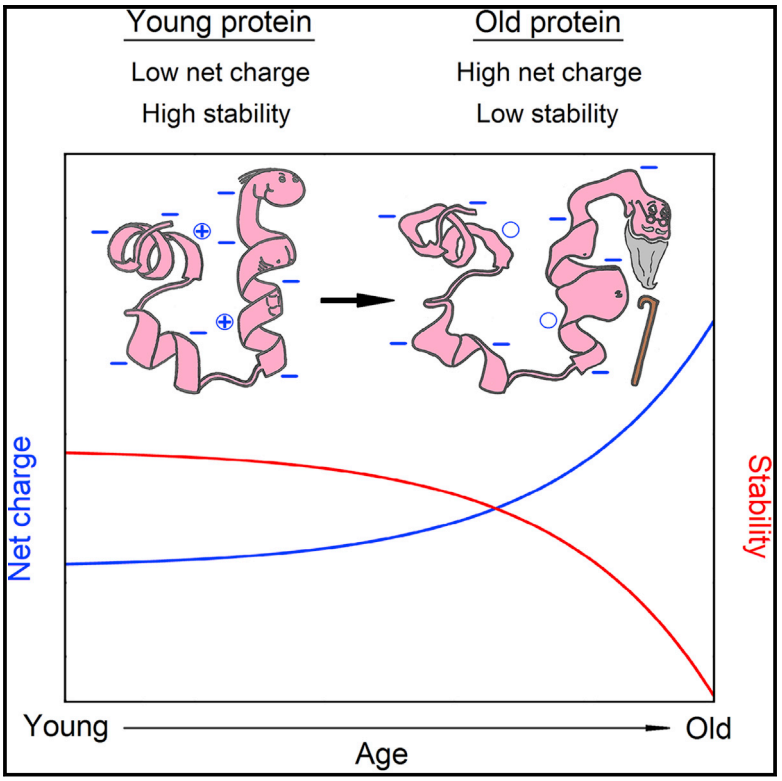
Theory

Structure

Highly Charged Proteins: The Achilles' Heel of Aging Proteomes

Graphical Abstract



Authors

Adam M. R. de Graff,
Michael J. Hazoglou, Ken A. Dill

Correspondence

adam.degraff@stonybrook.edu

In Brief

de Graff et al. show that random modification of side chain charge by oxidative damage could be a dominant source of protein stability loss in aging organisms. This provides a mechanism connecting damage to functional loss and sheds insight on the puzzle of how small levels of damage could affect aging.

Highlights

- Proteins undergo random damage from oxidation in aging
- Oxidative damage can change side chain charge, leading to protein stability loss
- Highly charged proteins are at particular risk of large oxidative stability loss
- Key pathways and aggregates of old cells are enriched in highly charged proteins

de Graff et al., 2016, Structure *24*, 329–336
February 2, 2016
http://dx.doi.org/10.1016/j.str.2015.11.006

CellPress

Structure
Theory

Highly Charged Proteins: The Achilles' Heel of Aging Proteomes

Adam M. R. de Graff,[1,*] Michael J. Hazoglou,[2] and Ken A. Dill[1,2,3]
[1]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA
[2]Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA
[3]Department of Chemistry, Stony Brook University, Stony Brook, NY 11794, USA
*Correspondence: adam.degraff@stonybrook.edu
http://dx.doi.org/10.1016/j.str.2015.11.006

SUMMARY

As cells and organisms age, their proteins sustain increasing amounts of oxidative damage. It is estimated that half of all proteins are damaged in old organisms, yet the dominant mechanisms by which damage affects proteins and cellular phenotypes are not known. Here, we show that random modification of side chain charge induced by oxidative damage is likely to be a dominant source of protein stability loss in aging cells. Using an established model of protein electrostatics, we find that short, highly charged proteins are particularly susceptible to large destabilization from even a single side chain oxidation event. This mechanism identifies 20 proteins previously established to be important in aging that are at particularly high risk for oxidative destabilization, including transcription factors, histone and histone-modifying proteins, ribosomal and telomeric proteins, and proteins essential for homeostasis. Cellular processes enriched in high-risk proteins are shown to be particularly abundant in the aggregates of old organisms.

INTRODUCTION

As cells age, their biomolecules accumulate oxidative damage from reactive by-products of respiration and other forms of metabolism (Adachi et al., 1998; Oliver et al., 1987; Sohal et al., 1993; Starkereed and Oliver, 1989). Proteins are particularly important targets (Smith et al., 1991) based on the following evidence. (1) In old organisms, up to 30% of proteins are carbonylated and at least 40%–50% of proteins likely have some form of oxidative damage (Starkereed and Oliver, 1989). (2) While oxidation also damages lipids and nucleotides of DNA and RNA, most cellular biomass is protein and most processes are mediated by proteins, including the repair of all other biomolecules. (3) In vivo and in vitro enzymatic activity of many protein species decrease with both acute and gradual increases in oxidative damage, resulting in decreased performance (Carney et al., 1991; Sharma and Rothstein, 1980). (4) The thermal stability of many proteins, known to be essential for their enzymatic activity, is reduced in aged organisms (Oliver et al., 1987; Sharma and Rothstein, 1980). (5) Loss of proteome stability is an early and universal feature of aging (Balch et al., 2008; Ben-Zvi et al., 2009), and is accompanied by loss of solubility and the aggregation of a vast number of proteins (Walther et al., 2015). (6) Major human diseases of aging, including Alzheimer's disease, Huntington's disease, amyotrophic lateral sclerosis, and cancer, have been linked to protein damage, decreased enzyme activity, and aggregation (Smith et al., 1991; Xu et al., 2011). (7) Oxidative protein damage appears to follow a universal trajectory with age across organisms (Figure 1). This universality is more readily explained in terms of random untargeted damage events across many different proteins in a proteome than in terms of specific failure of a small number of particular proteins or biochemical pathways.

There are many molecular mechanisms of oxidative damage to proteins (Table 1). Damage results from interaction with reactive oxygen species (ROS), reactive nitrogen species (RNS), and reactive lipid and glycolytic products (Shacter, 2000; Stadtman, 2006). Reactive species can cleave the polypeptide backbone, alter side chains, or covalently bond side chains to lipids, carbohydrates, or even other side chains. Damage to individual side chains is likely to be particularly relevant in aging (Stadtman, 2006). For one thing, the breakage of protein backbones can be readily handled by the cell's machinery for degrading proteins. For another, oxidative modification of individual side chains is estimated to be orders of magnitude more abundant than other forms of damage (Stadtman, 2006), with a large fraction of such modifications resulting in a change in side chain charge (Davies et al., 1987; Hipkiss, 2006; Rao and Moller, 2011; Stadtman, 2006).

Any plausible mechanism for the detrimental effects of oxidative damage on health must address a key question: How can such a small amount of oxidative damage, averaging less than one amino acid per protein molecule in aged organisms (Stadtman, 1992; Starkereed and Oliver, 1989), account for the large phenotypic changes associated with aging? Clearly the modification of single amino acids that are directly involved in catalytic function could decrease activity and, hence, fitness. However, catalytic sites account for only a small fraction of a protein, whereas the large majority of amino acids influence protein stability. Given the importance of stability to catalytic function and the low stability of many proteins (Supplemental Section 2), we propose that untargeted oxidative perturbations to protein *stability* could be the more abundant source of enzymatic activity loss in aging (Carney et al., 1991; Smith et al., 1991).

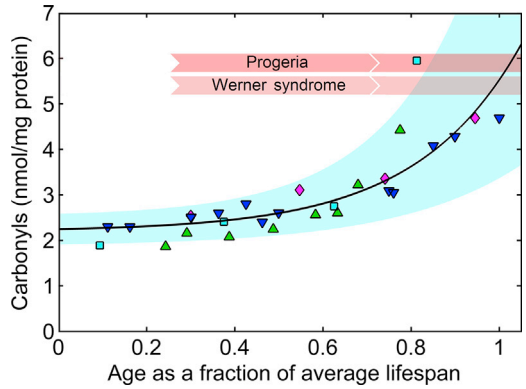


Figure 1. Age-Dependent Increase of Oxidative Damage to Proteins Across Organisms

The extent of protein damage, as measured by carbonyl content, with age in worms (purple diamonds) (Adachi et al., 1998), flies (green triangles) (Sohal et al., 1993), rats (cyan squares) (Starkereed and Oliver, 1989), and humans (blue downward triangles) (Oliver et al., 1987). The black line is the least-squares fit to the combined data. The cyan band represents the range of outcomes if the parameters of the fit are varied by ±15%. The horizontal pink bands correspond to the maximum levels observed toward the end of life in people with progeria and Werner syndrome, which are premature aging diseases (Oliver et al., 1987).

We hypothesize that random modification of side chain charge represents a dominant source of protein stability loss in aging cells and organisms. It is known that the most frequently oxidized side chains are those of the charged amino acids lysine and arginine (Petrov and Zagrovic, 2011; Requena et al., 2001; Stadtman, 2006), as well as the neutral amino acids cysteine, methionine, proline, threonine, and histidine (Hipkiss, 2006; Rao and Moller, 2011). Importantly, the main oxidation products of both charged amino acids are neutral (Hipkiss, 2006; Requena et al., 2001; Stadtman, 2006) and those of the neutral amino acids are often negative. For proteins with a net negative charge, such modifications have the effect of making the protein even more negative, with potentially destabilizing consequences. Oxidation can also perturb a protein's charge in the positive direction, as negatively charged amino acids can also be damaged to form neutral by-products, albeit less frequently than their positive counterparts (Davies et al., 1987; Hipkiss, 2006). For example, while cysteine is classified as a neutral amino acid, its near-neutral pK_a causes significant occurrence of its negative protonation state. This negative charge is removed when cysteine is converted to serine or forms a disulfide bridge, which could be particularly destabilizing to a protein with a net positive charge.

The most frequent cause of protein damage is thought to be metal-catalyzed oxidation, which preferentially damages proteins containing particular transition metals (Requena et al., 2001) (Table 1). Another important source of charge modification is protein deamidation, whereby asparagine and glutamine are converted to negatively charged aspartate and glutamate (Robinson and Robinson, 2001). In addition to the direct mechanisms of charge modification shown in Table 1, there are also indirect mechanisms. For example, the oxidation of guanine and methylation of cytosine in DNA, which become more abundant with age and are altered in cancer, can lead to mutations that change amino acid charge by changing the respective codons. In addition, a damaged side chain can change the protonation state of neighboring side chains or alter the binding of counterions. Lastly, protein charge can be modified by a change in pH, which not only varies between cellular compartments, but more importantly can also change with age, cancer, and disease (Henderson et al., 2014). Thus the proteins predicted to be the most destabilized by oxidative charge modification are also predicted to be sensitive to aging-related changes in pH. All of these changes are likely to occur mainly on protein surfaces, but not exclusively (Rao and Moller, 2011).

Table 1. Main Methods of Oxidative Modification to Amino Acid Side Chains

Method of Oxidation	Amino Acids Affected
Metal-catalyzed oxidation	Arg, Lys, His, Pro, Thr, Tyr, Cys, Met
$^1O_2$	Arg, Lys, His, Pro, Thr, Tyr, Cys, Met
$ONOO^-$	Tyr, Cys, Met
HOCl	Arg, Lys, Pro, Thr, Tyr, Cys, Met
Ozone	Arg, Lys, Pro, Thr, Cys, Met
γ-Ray	Arg, Lys, His, Pro, Thr, Tyr, Trp, Val, Leu, Cys, Met

All these methods can affect side chain charge (Shacter, 2000).

In summary, oxidative damage often randomly alters a protein's net charge. Given that all proteins generally have both positive and negative side chains that are susceptible to damage, oxidation can act to make positively charged proteins more positive and negatively charged proteins more negative, and vice versa. In this work, we quantify the effects of surface charge modification on protein stability in the context of aging. We find that short, highly charged proteins are particularly susceptible to large stability loss from such charge modifications, with the magnitude of this loss being comparable to the full native stabilities of those proteins. This work is particularly applicable to *random* charge modification, in contrast to those that occur as part of normal function, such as phosphorylation. This is because sites that have evolved to accommodate changes in charge as part of normal function are under high selection pressure for such change, whereas selection pressure is expected to be much weaker across all possible oxidization sites.

RESULTS

Net Charge on a Protein Decreases Its Folding Stability

We show here how particular subpopulations of the proteome can be significantly destabilized by only small charge changes, as can occur from random oxidative damage. It has been shown to be a good first approximation to assume that a protein's net charge is uniformly spread over its surface. This allows the electrostatic contribution ΔG to the folding stability to be written as (Ghosh and Dill, 2009)

$$\frac{\Delta G}{kT} = \frac{Q_d^2 l_b}{2R_d(1+\kappa R_d)} - \frac{Q_n^2 l_b}{2R_n(1+\kappa R_n)}, \qquad \text{(Equation 1)}$$

where Q_d and Q_n are the protein's net charge in the denatured and natively folded states, R_d and R_n are the associated radii of

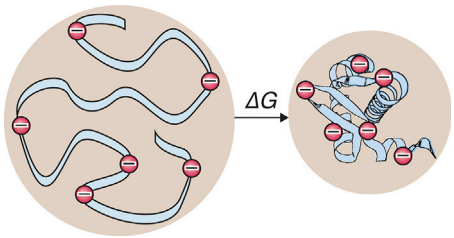


Figure 2. Protein-Folding Stability Is Reduced in Highly Charged Proteins

Shown here is a protein-folding process, and its corresponding folding free energy, ΔG. Proteins that have a net charge will have reduced stability, due to the compaction of the charge into a smaller space upon folding. See also Figure S1.

gyration, l_b is the Bjerrum length over which charge-charge interactions decay, κ is the inverse Debye screening length ($\kappa^2 = 2c_s l_b$, where c_s is the concentration of salt in solution), and kT is Boltzmann's constant multiplied by absolute temperature. More positive values of ΔG indicate greater stability. The average radius of gyration of native proteins is a known function of chain length, namely $R_n = 2.24N^{0.392}$ Å, while the denatured state radius is approximated here by $R_d = 1.927N^{0.598}$ Å (Kohn et al., 2004) (model predictions for a compact denatured state are given in Supplemental Section 3D). These radii are assumed to be unaffected by single oxidation events. Protein net charge Q_d and Q_n are estimated from the standard pK_a values of the side chains composing the protein sequence, with the exception of histidine in the native state, for which we use the average experimental pK_a measured across a set of folded proteins. The pK_as of the other side chains are generally sufficiently acidic or basic such that differences between native and denatured state values have little impact on the net charge (see Figure S1; Supplemental Sections 1A and 3D for further discussion). To describe the cytoplasm, we use biologically realistic values $l_b = 7.13$ Å and $\kappa = 0.03$ Å^{-1}, corresponding to a dielectric constant of 78.5 for water and a salt concentration of 0.1 M (Dill and Bromberg, 2011). While long-range charge interactions are weak under such conditions (roughly 0.15 kT per charge pair at a distance of 10 Å [Lee et al., 2002]), their large abundance makes them important in aggregate, especially for proteins with high net charge and low stability (stabilities often being only a few kT, see Figure S3).

Equation 1 expresses the principle that a greater net charge on a protein acts as an unfolding force because of the stronger charge-charge repulsions within the smaller volume of the folded state relative to the more expanded denatured ensemble (Dill and Stigter, 1995; Stigter et al., 1991), as shown in Figure 2. This electrostatics model has been previously shown to capture (1) experimental pH-salt phase diagrams for denaturing myoglobin, lysozyme, and RNase A (Ghosh and Dill, 2009), and (2) the observed dependence of folding stability on the square of the net charge (Dill and Stigter, 1995; Ghosh and Dill, 2009; Gitlin et al., 2006; Stigter et al., 1991). Most charges are located on protein surfaces, as we have assumed here, because burying charges in low-dielectric protein cores is very unfavorable (Tokuriki et al., 2007). Supplemental Section 3E discusses smaller contributions of spatial charge heterogeneity.

Small Changes in Charge Can Greatly Destabilize a Highly Charged Protein

By taking differences using Equation 1, it is readily shown that the charge perturbation from a single oxidation event—changing a protein's charge from Q to $Q \pm 1$—leads to the change of an average protein's folding stability by

$$\frac{\Delta\Delta G}{kT} = \frac{(\pm 2Q_d + 1)l_b}{2R_d(1+\kappa R_d)} - \frac{(\pm 2Q_n + 1)l_b}{2R_n(1+\kappa R_n)}. \qquad \text{(Equation 2)}$$

Equation 2 gives a key conclusion here: namely, that modifying just one charge on a protein can cause a large change in folding stability for a protein that is already highly charged, because the stability change $\Delta\Delta G$ scales with the net charge ($Q_n \approx Q_d$) (Supplemental Sections 1A and 3A). This effect is further amplified in smaller proteins due to their smaller radii of gyration.

Thus, the model predicts that changing one charge on a near-neutral protein has little effect on stability, while changing one charge on a highly charged protein can be strongly destabilizing. This is in good agreement with experiments. First, the magnifying effect of a protein's net charge on the degree of destabilization is consistent with results from charge ladder experiments, where stability is measured across a broad range of net charge (Gitlin et al., 2006). Second, the stability of the positively charged staphylococcal nuclease (Meeker et al., 1996; Schwehm et al., 2003) has been observed to be more sensitive to point mutations that remove negative side chains than to positive ones. Those studies show that $\Delta\Delta G$ scales linearly with ΔQ, consistent with Equation 2. Third, the extent of destabilization predicted by the model is similar to that observed in experimental studies of point mutations in which charged residues are exchanged for neutral ones that are good proxies for the oxidative by-products (Tokuriki et al., 2007), as shown in Figure 3 and Supplemental Section 4. It is therefore quite clear that, in general, single-charge mutations can significantly affect protein stability, a fact that has been used to design proteins with increased stability (Makhatadze et al., 2003).

Equation 2 makes another point. Random damage across a proteome should lead to little or no change in the *average* stability of the proteome, because half the events will be stabilizing (by bringing the net charge closer to neutrality) and half will be destabilizing (Figure 3 and Supplemental Sections 3B and 4). Even so, the consequences are not symmetrical. Consider two copies of a marginally stable protein. In a healthy proteome, these two copies are sufficiently stable. Now, consider two damage events: one stabilizes one copy of that protein and one destabilizes the other copy. Now we have only one stable protein, where before we had two. This is net destructive for the proteome (see Supplemental Section 3B for further discussion).

Despite the simplicity of Equation 2, its main predictions are likely to be quite robust to the details (Supplemental Section 3D). We can use it as a bioinformatics tool to search databases for proteins at high risk for oxidative destabilization. Given only an amino acid sequence, it can be used for rapid scanning of whole proteomes and comparisons across species (see Supplemental Sections 1B and 1C).

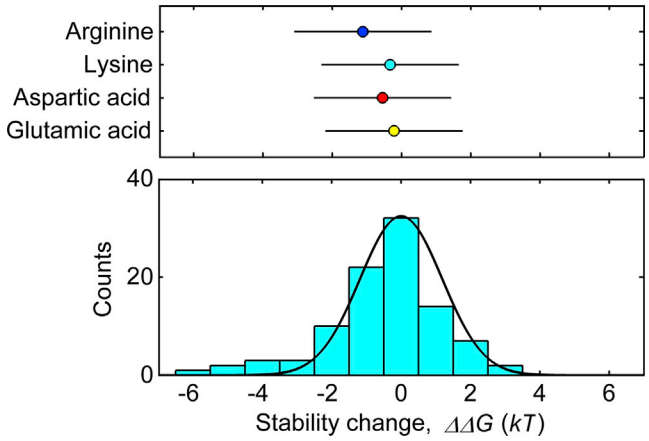


Figure 3. Distribution of Stability Change upon Single Charge Modifications

Experimental effect of the charge modification $Q \rightarrow Q \pm 1$ on protein stability from mutation data (histogram) compared with the model prediction (Equation 2) averaged across the human proteome (curve). Only mutations that turned charged residues into uncharged oxidation-product analogs (top) were used in the histogram. Error bars represent the SD of the data (see Supplemental Section 4, Figure S2, and Table S1).

Many Aging-Related Proteins Are at High Risk of Oxidative Destabilization

Which specific proteins are at high risk for electrostatic destabilization by a single oxidation event? Figure 4A, which shows key predictions from the model, contains three types of information. First, the color variation shows the biophysical model's distribution of protein destabilization according to Equation 2: the proteins expected to experience the highest degree of destabilization are shown as the reddest (high charge, short chains) and those at lowest risk are the bluest (near neutral, long chains). Second, the black line, computed from bioinformatics distributions, shows one SD of charge in the human proteome. Most proteins will be affected by less than 2 kT by a single oxidative damage event, as roughly two-thirds of the human proteome lies in the low-risk (blue) region below the black line. Third, Figure 4A contains 20 data points, indicating specific human proteins that are of known importance to aging or are involved in a process important to aging (Tacutu et al., 2013). They are plotted on the figure according to their charge and chain length (Table 2).

The electrostatic potential at the surface of these high-risk proteins differs markedly from their low-risk counterparts, as shown in Figure 5, by comparing the domains of two of the high-risk proteins identified by the model, telomerase reverse transcriptase (*TERT*) and nucleosome-remodeling factor subunit RbAp48 (*RBBP4*), with the low-risk protein ubiquitin. Both the TERT and RbAp48 domains have almost exclusively positive and negative surface potentials, respectively, indicating that any increase in their net charge will add to the electrostatic repulsion and destabilize their native state. Given that experimental stability data (Sawle and Ghosh, 2011) and modeling (Supplemental Section 2 and Figure S3) indicate that a significant fraction of natively folded proteins have stabilities of only 2–4 kT, the model suggests that the folded structure of these 20 proteins could be susceptible to losing most, if not all, of their stability from a single charge modification.

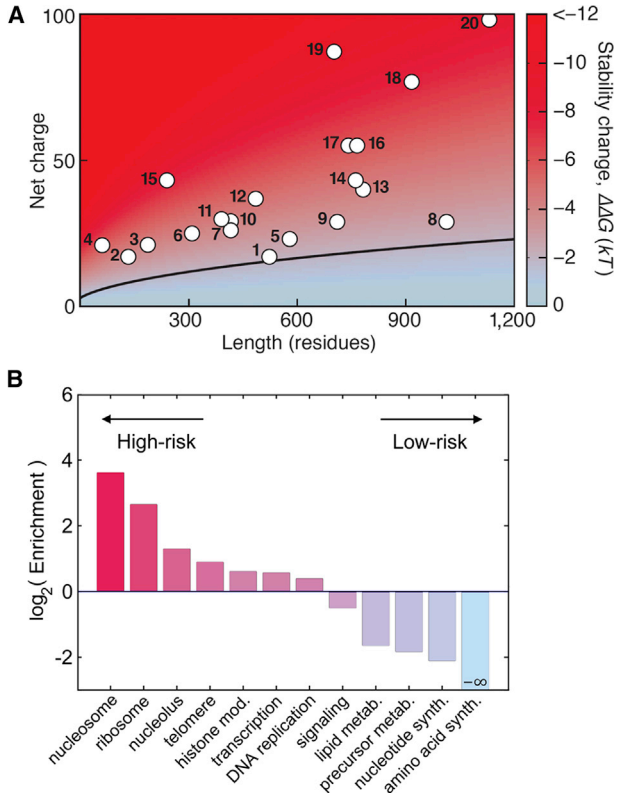


Figure 4. How a Single Charge Modification Affects the Stability of Proteins and Pathways

(A) Extent of destabilization from changing net protein charge from Q to $Q \pm 1$ depends on protein chain length and net charge, as given by Equation 2. Proteins can be partitioned into those whose stability is predicted to be robust to oxidation (blue) and those that are predicted to be heavily destabilized by oxidation (red). Black curve: one SD from neutrality within the human proteome (proteome-wide length and charge distribution is shown in Supplemental Sections 1B and 1C). The human proteins represented as white circles are both predicted to be particularly sensitive to oxidative destabilization and are of known importance to aging or involved in a process important to aging (Tacutu et al., 2013). See Supplemental Section 3D and Figure S3 for the sensitivity of the results to the radius of the denatured state.

(B) Certain cellular functions and pathways are enriched in highly charged proteins, and are thus predicted to be more sensitive to oxidative damage. For this enrichment, proteins are considered highly charged outliers if their net charge is greater than two SDs from neutrality (Supplemental Section 1B). Of the 14,079 human proteins verified at the protein level, 979 are outliers. Colors reflect the degree of enrichment within each GO cateogy (Table S2).

Next, we investigated how high-risk human proteins are distributed across categories of biological function. Proteins implicated in aging are often involved in one of the following dysfunctions: (1) altered packing of DNA around histones, (2) abnormal histone modification, (3) decreased telomere stability, (4) decreased transcriptional response to stress, and (5) decreased protein translation and degradation (Lepez-Otin et al., 2013). Consistent with our model, most of the protein categories conducting these processes are significantly enriched in high-risk proteins, as shown in Figure 4B and Table S2. Furthermore, most of these functions also tend to be conducted in the

Table 2. Human Proteins Predicted to Be Heavily Destabilized by a Small Change in Charge

	Gene Name	Function	Charge	Length (aa)
1	*HSF1*	transcription regulator	−17	529
2	*H2AFX*	histone	+17	143
3	*IGF1*	hormone	+20	195
4	*SHFM1*	proteasome	−21	70
5	*HSP90AA1*	protein folding	−23	585
6	*NFKBIA*	transcription factor binding	−25	317
7	*RBBP7*	histone binding	−26	425
8	*PARP1*	poly ADP ribosylation	+29	1,014
9	*MTA1*	histone deacetylase	+29	715
10	*RBBP4*	histone acetylase	−29	425
11	*TERF2IP*	telomere	−30	399
12	*MDM2*	E3 ubiquitin ligase	−37	491
13	*ELN*	structure	+40	786
14	*TOP1*	transcription regulator	+43	765
15	*RPS6*	ribosome	+43	249
16	*APP*	receptor binding	−55	770
17	*SIRT1*	histone deacetylase	−55	747
18	*BCLAF1*	transcription regulator	+77	920
19	*PJA2*	E3 ubiquitin ligase	−87	708
20	*TERT*	telomerase	+98	1,132

All proteins listed have identified relevance to aging or aging-related processes (Tacutu et al., 2013). aa, amino acids.

nucleus, suggesting that protection of high-risk proteins may be a major benefit to keeping the nucleus free of damaging agents.

DISCUSSION

A Causal Role for Protein Destabilization in the Aging Phenotype

Decreased protein stability can have many negative consequences, including reduced catalytic activity, altered binding to other biomolecules, increased aggregation or degradation, and decreased steady-state abundance (Sharma and Rothstein, 1980). Such effects are observed in many of the proteins in our high-risk categories.

The most enriched involve DNA binders such as histones, which tend to have high positive charge to complement DNA's negative phosphate backbone and interact with specific bases. In addition to causing the loss of specific charge-charge interactions with DNA, which alone could alter function, oxidation can also affect binding and turnover through a change in protein stability. Indeed, the abundance of certain histones has been observed to decrease with age, with overexpression leading to increased life span (Feser et al., 2010; Hu et al., 2014). Given their roles as regulators of gene expression and protectors of DNA against damage, it is not unexpected that their decreased abundance affects transcriptional activity, cell metabolism, and, in the case of histone H2AX, DNA repair (Feser et al., 2010; Hu et al., 2014) (Figure 4A and Table 2).

Telomeres are also maintained by highly charged histone-like proteins, and require telomerase to maintain their length. Oxidative stress is known to increase the rate of telomere shortening several-fold. Interestingly, both the histone-like TERF2-interacting telomeric protein 1 of the shelterin complex and telomerase reverse transcriptase are very highly charged (*TERF2IP* and *TERT*, Table 2). The present model therefore offers a plausible mechanism connecting telomere shortening to oxidative stress.

The tightness with which proteins bind to DNA is regulated by protein-modifying enzymes. Perhaps the most important in the aging field, the deacetylase SIRT1, is predicted to be among the highest-risk proteins in the human proteome. While SIRT1's fragility to oxidative damage has not, to our knowledge, been tested directly, its activity is known to decrease with age (Braidy et al., 2011) while its overexpression slows aging (Satoh et al., 2013). Similarly, the levels of the highly charged chromatin remodeling factors RbAp46/48 (*RBBP7*/*RBBP4*, Table 2) and MTA1 of the NuRD complex also decrease markedly with age, which is thought to play an important role in age-related memory loss (Alqarni et al., 2014).

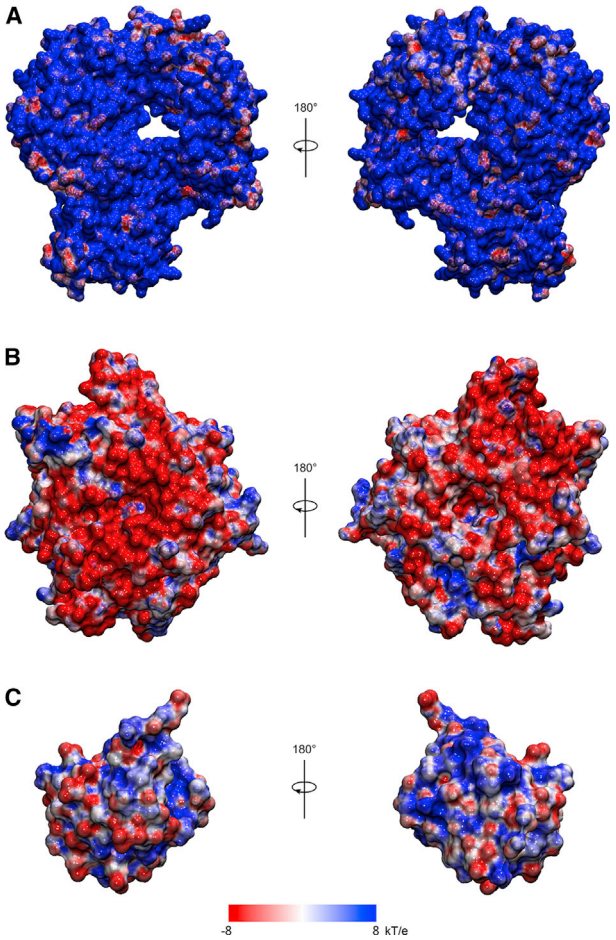


Figure 5. Sensitivity to Oxidative Destabilization in High-Risk Proteins Is Caused by the Strong Electrostatic Potential at Their Surface

The potential at the surface of (A) the positively charged telomerase reverse transcriptase (point 20 in Figure 4A and Table 2; PDB: 3KYL) and (B) the negatively charged nucleosome-remodeling factor subunit RbAp48 (point 10 in Figure 4A and Table 2; PDB: 2XU7) differ greatly from the weak potential at the surface of (C) the neutral protein ubiquitin (PDB: 1UBQ).

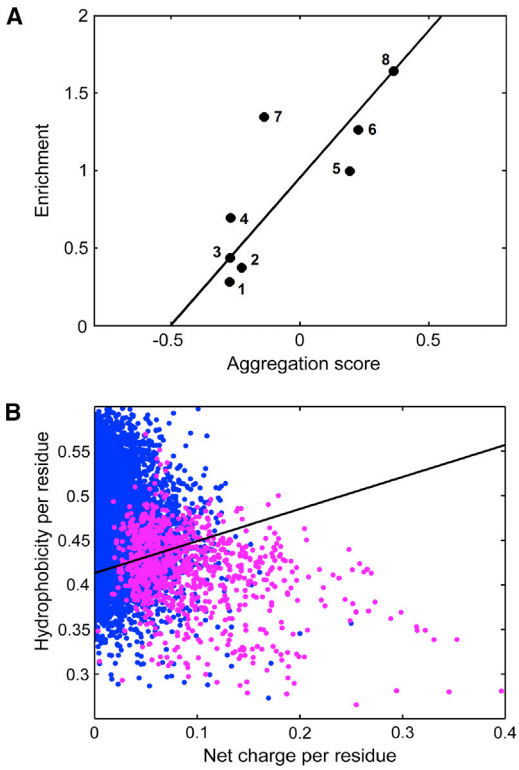


Figure 6. High-Risk Proteins Are More Likely to Aggregate

(A) The protein GO categories found to be enriched in the aggregates of aging *C. elegans* (aggregation score [Walther et al., 2015]) are the same categories predicted by the model to be enriched in high-risk proteins, defined as those more than two SDs from neutrality. Due to the low number of reviewed *C. elegans* proteins, the standard deviations from the human proteome were used (Supplemental Section 1B). GO categories in the figure are: 1, oxidation-reduction; 2, catalytic activity; 3, proteolysis; 4, metabolic process; 5, nucleus; 6, nucleic acid binding (includes RNA); 7, extracellular; 8, DNA binding. The black line is a guide to the eye.

(B) Proteins predicted to be at high risk of oxidative destabilization (more than two SDs from neutrality) are also more likely to have a less structured folded state prior to oxidation, as shown by the position of high-risk proteins (purple) on an Uversky plot (Uversky, 2002) relative to that of the entire human proteome (blue). The black line is the boundary previously found to delineate standard, folded proteins from those with unstructured tendencies (Uversky, 2002).

While several of the proposed mechanisms require further experimental testing, there is already direct evidence for the role of thermal stability in the declining inducibility of stress-resistance pathways such as the heat shock response. The heat shock response is activated by the binding of the transcription factor HSF-1 to DNA. Consistent with its high-risk status, evidence now suggests that HSF-1 loses its ability to provoke a strong stress response, in part because of the loss of its thermal stability when in complex with DNA, which limits the duration it is bound (Heydari et al., 2000). The heat shock response may be further sensitized to oxidative damage by the fact that HSF-1 binds DNA as a trimer, meaning that only one-third of HSF-1 needs to be damaged for most trimers to have an affected subunit.

Finally, the predicted sensitivity of numerous protein synthesis and degradation enzymes (*RPS6*, *MDM2*, *PJA2*, and *SHFM1*, Table 2) to oxidative damage is consistent with the observation that the rates of translation (Motizuki and Tsurugi, 1992) and degradation (Carney et al., 1991; Ryazanov and Nefsky, 2002; Smith et al., 1991; Starkereed and Oliver, 1989) decrease several-fold with age, with reduced catalytic capacity of their respective enzymes being partly responsible (Carney et al., 1991; Motizuki and Tsurugi, 1992). While free radical production often draws the spotlight when discussing protein damage levels, reduced protein turnover plays an equally important role (Ryazanov and Nefsky, 2002).

Aggregates Are Enriched in High-Risk Proteins

Further evidence for the hypothesis that highly charged proteins are the Achilles' heel of an aging proteome comes from examining protein aggregates in aging organisms, which are known to be enriched in oxidized proteins (Erjavec et al., 2007). First, aggregates in both aging budding yeast (Peters et al., 2012) and aging *Caenorhabditis elegans* (David et al., 2010; Reis-Rodrigues et al., 2012) are heavily enriched in ribosomal and nucleic acid-associated proteins, which are high-charge categories. Second, Figure 6A shows that the types of proteins that are abundant in the aggregates of aging cells (Walther et al., 2015) correlate very well with our predicted risk categories, based on net charge. Interestingly, the observation that highly charged proteins tend to aggregate goes against common wisdom that highly charged proteins do not aggregate (Lawrence et al., 2007).

Three key properties may help explain the overabundance of high-risk proteins in aggregates of aging organisms. First, high net charge is known to be a key predictor of unstable, disorder-prone proteins, as can be seen by the position of high-risk proteins on an Uversky plot (Uversky, 2002; Uversky et al., 2008) (Figure 6B and Supplemental Section 3C). To remain folded, such proteins may need to rely more heavily on the presence of their binding partners (Uversky et al., 2008), which can become problematic given that proper protein stoichiometry is lost with age (Walther et al., 2015). Second, disorder and low stability are strong indicators of a protein's likelihood of becoming damaged (Vidovic et al., 2014), and thus of having its stability perturbed in the first place. Third, based on sound principles of electrostatics, our model predicts that the high charge of these proteins amplifies the stability change resulting from a charge modification, a common outcome of oxidative damage, deamidation, and changes in pH (Henderson et al., 2014; Requena et al., 2001; Robinson and Robinson, 2001). These three properties converge to make highly charged proteins the weak link in aging proteomes.

In summary, we have proposed a mechanism by which oxidative damage affects proteins in aging cells. It predicts that random oxidative damage to protein side chains may be a major source of proteome-wide stability loss in aging cells (Stadtman, 1992, 2006; Starkereed and Oliver, 1989); that many proteins are hit non-specifically; that the protein structures at greatest risk for substantial destabilization from single oxidation events have high net charge and low stability; and that the stability losses from single charge modifications can be equal to the full native stabilities of these proteins. Reduced protein stability has various consequences, from decreased catalytic activity, to weakened protein-protein and protein-nucleic acid interactions, to increased aggregation, which are all hallmarks of aging cells

(Lepez-Otin et al., 2013). This mechanism therefore provides a hypothesis for searching databases for proteins that may be important to aging and aging-related diseases.

EXPERIMENTAL PROCEDURES

Protein Sequence Collection and Filtering

Proteome-wide sequence data were obtained from UniProt (Bateman et al., 2015). No post-translational modifications were applied, and only proteins that have been verified at the protein level (protein existence score of 1) were used. The resulting proteomes were then used to calculate the species-specific length and charge distributions shown in Figure S2, Table S1, and Supplemental Sections 1B and 1C. The analytical length and charge distribution determined for the human proteome was used to calculate the distribution of stability changes from single charge perturbations using Equation 2, as shown in Figure 3. It was assumed that all proteins had an equal chance of having their net charge perturbed from Q to $Q + 1$ as they did from Q to $Q - 1$.

Calculation of Protein Charge

The net charges of proteins in the denatured (Q_d) and native (Q_n) states were estimated from the standard pK_a values of the side chains composing their protein sequence (Lehninger et al., 2005), with the exception of histidine in the folded state. The average degree of protonation of histidine in the folded state, which deviates markedly from its standard value, was determined by taking the average of experimentally measured values across a set of folded proteins (Wisz and Hellinga, 2003). Native state pK_as of the other side chains are generally sufficiently acidic or basic that the native environment does not significantly affect their average protonation state, as shown by the histograms of experimental pK_as in Figure S1 and discussed in Supplemental Sections 1A and 3D.

Enrichment of High-Risk Proteins in Functional Categories

The enrichment of high-risk proteins within functional categories of the human proteome, shown in Figure 4B, was determined in several steps. First, the human proteome includes only the 14,079 proteins that have been verified at the protein level (protein existence score of 1) according to UniProt (Bateman et al., 2015). These verified proteins were then used to determine how the distribution of charge within proteins varies with protein length. As shown in Figure S2E of Supplemental Section 1B, the charge distribution is approximately Gaussian at each protein length, with the variance increasing linearly with length (Equation S4D). Proteins were deemed high-risk outliers if their charge was at least two SDs from neutrality. Second, the set of proteins belonging to each functional category was determined by gene ontology (GO) categories (Ashburner et al., 2000), as shown in Table S2. Enrichment was defined as the fraction of high-risk proteins within each GO category relative to their fraction in the total set of verified human proteins. p Values for these enrichments were determined using Fisher's exact one-tailed test (Table S2).

Comparison with Mutant Studies

The distribution of stability changes from single point mutations (Figure 3) was used as a proxy for the distribution expected for oxidative damage to charged side chains. Mutation data were obtained from the large dataset in Tokuriki et al. (2007), from which we extracted all single point mutations to amino acids that both: (1) are solvent-exposed in the wild-type protein (requiring at least 25% of the amino acid's surface area be accessible to the solvent [Tokuriki et al., 2007]) and (2) exchange charged amino acids for the neutral amino acids methionine, cysteine, alanine, or threonine. These four amino acids were chosen because their hydrophobicities most closely resemble those of the common oxidation by-products amino-adipic semialdehyde and glutamic semialdehyde (Petrov and Zagrovic, 2014), as discussed in Supplemental Section 4.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Modelling Material and Methods, three figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.11.006.

AUTHOR CONTRIBUTIONS

A.M.R.d.G. and K.A.D. conceived the idea. M.H. and A.M.R.d.G. retrieved data and performed bioinformatics analyses. A.M.R.d.G. and K.A.D. wrote the manuscript. All authors discussed the results and commented on the manuscript at all stages.

ACKNOWLEDGMENTS

We thank Drs. Thomas Nystrom, John van Drie, Kathlyn Parker, Kingshuk Ghosh, and Purushottam Dixit for insightful conversations, Dr. Sarina Bromberg for help with graphics, and the Laufer Center and NSF grant 1205881 for support.

Received: July 28, 2015
Revised: October 11, 2015
Accepted: November 11, 2015
Published: December 24, 2015

REFERENCES

Adachi, H., Fujiwara, Y., and Ishii, N. (1998). Effects of oxygen on protein carbonyl and aging in *Caenorhabditis elegans* mutants with long (age-1) and short (mev-1) life spans. J. Gerontol. A Biol. Sci. Med. Sci. *53*, B240–B244.

Alqarni, S.S.M., Murthy, A., Zhang, W., Przewloka, M.R., Silva, A.P.G., Watson, A.A., Lejon, S., Pei, X.Y., Smits, A.H., Kloet, S.L., et al. (2014). Insight into the architecture of the NuRD complex structure of the RbAp48-MTA1 subcomplex. J. Biol. Chem. *289*, 21844–21855.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Balch, W.E., Morimoto, R.I., Dillin, A., and Kelly, J.W. (2008). Adapting proteostasis for disease intervention. Science *319*, 916–919.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Ar-Ganiska, J., Bely, B., Bingley, M., et al. (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

Ben-Zvi, A., Miller, E.A., and Morimoto, R.I. (2009). Collapse of proteostasis represents an early molecular event in *Caenorhabditis elegans* aging. Proc. Natl. Acad. Sci. USA *106*, 14914–14919.

Braidy, N., Guillemin, G.J., Mansour, H., Chan-Ling, T., Poljak, A., and Grant, R. (2011). Age related changes in NAD plus metabolism oxidative stress and Sirt1 activity in Wistar rats. PLoS One *6*, e19194.

Carney, J.M., Starkereed, P.E., Oliver, C.N., Landum, R.W., Cheng, M.S., Wu, J.F., and Floyd, R.A. (1991). Reversal of age-related increase in brain protein oxidation, decrease in enzyme activity, and loss in temporal and spatial memory by chronic administration of the spin-trapping compound n-tert-butyl-alpha-phenylnitrone. Proc. Natl. Acad. Sci. USA *88*, 3633–3636.

David, D.C., Ollikainen, N., Trinidad, J.C., Cary, M.P., Burlingame, A.L., and Kenyon, C. (2010). Widespread protein aggregation as an inherent part of aging in C. elegans. PLoS Biol. *8*, e1000450.

Davies, K.J.A., Delsignore, M.E., and Lin, S.W. (1987). Protein damage and degradation by oxygen radicals .2. Modification of amino acids. J. Biol. Chem. *262*, 9902–9907.

Dill, K.A., and Bromberg, S. (2011). Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience, Second Edition (Garland Science), pp. 1–756.

Dill, K.A., and Stigter, D. (1995). Modeling protein stability as heteropolymer collapse. Adv. Protein Chem. *46*, 59–104.

Erjavec, N., Larsson, L., Grantham, J., and Nystrom, T. (2007). Accelerated aging and failure to segregate damaged proteins in Sir2 mutants can be suppressed by overproducing the protein aggregation-remodeling factor Hsp104p. Genes Dev. *21*, 2410–2421.

Feser, J., Truong, D., Das, C., Carson, J.J., Kieft, J., Harkness, T., and Tyler, J.K. (2010). Elevated histone expression promotes life span extension. Mol. Cell *39*, 724–735.

Ghosh, K., and Dill, K.A. (2009). Computing protein stabilities from their chain lengths. Proc. Natl. Acad. Sci. USA *106*, 10649–10654.

Gitlin, I., Carbeck, J.D., and Whitesides, G.M. (2006). Why are proteins charged? Networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis. Angew. Chem. Int. Ed. Engl. *45*, 3022–3060.

Henderson, K.A., Hughes, A.L., and Gottschling, D.E. (2014). Mother-daughter asymmetry of pH underlies aging and rejuvenation in yeast. Elife *3*, e03504.

Heydari, A.R., You, S.H., Takahashi, R., Gutsmann-Conrad, A., Sarge, K.D., and Richardson, A. (2000). Age-related alterations in the activation of heat shock transcription factor 1 in rat hepatocytes. Exp. Cell Res. *256*, 83–93.

Hipkiss, A.R. (2006). Accumulation of altered proteins and ageing: causes and effects. Exp. Gerontol. *41*, 464–473.

Hu, Z., Chen, K.F., Xia, Z., Chavez, M., Pal, S., Seol, J.H., Chen, C.C., Li, W., and Tyler, J.K. (2014). Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. Genes Dev. *28*, 396–408.

Kohn, J.E., Millett, I.S., Jacob, J., Zagrovic, B., Dillon, T.M., Cingel, N., Dothager, R.S., Seifert, S., Thiyagarajan, P., Sosnick, T.R., et al. (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. Proc. Natl. Acad. Sci. USA *101*, 12491–12496.

Lawrence, M.S., Phillips, K.J., and Liu, D.R. (2007). Supercharging proteins can impart unusual resilience. J. Am. Chem. Soc. *129*, 10110.

Lee, K.K., Fitch, C.A., and Garcia-Moreno, B. (2002). Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein. Protein Sci. *11*, 1004–1016.

Lehninger, A.L., Nelson, D.L., and Cox, M.M. (2005). Lehninger Principles of Biochemistry, Fourth Edition (W.H. Freeman).

Lepez-Otin, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194–1217.

Makhatadze, G.I., Loladze, V.V., Ermolenko, D.N., Chen, X.F., and Thomas, S.T. (2003). Contribution of surface salt bridges to protein stability: guidelines for protein engineering. J. Mol. Biol. *327*, 1135–1148.

Meeker, A.K., Garcia-Moreno, B., and Shortle, D. (1996). Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. Biochemistry *35*, 6443–6449.

Motizuki, M., and Tsurugi, K. (1992). The effect of aging on protein synthesis in the yeast *Saccharomyces cerevisiae*. Mech. Ageing Dev. *64*, 235–245.

Oliver, C.N., Ahn, B.W., Moerman, E.J., Goldstein, S., and Stadtman, E.R. (1987). Age-related changes in oxidized proteins. J. Biol. Chem. *262*, 5488–5491.

Peters, T.W., Rardin, M.J., Czerwieniec, G., Evani, U.S., Reis-Rodrigues, P., Lithgow, G.J., Mooney, S.D., Gibson, B.W., and Hughes, R.E. (2012). Tor1 regulates protein solubility in *Saccharomyces cerevisiae*. Mol. Biol. Cell *23*, 4679–4688.

Petrov, D., and Zagrovic, B. (2011). Microscopic analysis of protein oxidative damage: effect of carbonylation on structure, dynamics, and aggregability of villin headpiece. J. Am. Chem. Soc. *133*, 7016–7024.

Petrov, D., and Zagrovic, B. (2014). Microscopic analysis of protein oxidative damage: effect of carbonylation on structure, dynamics, and aggregability of villin headpiece (vol 133, pg 7016, 2011). J. Am. Chem. Soc. *136*, 2175–2176.

Rao, R.S.P., and Moller, I.M. (2011). Pattern of occurrence and occupancy of carbonylation sites in proteins. Proteomics *11*, 4166–4173.

Reis-Rodrigues, P., Czerwieniec, G., Peters, T.W., Evani, U.S., Alavez, S., Gaman, E.A., Vantipalli, M., Mooney, S.D., Gibson, B.W., Lithgow, G.J., et al. (2012). Proteomic analysis of age-dependent changes in protein solubility identifies genes that modulate lifespan. Aging Cell *11*, 120–127.

Requena, J.R., Chao, C.C., Levine, R.L., and Stadtman, E.R. (2001). Glutamic and aminoadipic semialdehydes are the main carbonyl products of metal-catalyzed oxidation of proteins. Proc. Natl. Acad. Sci. USA *98*, 69–74.

Robinson, N.E., and Robinson, A.B. (2001). Deamidation of human proteins. Proc. Natl. Acad. Sci. USA *98*, 12409–12413.

Ryazanov, A.G., and Nefsky, B.S. (2002). Protein turnover plays a key role in aging. Mech. Ageing Dev. *123*, 207–213.

Satoh, A., Brace, C.S., Rensing, N., Cliften, P., Wozniak, D.F., Herzog, E.D., Yamada, K.A., and Imai, S. (2013). Sirt1 extends life span and delays aging in mice through the regulation of Nk2 homeobox 1 in the DMH and LH. Cell Metab. *18*, 416–430.

Sawle, L., and Ghosh, K. (2011). How do thermophilic proteins and proteomes withstand high temperature? Biophys. J. *101*, 217–227.

Schwehm, J.M., Fitch, C.A., Dang, B.N., Garcia-Moreno, B., and Stites, W.E. (2003). Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effect. Biochemistry *42*, 1118–1128.

Shacter, E. (2000). Quantification and significance of protein oxidation in biological samples. Drug Metab. Rev. *32*, 307–326.

Sharma, H.K., and Rothstein, M. (1980). Altered enolase in aged *Turbatrix aceti* results from conformational changes in the enzyme. Proc. Natl. Acad. Sci. USA *77*, 5865–5868.

Smith, C.D., Carney, J.M., Starkereed, P.E., Oliver, C.N., Stadtman, E.R., Floyd, R.A., and Markesbery, W.R. (1991). Excess brain protein oxidation and enzyme dysfunction in normal aging and in Alzheimer disease. Proc. Natl. Acad. Sci. USA *88*, 10540–10543.

Sohal, R.S., Agarwal, S., Dubey, A., and Orr, W.C. (1993). Protein oxidative damage is associated with life expectancy of houseflies. Proc. Natl. Acad. Sci. USA *90*, 7255–7259.

Stadtman, E.R. (1992). Protein oxidation and aging. Science *257*, 1220–1224.

Stadtman, E.R. (2006). Protein oxidation and aging. Free Radic. Res. *40*, 1250–1258.

Starkereed, P.E., and Oliver, C.N. (1989). Protein oxidation and proteolysis during aging and oxidative stress. Arch. Biochem. Biophys. *275*, 559–567.

Stigter, D., Alonso, D.O.V., and Dill, K.A. (1991). Protein stability - electrostatics and compact denatured states. Proc. Natl. Acad. Sci. USA *88*, 4176–4180.

Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., and de Magalhaes, J.P. (2013). Human ageing genomic Resources: Integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res. *41*, D1027–D1033.

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D.S. (2007). The stability effects of protein mutations appear to be universally distributed. J. Mol. Biol. *369*, 1318–1332.

Uversky, V.N. (2002). Natively unfolded proteins: a point where biology waits for physics. Protein Sci. *11*, 739–756.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D(2) concept. Annu. Rev. Biophys. *37*, 215–246.

Vidovic, A., Supek, F., Nikolic, A., and Krisko, A. (2014). Signatures of conformational stability and oxidation resistance in proteomes of pathogenic bacteria. Cell Rep. *7*, 1393–1400.

Walther, D.M., Kasturi, P., Zheng, M., Pinkert, S., Vecchi, G., Ciryam, P., Morimoto, R.I., Dobson, C.M., Vendruscolo, M., Mann, M., et al. (2015). Widespread proteome remodeling and aggregation in aging *C. elegans*. Cell *161*, 919–932.

Wisz, M.S., and Hellinga, H.W. (2003). An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins *51*, 360–377.

Xu, J., Reumers, J., Couceiro, J.R., De Smet, F., Gallardo, R., Rudyak, S., Cornelis, A., Rozenski, J., Zwolinska, A., Marine, J.-C., et al. (2011). Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. Nat. Chem. Biol. *7*, 285–295.

Structure, Volume *24*

Supplemental Information

Highly Charged Proteins: The Achilles' Heel of Aging Proteomes

Adam M. R. de Graff, Michael J. Hazoglou, and Ken A. Dill

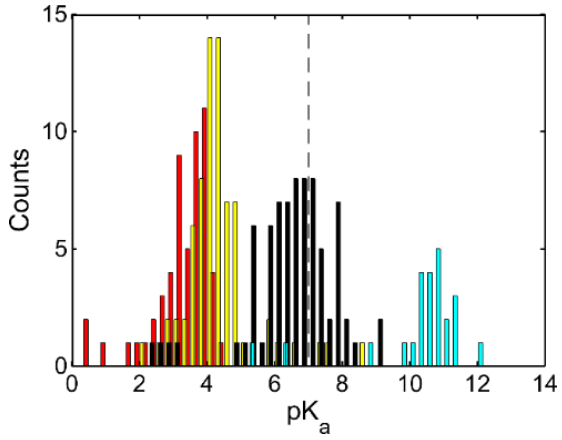


Figure S1, related to Figure 2. Experimentally measured pK_a values of side chains in folded state of proteins. From left to right: aspartic acid (red), glutamic acid (yellow), histidine (black), and lysine (cyan) (see Supplemental Section 1A).

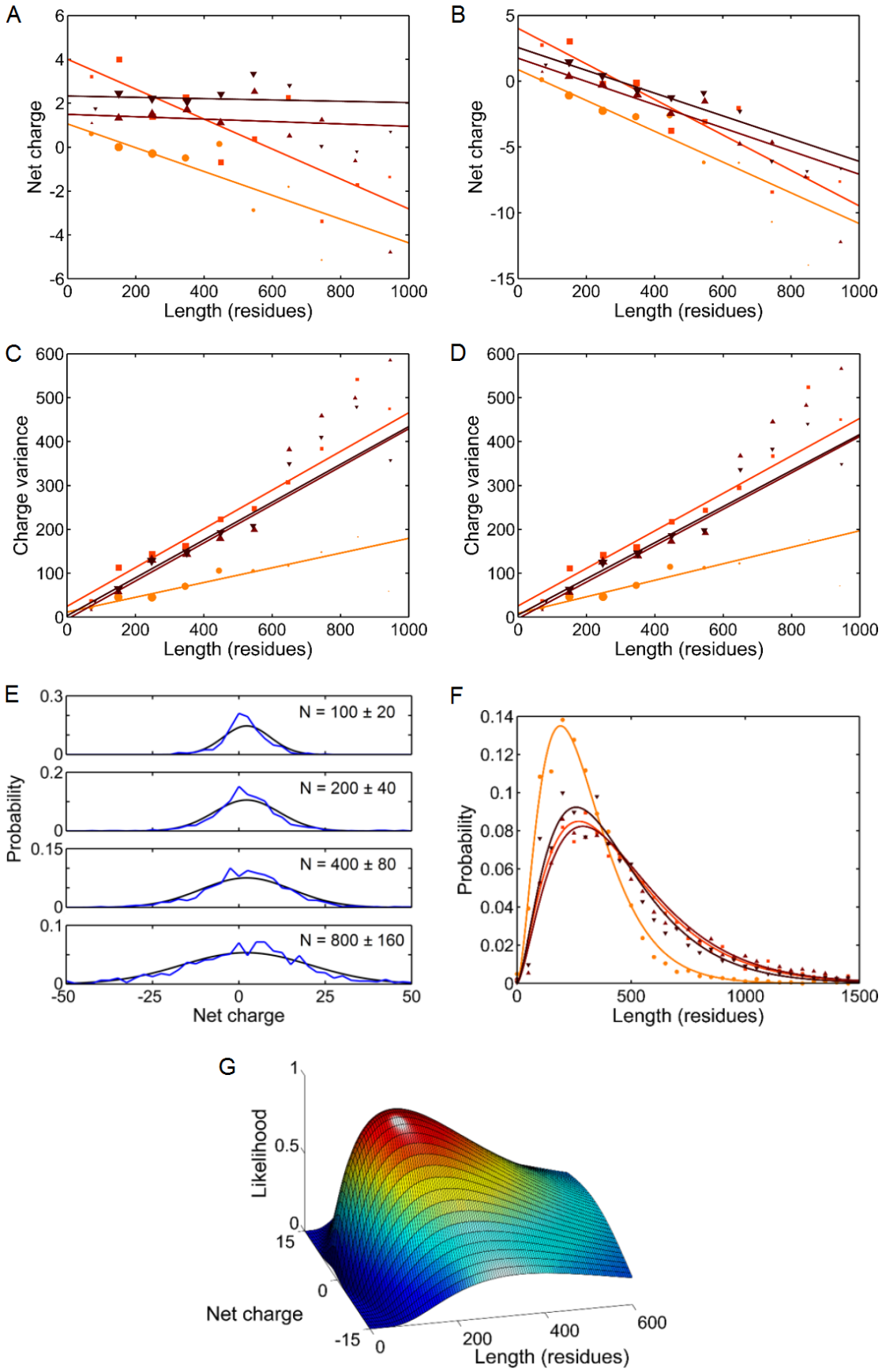

A
Net charge
Length (residues)
B
Net charge
Length (residues)
C
Charge variance
Length (residues)
D
Charge variance
Length (residues)
E
Probability
N = 100 ± 20
N = 200 ± 40
N = 400 ± 80
N = 800 ± 160
Net charge
F
Probability
Length (residues)
G
Likelihood
Net charge
Length (residues)

Figure S2, related to Figure 3. Distribution of protein charge and length. (A) Average folded protein charge as a function of protein chain length in *E. coli* (light orange circles), *S. cerevisiae* (orange squares), *M. musculus* (brown upward triangles), and *H. sapiens* (black downward triangles). Symbol areas are proportional to the number of proteins contributing to the average of each bin. (B) Average unfolded protein charge as a function of chain length. (C) Variance in the net charge of folded proteins as a function of protein chain length. (D) Variance in the net charge of unfolded proteins as a function of chain length (see Supplemental Section 1B). (E) Distribution of protein charge in the human proteome within slices of various lengths (blue) compared with Gaussians having averages and variances predicted from the length, given by Eqs. S2D and S4D respectively. (F) Experimental length distributions (symbols) and their respective fits to Gamma distributions (see Table S1). (G) The distribution $P_{Hsap}(N, Q)$ of the human proteome as a function of protein chain length *N* and net charge Q (see Supplemental Section 1C).

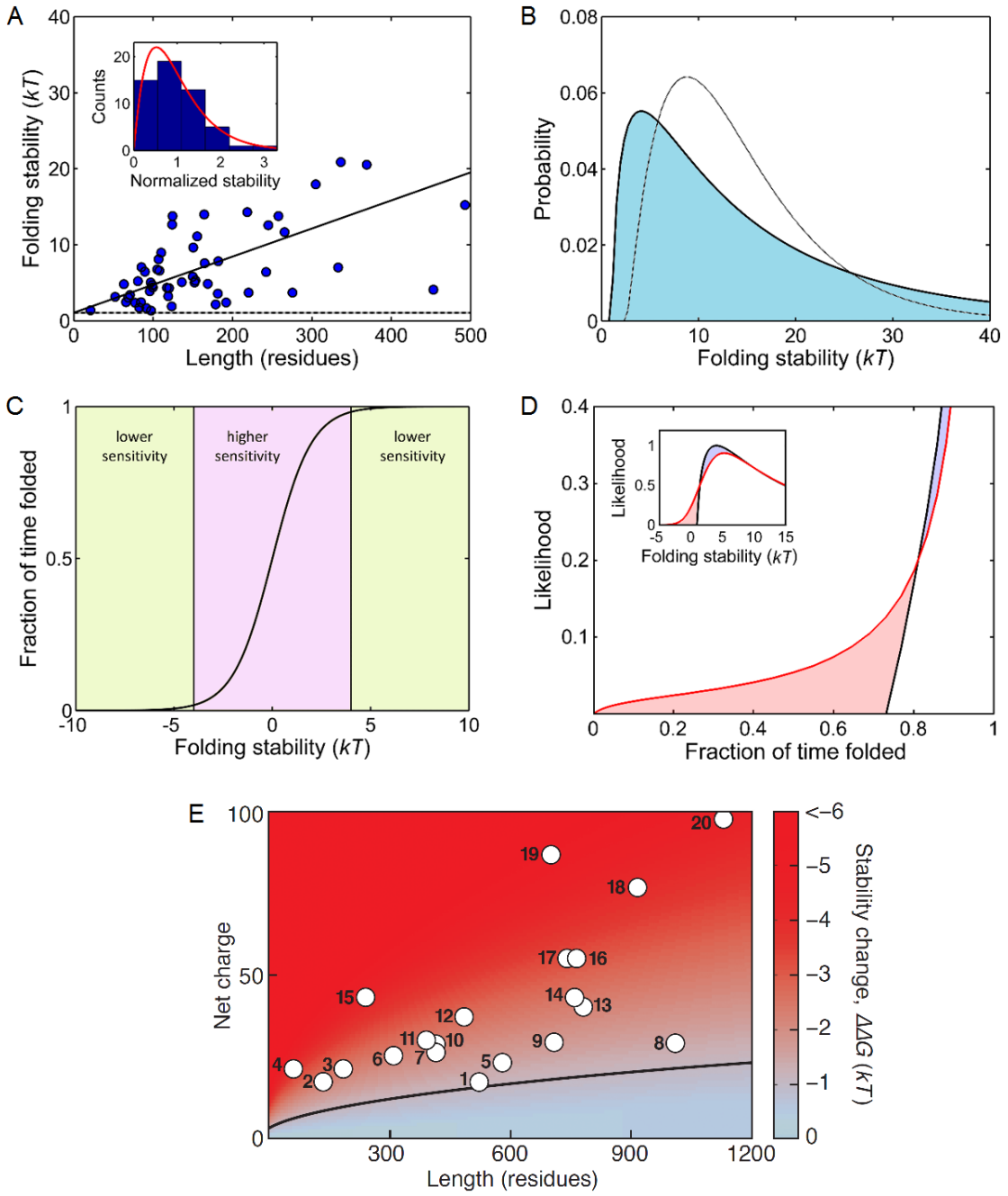


Figure S3, related to Figure 4. Many proteins are barely stable and can lose their stability from a single oxidation event. (A) Experimentally measured folding stability of proteins (circles), together with the length-dependent average stability $\overline{\Delta G}(N)$ (solid line) and the length-independent minimum stability ΔG_{min} (dashed line). (Inset) Histogram of experimental stabilities normalized to $\overline{\Delta G}(N)$ compared with a normalized Gamma distribution (average equal to one, red curve). See Supplemental Section 2A. (B) Predicted folding stability distribution $P_{Hsap}(\Delta G)$ of the human proteome based solely on the chain length distribution $P_{Hsap}(N)$ (grey dotted curve, scaled using Eq. S7) compared to the prediction after the natural variation in stability at constant protein chain length (the scatter about the line in Fig. S3A) is properly accounted for using Eq. S8 (solid curve and shaded region) (see Supplemental Section 2B). (C) The fraction of time a reversibly folding protein is folded depends in its stability. This fraction is particularly sensitive to the stability between roughly ±4 kT. (D) (Inset) Comparison of the stability distribution of the

human proteome when all proteins are unoxidized (black) and singly oxidized (red). (Main) The same data as the inset, but plotted as a function of the fraction of time each protein is folded, as given by Eq. S11. While small in number, the tail of the oxidized distribution at low stability contributes disproportionately to the unfolded protein load in a cell (see Supplemental Section 3B). (E) The stability change as a function of protein chain length and net charge, predicted using a more compact denatured state than that of Fig. 4A. Proteins can be partitioned into those whose stability is predicted to be robust to oxidation (blue) and those that are predicted to be heavily destabilized by oxidation (red). Black curve: one standard deviation from neutrality within the human proteome (Supplemental Section 1B). The human proteins shown as white circles are both predicted to be particularly sensitive to oxidative destabilization and are either of known importance to aging or are involved in a process important to aging (Table 2). See Supplemental Section 3D.

Table S1, related to Figure 3. Parameters for species-specific protein length distributions.

Species	k	θ (aa)	$\mu = k\theta$ (aa)
E. coli	2.82	105	295
S. cerevisiae	2.49	182	453
M. musculus	2.57	183	472
H. sapiens	2.59	163	421

Parameter values were determined by fitting each proteome to a Gamma distribution, as shown in Eq. S6. Only proteins whose existence has been experimentally verified at the protein level were used. *aa* = amino acids.

Table S2, related to Figure 4. Enrichment in charged proteins varies greatly between functional categories.

Function	GO term	High-charge	Total	Enrichment	p Value
Nucleosome	0000786	51	60	12.2	$< 2.2 \times 10^{-16}$
Ribosome	0005840	85	195	6.3	$< 2.2 \times 10^{-16}$
Nucleolus	0005730	133	781	2.4	$< 2.2 \times 10^{-16}$
Telomere organization	0032200	8	62	1.9	0.064
Histone modification	0016570	31	293	1.5	1.3×10^{-2}
Transcription	0006351	188	1828	1.5	7.8×10^{-9}
DNA replication	0006260	18	198	1.3	0.15
Signaling	0023052	206	4181	0.7	1.6×10^{-10}
Lipid metabolism	0006629	21	944	0.3	1.3×10^{-11}
Precursor metab. & energy	0006091	7	358	0.3	1.3×10^{-5}
Nucleotide synthesis	0009165	3	186	0.2	7.9×10^{-4}
Amino acid synthesis	0008652	0	94	0	1.1×10^{-3}

Proportion of proteins that are highly charged outliers were determined for several GO categories. Highly charged outliers are defined as proteins having a net charge at least two standard deviations from neutrality. There was a total of 979 highly charged outliers among the 14,079 human proteins whose existence has been experimentally verified at the protein level. p Values were found using Fisher's exact one-tailed test.

Supplemental Modelling Material and Methods

1A. Protein charge can be calculated from folding-dependent pK_a values

Protein charge plays an important role in protein folding. To estimate the contribution of charge-charge interactions to the stability of the folded state, it is necessary to know the charge of the protein in both the folded and denatured states, as shown in Eq. 1. While the charge of the denatured state can be estimated from standard pK_a values (Lehninger et al., 2005), this is not appropriate for the folded state, particularly for side chains with pK_a values close to the pH of their surroundings. We derive effective pK_a values for folded proteins from experimental pK_a data (Wisz and Hellinga, 2003).

A histogram of these experimental pK_a values, as shown in Fig. S1, demonstrates that the pK_a values of very acidic or basic side chains are clustered around their standard values, with rare outliers in which glutamic acid is protonated or lysine is deprotonated tending to cancel each other out. We therefore model all charged amino acids using their standard pK_a values. In the experimental dataset (which does not included cysteine), only the pK_a values of histidine are close enough to 7 for it to contribute significantly to charge differences between the folded and denatured states. Given that the probability of protonation is

$$P = \frac{1}{1+10^{(pH-pKa)}} \quad \text{(S1)}$$

the overall protonation probability can be found by summing Eq. S1 over the experimental data. This summation results in the prediction that 40% of histidines are protonated in folded proteins, differing substantially from the value of 10% predicted from its standard pK_a ~ 6. All charges presented in this work were calculated using histidine protonation fractions of 0.40 and 0.10 for folded and denatured proteins respectively, equivalent to *effective* pK_a values of 6.82 and 6.05. Note that it is incorrect to use the *average* pK_a value to compute protein charge, as the protonation probability (Eq. S1) is not a linear function of the pK_a.

The pK_a of cysteine (~8) is also close enough to 7 for there to be two significant protonation states. However, the protonation state of cysteine is harder to model than histidine, most notably because cysteine side chains can form disulfide bridges. As it is not possible to properly account for this without knowing each protein's structure, we chose to model both the folded and denatured states using cysteine's standard pK_a value of 8. In summary, we use standard pK_a values for both the folded and unfolded states of all amino acids except for histidine, for which we use a folded pK_a of 6.82 and an unfolded pK_a of 6.05.

1B. The average and variance of protein charge depend linearly on protein chain length

Our model predicts that the stability change caused by oxidation of a charged side chain depends on the charge and length of a protein. Short and highly charged proteins are predicted to be at particularly high risk of large destabilization (Fig. 4A). In order to determine whether proteomes contain many of these susceptible proteins, we determined the charge and length distributions of several well-studied proteomes.

We collected the sequences of all proteins whose existence has been verified at the protein level for the species *E. coli*, *S. cerevisiae*, *M. musculus*, and *H. sapiens*. All proteomes display a broad distribution of net charge that can be well-characterized by a length-dependent average and a length-dependent variance, both displaying a linear dependence on chain length *N*. In the folded state, the average protein charge is given by

$$\mu_{Ecoli} = 1.1 - 0.0054N \quad \text{(S2A)}$$
$$\mu_{Scer} = 4.0 - 0.0068N \quad \text{(S2B)}$$
$$\mu_{Mmus} = 1.5 - 0.0005N \quad \text{(S2C)}$$
$$\mu_{Hsap} = 2.3 - 0.0003N \quad \text{(S2D)}$$

as shown in Fig. S2A. The enhanced positive charge of histidine in the folded state causes the charge of folded mouse and human proteins to be virtually independent of length. In contrast, applying traditional pK_a values to the unfolded state results in an average charge that becomes increasingly negative with increasing chain length in all four species (Fig. S2B), as given by

$$\mu_{Ecoli} = 0.9 - 0.0117N \quad \text{(S3A)}$$
$$\mu_{Scer} = 4.0 - 0.0135N \quad \text{(S3B)}$$
$$\mu_{Mmus} = 1.7 - 0.0088N \quad \text{(S3C)}$$
$$\mu_{Hsap} = 2.6 - 0.0086N \quad \text{(S3D)}$$

As with protein stability, the average charge does not tell the whole story, as it lacks information about variability. By plotting the variance of the net charge as a function of chain length for the folded and unfolded states (Figs. S2C and S2D respectively), we find that the charge variance σ^2 grows linearly with chain length over the range of lengths containing the majority of each organism's proteome. Slightly enhanced variance is seen in proteins containing more than 600 residues. The charge variance of the folded state goes as

$$\sigma^2_{Ecoli} = 11.7 + 0.17N \quad \text{(S4A)}$$
$$\sigma^2_{Scer} = 24.8 + 0.44N \quad \text{(S4B)}$$
$$\sigma^2_{Mmus} = -4.5 + 0.43N \quad \text{(S4C)}$$
$$\sigma^2_{Hsap} = 3.4 + 0.43N \quad \text{(S4D)}$$

while that of the unfolded state goes as

$$\sigma^2_{Ecoli} = 8.8 + 0.19N \quad \text{(S5A)}$$
$$\sigma^2_{Scer} = 25.8 + 0.43N \quad \text{(S5B)}$$
$$\sigma^2_{Mmus} = -3.0 + 0.41N \quad \text{(S5C)}$$
$$\sigma^2_{Hsap} = 5.3 + 0.41N \quad \text{(S5D)}$$

No systematic difference is observed between the variance of the folded and unfolded states, but there is a marked difference between the prokaryotic (*E. coli*) proteome and the three eukaryotic ones. The charge variance of *E. coli*'s proteins is ~50% lower, and thus its standard deviation ~25% lower, than eukaryotic proteins of the same length. All else being equal, our model predicts that this reduction makes the *E. coli* proteome much less susceptible to oxidative destabilization, a prediction that would be interesting to verify experimentally. Such robustness would be greatly beneficial in metabolically active organisms such as *E. coli* that are likely creating free radicals at a much higher rate per unit mass than eukaryotic organisms.

In order to show that the average and variance alone are sufficient to describe the distributions, the shape of the charge distribution was found for human proteins of similar length. For each length, the net charge closely follows a Gaussian distribution, as shown in Fig. S2E.

As Gaussians are entirely described by their mean and variance, these two quantities are therefore sufficient to accurately capture the shape of the proteome-wide charge distribution $P(Q|N)$ (the probability of *Q* given *N*). Knowing this distribution enables us to determine how many proteins are in the highly charged tails of the distribution predicted to be heavily destabilized by oxidative damage and thus potentially problematic in aged organisms.

1C. Protein chain length is Gamma-distributed

Chain length plays a central role in determining the folding stability of proteins (Supplemental Section 2A) and their sensitivity to oxidative destabilization (Eq. 2). Longer protein chains allow a net charge to be spread out over a larger volume thus reducing charge-charge repulsion, yet longer proteins chains also have the potential for much higher net charge, as shown in the previous section. It is therefore critical to know the distribution of chain lengths in order to assess the effect of oxidation on protein stability.

Using the same set of protein sequences used to calculate the charge distributions in the previous section, we calculated the length distribution $P(N)$ of the proteomes of *E. coli*, *S. cerevisiae*, *M. musculus*, and *H. sapiens*. All four distributions are well-characterized by a Gamma distribution

$$P(N) = N^{k-1}e^{-\frac{N}{\theta}}/(\Gamma(k)\theta^{k}) \quad \text{(S6)}$$

with a peak at ~200 amino acids for *E. coli* and ~250 for the three eukaryotes, as shown in Fig. S2F. Parameters *k* and *θ* for each species are shown in Table S1. Much like with net charge, the three eukaryotic proteomes cluster together, distinct from that of *E. coli*. While the shorter *E. coli* proteome may allow a greater fraction of the proteome to fold without the help of chaperones, it is likely to result in lower average protein stability, as stability generally increases linearly with protein chain length (Fig. S3A). It is possible that this lower stability imposed

evolutionary pressure for *E. coli* proteins to have lower net charge per residue, as observed in the previous section.

By combining Figs. S2E and S2F, a species-specific distribution $P(N,Q) = P(N)P(Q|N)$ can be obtained that captures both the length and charge distribution of each species' proteome (Fig. S2G). Due to the narrower charge distribution of short proteins, the distribution $P_{Hsap}(N,Q)$ has a peak density at the surprisingly short length of ~130 amino acids and Q = +2, as shown in Fig. S2G. Comparison of the distribution $P_{Hsap}(N,Q)$ with the extent of oxidative destabilization $\Delta\Delta G$ shown in Fig. 4A allows us to determine if there are regions of the proteome that are both heavily destabilized by oxidation and contain many protein species. Interestingly, all four proteomes examined lie largely within the region of greatest robustness to charge perturbation, shown by the blue region in Fig. 4A. This region broadens with chain length at a rate similar to that of $P_{Hsap}(N,Q)$. Only the tails of the distribution $P_{Hsap}(N,Q)$ lie outside of the robust region. It is these highly charged tails that are predicted to contain many proteins that are potentially problematic in aging organisms, as shown in Figs. 4A and 4B.

2A. Average folding stability increases linearly with protein chain length and has Gamma-distributed variability among proteins of equal length

The effect of oxidation on the structural integrity of a protein depends on the protein's folding stability prior to oxidation. In order to determine if certain chain lengths are particularly stable or unstable, we use high-quality experimental data to characterize the distribution of stabilities of mesophilic organisms and its dependence on chain length.

Previous theoretical work has suggested that stability should scale linearly with chain length (Ghosh and Dill, 2009; Zeldovich et al., 2007). This dependence can be seen by plotting high-quality experimental stability data collected across multiple mesophilic organisms (Sawle and Ghosh, 2011), as shown in Fig. S3A. However, the scatter about the average is substantial

and must be accounted for to properly characterize the proteome-wide stability distribution. Insight into the reasons for this scatter can be found from a theoretical study that investigated how the variability of protein sequence and selection pressure for stable proteins can lead to broad stability distributions (Zeldovich et al., 2007). In their study, Zeldovich *et al.* found that the number of possible protein sequences with folding stabilities greater than a certain cutoff decreases rapidly with increasing cutoff, while at very low stabilities the number of expected proteins in a proteome should decreases even more rapidly due to evolutionary selection against unstable proteins (Zeldovich et al., 2007). Together, they predict that evolutionary selection will converge to a stability distribution similar in shape to a Gamma distribution. To test this hypothesis, we performed a maximum-likelihood fitting of the data in Fig. S3A to a Gamma distribution with a width that scales linearly with chain length *N* and a constant minimum stability ΔG_{min}. This resulted in an average stability $\overline{\Delta G}(N)$ given by

$$\overline{\Delta G}(N) = (1.06 + 0.0369N)\ \mathrm{kT} \tag{S7}$$

as shown by the sloped black line in Fig. S3A. All of the experimental stability data can now be collapsed onto a single Gamma distribution by subtracting $\Delta G_{min} = 1.06\ \mathrm{kT}$ from the experimental data and dividing the remainder by the length-dependent term in Eq. S7. This results in a normalized distribution of stabilities shown in the inset of Fig. S3A that is indeed consistent with a Gamma distribution. Organismal protein stabilities can therefore be modeled by a distribution $P(\Delta G|N)$ that is a Gamma-distributed for proteins of fixed length *N*. This distribution is helpful in interpreting the effects of oxidative damage, as it warns that even though big proteins are more stable *on average*, there are proteins of all sizes that are barely stable and thus susceptible to oxidation-induced unfolding. For this reason, we judge a protein's susceptibility to oxidative destabilization in this work based solely on $\Delta\Delta G$ rather than on a more complex combination of $\Delta\Delta G$ and $\overline{\Delta G}$.

2B. Estimating the stability distribution of a proteome from its length distribution

For very few proteins are there yet high quality stability measurements. However, from these measurements, it has been found that the average protein stability scales linearly with chain length and displays Gamma-distributed scatter about this average, as shown in Fig. S3A. These observations allow the stability distribution of the human proteome $P_{Hsap}(\Delta G)$ to be found from the probability distribution $P(\Delta G|N)$ (Fig. S3A) and the length distribution $P_{Hsap}(N)$ (Fig. S2F) through the equation

$$P_{Hsap}(\Delta G) = \int_{N=0}^{\infty} P(\Delta G|N)\, P_{Hsap}(N) dN \qquad \text{(S8)}$$

The resulting stability distribution $P_{Hsap}(\Delta G)$, shown by the solid curve and shaded region in Fig. S3B, can be compared to the distribution had the scatter about the line in Fig. S3A been ignored (grey curve in Fig. S3B). The marked difference between the two curves demonstrates the importance of variability in assessing the stability of proteomes. The results further indicate that a startling fraction of human proteins are perched near the edge of stability. With an average stability predicted to be ~10 kT and a peak density at only ~4 kT, the stability of many proteins are similar in magnitude to the destabilizing effect of oxidation shown in Fig. 4A. These proteins are therefore expected to lose their native state upon one or two oxidation events. They are also particularly sensitive to other age-related phenomena such as DNA mutation and protein mistranslation resulting in an amino acid substitution.

3A. Mean-field derivation of the stability change upon side chain oxidation

The effect of oxidative damage on protein stability can be determined by starting with Eq. 1 for the dependence of folding stability on net charge (Ghosh and Dill, 2009). The electrostatic contribution to the stability change from oxidation can be found from Eq. 1 by altering the charge from $Q \rightarrow Q + \Delta Q$ and taking the difference, namely

$$\frac{\Delta\Delta G}{kT} = \frac{\Delta G(Q+\Delta Q)}{kT} - \frac{\Delta G(Q)}{kT} = \frac{(2Q_d\Delta Q+\Delta Q^2)l_b}{2R_d(1+\kappa R_d)} - \frac{(2Q_n\Delta Q+\Delta Q^2)l_b}{2R_n(1+\kappa R_n)} \quad \text{(S9)}$$

where *kT* is the thermal energy scale. It is assumed that $\Delta Q = Q_{ox} - Q$ is the same for both the native and denatured states of a protein (*ox* = oxidized). This assumption is likely to be true unless the oxidized side chain modifies the pK_a value of surrounding amino acids sufficiently to change their protonation states. Given that histidine is the amino acid most likely to have its protonation state modified and makes up only 2.3% of amino acids in the human proteome, our assumption is likely to be true in most cases.

For single amino acid oxidation events, the three possible outcomes are

$$\frac{\Delta\Delta G}{kT} = \frac{(-2Q_d+1)l_b}{2R_d(1+\kappa R_d)} - \frac{(-2Q_n+1)l_b}{2R_n(1+\kappa R_n)} \quad \text{for } \Delta Q = -1 \quad \text{(S10A)}$$

$$\frac{\Delta\Delta G}{kT} = 0 \quad \text{for } \Delta Q = 0 \quad \text{(S10B)}$$

$$\frac{\Delta\Delta G}{kT} = \frac{(2Q_d+1)l_b}{2R_d(1+\kappa R_d)} - \frac{(2Q_n+1)l_b}{2R_n(1+\kappa R_n)} \quad \text{for } \Delta Q = +1 \quad \text{(S10C)}$$

The stability change can be seen to depend linearly on the net charge Q of the protein prior to oxidation. Alterations that cause the net charge to become further from neutrality are predicted to reduce protein stability, as net charge resists compression into the smaller volume occupied by the folded state (Fig. 2), while those that bring the net charge closer to neutrality are predicted to have a stabilizing effect. Equations 2 and S10A-C represent the *average* effect of charge modification on stability, with individual proteins expected to show variability about this average.

3B. Even if the average destabilization from oxidative damage were zero, oxidation would still destabilize the proteome

The model makes the counter-intuitive prediction that oxidative damage can *stabilize* a protein by modifying side chain charge. While an oxidation event that increases the magnitude of the net charge (by either taking away a positive charge from a negative protein or by taking away a

negative charge from a positive protein) is predicted to destabilize a protein on average, *stabilization* results when oxidation brings a protein's net charge closer to zero. The average stability is predicted to remain unchanged upon damage if protein charge is modified from $Q \rightarrow Q \pm 1$ with equal probability. However, stability changes $+\Delta\Delta G$ and $-\Delta\Delta G$ do *not* have equal and opposite consequences on protein folding propensity. This can be seen in reversible folders at equilibrium, where the fraction of copies of a given protein that are expected to be properly folded can be found from the folding stability through the relation

$$f_n = 1/(1 + e^{-\Delta G/kT}) \quad \text{(S11)}$$

where ΔG is the folding stability. If half the copies of a particular protein are stabilized by $+\Delta\Delta G$ while the other half are destabilized by $-\Delta\Delta G$, the saturation of Eq. S11 at high stability (ΔG > ~4 kT, Fig. S3C) and rapid drop at low ΔG causes there to be fewer folded proteins after damage, despite maintaining the same average stability. Said differently, since a healthy proteome is largely folded, the only changes observable from random oxidation will be the destabilizing modifications of the most marginally stable proteins (ΔG < ~4 kT). The great importance of these proteins can be seen by taking the stability distribution of the undamaged and singly oxidized proteomes (where $Q \rightarrow Q \pm 1$ with equal probability), as shown in the inset of Fig. S3D, and replotting it versus the fraction of time each protein is folded. While small in number, the marginally stable proteins disproportionately contribute to the unfolded protein load, which is ultimately what a cell's chaperones and proteasomes sense and must fight against. For that reason, we focus on the destabilizing aspect of oxidation in this work.

3C. Proteins sensitive to oxidative destabilization are more prone to disorder before oxidation

The model presented here predicts that proteins which are highly charged per unit length are the most susceptible to oxidation-induced destabilization. Equation 1 of the model also predicts

that high charge density should make it harder for these proteins to fold even in the absence of damage. This prediction can be tested by determining where these proteins (with net charge greater than $\pm 2\sigma_{Hsap}(N)$) lie on an Uversky plot (Uversky, 2002). This plot, which describes a protein in terms of its net charge per amino acid and its average hydrophobicity per amino acid, has been shown to predict whether monomeric proteins contain a stable folded state or are intrinsically disordered. Note that only arginine, lysine, aspartic acid, and glutamic acid are counted as charged amino acids in Uversky plots. The hydrophobicity of each amino acid is defined using a normalized Kyte-Doolittle scale such that all hydrophobicities vary between zero (least hydrophobic) and one (most hydrophobic).

Of the 14,079 proteins in the human genome whose existence has been verified at the protein level (blue dots), 80% lie in the stable, folded region of the Uversky plot (Fig. 6B). This number decreases dramatically to only 30% for the set of proteins predicted by our model to be the most susceptible to oxidative destabilization (purple). While this finding qualitatively supports the prediction that proteins susceptible to oxidative destabilization are less stable in the absence of oxidation, it does not mean that 70% of our predicted proteins lack a stable folded state. The reason for this is that the Uversky plot (and our model) treats proteins as monomers in the absence of binding partners. Indeed many proteins in the human proteome do not function as monomers but instead operate as part of multi-protein and RNA-protein complexes. The SHFM1 protein is a case in point: its net charge per amino acid of 0.31 and average hydrophobicity per amino acid of 0.35 places it in the intrinsically disordered region of the Uversky plot (Fig. 6B), yet it has been shown to have a stable structure when part of a larger complex.

The take-away message here is that many of the proteins predicted to be greatly destabilized by oxidative charge modification are prone to disorder even before oxidation. The experimental observation that proteins containing many disorder-promoting residues and low

stability are bigger targets for oxidative damaged (Vidovic et al., 2014) serves to add further support that these proteins are more likely to malfunction in old age.

3D. Sensitivity of the predicted set of susceptible proteins to the denatured state radius

The extent of destabilization predicted by Eq. 2 depends not only on the radius of gyration of the native state, which is relatively easy to predict due to the compact nature of folded proteins, but also on the radius of the denatured state. However, there is great variation in the nature of the denatured state and its effective radius of gyration. Due to insufficient experimental data under physiological conditions, we have modeled the denatured state in the main text with radii collected using chemical denaturant experiments (Kohn et al., 2004). This is most certainly an overestimate of the denatured state radius under physiological conditions.

The denatured state for protein sequences with charge and hydrophobicity profiles typical of proteins with a stable, folded native state are more accurately captured by a swollen, molten globule-like ensemble. The radius of this ensembles have been observed to possess a similar length dependence as the native state but with a radius roughly 50% greater (see NU_{PMG} in Fig. 3 of Uversky (Uversky, 2002)). Inserting this relation for R_d into Eq. 2 decreases $\Delta\Delta G$ by roughly 40%, as can be seen by comparing Fig. 4A with Fig. S3E. However, it barely affect the length- and charge-dependence of $\Delta\Delta G$ and hence does not affect the model's primary prediction that short, highly charged proteins are the most susceptible to oxidative destabilization.

3E. Estimating the contribution of local charge-ordering to oxidation-induced destabilization

Our model is based on a mean-field approximation describing the charge-charge interaction energy of side chains in which it is assumed that the net charge is distributed uniformly over the

surface of the protein. While the diverse conformational ensemble associated with the denatured state is likely to have a self-averaging effect similar to our model approximation, such diversity does not exist in the native state, which contains well-defined amino acid positions and hence significant spatial charge-charge correlations. The total interaction energy of the charged amino acids in the native state is therefore more than just the mean-field term. It has been found experimentally that charged amino acids in folded proteins are spatially ordered in such a way that charges of one sign are more likely to be closely surrounded by charges of the opposite sign (Wada and Nakamura, 1981). This spatial correlation stabilizes the native state beyond that which is predicted from the mean-field model. Spatial correlations are absent in our derivation of $\Delta\Delta G$. However, the effect of such spatial correlations, specifically in the form of salt bridges, can be estimated by collecting together data from the literature.

Firstly, experimental mutation studies have been conducted to probe the stabilizing effect of salt bridges. In one such study, a roughly linear decrease in the stabilizing effect of salt bridges was found with increasing solvent exposure, with salt bridges making a negligible contribution *on average* when more than 50% of the salt bridge atoms were exposed to a solvent containing physiological salt concentrations (Takano et al., 2000). As oxidative damage primarily attacks exposed side chains, this evidence suggests that oxidation of charged side chains does not systematically destabilize proteins, consistent with our mean-field model prediction.

Secondly, a rough estimate of the magnitude of the stability contribution from spatial correlations was made by Wada *et al.* (Wada and Nakamura, 1981). Using protein crystal structures, they created a histogram of distances between like and unlike charges. They found structural ordering extending out to several times the Bjerrum length, which characterizes the charge-charge screening distance, although the dominant peak in the correlation was confined to roughly one Bjerrum length, or ~7 Å. They found an average interaction energy of ~66 kT per 150 amino acid protein, or ~4 kT per charged side chain (on average 22% of side chains are

charged, excluding histidine). These unrealistically high values are due to their modeling of charge-charge interactions using a dielectric constant of 4, whereas it is now known that the interaction of surface charges is much better captured using a dielectric constant of 80, close to that for water. A rough estimate of the contribution of local charge ordering to the stability of a protein can therefore be determined by renormalizing the results of Wada *et al.* by a factor of 4/80, which predicts an average stabilizing effect of only 0.2 kT per charged side chain.

However, perhaps the most direct estimate of a systematic local contribution to oxidation-induced destabilization comes from experimental mutation studies (Tokuriki *et al.*) that mutate charged amino acids to ones having similar hydrophobicity to those of the major oxidation products (Petrov and Zagrovic, 2011, 2014). While discussed in greater detail in the following section, the mutation data suggests that oxidation of charged residues is associated with an average destabilization of ~0.5 kT per oxidation event.

In summary, all three sources lead to estimates for the contribution of local charge-ordering that are no greater than ~0.5 kT. Most importantly, however, is that local contributions by their very nature *depend very little* on the overall size and charge of the rest of the protein. Any local contribution therefore acts as a small systematic bias that contributes equally to all proteins independent of their length and charge. Local effects would therefore be expected to have very little effect on the central prediction of our model, namely that short and highly charged proteins are the most susceptible to oxidative destabilization.

4. Predicted distribution of stability changes from oxidation is similar to that from experimental point mutation studies

In addition to being able to capture the change in stability observed in pH-induced unfolding (Ghosh and Dill, 2009) and charge-ladder experiments (Gitlin et al., 2006), a further test of Eq. 2's validity comes from experimental studies on the destabilizing effect of single point mutations.

While it would be ideal to compare our model with actual oxidation data composed of known oxidation sites, oxidation levels, and stability changes, these much-needed experiments have yet to be conducted. However, it is known that a subset of amino acids act as good proxies to common oxidation products due to their similar hydrophobicity (Petrov and Zagrovic, 2011, 2014). Starting from a curated mutation dataset (Tokuriki et al., 2007), we kept only those mutations that: (1) mutate charged amino acids to amino acids that are good proxies to oxidation products and (2) involve solvent-exposed residues accessible to oxidation. The resulting distribution of stability changes can be compared with the distribution predicted by Eq. 2 when applied to the human proteome, as shown in Fig. 3. If the compact denatured state were instead used, the standard deviation of the predicted distribution would be roughly half of that shown in Fig. 3.

For the purpose of predicting the effects of oxidation on stability, not all mutations should be included. The most abundant carbonylated products of side chain oxidation are known to be amino-adipic semialdehyde and glutamic semialdehyde (Petrov and Zagrovic, 2011). Methionine, cysteine, alanine, and threonine have been shown to be good proxies for these oxidation products due to their similar hydrophobicity (Petrov and Zagrovic, 2011). We therefore selected only those experiments that mutated a charged amino acid to one of these four proxies, resulting in the combined distribution of stability changes shown in Fig. 3. Two main observations can be made. First, we find average stability changes of -1.1 kT, -0.3 kT, -0.5 kT, and -0.2 kT for arginine, lysine, aspartic acid, and glutamic acid respectively, with an overall average of -0.5 kT. While our model predicts an equal number of stabilizing and destabilizing mutations due to our assumption that $Q \rightarrow Q + 1$ and $Q \rightarrow Q - 1$ are equally likely, the existence of a small systematic bias of -0.5 kT would not change the conclusions of this work. Second, the width of the predicted distribution is similar to the experimental one, suggesting that the strength of the perturbations predicted by the model is indeed realistic. Note that the comparison of the average and variance in the stability change to experimental data is

only appropriate if the experimental set of proteins has charge characteristics typical of the human proteome as a whole. Fortuitously, the experimental set of proteins has an average net charge per unit length very close to that of the whole human proteome. Furthermore, the set contains a similar number of mutations that bring the net charge closer to neutrality as those that bring it farther away. The experimental dataset therefore serves as a fair comparison for the model's predictions.

REFERENCES

Takano, K., Tsuchimori, K., Yamagata, Y., and Yutani, K. (2000). Contribution of salt bridges near the surface of a protein to the conformational stability. Biochemistry *39*, 12375-12381.

Wada, A., and Nakamura, H. (1981). Nature of the charge-distribution in proteins. Nature *293*, 757-758.

Zeldovich, K.B., Chen, P.Q., and Shakhnovich, E.I. (2007). Protein stability imposes limits on organism complexity and speed of molecular evolution. Proceedings of the National Academy of Sciences of the United States of America *104*, 16152-16157.