡

*Department of Cellular and Molecular Pharmacology, †Howard Hughes Medical Institute, 513 Parnassus Avenue, University of California, San Francisco, CA 94143-0450

Edited by David S. Eisenberg, University of California, Los Angeles, CA, and approved August 25, 2000 (received for review June 2, 2000)

Glutamine/asparagine (Q/N)-rich domains have a high propensity to form self-propagating amyloid fibrils. This phenomenon underlies both prion-based inheritance in yeast and aggregation of a number of proteins involved in human neurodegenerative diseases. To examine the prevalence of this phenomenon, complete proteomic sequences of 31 organisms and several incomplete proteomic sequences were examined for Q/N-rich regions. We found that Q/N-rich regions are essentially absent from the thermophilic bacterial and archaeal proteomes. Moreover, the average Q/N content of the proteins in these organisms is markedly lower than in mesophilic bacteria and eukaryotes. Mesophilic bacterial proteomes contain a small number (0–4) of proteins with Q/N-rich regions. Remarkably, Q/N-rich domains are found in a much larger number of eukaryotic proteins (107–472 per proteome) with diverse biochemical functions. Analyses of these regions argue they have been evolutionarily selected perhaps as modular ''polar zipper'' protein–protein interaction domains. These data also provide a large pool of potential novel prion-forming proteins, two of which have recently been shown to behave as prions in yeast, thus suggesting that aggregation or prion-like regulation of protein function may be a normal regulatory process for many eukaryotic proteins with a wide variety of functions.

Amyloids are self-propagating, β-sheet-rich, fibrillar protein aggregates associated with a number of diseases (1). A variety of proteins with no apparent sequence similarities can form amyloids or related β-sheet-rich aggregates. For an important subset, however, aggregation is mediated by the presence of a glutamine/asparagine (Q/N)-rich domain. These aggregation-prone Q/N-rich sequences have been observed in two distinct contexts. First, expanded pure glutamine repeats are found in several proteins that cause neurodegenerative diseases (2, 3). Second, Q/N-rich domains are found in two *Saccharomyces cerevisiae* proteins, Sup35p and Ure2p, that are responsible for the prion-like inheritance of the yeast non-Mendelian factors [*PSI*$^+$] and [*URE3*], respectively (4–6). In both of these phenomena, an altered conformation of a protein results in the loss of its normal function and the acquisition of the ability to convert the normally functioning form of the protein to the abnormal (prion) form.

In the case of several neurodegenerative diseases involving the deposition of protein aggregates, abnormally long pure glutamine repeats result from the unstable expansion of CAG codons at the DNA level. In all of these diseases, which include Huntington's and Kennedy's diseases and several spinocerebellar ataxias, expanded glutamine repeats in the affected proteins form intranuclear neuronal inclusions (2, 3). Codon expansion beyond a threshold repeat length, typically 37 uninterrupted glutamines, causes aggregation and the onset of disease, both repeat length and disease severity often increasing with successive generations (7, 8).

In contrast to the codon expansion-related neurodegenerative diseases, the Q/N-rich regions of Sup35p and Ure2p are highly interspersed with other amino acids. Several lines of evidence suggest a causative relationship between Q/N content and the formation of prion aggregates. First, these Q/N-rich sequences form the minimal domains required to cause self-propagating aggregation and prion-like inheritance in both of these proteins, although they bear no other sequence similarity (5, 9–12). These Q/N-rich regions drive prion formation by causing self-propagating aggregation and loss of function of the normally soluble forms of these proteins. Second, Q/N content is highly conserved across Sup35p sequences from a number of distantly related yeasts and, as with the *S. cerevisiae* protein, these domains are necessary and sufficient for prion-based inheritance (13–15). Finally, mutations of many glutamines and asparagines in *S. cerevisiae* Sup35p, most often to charged residues, lead to prion curing and protein solubilization (4).

A number of observations from structural studies have led to a model that explains why Q/N-rich regions have such a high propensity to aggregate and/or form fibrils. X-ray diffraction studies of oligoglutamine fibrils have demonstrated that they, like other unrelated amyloids, appear to consist of a cross-β-helix formed with β strands radiating from the helical axis (16, 17). Although no high-resolution data are available for any amyloid fibrils, Perutz and coworkers have created a sterically reasonable model of polar side-chain interactions with main-chain amides in the context of antiparallel β-sheet structures (16).

High Q/N content can be sufficient to cause protein aggregation, and a subset of these Q/N-rich aggregating proteins is capable of stably propagating in a prion-like manner. Therefore, it should be possible to predict novel prion- and/or aggregate-forming proteins on the basis of their Q/N content. In this study, an algorithm that identifies proteins with Q/N-rich regions was devised. By using this algorithm, all of the translated ORFs containing Q/N-rich regions across the available completed proteomic sequences were identified. These data have provided evidence for a conserved function of Q/N-rich domains in eukaryotes, perhaps as a mediator of specific protein–protein interactions. Furthermore, they provide a pool of potential yeast prions and/or amyloid-forming proteins, a few of which have already been tested and shown to act as prions and/or aggregates (13, 18).

Methods

Sequences and Homology Identification. The complete proteomic sequences and the GENPEPT nonredundant protein sequence database for all of the organisms examined in this study were obtained as FASTA format files from either the National Center for Biotechnology Information (NCBI) or the European Bioinformatics Institute. The incomplete proteomic sequences were obtained from SwissProt. Homologies were identified by using PSI-BLAST at the NCBI web site by using the default settings (19).

This paper was submitted directly (Track II) to the PNAS office.

‡To whom reprint requests should be addressed. E-mail: jsw1@itsa.ucsf.edu.

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.

Because of the high Q/N content of the proteins being studied, a filter for low-complexity regions was always used, thereby reducing the probability of predicting false homologues.

Biochemical functions for the yeast proteins identified were assigned on the basis of their classifications at the Yeast Protein Databank (YPD). When none was listed at YPD, putative functions were assigned on the basis of homologies determined by PSI-BLAST by using full-length proteins as query sequences.

Identification of Q/N-Rich Regions. The program DIANA moves through each ORF by single amino acid steps by using a window of arbitrary size determined by the user, returning the sequence of the most Q/N-rich region along with a number of statistics including its amino acid composition, the ORF title, size, and the amino acid composition of the full length protein. These records are then imported into a database for further sorting. When multiple regions occur within a protein that are equally Q/N rich, the algorithm returns the last one. Although high Q/N content was the target of the search described here, DIANA can be modified to identify regions rich in any amino acid or combination of amino acids. The code for DIANA can be obtained at http://itsa.ucsf.edu/~mmichel/DIANA. DIANA was written in PERL 5 and executed on an SGI Octane IRIX6 workstation.

Probability of Occurrence of Q/N Repeats and Q/N-Rich Regions. The probability of random occurrence of a Q/N-rich region was calculated by using the Poisson distribution:

$$f(i) = \frac{e^{-m}m^{i}}{i!},$$

where $f(i)$ is the probability of an event happening i times, and m is the mean percent occurrence of glutamine plus asparagine per predicted ORF in the yeast proteome. The value of i used is 30, the definition of a Q/N-rich region in this study. The value of m is 7.68, the average Q/N content per 80 mer. The resulting value of $f(i)$ implies that one in every 6.336×10^{10} 80 mers will have a Q/N content of at least 30. $f(i)$ was then multiplied by $n = 2.4477283 \times 10^{6}$, the number of consecutive 80 mers in yeast moving in single amino acid increments through each predicted ORF. The resulting value implies that the probability of a Q/N-rich domain occurring randomly given the Q/N content of *S. cerevisiae* is one in every 13,629 proteomes.

Results and Discussion

The Frequency of Q/N-Rich Regions in Eukaryotic, Archaeal, and Bacterial Proteomes. To predict candidates for novel prions or amyloidogenic proteins, an algorithm named DIANA (Defined Interval Amino acid Numerating Algorithm) was designed to identify proteins containing regions of consecutive amino acids with exceptionally high Q/N content. DIANA moves a window of arbitrary size through each translated ORF of a proteome by single amino acid steps. At each step, the total Q/N content within the window is determined. A database is then created containing the sequence of the most Q/N-rich region per protein, its amino acid composition, and the size and amino acid composition of the full length protein.

We used DIANA to examine the predicted translated ORFs for the three eukaryotic, six thermophilic archaeal, two thermophilic bacterial, and 20 mesophilic bacterial genomes available at the National Center for Biotechnology Information at the time of analysis (Fig. 1). In addition, the incomplete proteomic sequences from humans, mice, the plant *Arabidopsis thaliana*, and the GENPEPT nonredundant protein sequence database were examined. The sequences in these incomplete databases are biased toward commonly studied classes of proteins, and therefore they are not included in the following analyses unless specifically stated.

For the purpose of the present analyses, "Q/N-rich regions" are defined as 80 consecutive residues containing at least 30 glutamines and/or asparagines. We chose a window length of 80 amino acids because it is somewhat smaller than the known prion-forming domain of Sup35p (9–12), comparable to the size of prion-forming domain of Ure2p (5) and about twice the size of the minimum glutamine expansion needed to cause disease (2, 3). Therefore, long nonrepeating sequences rich in glutamine and asparagine can be identified without biasing the search against more concentrated and compact pure glutamine or asparagine repeats. For a region to be considered "Q/N-rich," we required that it have a minimum Q/N content of 30. This value was chosen because it is only one less than that of the most Q/N-rich 80 mer of the *Pichia pastoris* Sup35p homologue (13). Furthermore, this value lies beyond the expected distribution of Q/N content for the most Q/N-rich 80 mers for all of the organisms examined (Fig. 2 *a–d*; see below). This conservative definition of a Q/N-rich region allows us to focus on only the most promising candidates for proteins capable of forming prions and/or amyloids.

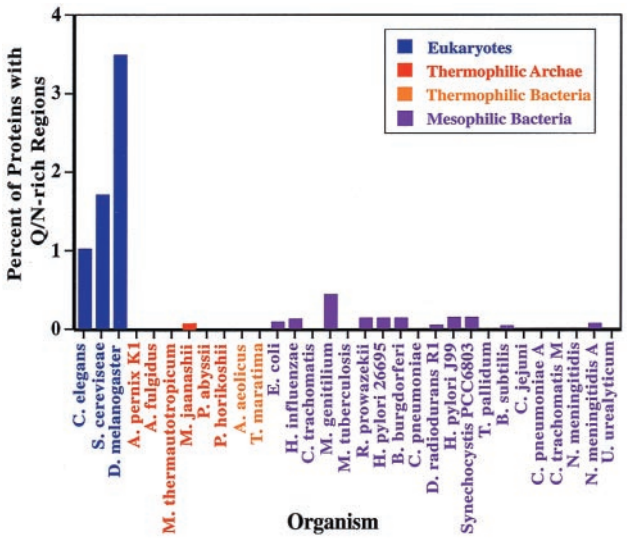

Fig. 1. Many eukaryotic proteins contain Q/N-rich regions. The percent of ORFs with a region of 80 consecutive amino acids containing a Q/N content of at least 30 was determined for the indicated proteomes.

Even with these relatively stringent criteria, a surprisingly large number of proteins with Q/N-rich regions were identified in eukaryotes (Figs. 1 and 2*a*). In *S. cerevisiae*, 107 polypeptides with Q/N-rich regions (1.69% of all ORFs) were identified, 143 were found in *Caenorhabditis elegans* (1.00% of all ORFs), and 472 were found in *Drosophila melanogaster* (3.47% of all ORFs).

A number of the eukaryotic sequences examined contained sequences with exceptionally high Q/N content. The region identified with the greatest total Q/N content was found in an adenylyl cyclase from *Dictyostelium discoideum*, with a total Q/N content of 72. This region also contains the highest glutamine content, with 71 glutamines and the longest uninterrupted glutamine repeat with 45 tandem glutamines. The proteins with the most asparagine-rich region and the longest asparagine repeat are also found in *Dictyostelium.* A prespore-specific protein-containing Q/N-rich region with 68 asparagines is the most asparagine-rich domain identified; the longest uninterrupted asparagine repeat consisted of 49 tandem asparagines and is found in a protein–tyrosine phosphatase.

All of the eukaryotic proteomes examined contained a subset of sequences with relatively pure glutamine or asparagine repeats reminiscent of the codon repeat disorders. Interestingly, none of these regions can be attributed solely to trinucleotide expansion, as all exhibit substantial codon variation. However, roughly three-fourths of the Q/N-rich sequences have a com-

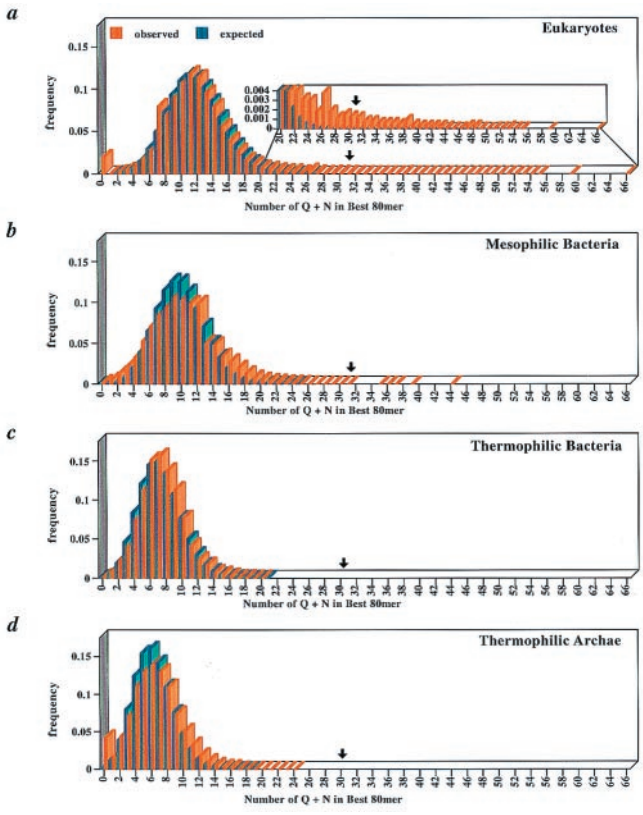

Fig. 2. Expected and observed distributions of Q/N content of the best 80 mers across mesophilic and thermophilic proteomes. A comparison of the expected and observed distributions of Q/N contents across the most Q/N-rich regions of (*a*) eukaryotes, (*b*) mesophilic bacteria, (*c*) thermophilic bacteria, and (*d*) archaea. Arrows indicate the minimum Q/N content required for a domain to be considered Q/N rich. (*a Inset*) Magnification of the observed and expected distributions of Q/N-rich regions. Expected values were calculated on the basis of the average Q/N content per best 80 mer for each of the four types of organisms.

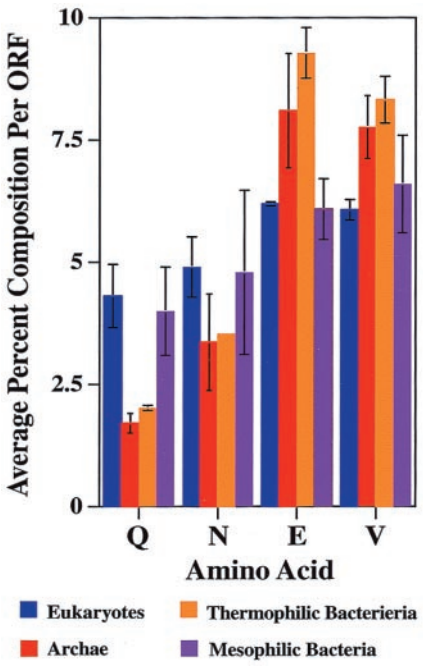

Fig. 3. Thermophiles have a lower Q/N content and higher glutamate and valine content than mesophiles. The average percent amino acid content per ORF per proteome was determined for the completed proteomes. Shown are glutamine (Q), asparagine (N), glutamate (E), and valine (V). The other amino acids showed little variation. Error bars indicate the observed variance among the indicated class.

position more similar to the yeast prion-forming domains in that they are interspersed with other amino acids.

In eleven of twenty of the mesophilic bacterial proteomes queried, Q/N-rich regions were present in a small number of proteins (one to four) (Figs. 1 and 2*b*). The other nine proteomes contain no Q/N-rich regions. In contrast to the eukaryotic Q/N-rich proteins that are predominantly cytoplasmic, a number of bacterial outer-membrane proteins were identified. None of these bacterial proteins contain long uninterrupted glutamine or asparagine repeats. Interestingly, over half of the bacterial Q/N-rich regions contain imperfect repeats of glutamine and/or asparagine in combination with other amino acids. Imperfectly repeating sequences have been suggested to promote and stabilize prion formation in the yeast protein Sup35p, although at present there is no evidence for prion-like behavior in bacteria (20).

In marked contrast to the eukaryotes and mesophilic bacteria, Q/N-rich regions are nearly absent from the proteomes of all of the thermophilic organisms tested, including all six archaea and both thermophilic bacteria (Figs. 1 and 2 *c* and *d*). Among these eight proteomes, only one predicted protein sequence was found that contained a Q/N-rich region. Furthermore, short tandem repeats of glutamine or asparagine residues rarely occur in thermophiles. The longest glutamine or asparagine repeat in three of the complete thermophilic proteomes examined was four residues long and only three residues long in the other five proteomes, whereas the longest consecutive glutamine or asparagine repeat is thirty-seven residues long and seven across all of the complete eukaryotic and mesophilic bacterial proteomes, respectively. These observations suggest that tandem Q/N repeats may be specifically selected against in thermophiles, whereas there is a strong positive selection for both Q/N-rich regions and tandem repeats in eukaryotes (see below).

Glutamine and Asparagine Are Less Abundant in Proteins from Thermophiles. One possible explanation for the lack of Q/N-rich domains in thermophiles might be that these amino acids are underrepresented in these organisms. Indeed, Haney and coworkers have noted that a set of proteins from thermophilic *Methanococcus* species tend to have fewer polar residues (serine, threonine, glutamine, and asparagine) and more charged residues (especially glutamate, aspartate, and lysine) than homologous proteins from mesophilic *Methanococcus* species (20). However, studies comparing the relative amino acid frequencies in proteins across entire proteomes of thermophiles and mesophiles have not previously been reported.

We determined the relative abundance of the different amino acids for all of the completely sequenced proteomes to determine whether the difference in the number of Q/N-rich domains found in thermophiles and mesophiles could simply result from differences in overall amino acid compositions (Figs. 2 and 3). Interestingly, the amino acid that differs the most in abundance between mesophiles and thermophiles is glutamine, with its relative frequency in thermophilic archaea and bacteria being less than half of that of the mesophilic bacteria and eukaryotes (1.85% vs. 4.31%). Asparagine is also on average less abundant in thermophiles, although the asparagine content shows greater variability in the archaea. By contrast, in thermophiles, glutamate is significantly enriched and valine is modestly enriched, 1.5- and 1.3-fold, respectively. There were no systematic differences in the relative abundance of the other amino acids (data not shown). Thus, although Q/N-rich regions would not have been expected to occur at random in either thermophilic or mesophilic organism, their chance occurrence in thermophiles is even less likely (Fig. 2 *a–d* and see below).

Why would thermophiles have evolved to have lower Q/N content than mesophiles? One possibility is chemical lability. Glutamine/asparagine deamination and glutamate/aspartate succinamide formation occur readily in peptides at 100°C (22). However,

there is evidence that these modifications rarely occur in well-ordered proteins. Thus, chemical instability should be cited as a reason for reduced Q/N content only if considered in the context of structure. Because Q/N-rich regions tend to form relatively unstructured loops in unliganded native proteins, they would be expected to be particularly susceptible to deamination (23). In addition to promoting chemical stability at high temperatures, the lower Q/N content could promote fold stability. Thermophilic proteins have been shown to have fewer loops than their mesophilic counterparts, presumably to attain a more compact structure to achieve greater thermostability (24). Finally, lower Q/N content might reduce protein aggregation. Although Q/N-rich regions are aggregation prone at lower temperatures as well, this property is likely to be exacerbated in thermophiles. In contrast, the presence of charged residues such as glutamate has been shown to reduce or prevent the aggregation of Q/N-rich regions (4).

Amino Acid Conservation with Codon Variation Suggests a Conserved Function for Q/N-Rich Regions. Although there are several plausible reasons why Q/N-rich regions and even short tandem repeats would be disfavored in thermophiles, the question remains whether there is appreciable selection for Q/N-rich regions in proteins from mesophilic organisms. The Poisson distribution function is used to predict the frequencies with which independent events will occur given a known mean frequency of occurrence. Deviation from this expected distribution is an indicator of nonrandom distribution (25). Poisson distributions were calculated for the eukaryotes, archaea, and mesophilic and thermophilic bacteria by using the average Q/N content of the best 80 mer from each protein within the proteomes as the expected mean. The distributions predicted from the Poisson function were then compared with the actual distribution of the Q/N contents of the best 80 mers across all of the proteomes examined (Fig. 2). In every case, the bulk of the distribution closely follows the expected distribution. However, for the mesophilic organisms, especially the eukaryotes, there are exceptionally long tails extending far beyond the expected values. Given the average Q/N content of *S. cerevisiae*, a consecutive 80 amino acid regions with a Q/N content of 30 would randomly occur only once in every 13,629 proteomes of the same size. Furthermore, the probability of a region having a Q/N content of 66 like the Q/N-rich region of Snf5p is vanishingly small (1 in 5.64×10^{32} proteomes). The probabilities of these very Q/N-rich regions occurring are so miniscule that it is extraordinarily unlikely that the presence of these domains is merely because of chance.

An alternative explanation for the presence of these domains is that they are the result of codon expansion rather than functional selection at the protein level. However, few of the polypeptides identified in this study contain long uninterrupted tracts of polyglutamine or polyasparagine. Moreover, the small number of repeat sequences that are present have codon variation. For example, the 37 consecutive glutamines of Snf5p are encoded by interspersed CAA and CAG codons with no more than eight consecutive CAGs and nine consecutive CAAs. This observation suggests that they are conserved sequences and are not merely trinucleotide expansions resulting from aberrant DNA metabolism. Finally, the presence of a large number of Q/N-rich regions that are enriched in both glutamine and asparagine suggests there is positive evolutionary selection for the chemical nature of their amido side chains.

Have these Q/N-rich regions been conserved because they play a common role in a particular class of proteins? The known or predicted biochemical functions of all of the proteins containing Q/N-rich regions in *S. cerevisiae* are listed in Table 1. (see supplementary material at www.pnas.org for a complete list of the ORFs identified.) Interestingly, proteins with a broad range of biochemical functions were found to contain Q/N-rich domains, although the identified proteins can be placed in functional clusters including transcription and translation factors, nucleoporins, DNA- and RNA-binding proteins, and proteins involved in vesicular trafficking. These observations indicate that the function of Q/N-rich domains is not specific to any single class of protein, but that their presence has been conserved across several of the diverse protein classes in which they occur.

The extreme improbability of these Q/N-rich regions occurring randomly, the lack of evidence for codon expansion, and the observation that their occurrence is conserved in a broad spectrum of proteins strongly support the conclusion that the Q/N content of these regions is because of positive evolutionary selection. Furthermore, because these regions are compact and can be found fused internally or at either terminus to a variety of different unrelated proteins, they are likely to behave as functionally conserved modular domains.

Identification of Potential Prion- and Amyloid-Forming Proteins. The two extensively characterized yeast prion phenomena were not discovered in directed screens for novel prions; rather, their existence was deduced only after decades of careful observation of their non-Mendelian inheritance (26, 27). Given the past lack of methodology for systematically identifying prions and the large pool of candidate Q/N-rich regions, it seems probable that significantly more prions remain to be identified.

It is likely that many of the proteins containing Q/N-rich regions will aggregate when overexpressed. Which of these will propagate stably, thus allowing prion-based inheritance, is less clear and needs be established experimentally on a case-by-case basis. Recently, Santoso and coworkers and, independently, Sondheimer and Lindquist described two similar genetic systems that should now allow the rapid testing of Q/N-rich domains for the ability to support stable prion-like inheritance (13, 18).

In our laboratory, two of the *S. cerevisiae* proteins identified by DIANA were tested for their ability to behave as prions and/or aggregates. The first protein examined, New1p, was previously uncharacterized but contains a Q/N-rich region with very similar imperfect repeats to those in the prion-forming domain of Sup35p, including the identical pattern YQQGGYXYN. The second protein tested, Pan1p, is involved in actin cytoskeletal organization that, apart from its Q/N richness, bears no similarity to either known yeast prion (28). Both were found to aggregate when overexpressed in *S. cerevisiae*, but with the experimental system used, only New1p was observed to behave as a prion (13) (L. Z. Osherovich and J.S.W., unpublished data).

DIANA also identified other sequences with strong similarities beyond high Q/N content to the prion-forming domain of the Sup35 protein. These proteins include YBR016, YDR210W, and Rnq1p. Indeed, while this study was being completed, Sondheimer and Lindquist demonstrated that a chimera consisting of the Q/N-rich domain of Rnq1p fused to the nonprion-forming domain of Sup35p is capable of exhibiting prion-like behavior (18). Moreover, the endogenous protein acts as a prion. In addition, they showed that a number of proteins with sequence similarities to Sup35p are capable of causing the aggregation of GFP *in vivo*. There are also many promising candidates whose Q/N-rich domains strongly resemble the Q/N-rich domain of Ure2p, including YIL130 and YLR278. Most of these domains contain long stretches of asparagine interspersed with other amino acids, particularly serine.

Although proteins with statistically significant similarities to Sup35p or Ure2p are especially promising candidates, it is likely that a wide array of Q/N-rich sequences can behave as prions, because Ure2p and Sup35p are not homologous to each other, and their prion-forming domains have very dissimilar amino acid compositions. For example, YOR197 is a predicted ORF of unknown function whose Q/N-rich region contains imperfect repeats of QQYG. These repeats are reminiscent of the imperfect PQGGYQQYN repeats present in Sup35p that have been shown to promote conversion to the prion state (20). In addition

Table 1. Proteins containing Q/N-rich regions in *S. cerevisiae* fall into a broad spectrum of functional clusters

Cluster	Protein	Q/N
Nonkinase signaling proteins		
	GPR1	55
	RPI1	35
	PHO81	34
Kinases		
	NPR1, ARP1	52
	CBK1	49
	YCK1	40
	YCK2	36
	HRR25	36
	APG13	36
	YAK1	33
	SKY1	32
	KSP1	32
	SCH9	31
	SKS1	31
	YBR059	38
DNA/RNA binding		
	MPT5, HTR1	31
	YPR042	51
	YBR180	34
	YGL014	34
	YLL013	33
RNA processing		
	NAB1	48
	LSM4	38
	PCF11	37
	NAB3	34
	PUB1	33
	RNA15	30
	PSP2	33
	NGR1	31
	YBL051	30
Nucleoporins		
	NUP116	36
	NUP100	32
	NUP57	31
	NUP49	30
Transcription		
	SNF5	66
	CYC8	54
	MCM1	49
	IXR1, ORD1	49
	MED2	48
	GAL11	47
	SWI11	43
	TAF61	43
	PDR1	38
	URE2	38
	DAL81	38
	MOT3, HMS1	38
	RLM1	36
	SNF2, GAM1	36
	SWI4	35
	AZF1	35
	CCR4	34
	CRZ1	32
	PDC2	32
	DAT1	31
	SFP1	30
	SPT20	30
	CDC39	30
	YIL130	55
	YEL007	46
	YDR409	37
	POP2	37
	YMR263	35
	EPL1	32
	TBS1	31
	YLR373	30
Translation		
	VAR1	40
	SUP35	38
	NEW1, YPL226	32
	TIF4632	30
Vessicular trafficing/secretory pathway		
	ENT2	48
	SEC61	45
	YAP1801	39
	PAN1	38
	SLA2	38
	ENT1	36
	SCD5	36
	FAB1	32
	ANP1	31
	YKL054	53
Outer membrane proteins		
	YDR213	44
	YPR022	42
	PSP1	31
	YLR177	30
Other		
	RNQ1	42
	CDC27	35
	MAD1	34
	GRR1	32
	JSN1	31
	SLF1	30
	VAC7	30
	YKR096	38
	YGL066	37
	YKL088	32
Unknown		
	MSS11	52
	HOT1	33
	YBR016	41
	YBL081	38
	YIL105	36
	YBR238	33
	YBL029	31
	YOR197	30
	YML053	30

S. cerevisiae proteins containing Q/N-rich regions were placed into functional groups on the basis of either known biochemical function or sequence homology to proteins of known function (indicated by italics).

to the presence of repeats, YOR197 is intriguing because, as in huntingtin, a proline-rich domain follows its Q/N-rich domain.

It will also be interesting to study the function and aggregation properties of some proteins independently of whether they also prove to be novel prions. For example, YGL066 is an uncharacterized protein with significant sequence homology ($E = 2 \times 10^{-6}$) to human ataxin-7, a protein that causes a human spinocerebellar ataxia in its mutant glutamine-expanded form (29). This homology is external to the Q/N-rich region, suggesting a true conserved biochemical function rather than a mere reflection of similar amino acid content in their Q/N-rich regions. Studying the function of this protein in yeast could help elucidate the currently unknown function of its disease-causing human counterpart.

The growing number of known yeast prions together with the large pool of Q/N-rich candidates raises the possibility that the formation of prion-like aggregates is a conserved and beneficial

cellular process that occurs in a broad range of proteins. Consistent with this proposal, recent studies reveal that the ability of Sup35p to support prion-based inheritance is strongly conserved across diverse species of budding yeast (13–15). In addition to the established role of [*PSI*$^+$] in regulating translation termination (6), Tuite and coworkers recently showed that in some strains, [*PSI*$^+$] yeast show an increased tolerance to thermal stress and high ethanol concentrations compared with isogenic strains lacking the [*PSI*$^+$] prion (30). Moreover, the [Het-s] prion phenomenon has a well-characterized role in regulating heterokaryon incompatibility in the fungus *Podospora anserina* (31).

Why might prion-based inheritance be advantageous to an organism? Prion formation provides a mechanism for a protein to inhibit its own activity by specifically self aggregating in response to overexpression or to other changes in the environment (30, 32). Potentially, a major advantage of prion-based functional inhibition would be that, unlike protein inactivation caused by DNA mutations, the inhibited prion-like state can be propagated indefinitely while retaining the ability to revert to the original functional state. Finally, because prion-forming domains and other Q/N-rich domains occur in proteins with a variety of functions, it is possible that a broad range of proteins can be regulated by the aggregation of Q/N-rich domains. In this regard, it is significant that prion-like aggregate formation is highly self specific, thereby allowing multiple different prion-states to propagate independently within a single cell (13).

The challenge now is to determine experimentally which of these Q/N-rich domains can behave as prions and/or aggregates. This information should help both in refining prion prediction algorithms and in understanding the physiologic role for prions and/or other self-specific protein aggregates.

Q/N-Rich Domains as Modulators of Specific Protein–Protein Interactions? It is clear that there has been a positive selective pressure for certain proteins to contain Q/N-rich domains. What function besides prion-based inheritance do these domains impart that would make their presence advantageous?

Perutz has suggested that Q/N-rich regions (and possibly regions rich in other polar residues) might behave as modular mediators of protein–protein interactions termed "polar zippers" because of the capacity of their side chains to form hydrogen bond networks (16). Supporting this hypothesis, there have been several recent experimental reports of domains containing Q/N-rich regions mediating specific protein–protein interactions. First, the glutamine-rich region in transcription factor Sp1 was shown by functional mapping *in vivo* to bind the dTAFII110 component of the *Drosophila* TFIID complex (33). Second, Pan1p and Sla1p, two proteins that act in complex to promote cytoskeletal organization in yeast, were shown by coimmunoprecipitation and two-hybrid analysis to interact via their domains containing Q/N-rich regions (28). Finally, The Sla1p Q/N-rich domain and Sup35p prion-forming domains interact in two-hybrid studies (34). Moreover, this interaction appears to influence the rate of conversion to the prion state by Sup35p. It will be important to directly demonstrate the role of glutamine and asparagine residues in stabilizing these interactions.

The roles of Q/N-rich domains in mediating both protein interactions and in promoting aggregation could be different manifestations of the same modular functionality. Indeed, it is possible that prion and/or aggregate formation is a conserved function of certain Q/N-rich regions that results from their ability to mediate protein–protein interactions in a self-propagating manner. By empirically examining the roles of the Q/N-rich domains identified in this study, the possible intersection of their functions as mediators of protein–protein interactions and the formation of both prion and nonprion aggregates can be elucidated.

Summary

By identifying all of the Q/N-rich regions across a number of completely sequenced proteomes, we have revealed that Q/N-rich domains are ubiquitous and modular, are associated with a variety of types of proteins, and have been evolutionarily conserved in a number of these proteins. This conservation suggests an independent functional role for these domains. Growing evidence suggests that Q/N-rich domains function as mediators of specific protein–protein interactions. We argue that their conserved ability to form self-specific prion and nonprion aggregates might be simply an extension of their ability to act as interaction domains by forming extensive, self-propagating, hydrogen-bonding networks.

In addition to providing evidence for conserved modular functions for Q/N-rich regions, the identification of Q/N-rich domains in eukaryotes has proven useful for identifying two novel prions and a nonprion-aggregating protein (13, 18) (L. Z. Osherovich and J.S.W., unpublished data). By using a new genetic system and biochemical methods, additional prions and amyloidogenic proteins can now be rapidly identified, enabling refinement of the criteria used to predict prion-like or amyloidogenic behavior. This information will facilitate the analysis of prion-based inheritance in normal cellular physiology. Finally, as the Human Genome Project is completed, the algorithm described here in combination with other experimental observations may be useful for identifying new proteins prone to forming disease-causing aggregates.

We thank P. Babbitt, P. Bosque, F. Cohen, A. DePace, I. Kuntz, J. Nguyen, L. Osherovich, A. Santoso, H. Sparrer, R. Stroud, and members of the Weissman, Cohen, and Babbitt groups for helpful discussions. This work was supported by the Searle Scholars Program, the David and Lucile Packard Foundation, the National Institutes of Health, and a National Science Foundation predoctoral fellowship (M.D.M.).

1. Koo, E. H., Lansbury, P. T., Jr. & Kelly, J. W. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9989–9990.
2. Perutz, M. F. (1999) *Trends Biochem. Sci.* **24,** 58–63.
3. Ross, C. A., Margolis, R. L., Becher, M. W., Wood, J. D., Engelender, S., Cooper, J. K. & Sharp, A. H. (1998) *Prog. Brain Res.* **117,** 397–419.
4. DePace, A. H., Santoso, A., Hillner, P. & Weissman, J. S. (1998) *Cell* **93,** 1241–1252.
5. Maddelein, M. L. & Wickner, R. B. (1999) *Mol. Cell. Biol.* **19,** 4516–4524.
6. Serio, T. R. & Lindquist, S. L. (1999) *Annu. Rev. Cell. Dev. Biol.* **15,** 661–703.
7. Stine, O. C., Pleasant, N., Franz, M. L., Abbott, M. H., Folstein, S. E. & Ross, C. A. (1993) *Hum. Mol. Genet.* **2,** 1547–1549.
8. Klockgether, T. & Evert, B. (1998) *Trends Neurosci.* **21,** 413–418.
9. Ter-Avanesyan, M. D., Kushnirov, V. V., Dagkesamanskaya, A. R., Didichenko, S. A., Chernoff, Y. O., Inge-Vechtomov, S. G. & Smirnov, V. N. (1993) *Mol. Microbiol.* **7,** 683–692.
10. Patino, M. M., Liu, J. J., Glover, J. R. & Lindquist, S. (1996) *Science* **273,** 622–626.
11. Ter-Avanesyan, M. D., Dagkesamanskaya, A. R., Kushnirov, V. V. & Smirnov, V. N. (1994) *Genetics* **137,** 671–676.
12. Derkatch, I. L., Chernoff, Y. O., Kushnirov, V. V., Inge-Vechtomov, S. G. & Liebman, S. W. (1996) *Genetics* **144,** 1375–1386.
13. Santoso, A., Chien, P., Osherovich, L. Z. & Weissman, J. S. (2000) *Cell* **100,** 277–288.
14. Chernoff, Y. O., Galkin, A., Lewitin, E., Chernova, T. A., Newnam, G. P. & Belenkiy, S. M. (2000) *Mol. Microbiol.* **35,** 865–876.
15. Kushnirov, V. V., Kochneva-Pervukhova, N., Chechenova, M. B., Frolova, N. S. & Ter-Avanesyan, M. D. (2000) *EMBO J.* **19,** 324–331.
16. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5355–5358.
17. Sunde, M. & Black, C. (1997) *Adv. Protein Chem.* **50,** 123–159.
18. Sondheimer, N. & Lindquist, S. (2000) *Mol. Cell* **5,** 163–172.
19. Liu, J. J. & Lindquist, S. (1999) *Nature (London)* **400,** 573–576.
20. Haney, P. J., Badger, J. H., Buldak, G. L., Reich, C. I., Woese, C. R. & Olsen, G. J. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3578–3583.
21. Daniel, R. M., Dines, M. & Petach, H. H. (1996) *Biochem. J.* **317,** 1–11.
22. Wright, P. E. & Dyson, H. J. (1999) *J. Mol. Biol.* **293,** 321–331.
23. Thompson, M. J. & Eisenberg, D. (1999) *J. Mol. Biol.* **290,** 595–604.
24. Bulmer, M. (1979) *Principles of Statistics* (Dover, New York).
25. Cox, B. S., Tuite, M. F. & McLaughlin, C. S. (1988) *Yeast* **4,** 159–178.
26. Wickner, R. B. (1994) *Science* **264,** 566–569.
27. Tang, H. Y., Xu, J. & Cai, M. J. (2000) *Mol. Cell. Biol.* **20,** 12–25.
28. David, G., Abbas, N., Stevanin, G., Dürr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E., *et al.* (1997) *Nat. Genet.* **17,** 65–70.
29. Eaglestone, S. S., Cox, B. S. & Tuite, M. F. (1999) *EMBO J.* **18,** 1974–1981.
30. Coustou, V., Deleu, C., Saupe, S. & Begueret, J. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 9773–9778.
31. Chernoff, Y. O., Derkach, I. L. & Inge-Vechtomov, S. G. (1993) *Curr. Genet.* **24,** 268–270.
32. Gill, G., Pascal, E., Tseng, Z. H. & Tjian, R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 192–196.
33. Bailleul, P. A., Newnam, G. P., Steenbergen, J. N. & Chernoff, Y. O. (1999) *Genetics* **153,** 81–94.
34. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.