doi:10.1016/j.jmb.2007.01.070 J. Mol. Biol. (2007) 368, 273–282

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Evolution of Budding Yeast Prion-determinant Sequences Across Diverse Fungi

Luke B. Harrison[1], Zhan Yu[1], Jason E. Stajich[2] Fred S. Dietrich[2] and Paul M. Harrison[1]*

[1]*Department of Biology McGill University Stewart Biology Building 1205 Docteur Penfield Ave Montreal, QC Canada H3A 1B1*

[2]*Department of Molecular Genetics and Microbiology Duke University, Durham NC 27710, USA*

**Corresponding author*

Prions are transmissible self-replicating alternative states of proteins. Four prions ([PSI+], [URE3], [RNQ+] and [NU+]) can be inherited cytoplasmically in *Saccharomyces cerevisiae* laboratory strains. In the case of [PSI+], there is increasing evidence that prion formation may engender mechanisms to uncover hidden genetic variation. Here, we have analysed the evolution of the prion-determinant (PD) domains across 21 fungi, focusing on compositional biases, repeats and substitution rates. We find evidence for constraint on all four PD domains, but each domain has its own evolutionary dynamics. For [PSI+], the Q/N bias is maintained in fungal clades that diverged one billion years ago, with purifying selection observed within the *Saccharomyces* species. The degree of Q/N bias is correlated with the degree of local homology to prion-associated repeats, which occur rarely in other proteins (<1% of sequences for the proteomes studied). The evolutionary conservation of Q/N bias in Sup35p is unusual, with only eight other *S. cerevisiae* proteins showing similar, phylogenetically deep patterns of bias conservation. The [URE3] PD domain is unique to *Hemiascomycota*; part of the PD domain shows purifying selection, whereas another part engenders bias changes between clades. Also, like for Sup35p, the [RNQ+] and [NU+] PD domains show purifying selection in *Saccharomyces* species. Additionally, in each proteome, we observe on average several hundred yeast-prion-like domains, with fewest in fission yeast. Our findings on yeast prion evolution provide further support for the functional significance of these molecules.

Keywords: prion; amyloid; bias; yeast; evolution

Introduction

Prions are alternative, propagating, transmissible states of normal cellular proteins. Prions were originally defined as the causative agent in mammalian neurodegenerative diseases, including scrapie in sheep and Creutzfeldt-Jakob disease in humans.[1] In *Saccharomyces cerevisiae*, prions were first identified as cytoplasmic elements inherited in a non-Mendelian fashion.[2–4] There are four known: [PSI+], [URE3], [NU+] and [RNQ+].[2,3,5–7] [PSI+] arises from the propagation of a misfolded form of Sup35p, part of the translation termination complex. Thus, formation of [PSI+] prions reduces the efficiency of translation termination and increases levels of nonsense-codon readthrough.[3,8,9] Such readthrough has been demonstrated to be a potential mechanism to uncover cryptic genetic variation.[10,11] [URE3], the prion form of the nitrogen catabolism protein Ure2p, functions to upregulate poor nitrogen source usage, even when rich sources are available.[2,4,5] Two other prions, [RNQ+] and [NU+] can function to allow [PSI+] induction by Sup35p overexpression.[12,13]

A defining characteristic of the known yeast prions is a region with a pronounced bias for glutamine (Q) and/or asparagine (N) residues.[7,14] Mutation of these residues diminishes or abolishes prion formation.[15,16] Previously, we have shown that, for three of the prions ([PSI+], [URE3] and [RNQ+]), Q/N-biased regions defined using a simple binomial probability algorithm are congruent with prion-determinant (PD) domains found in experiments.[14] For the fourth prion [NU+], the algorithmically

Abbreviations used: PD, prion determinant; LPS, lowest probability sub-sequence; YPL, yeast-prion-like.

E-mail address of the corresponding author: paul.harrison@mcgill.ca

defined Q/N-biased region coincides with a region necessary for [NU+] propagation.[7] Also, repetitive sequence motifs in the PD domain are structurally important in prion formation.[17–19] Specifically, decreasing/increasing the copy number of [PSI+] prion peptide repeats retards/induces prion formation.[17–19] PD domains for [PSI+] and [URE3] form β–sheet amyloid fibrils.[20,21] Crystallographic data for a [PSI+] peptide repeat[21] indicates stabilization by polar-zipper side-chain hydrogen bonds amongst glutamine and asparagine residues, and aromatic π interactions between tyrosine residues.

There is a growing body of evidence for conservation of prion-forming ability in yeasts. Homologues of *S. cerevisiae* PD domains from other yeasts can also form prions, either in *S. cerevisiae*, or in cells of their own species.[22–25] Also, "scrambled" forms of the Ure2p and Sup35p PD domains that maintain the amino acid composition, can form prions in *S. cerevisiae*, indicating that prion formation is primarily dependent on the composition of PD domains.[26,27]

Here, we survey the evolution of four known prion-determinant domains from budding yeast, across 21 diverse fungi. We demonstrate that prion-associated biases are maintained in fungi that are estimated to have diverged from each other about one billion years ago; furthermore, there is evidence for purifying selection, to varying degrees, on different prion domains and sub-domains.

Results and Discussion

What are the evolutionary dynamics of PD domains in yeasts and other fungi? Is there evidence of evolutionary constraint? We hypothesize that there might be such constraint, and so we examined PD evolution in two ways: (i) the conservation of the degree of Q/N compositional bias in the PD domains, using a method previously described, which calculates the lowest probability sub-sequences (LPS) within a given sequence;[14] and (ii) ratios of non-synonymous and synonymous codon substitution rates (i.e. K_a/K_s values), which is an indicator of positive selection (if K_a/K_s is significantly >1) or purifying selection ($K_a/K_s<1$). We studied the four known prions (Table 1), and yeast-prion-like biases, in 21 recently sequenced fungal genomes (Figure 1).

[PSI+]/Sup35p

Firstly, for Sup35p, we examined the degree of Q/N compositional bias in the PD domain across a diverse set of fungi, from the *Basidiomycota* and *Euascomycota*, as well as the *Hemiascomycota*. A maximum-likelihood phylogenetic tree was calculated for Sup35p orthologs (Figure 2). This tree generally follows a reference phylogeny that was derived for fungi (Supplementary Data, Figure 1). The degree of compositional bias for glutamine and asparagine was calculated as described, using the LPS method.[14] The biases were marked on the tree with binomial *P*-values (see the legend to Figure 2 and Methods for details). For each species, the number of regions in the proteome that are as biased as the PD domain homolog is also an indicator of bias maintenance (Supplementary Data, Table 2). Strikingly, the Q/N bias is conserved even in the evolutionarily distant fungal groups *Basidiomycota* and *Euascomycota*, with only about one region in every 100 sequences or so, at least as biased as the Sup35p PD domain (Figure 2 and Supplementary Data, Table 2). The last common ancestor with *Basidiomycota* and *Euascomycota* has been estimated using molecular clock analysis to be ~1 to 1.2 billion years ago.[28] The biases are strongest not in the *S. cerevisiae* sequence, but in Sup35p from the other hemi-ascomycetes *Candida glabrata*, *Saccharomyces castellii* and *Candida albicans*. The latter has been shown experimentally to form prions in its own cells.[23]

How unusual is Sup35p's maintenance of Q/N bias in the *Basidiomycota* and *Euascomycota*? To check this, we took the list of *S. cerevisiae* proteins which have as much Q/N bias as the Sup35p prion determinant (totalling 52 sequences). We then searched for orthologs that maintain a yeast-prion-like (YPL) region in *Basidiomycota* and *Euascomycota*

Table 1. Domains in the prion sequences, with analysis of K_a/K_s

Name	Location of PD	Aligned range[a]	Mean pairwise K_a/K_s (+/−s.d.)[b]	$2\Delta L$[c]
Sup35p/[PSI+] PD domain	1–123	1–123	0.146 (+/−0.123)	110 (4)
Sup35p/[PSI+] C-terminal globular domain	–	237–685	0.027 (+/−0.015)	434 (4)
Ure2p/[URE3] PD domain	1–65	7–45	0.024 (+/−0.029)	212 (12)
Ure2p/[URE3] C-terminal globular domain	–	109–351	0.019 (+/−0.012)	1876 (12)
Rnq1p/[RNQ+] PD domain	153–405	153–405	0.084 (+/−0.029)	358 (4)
New1p/[NU+] PD domain	1–153	2–153	0.015 (+/−0.012)	142 (6)

[a] Start and end points of the domains analysed in the relevant *S. cerevisiae* protein sequence; these serve to indicate the subsequences used to calculate K_a/K_s, guided by MUSCLE multiple alignments (Supplementary Data, Table 4).
[b] Pairwise maximum-likelihood–derived K_a/K_s values were calculated using PAML[32] (see Methods for details).
[c] For all of the sequences listed for each prion above, branch-specific K_a/K_s values were calculated using maximum likelihood, and the package PAML.[32] The log likelihood ratio test for selection (K_a/K_s significantly < or >1.0) was performed as described.[33] The test statistic, $2\Delta L$ for the log likelihood ratio test is tabulated, with the number of degrees of freedom in parentheses (all results were significant at $P<0.001$).

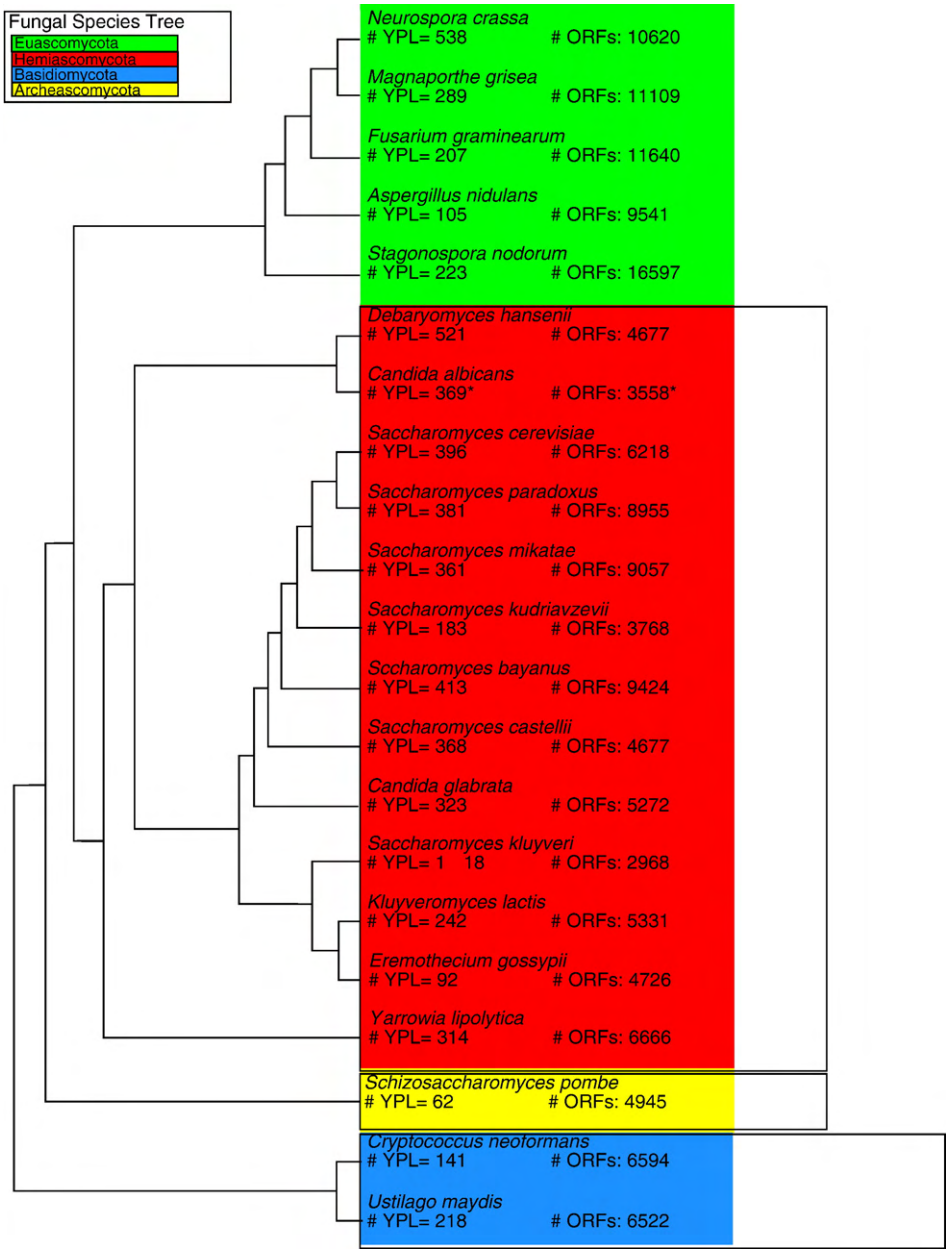

Figure 1. The 21 proteomes analyzed. The reference species tree on which this schematic Figure was based is in Supplementary Data, Figure 1. Listed for each proteome is: the organism name; the number of ORFs that have yeast-prion-like Q/N-rich regions (binomial P-value $\leq 1 \times 10^{-10}$) for the LPS method and the total number of ORFs. The blocks of colour separate the clades *Basidiomycota, Archaeascomycota, Hemiascomycota* and *Euascomycota*. Additional information for these proteomes is listed in Supplementary Data, Table 1, with details of the genome sequencing projects given in Supplementary Data, Table 5.

(Table 2A). Up to 29% of the 52 sequences maintain a YPL region in one of these clades (Table 2A). However, there are only eight other proteins (in addition to Sup35p), that maintain a YPL region in both the *Basidiomycota* and *Euascomycota* (Table 2B). Two of these are linked to clathrin assembly; others are associated with endocytosis and stress tolerance. The latter is notable from the point of view of the potential role of any yeast prions in adaptive evolution. There is growing evidence for such a role for the [PSI+] prion; in *S. cerevisiae* cells, [PSI+] prions cause revelation of phenotypic diversity in response to an array of different environmental conditions,[11] and this has been linked to nonsense-codon readthrough in specific instances.[10]

Since, for [PSI+], variation in copy number of the prion-associated repeats is linked to the efficiency of prion propagation,[17–19] we also analysed the repeat structure of Sup35p. We defined a consensus repeat sequence using the program Radar†[29] (see Methods). We searched for repeat homologies to this consensus, using BLAST with parameters recommended for the examination of short, nearly exact matches,[30] and found only two other matches in >6000 *S. cerevisiae*

† http://www.ebi.ac.uk/Radar/

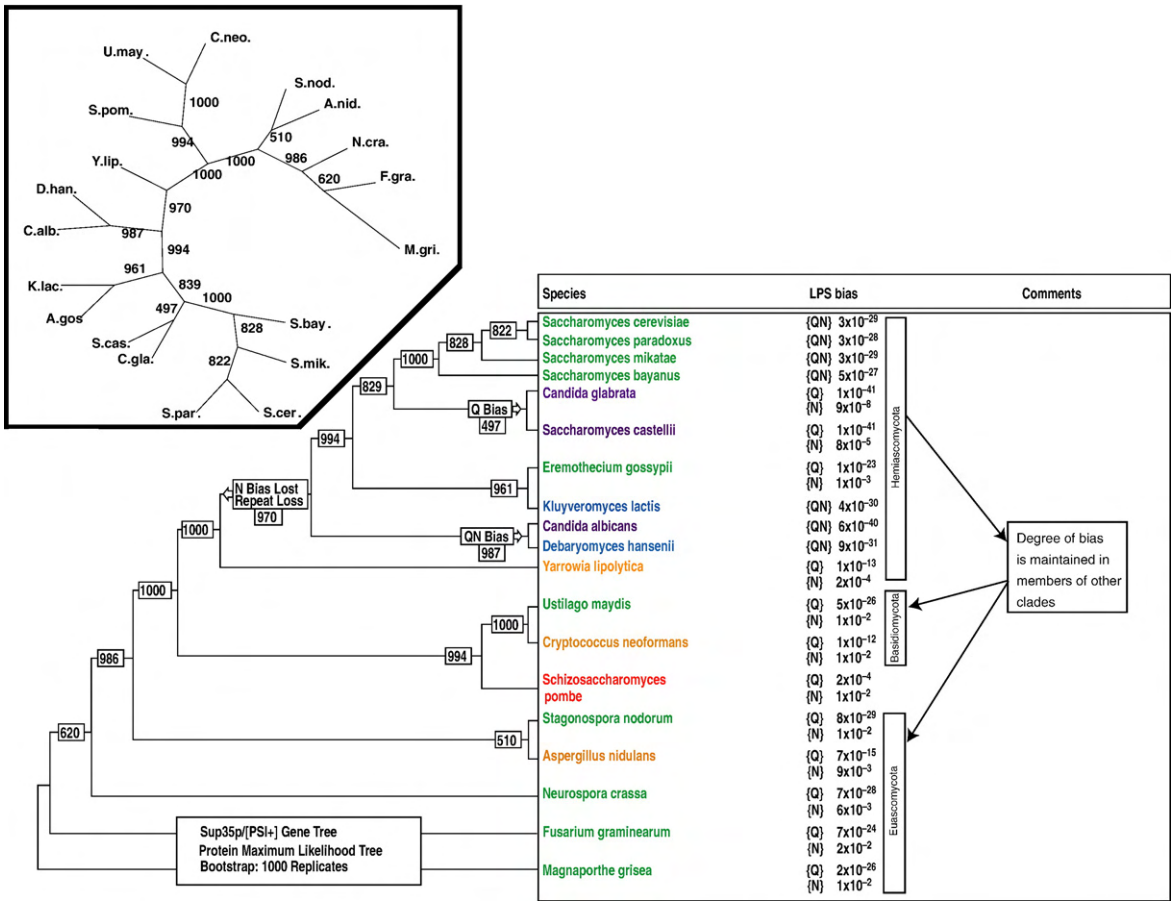

Figure 2. An unrooted maximum-likelihood majority rule consensus tree for Sup35p, which yields the [PSI+] prion. Maximum likelihood trees for protein distance produced using a majority rule consensus approach, based on 1000 bootstrap replicates using the PHYLIP *bootseq*, *proml* and *consense* programs.[40] The branches are labelled with their bootstrap values, with additional labels summarizing bias trends. For each sequence, the following are given: species name; LPS bias type ({Q}, {N} or {QN}) and *P*-value. Schematic comments in boxes point out overarching trends. A radial representation is shown as an inset. A colour-coding in the typeface of the species name is used to indicate the degree of bias in the prion region: purple for bias *P*-value $\leq 10^{-40}$; blue for 10^{-30} to 10^{-39}; green, 10^{-20} to 10^{-29}; orange, 10^{-10} to 10^{-19}; red $>10^{-10}$.

proteins. Furthermore, there are on average only three other matching open-reading frames (ORFs) for this repeat in the whole proteomes of any of the organisms examined (Supplementary Data, Table 6). Thus, the defined repeat is clearly unusual enough to be discriminatory of the Sup35 prion sequence. This discrimination is arguably a further indication of the functional importance of the repeat structure. Also, we examined the relationship between the number of Sup35p repeat homologies detected in orthologs, and the degree of bias for Q and N. After removing redundant sequences, we find that the degree of bias in the Sup35p PD domain is correlated with the approximate number of prion-associated repeat homologies (Figure 3(a)). This variation in copy number of prion-associated repeats may indicate species-specific variations in efficiency of prion propagation,[17–19] linked to imperfect repeat duplication in the PD domain. The importance of repeat duplication is also supported by evolutionary simulations of the Sup35p PD (Supplementary Data, Figure 3). If governed by a simple Jukes–Cantor mutational model, the bias of the Sup35 PD domain would be unlikely to be maintained in the *Basidiomycota* and *Euascomycota* ($P \leq 0.001$; Supplementary Data, Figure 3). We also examined how changes in the PD domain may be correlated with changes in the Sup35p globular domain. We found that change in the degree of Q/N bias is correlated with the pairwise sequence identity for the alignment of the C-terminal Sup35p globular domains (Supplementary Data, Figure 5). This may be due to correlated mutation, due to an interaction between the PD and globular domains.

Selection pressures on coding sequences are reflected in the ratios of the number of non-synonymous substitutions per synonymous site to the number of synonymous substitutions per synonymous site, termed K_a/K_s. Values significantly <1.0 indicate purifying or negative selection. Here, calculated K_a/K_s values indicate purifying selection for both the PD domain and the C-terminal globular domain of Sup35p in *Saccharomyces* species (Table 1 and Supplementary Data, Table 3). This is demonstrated both by the mean pairwise K_a/K_s values calculated, and the value of $2\Delta L$, the test statistic for

Table 2. Maintenance of bias for the Sup35p/[PSI+] prion

A. *Numbers of ORFs that are as biased as the Sup35/[PSI+] prion determinant in S. cerevisiae, and that maintain their bias in other clades*

Species→	Archaeascomycota (*Schizosaccharomyces pombe*)				Basidiomycota (*Cryptococcus neoformans, Ustilago maydis*)				Euascomycota (*Neurospora crassa, Magnaporthe grisea, Fusarium graminareum, Aspergillus nidulans, Stagnosporum nodorum*)			
BBH alignment	Masked		Not masked		Masked		Not masked		Masked		Not masked	
ORF category	Has YPL	Total	Has YPL	Total	Has YPL	Total	Has YPL	Total	Has YPL	Total	Has YPL	Total
Total = 52[a]	0	8 (15%)	0	20 (38%)	2 (4%)	2 (4%)	10 (19%)	14 (27%)	4 (8%)	6 (12%)	15 (29%)	23 (44%)

B. *The nine ORFs that maintain yeast-prion-like bias ($P \leq 10^{-10}$) in at least one of the Basidiomycota and the Euascomycota*

ORF name	Short name	Description
YDR172W	SUP35	Translation termination factor eRF3; makes prion [PSI+]
YGL181W	GTS1	Protein also contains a zinc-finger in the N terminus; regulates ultradian rhythm, cell size, cell cycle, lifespan, sporulation, heat tolerance, and multi-drug transport
YHR161C	YAP1801	Protein involved in clathrin cage assembly; binds Pan1p and clathrin; homologous to Yap1802p, member of the AP180 protein family
YIR006C	PAN1	Part of actin cytoskeleton-regulatory complex Pan1p-Sla1p-End3p, associates with actin patches on the cell cortex; promotes protein-protein interactions essential for endocytosis; previously thought to be a subunit of poly(A) ribonuclease
YLR206W	ENT2	Epsin-like protein required for endocytosis and actin patch assembly and functionally redundant with Ent1p; contains clathrin-binding motif at C terminus
YMR043W	MCM1	Transcription factor involved in cell-type-specific transcription and pheromone response; plays a central role in the formation of both repressor and activator complexes
YMR047C	NUP116	Subunit of the nuclear pore complex (NPC) that is localized to both sides of the pore; contains a repetitive GLFG motif that interacts with mRNA export factor Mex67p and with karyopherin Kap95p; homologous to Nup100p
YMR173W	DDR48	DNA damage-responsive protein, expression is increased in response to heat-shock stress or treatments that produce DNA lesions
YNL161W	CBK1	Serine/threonine protein kinase that regulates cell morphogenesis pathways; involved in cell wall biosynthesis, apical growth, proper mating projection morphology, bipolar bud site selection in diploid cells, and cell separation

Each of these ORFs is classed as a YPL in at least one of the Euascomycota and Basidiomycota studied. The standard long and short ORF names are given, along with a description adapted from text on the Saccharomyces Genome Database website (http://www.yeastgenome.org).

The total number of orthologs detected in at least one of the species in the clades are tabulated. Orthology was determined using the standard bi-directional best hits (BBH) approach, using a BLAST *e*-value of 1×10^{-4} and requiring alignment to a majority (≥ 0.5) of either aligned sequence. These were counted up, with or without masking of the biased regions under consideration, and with or without the requirement for maintenance of the yeast-prion-like (YPL) Q+N biased region ($P \leq 10^{-10}$).

[a] Total number of ORFs in *S. cerevisiae* that have a region as biased for Q and/or N as the Sup35 prion determinant.

the log likelihood ratio test for selection, for the whole set of multiply aligned prion-determinant sequences.[31] The statistic $2\Delta L$ is an indication of whether the examined sequences are evolving under significant selection (either purifying or diversifying).[31] Purifying selection had been previously noted in a pairwise comparison of Sup35p PD domains from *S. cerevisiae* and *Saccharomyces paradoxus*.[32]

It is possible that the degree of PD domain conservation for Sup35p and other prions, is linked to variation in the conservation of other factors linked to the efficiency of prion propagation, for example the chaperone Hsp104.[8] However, this chaperone has orthologs in all of the fungi examined. Certainly, the lack of conservation of the biases for three of the PD domains ([RNQ+], [NU+] and [URE3]) does not correlate with the presence or absence of this chaperone.

[URE3]/Ure2p

The evolutionary behaviour of the Ure2p PD domain is markedly different from that of Sup35p, showing more variation (Figure 4). Unlike for Sup35p, the Ure2p PD domain is not conserved outside *Hemiascomycota*. Its characteristic N bias is conserved in other close *Saccharomyces* species, but switches to a predominant Q bias in the *Kluyveromyces* and *Yarrowia lipolytica*, and a mingled QN bias in other parts of the tree (Figure 4). Remarkably, this bias switch is partitioned chiefly from about residue 45 onwards in the prion domain (denoted here as Ure2p-PD_{45+}; see Figure 5), with the first 45 residues under purifying selection (denoted Ure2p-$PD_{1–45}$), even where a dramatic overall switch of bias occurs (Table 1 and Supplmentary Data, Table 3). The equivalent Q-rich PD domain of *Kluyveromyces lactis* has been shown not to form [URE3] prions heterologously in *S. cerevisiae*.[25] From examining the branching order of the relevant trees (Figures 1 and 4), the most parsimonious explanation is that the bias for N arose in the *Saccharomyces* lineage, with a Q-rich or Q/N-rich ancestral domain. The purifying selection on Ure2p-$PD_{1–45}$, occurs to similar levels as for the C-terminal globular domain, throughout the *Hemiascomycota*

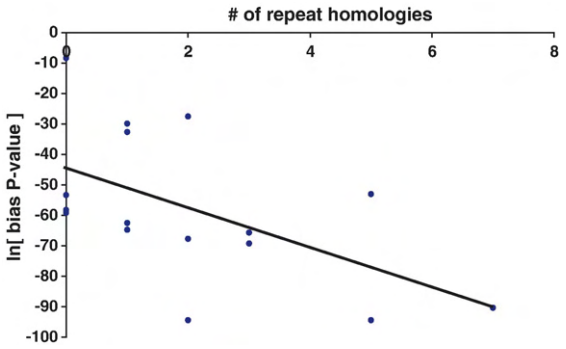

Figure 3. Comparisons of the degree of bias in the Sup35 prion determinant, local repeat homology and sequence divergence. The number of prion-associated homologies in the Sup35p PD domain is correlated with ln[LPS *P*-value] for the species studied, with Pearson $r^2 = 0.31$ ($r^2 = 0.29$ if sequences with number or repeat homologies = 0 are excluded; both r^2 values have $P < 0.05$). Parameters for blastp that were recommended for examination of short, nearly exact matches [-F F -W 2 -M PAM30 -G 9 -E 1 -e 10.0],[30] and yielded only two other matches to the Sup35 repeat consensus in >6000 *S. cerevisiae* ORFs. To remove redundancy, Sup35 sequences were discarded that could align significantly to the *S. cerevisiae* PD sequence for K_a/K_s analysis [*blastp* e-value $\leq 10^{-4}$, identity >50%].

(Table 1 and Supplementary Data, Table 3). This purifying selection indicates that there is evolutionary pressure against substitutions in the PD domain sequences, at specific positions. This is seemingly at odds with experimental work indicating that scrambled PD domains (with randomized sequences that maintain the native PD sequence composition) can form prions also.[26,27] However, it is possible that there is evolutionary pressure to maintain a specific repertoire of prion strain conformations that would cause such purifying selection. Even so, why, in contrast, would the type of PD domain bias change so substantially across Ure2p-PD_{45+} from Q bias, to N bias to mingled QN bias in different parts of the Ure3p tree? The variety of bias in the Ure2p-PD_{45+} region homologs (in one case, in place of the N bias is a stretch of polyglutamine) is certainly consistent with the experimental results on scrambled PD domains.[26,27] In light of these results, it would be interesting to test the prion-forming ability of hybrid PD domains comprising different Ure2p-$PD_{1–45}$ and Ure2p-PD_{45+} homologs, from different species. Simply, however, the purifying selection on Ure2p-$PD_{1–45}$ may arise because of other functional roles for the PD domains that have not been elucidated.

[RNQ+]/Rnq1p

This prion is much less studied, but is able to serve as [PIN+], the prion required for *de novo* appearance of [PSI+], under Sup35p overexpression.[6,12] The Rnq1p PD domain is conserved to only a limited extent outside of *S. cerevisiae* (Supplementary Data, Figure 2(a)). We found orthologs for Rnq1p in only nine of the 21 species studied, all of them hemiascomycetes. However, Rnq1p does not have a known, functionally characterized globular domain beyond the boundaries of the prion determinant; thus, this orthology detection relies on significant sequence alignment of residues 1–153 of Rnq1p, a part of the sequence which may not be functionally important and thus absent in Rnq1p sequences of other species. It is also possible that, even if [PSI+] formation is conserved in other fungi, the identity of [PIN+] may be more variable.[6,12] A high degree of Q/N bias (P-value $< \sim 1 \times 10^{-40}$) arises in *Saccharomyces* species and *C. glabrata* (Supplementary Data, Figure 2(a)). K_a/K_s analysis on the part of Rnq1p PD domain that can be unambiguously aligned indicates that it is under purifying selection in *Saccharomyces* species (Table 1 and Supplementary Data, Table 3).

[NU+]/New1p

[NU+], like [RNQ+], can serve as the [PIN+] prion, but only when it is itself overexpressed.[12,13] The mild overall Q/N bias is maintained in most of the *Hemiascomycota* (Supplementary Data, Figure 2(b)), through maintenance of a Q/N rich region equivalent to the $(NYN)_n$ repeat that was shown to be necessary for [NU+] prion formation in *S. cerevisiae*.[13] As for Ure2p, the PD domain is absent outside of *Hemiascomycota*. Purifying selection is evident for the segment of the PD domain outside of the $(NYN)_n$ repeat, but only in *Saccharomyces* species (Table 1 and Supplementary Data, Table 3).

Codon adaptation index

Codon adaption index (CAI) has been suggested as an indicator of the evolutionary and functional significance of a coding sequence. High values of CAI arise for highly expressed proteins, and low values for lowly expressed or spurious sequences.[33] We compared the CAI values for the four prion sequences (Supplementary Data, Figure 4) to the overall CAI distribution of >4000 verified yeast ORFs (from the Saccharomyces Genome Database‡). (This was examined previously for Sup35p[32]). The four prion sequences have values in the top 95% of ORFs, either considering the whole ORF sequence, or shorter sequence truncated at either end of the prion determinant (Supplementary Data, Figure 4). These additional values for truncated sequences were calculated using "CAI Calculator 2".[34] Most verified yeast ORFs have very low CAI values, i.e. there is a sharp peak in the overall CAI distribution at (0.10–0.15) (median = 0.15; mean = 0.19(±0.14)). Two of the prions (Sup35p and New1p) occur in the top 20% of CAI values, either for the whole sequence or just for the prion determinant subsequence (Supplmentary Data, Figure 4), and have thus undergone some degree of codon choice optimization, presumably linked to relatively high expression levels.

‡ http://www.yeastgenome.org

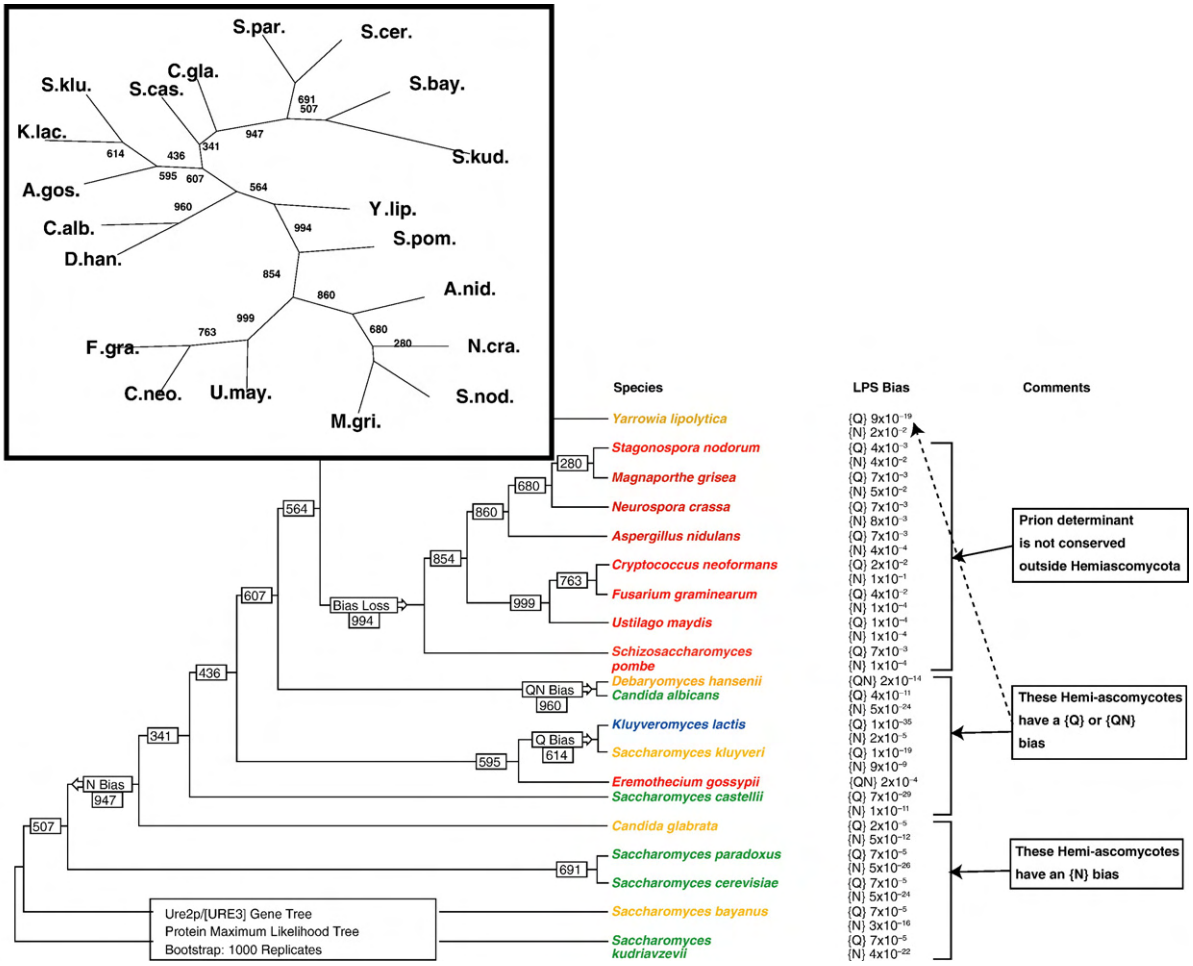

Figure 4. An unrooted maximum-likelihood majority rule consensus tree for Ure2p, which yields the [URE3] prion. This tree is derived as described in the legend to Figure 2.

```
residue # in S.cerevisiae          ....1.....................................37
K.lactis                           MQQDMQNGGP-------GNTISNLSSALRQVNLG--NSNTTTDQSNISID
S.cerevisiae                       ----MMNNNG--------NQVSNLSNALRQVNIGNRNSNTTTDQSNINFE
S.paradoxus                        ----MMNNNG--------NQVSNLSNALRQVNIGNRNSNTTTDQSNINFE
S.castellii                        ----MNNLNGH------TNQISNLSNALQQVNID--NSNTTTDQSNINFD
S.pombe                            ----------------------------------------------------

residue # in S.cerevisiae          ..........................61....................
K.lactis                           FN-----QQQLLEEANQGSINAYNAQQQQQEHLQQQAQQQQLHMQQLQQAQ
S.cerevisiae                       FSTGVNNNNNNNSSSNNNN-------VQNNN--------------------
S.paradoxus                        FSAGVNNNNNNSSSSNNNNNNNNN--AQNNN--------------------
S.castellii                        VSSNENSNHNSQHNSDSNILEQN---YSNTE--------------------
S.pombe                            ----------------------------------------------------

residue # in S.cerevisiae          .....63.....................................100
K.lactis                           QQQAQQQAHQQQQQVHQVQHQHVQQDHMPIGQSQQQAMYQGPNPIDSSRI
S.cerevisiae                       -----SGRNGSQNNDNENNIKNTLEQ-----HRQQQQAFSDMSHVEYSRI
S.paradoxus                        -----SGRNGSQSNDNGNNIKDTLEQ-----HRQQQQAFSDMSHVEYSRI
S.castellii                        -----SNRQ-EQLQQQQQQQQQQQQQ-----QQQQQQQHAPFGDIEYSRI
S.pombe                            ----------------------------------------------------

red = under purifying selection
purple = N bias
green = Q bias
blue = Q+N bias
```

Figure 5. Purifying selection and bias change in the Ure2p PD domain. An alignment of representative sequences that indicate the bias switching in part of the Ure2p PD domain (from about residue 45 onwards, denoted Ure2p-PD_{45+} in the text). The part of the Ure2p domain that is under purifying selection (denoted Ure2p-$PD_{1–45}$) is on red. The bias-changing area is coloured green, blue or purple.

Proteome-wide trends in Q/N biased regions

The search for other prion phenomena is on-going, with recent evidence reported for multiple further prions linked to Q/N bias.[35] We calculated the number of "yeast-prion-like" (YPL) domains in the fungal proteomes, as previously for a smaller, more diverse group of eukaryotes, using somewhat different thresholds.[14] Typically, a fungal proteome has several hundred of these domains, with the exception of a notable relative scarcity in *Schizosaccharomyces pombe* (Figure 1), as noted previously.[14] Also, with regard to prion bias loss, this is the only species, out of all the fungi studied that exhibits substantial loss/degeneration of PD domains for all four prions. A complete list of these YPL regions is provided (Supplementary Data, Table 7).

Conclusions

We have shown that each of the known prion domains has idiosyncratic evolutionary dynamics, with varying degrees of evolutionary constraint, as judged from K_a/K_s analysis. Furthermore, the maintenance of bias in fungal clades that diverged about one billion years ago, particularly for the Sup35p PD domain, may be an indicator of compositional selection, as has been postulated for protamines;[36] however, further work on the neutral evolution of biased domains would be instructive for addressing this question more completely. This maintenance of bias for the Sup35p is shown to be unusual, with only eight other *S. cerevisiae* proteins showing the same, phylogenetically deep pattern of bias conservation across fungi. The characteristics maintained in PD domains (Q/N bias and repeats) are linked to the β–sheet formation that underlies the prion phenomenon. However, it is also possible that β–sheet –forming PD domains mediate another distinct function yet to be elucidated, not directly related to prion formation *per se*; indeed, some Sup35p deletion-mutant phenotypic behaviours cannot be explained as either [*psi*-] or [*PSI*+].[11]

Our data can be used to home in on further potential prions.

Methods

Calculation of compositional bias

Degree of compositional bias for Q and N residues was calculated using the lowest probability subsequence (LPS) method.[14] Where a better *P*-value arises for Q+N in combination (than either Q or N alone), the bias is considered a mingled QN bias. The LPS method searches for the lowest probability subsequence in a sequence, given a reference composition, with the degree of bias indicated by a binomial *P*-value. Yeast-prion-like regions are defined using a *P*-value threshold of 1×10^{-10}.[14] Neither the amount of bias in the prion domains, nor the total number of yeast-prion-like regions observed, is correlated with the overall percentage composition of the proteomes that is Q or N.

Reference phylogeny for fungal genomes

A reference phylogeny for fungal genomes (Supplementary Data, Figure 1) was generated from 60 orthologous genes in 31 fungal genomes, using gene prediction software to supplement existing annotations§. Orthologs were identified by FASTA[37] (*e*-value ≤ 0.01) using the bi-directional best hits approach. Pairwise orthologs were combined into groups using single linkage clustering and groups with exactly one gene per genome were retained. MUSCLE multiple alignments[38] were concatenated for a random sample of 60 of these sets. The phylogenetic tree was constructed using PHYML[39] with the WAG substitution model, and 100 bootstrap replicates. This phylogeny clusters all of the *Saccharomyces* clade together (*Saccharomyces kluyveri* is a *Kluyveromyces*).

Individual molecular phylogenies

Maximum likelihood PHYLIP v3.6b[40] was used to make maximum-likelihood majority-rule consensus trees based on MUSCLE multiple alignments[38] of orthologous proteins to the *S. cerevisiae* Sup35p, that were collated using the bi-directional best hits approach, and *blastp* using *e*-value $\leq 10^{-4}$.[30] In addition to default low-complexity masking, the entire LPS region defined for Sup35p was masked. If a proteome was found to be lacking an ortholog in its proteome set, a *tblastn* search[30] was conducted against the corresponding genome to insure mis-annotation was not responsible (two such mis-annotations were detected in the genomes analysed). Maximum likelihood trees for protein distance were produced using a majority rule consensus approach based on 1000 bootstrap replicates using the PHYLIP *bootseq*, *proml* and *consense* programs.

K_a/K_s calculations

Sets of prion sequence orthologs were analysed for K_a/K_s: (i) in a pairwise fashion; and (ii) using phylogenetics trees. Firstly, MUSCLE[38] multiple alignments were filtered for pairs of sequences that align to each other with $\geq$70% sequence identity over $\geq$25 residues. Maximum-likelihood–derived K_a/K_s values were calculated using PAML,[41] for these sequence pairs. The proteomes whose PD sequences meet this criterion are as follows: for Sup35p (*S. cerevisiae*; *S. paradoxus*; *Saccharomyces mikatae*; *Saccharomyces bayanus*); for Rnq1p (*S. cerevisiae; S. paradoxus; Saccharomyces kudravzievii; S. bayanus*); for New1p (*S. cerevisiae; S. paradoxus; S. kudravzievii; S. bayanus; S. mikatae*); for Ure2p (*S. cerevisiae; S. paradoxus; S. kudravzievii; S. bayanus; S. castellii; S.kluyverii; K. lactis; Eremothecium gossypii*). More K_a/K_s details are available in Supplementary Data, Table 3.

Secondly, branch-specific K_a/K_s values were calculated for maximum-likelihood–derived trees, with PAML.[41] The log likelihood ratio test for selection (K_a/K_s significantly< or >1.0) was performed as described,[31] through calculation of the test statistic, $2\Delta L$.

Definition of consensus Sup35p repeat

We defined a consensus for the prion-associated repeat sequence in Sup35p using the website of the program Radar‖,[29] with default settings. It was defined

§ http://fungal.genome.duke.edu
‖ http://www.ebi.ac.uk/Radar/

as *QGGYQQYNPXGGYQQN*. At each position invariant residues (labelled in bold) and most common residues (labelled in italics) defined this consensus, with variable positions labelled X. To find homologies to this consensus, we used BLASTP, with parameters that have been recommended for examination of short, nearly exact matches [-F F -W 2 -M PAM30 -G 9 -E 1 -e 10.0].[30] These parameters yield only two other matches to the Sup35 repeat consensus in >6000 *S. cerevisiae* ORFs. The number of matching ORFs to this consensus is no more than 1% of the whole proteome for the fungi studied (Supplementary Data, Table 6).

Acknowledgements

This work was made possible by grants from the National Science and Engineering Research Council of Canada, and from McGill University.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.01.070

References

1. Prusiner, S. (1998). Prions. *Proc. Natl Acad. Sci. USA*, **95**, 13363–13383.
2. Lacroute, F. (1971). Non-Mendelian mutation allowing ureidosuccinic acid uptake in yeast. *J. Bacteriol.* **106**, 519–522.
3. Cox, B. (1965). [PSI], a cytoplasmic suppressor of super-suppression in yeast. *Heredity*, **20**, 505–521.
4. Wickner, R. (1994). [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science*, **264**, 528–530.
5. Wickner, R., Edskes, H., Roberts, B., Baxa, U., Pierce, M., Ross, E. & Brachmann, A. (2004). Prions: proteins as genes and infectious entities. *Genes Dev.* **18**, 470–485.
6. Sondheimer, N. & Lindquist, S. (2000). Rnq1: an epigenetic modifier of protein function in yeast. *Mol. Cell*, **5**, 163–172.
7. Santoso, A., Chien, P., Osherovich, L. & Weissman, J. (2000). Molecular basis of a yeast prion species barrier. *Cell*, **100**, 277–288.
8. Shorter, J. & Lindquist, S. (2005). Prions as adaptive conduits of memory and inheritance. *Nature Rev. Genet.* **6**, 435–450.
9. Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M. & Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.* **316**, 409–419.
10. True, H., Berlin, I. & Lindquist, S. (2004). Epigenetic regulation of translation reveals hidden genetic variation to produce comlex traits. *Nature*, **431**, 184–187.
11. True, H. & Lindquist, S. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, **407**, 477–483.
12. Derkatch, I., Bradley, M., Zhou, P., Chernoff, Y. & Liebman, S. (2001). Prions affect the appearance of other prions: the story of [PIN+]. *Cell*, **106**, 171–182.
13. Osherovich, L. & Weissman, J. (2001). Multiple Gln/Asn-rich prion domains confer susceptibility to induction of the yeast [PSI(+)] prion. *Cell*, **106**, 183–194.
14. Harrison, P. & Gerstein, M. (2003). A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes.*Genome Biol.* **4**, R40; (E-pub).
15. DePace, A., Santoso, A., Hillner, P. & Weissman, J. (1998). A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion. *Cell*, **93**, 1241–1252.
16. Maddelein, M. & Wickner, R. (1999). Two Prion-inducing regions of Ure2p are non-overlapping. *NMol. Cell. Biol.* **19**, 4516–4524.
17. Liu, J. & Lindquist, S. (1999). Oligopeptide-repeat expansions modulate protein-only inheritance in yeast. *Nature*, **400**, 573–576.
18. Osherovich, L., Cox, B., Tuite, M. & Weissman, J. (2004). Dissection and design of yeast prions. *PLoS Biol.* **2**, E86.
19. Parham, S., Resende, C. & Tuite, M. (2001). Oligopeptide repeats in the yeast protein Sup35p stabilize intermolecular prion interactions. *EMBO J.* **20**, 2111–2119.
20. Baxa, U., Cheng, N., Winkler, D., Chiu, T., Davies, D., Sharma, D. *et al.* (2005). Filaments of the Ure2p prion protein have a cross-beta core structure. *J. Struct. Biol.* **150**, 170–179.
21. Nelson, R., Sawaya, M., Balbirnie, M., Madsen, A., Riekel, C., Grothe, R. & Eisenberg, D. (2005). Structure of the cross-beta spine of amyloid-like fibrils. *Nature*, **435**, 773–778.
22. Nakayashiki, T., Ebihara, K., Bannai, H. & Nakamura, Y. (2001). Yeast [PSI+] "prions" that are crosstransmissible and susceptible beyond a species barrier through a quasi-prion state. *Mol. Cell*, **7**, 1121–1130.
23. Tanaka, M., Chien, P., Yonekura, K. & Weissman, J. (2005). Mechanism of cross-species prion transmission: an infectious conformation compatible with two highly divergent yeast prion proteins. *Cell*, **121**, 49–62.
24. Talarek, N., Maillet, L., Cullin, C. & Aigle, M. (2005). The [URE3] prion is not conserved among *Saccharomyces* species. *Genetics*, **171**, 23–34.
25. Baudin-Bailleu, A., Fernandez-Bellot, E., Reine, F., Coissac, E. & Cullin, C. (2003). Conservation of the prion properties of Ure2p through evolution. *Mol. Biol. Cell*, **14**, 3449–3458.
26. Ross, E., Baxa, U. & Wickner, R. (2004). Scrambled prion domains form prions and amyloid. *Mol. Cell. Biol.* **24**, 7206–7213.
27. Ross, E., Edskes, H., Terry, M. & Wickner, R. (2005). Primary sequence independence for prion formation. *Proc. Natl Acad. Sci. USA*, **102**, 12825–12830.
28. Hedges, S. (2002). The origin and evolution of model organisms. *Nature Rev. Genet.* **3**, 838–849.
29. Heger, A. & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Struct. Funct. Genet.* **41**, 224–237.
30. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
31. Yang, Z. (1998). Likelihood ratio tests for detecting

positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573.
32. Jensen, M., True, H., Chernoff, Y. & Lindquist, S. (2001). Molecular population genetics and evolution of prion-like protein in *S. cerevisiae*. *Genetics*, **159**, 527–535.
33. Sharp, P. & Li, W. (1987). The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.* **15**, 1281–1295.
34. Wu, G., Culley, D. & Zhang, W. (2005). Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology*, **151**, 2175–2187.
35. Ter-Avanesyan, M., Derkatch, I., Baskakov, I. & Kushnirov, V. (2006). Unraveling prion structures and biological functions. *Genome Biol.* **6**, 366.
36. Rooney, A., Zhang, J. & Nei, M. (2000). An unusual form of purifying selection in a sperm protein. *Mol. Biol. Evol.* **17**, 278–283.
37. Lipman, D. J. & Pearson, W. (1988). Improved tool for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
38. Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797.
39. Guindon, S. & Gascuel, O. (2003). A simple fast accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704.
40. Felsenstein, J. (1989). PHYLIP–Phylogeny Inference Package (version 3.2). *Cladistics*, **5**, 164–166.
41. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.

Edited by F. E. Cohen

(Received 23 October 2006; received in revised form 1 January 2007; accepted 25 January 2007)
Available online 3 February 2007