Marginal protein stability drives subcellular proteome isoelectric point

Kaiser Loell[a,b] and Vikas Nanda[a,b,1]

[a]Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854; and [b]Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers University, Piscataway, NJ 08854

Edited by David Baker, University of Washington, Seattle, WA, and approved October 3, 2018 (received for review May 26, 2018)

There exists a positive correlation between the pH of subcellular compartments and the median isoelectric point (pI) for the associated proteomes. Proteins in the human lysosome—a highly acidic compartment in the cell—have a median pI of ∼6.5, whereas proteins in the more basic mitochondria have a median pI of ∼8.0. Proposed mechanisms reflect potential adaptations to pH. For example, enzyme active site general acid/base residue p*K*s are likely evolved to match environmental pH. However, such effects would be limited to a few residues on specific proteins, and might not affect the proteome at large. A protein model that considers residue burial upon folding recapitulates the correlation between proteome pI and environmental pH. This correlation can be fully described by a neutral evolution process; no functional selection is included in the model. Proteins in acidic environments incur a lower energetic penalty for burying acidic residues than basic residues, resulting in a net accumulation of acidic residues in the protein core. The inverse is true under alkaline conditions. The pI distributions of subcellular proteomes are likely not a direct result of functional adaptations to pH, but a molecular spandrel stemming from marginal stability.

molecular spandrels | neutral evolution | protein folding | acid–base equilibria

Of amino acid side-chain functional groups in proteins, several can occur in multiple protonation states, depending on their acid dissociation constant (p*K*) and the acidity/basicity (pH) of the surrounding environment. This leads to a biochemical property at the protein level known as the isoelectric point (pI), the pH at which the protein exhibits no net charge. The pI may be approximated based on a protein sequence (1), or more accurately estimated given knowledge of its 3D structure (2). It can be measured experimentally using methods that assess electrophoretic mobility in a pH gradient (3). Estimates of this biochemical parameter are central to protocols for protein isolation and purification or defining solution conditions to promote solubility or crystallization.

A number of proteome-scale studies of eukaryotes have identified interesting features of pI distributions as a function of cellular environment or subcellular location (4–13). In particular, the aggregate pIs of proteins making up subcellular proteomes tend to positively correlate with the local pH of that compartment (4–6). Acidic compartments such as the lysosome have proteins with higher frequencies of aspartate and glutamate relative to the entire proteome, whereas proteins in basic compartments such as the mitochondrion have higher frequencies of basic amino acids—lysine or arginine. This implies functional and evolutionary links exist between environmental pHs and protein pIs.

The nature of such evolutionary connections is not clear. Many proteins become insoluble as the pH approaches the pI. One might expect the opposite of what is observed: pIs would be distributed far from a subcellular compartment's pH to prevent protein aggregation. On the other hand, the folded states of proteins are often most stable at pH values near their pI, a factor that also correlates with their optimal pH of function (14, 15). This suggests a hypothesis where selection for functional optimality, and therefore optimal stability, results in protein pIs matching subcellular pH. Such selection could apply broadly across many proteins, resulting in proteome-wide effects (12).

However, rather than exhibiting high stability under physiological conditions, the majority of proteins are marginally stable, with free energy differences of only 5 kcal/mol to 15 kcal/mol between the folded and unfolded states (16). Neutral evolution theory posits most diversity can be explained by the accumulation of random mutations that have minimal impact on fitness (17). Models of protein evolution demonstrate that proteome-wide marginal stability can be understood as neutral, rather than positive selection for instability (18, 19). Given the assumptions that mutation is a stochastic process, and most mutations either have little impact or are deleterious to the free energy of folding (17, 20–22), stabilities of proteins will be distributed near a critical free energy threshold, below which they cease to fold and therefore function (Fig. 1). This promotes a state of marginal stability. The question then arises as to how neutral selection would constrain the distribution of pIs of marginally stable proteins.

We model an evolutionary process resulting in marginally stable protein population that demonstrates pI distributions contingent on environmental pH. A simple spherical protein model links pH effects on protonation state of titratable amino acids to the energetic cost of burying ions in the low dielectric interior of a protein. The model exhibits marginal stability under a simulated neutral evolution trajectory and links the pH and pI through electrostatic contributions to the energetics of folding.

Results and Discussion

Sequence Analysis. Examination of curated protein sequence sets was consistent with observed correlations between pI and subcellular location (4–6). Subcellular compartment proteomes were established for human, mouse, and rat sequences using

Significance

The rules of protein design are often derived from analyzing the properties of natural proteins. However, unlike designed proteins, natural proteins are the product of long evolutionary histories, confounding efforts to identify which properties are directly related to function versus those that are a by-product of the evolutionary process. We critically examine the observation that the isoelectric point—the pH at which proteins have no net charge—correlates with the pH of the subcellular environment occupied by that protein. This correlation is not due to adaptive selection, but is a by-product of neutral evolution. This insight can be used to understand the evolution of proteins at a range of pH and design molecules for such environments.

Author contributions: K.L. and V.N. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: nanda@cabm.rutgers.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1809098115/-/DCSupplemental.

Published online November 1, 2018.

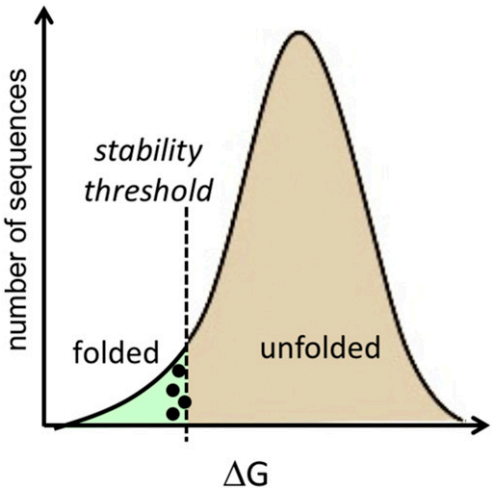

Fig. 1. Normal distribution representing the free energy of folding for protein sequence space. The majority of sequences are above a critical threshold for stability (i.e., $\Delta G \leq 0$) and hence do not result in functional molecules. It is proposed that existing proteins (black dots) are marginally stable due to neutral evolution processes (18).

gene ontology (GO) information from the ENSEMBL sequence database (23). Only proteins corroborated by manual annotations from the SWISS-PROT database (24) were included. Median pI values grouped by subcellular location were consistent across the three species (Table 1). The largest difference in median pI was the mitochondrion with its basic outer lumen (Fig. 2). Other compartments such as the nucleus and cytosol showed only modest shifts in pI relative to the lysosome. We compared database-derived ontology specifications of subcellular localization to those derived from sensitive proteomics analyses. Subcellular location of proteins can be determined by fractionation methods (25) that separate organelles based on density with high-resolution mass spectrometry approaches (26, 27). The pI distributions and medians were consistent with those from lysosome and mitochondria proteomes determined from the sequence database analysis (Fig. 2 *B* and *C*). Relative amplitudes of the acidic and basic components of the bimodal pI distribution correlated with organelle pH, demonstrating a bias for acidic over basic residues in the lysosomal proteome and the inverse for the mitochondrial proteome.

Model Construction. An idealized protein model was constructed to capture the energetic linkage between amino acid composition, stability, and pH. Amino acids were represented as evenly spaced points on a cubic lattice within a cutoff radius r from the protein center of mass, resulting in an approximate spherical shape (Fig. 3*A*). These were classified based on distance d from the center as exposed ($r - 1 < d \leq r$) or buried ($d \leq r - 1$), resulting in a single layer of exposed amino acids surrounding a buried core. A cutoff of $r = 4$ lattice units produced a model with 257 amino acids. The number of exposed and buried amino acids were approximately equal: 134 exposed, 123 buried. This was a reasonable reflection of natural proteins of similar size. For example, the structure of green fluorescent protein (GFP) consists of 229 residues excluding the chromophore, of which 119 are buried—defined as fraction accessible surface area <0.3 for GFP (28) calculated by the DSSP software package (29).

The model simulates pH effects on stability as specified by protonation states of titratable residues: aspartic acid (D), glutamic acid (E), histidine (H), cysteine (C), tyrosine (Y), arginine (R), and lysine (K). There is a significant energetic cost associated with desolvating a buried ion during folding. The total free energy of folding was determined by the transfer energy of burying residues in the core and the ionization energies of titratable groups in the folded and unfolded states (30) (Fig. 3*B* and Eq. **1**),

$$\Delta G_{folding}(\mathrm{pH}) = \Delta G_{transfer} + \Delta G^{F}_{ionization}(\mathrm{pH}) - \Delta G^{U}_{ionization}(\mathrm{pH}) \quad [1]$$

$$\Delta G_{transfer} = \sum \left[\Delta G_{conf} + f_{burial}(i)\Delta G_{burial}(i)\right]; \quad f_{burial} = 0 \text{ if buried, 1 if exposed} \quad [2]$$

$$\Delta G^{F}_{ionization}(\mathrm{pH}) - \Delta G^{U}_{ionization}(\mathrm{pH}) = \sum f_{burial}(i) \cdot RT \ln \frac{1 + 10^{z(pH - pK^{U}(i))}}{1 + 10^{z(pH - pK^{F}(i))}}. \quad [3]$$

The pH-independent component, $\Delta G_{transfer}$, comprises a conformational entropy term (ΔG_{conf}) associated with restricting conformations of each amino acid in the folded state: 3.4 cal/K·mol per residue, 310 K (31). The other component is the energetic cost of burial (Eq. **2**: ΔG_{burial}) based on experimentally measured partition coefficients between an aqueous and organic phase (32). Here, z is the formal charge of an ionized titratable amino acid. In neglecting chain connectivity, we are not sampling an ensemble of conformations and assessing free energy of folding relative to a native state. Instead, by focusing solely on the energetics of burial, we treat a folded protein as the phase separation of buried groups between aqueous and hydrophobic environments.

The pH-dependent free energy of ionization is estimated assuming ideal values for pK in the unfolded state or exposed folded state (R = 12.5, K = 10.5, D = 3.9, E = 4.1, and H = 6.0), and a shift in pK of 3.0 units for buried groups (Eq. **3**). This is within the range of observed pK shifts for amino acids in protein cores and would result in a folding penalty of 4.2 kcal/mol per buried residue at a pH when it is fully ionized (33, 34).

Evolutionary Trajectories. Single-sequence trajectories were calculated using constant-temperature Monte Carlo sampling with a Metropolis criteria (35, 36) for accepting deleterious mutations. All amino acid transitions were assigned equal probability, allowing for rapid sampling of sequence space. Each trajectory was run for 40,000 iterations, ample for convergence as assessed by leveling of ΔG of folding. Distributions were generated from 10,000 trajectories for each simulation pH value.

The results of typical simulations are shown in Fig. 4, where randomized starting amino acid compositions resulted in highly unfavorable folding energies, due to poor energies of burial. Within ~500 iterations, computed free energies were below the selection threshold of 0.0 kcals/mol. The trajectories satisfied the property of marginal stability, maintaining a median $\Delta G_{folding}$ around −2 kcal/mol (Fig. 4*A*) and ranging as low as −10 kcal/mol (*SI Appendix*, Fig. S1).

Table 1. Median pIs calculated from annotated subcellular proteomes

Subcellular compartment	pH	Median pI		
		Mouse	Rat	Human
Cytosol	7.2	6.2	6.2	6.2
Lysosome	4.7	6.5	6.7	6.6
Golgi	6.4	6.6	6.5	6.5
Cytoplasm	7.2	6.7	6.8	6.6
Nucleus	7.2	6.9	6.8	6.9
ER	7.2	7.0	7.2	7.0
Mitochondrion	8.0	8.3	8.3	8.0

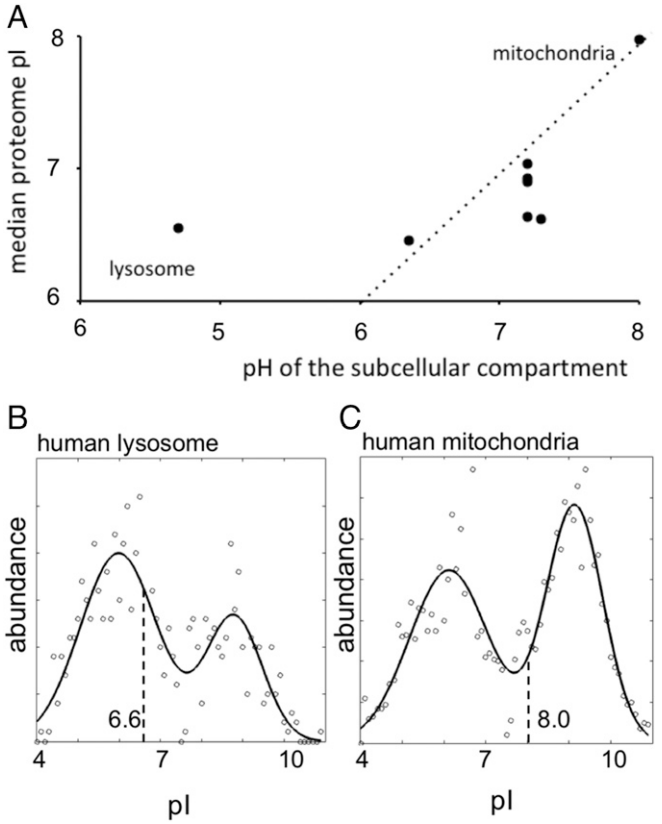

Fig. 2. Correlation between pH and pI. (*A*) Median subcellular proteome pIs for human lysosome, mitochondria, cytoplasm, nucleus, Golgi, and ER compared with reported pH (5) (see Table 1 for values). Dotted line where pH = pI is provided for reference. Compartment-associated pI show bimodal distribution in (*B*) lysosome and (*C*) mitochondria, with shifts in the relative magnitudes of low-pH and high-pH components determining the median pI.

Stability was maintained by the inclusion of hydrophobic residues in the core, with the mean hydrophobicity steadily increasing in the first phase of optimization (Fig. 4*B*). Simulated hydrophobicities approach that of globular proteins such as lysozyme, suggesting reasonable steady-state amino acid compositions for buried residues in the simulation. Dissection of one simulation end-point run at pH 5.5 shows a model with varied composition on the surface, but a significantly hydrophobic core (Fig. 4*C*).

In addition to hydrophobic groups, approximately 1 in 10 core residues were titratable. If ionized, they would significantly destabilize the folded state, and would need to be compensated for by the favorable burial of hydrophobic groups. In this example (Fig. 4*C*), most buried titratable residues are D or E, which have p*K*s close to the environmental pH.

pI vs. pH. Simulations were run at fixed pH from 4.0 to 11.0. The median pI of the ensemble of trajectories positively correlated with pH (Fig. 5*A*). Calculated pIs exhibited a bimodal distribution, due to the uneven distribution of titratable residue p*K*s across the range of pH sampled (Fig. 5*B*) (37, 38). Changes in the median pI were due to shifts in the population from acid to basic sequences. In the models themselves, the shifts in pI were due to an increased number of buried titratable groups with p*K*s near the environmental pH (Fig. 5*C*). Proteins evolving under acidic pH had a large number of acids in the core. Under basic conditions, more bases were observed. At neutral pH, the frequency of histidine was higher. No pH-dependent biases in surface residue compositions were observed.

In constructing a parsimonious model, we incorporated a minimal set of energetic terms: $\Delta G_{ionization}$, ΔG_{burial}, and ΔG_{conf}. We examined sensitivity of model behavior to key parameters. $\Delta G_{ionization}$ was modulated by varying ΔpK from 0.5 to 4.0 pH units. At small shifts, pI was largely insensitive to pH, whereas, for the large shifts, the correlation was pronounced (*SI Appendix*, Fig. S2*A*). Based on lysosomal pH = 5, mitochondrial pH = 8, and $\Delta pK = 3$, model calculations underestimated median pI of these subcellular environments by approximately 1 pH unit (*SI Appendix*, Fig. S2 *B* and *C*). The pK shifts of 0.5 or 1.0 pH units better approximate the acidic lysosome, whereas larger shifts are more consistent with the basic mitochondria. Applying a uniform pK shift to all residues in the model neglects the heterogeneity of local buried protein environments (30, 33, 34, 39) and the bias of protein interiors toward a positive potential which would preferentially stabilize buried acids over bases (40).

The model fails to recapitulate the pH–pI correlation if the minimal set of energetic terms are either removed or assigned nonrealistic values. Removing or doubling the magnitude of ΔG_{burial} in Eq. **2** both resulted in a complete loss of correlation except at the most acidic and basic simulation pH values (*SI Appendix*, Fig. S3 *A–D*). Similarly, eliminating or doubling ΔG_{conf} resulted in a median pI of 7.2 to 7.3 at all simulation pH values. Using a combined backbone and side-chain estimate of ΔG_{conf} = 4.7 cal/mol·K per residue based on NMR-constrained native and denatured ensembles (41) produced a similar correlation to the backbone-only ΔG_{conf} used in this study (*SI Appendix*, Fig. S3*E*).

Varying the threshold of $\Delta G_{folding}$ for selection from 0 kcals/mol to −5 kcals/mol shifted the distribution of marginal stabilities, but did not perturb the distribution of pIs (*SI Appendix*, Fig. S1). Imposing large positive or negative thresholds selected for unstable or highly stable sequences, respectively. In both cases, a diminished pH–pI correlation was observed with increasing the threshold (Fig. 6). The relationship between pI and pH was most pronounced near the $\Delta G = 0$ kcals/mol threshold, indicating this effect is associated with marginal stability. However, natural proteins do not exhibit stabilities of <−50 kcals/mol. Such high stabilities are possible in this model because we do not consider free energies relative to an ensemble of chain conformations. Highly stable proteins in our model

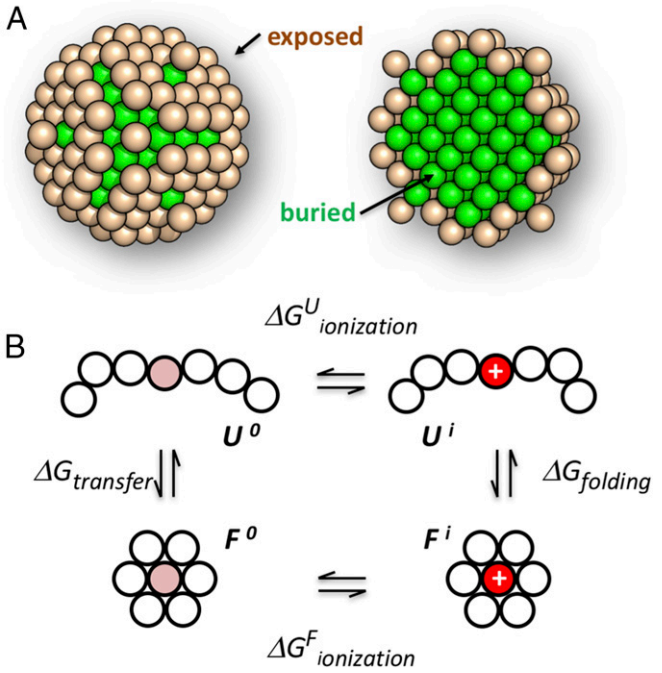

Fig. 3. (*A*) Protein model consists of residues on a rectangular lattice partitioned into exposed (brown) and buried (green) positions. (*B*) Energetic model for protein evolution based on the thermodynamic cycle of coupled protein folding and ionization of titratable residues.

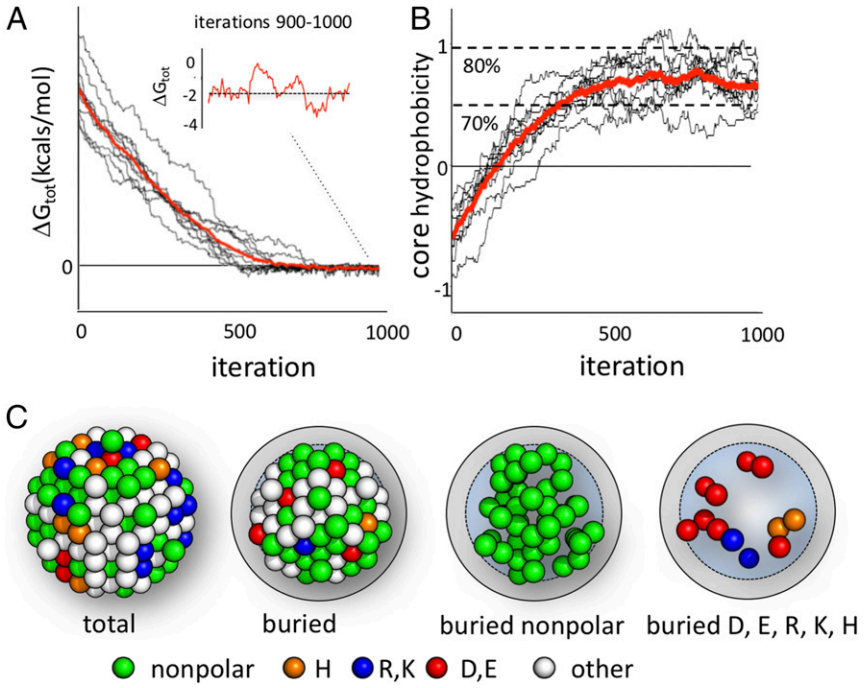

Fig. 4. (*A*) Free energy of folding (Eq. **1**) for 10 sequence simulation trajectories run at a fixed pH of 5.5. Average of the trajectories is in red. First 1,000 of 40,000 iterations are shown. (*B*) Mean hydrophobicity of protein cores calculated using the Kyte and Doolittle (56) scale for the 10 trajectories, and compared with core hydrophobicity of hen egg white lysozyme (57) at 70% or 80% buried as determined from solvent accessibility estimations by DSSP (29). (*C*) Example of a trajectory endpoint showing positions of titratable, nonpolar (F, L, V, I, M), and other amino acid types on the surface and core.

have extensively hydrophobic cores that would be unlikely to form unique native states (42). Despite this limitation, natural proteins are marginally stable, due to biophysical constraints imposed by topology, folding, and evolution (43), and are likely subject to neutral evolution effects modeled here.

Smaller and larger models were also tested. For larger radii of 5 or 6 residues, the model behaved similarly, given that the ratio of surface to buried residues was largely unchanged (*SI Appendix*, Fig. S4). However, at $r = 3$ residues, the pI was largely insensitive to simulation pH, due to the three-to-one ratio of surface to core positions, with surface titratable residues dominating the observed pIs.

Comparison with Natural Proteins. A key prediction of the model is that the cores of proteins in acidic environments will preferentially contain acidic residues D + E over other titratable residues. Protein cores in basic environments will preferentially contain K + R. Residues on the surface do not have large changes in accessibility upon folding, and thus would not be expected to show similar biases based on environmental pH. An analysis of 40 lysosomal and 200 mitochondrial protein structures in the Protein Data Bank (PDB) support this prediction (Fig. 7 and *SI Appendix*, Table S1). Buried/core K and R residues were overrepresented in mitochondrial proteins over lysosomal proteins, while E + D were underrepresented.

In exposed/surface positions, the log odds ratios were reversed, with acid residues slightly overrepresented on mitochondrial protein surfaces, and basic residues underrepresented. The underlying mechanism driving this bias is uncertain. In the basic outer lumen of mitochondria, exposed D + E residues would likely be charged, perhaps providing charge stability to proteins and compensating the aggregation tendencies facilitated by the convergence of pI and pH (9). Notably, most of the adaptive mechanisms hypothesized involve interaction of surface titratable groups with the subcellular environment, counter to the observation that most of the pI bias comes from buried residues (15).

Conclusions

The correlation between the pI distribution of a subcellular proteome and local pH can be understood from a simple protein model of neutral sequence evolution. This correlation does not arise out of an intrinsic fitness advantage, but rather as a by-product of marginal stability. Marginal stability has advantages for proteostasis, where degradation of proteins can be highly regulated (44), without requiring forced unfolding of overly stable proteins. It is appreciated that beneficial features of marginal stability are not a direct result of positive selection, but rather are due the large available space of unstable sequences (18, 45). The term for biological features mistakenly attributed to direct selection is "spandrels," coined by Gould and Lewontin

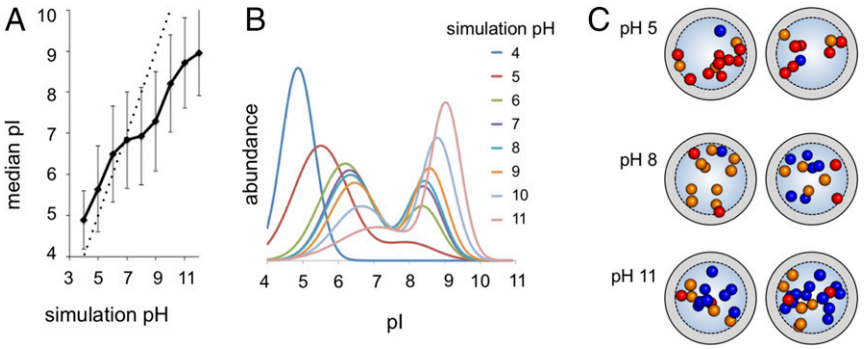

Fig. 5. (*A*) Median pI as a function of simulation pH; pI = pH reference shown as dotted line. (*B*) Shifts in median pI are attributed largely to changes in the magnitude of acidic and basic peaks in a biomodal pI distribution. (*C*) Examples of final distributions of core titratable groups; two independent examples shown. E + D, red spheres; H, orange spheres; K + R, blue spheres.

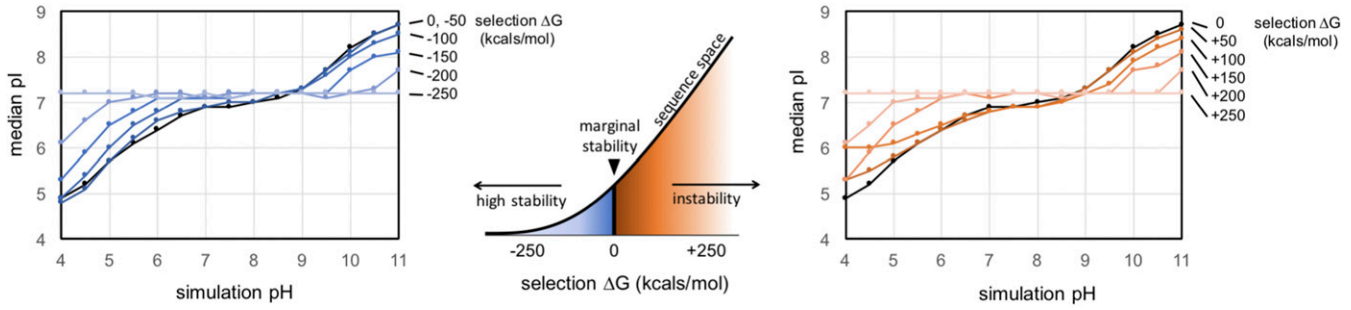

Fig. 6. Median pI sampled as a function of simulation pH at a series of favorable and unfavorable selection thresholds for $\Delta G_{folding}$. For selection DG values tending toward high or low stability, the median pI approaches that of a random sequence (7.2 to 7.3). Under strong selection for stability, the core is highly hydrophobic, and surface amino acids dominate the pI. Under weak selection for stability, protein sequences are random in both the core and surface. Optimal pH–pI correlations are observed under marginal stability conditions.

(46). Likewise, while the pH adaptation of pI may be advantageous to subcellular-specific proteomes (12, 15), the models presented here provide a simple explanation for this convergence: Fixation of titratable residues in the cores of proteins are likely constrained by marginal stability.

Important features of proteins and their evolution were not fully captured by our model. Neutral theory proposes the majority of fixed mutations are selectively neutral (17). However, it has been noted that buried charges are more likely to be conserved than those on protein surfaces, due to surrounding networks of stabilizing backbone and side-chain interactions (34, 40). The preferential stabilization of buried acids due to interactions with backbone amides and a net positive potential can explain the observed preference for D and E over R and K in protein cores, with environmental pH enhancing this preference in acidic subcellular environments. Our model does not capture such compensatory effects. An explicit structural model including pairwise interactions, likely in an atomistic context, would be necessary. Such an approach was taken by Chan and Warwicker (4) using natural protein structures. They noted that the spatial distribution of charges and their contributions to stability correlate more directly with pH than pI, where stability is attributed primarily to the location and titration state of histidine. A more realistic neutral evolution model would consider population dynamics (47, 48) and environmental robustness (49), both expected to produce stability distributions with a greater margin relative to the threshold, facilitating protein evolvability.

Biasing core positions in protein design libraries to match p*K* and environmental pH could improve screen hit rates. Inverting this bias could be used to screen libraries for pH switches, as has been achieved with de novo design (50, 51). The pH switches are found in naturally occurring lysosomal proteins (52) and serve as instructive exemplars for design. Identifying which features of proteins and proteomes are the result of direct selection or are spandrels is important for understanding molecular evolution (53, 54), and for identifying properties to optimize in engineering and design of synthetic proteins.

Materials and Methods

Sequence Database Analysis. Sequences were obtained for proteins associated with subcellular compartments from *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus* datasets with the BIOMART tool and Gene Ontology (GO) annotations for the subcellular location and the ENSEMBL sequence database: lysosome, GO:0005764; Golgi, GO:0005794; cytoplasm, GO:0005737; nucleus, GO:0005634; ER, GO:0005783; and mitochondrion, GO:0005829 (23). Sequence sets were corroborated with experimentally reviewed sequence data from SWISS-PROT (24), obtained by searching for sequences with location or GO matching the target organelles and curated proteins from lysosome (26) and mitochondrion (27) proteomics data sets.

Calculating pI and Distributions. The pIs of natural and model proteins were estimated from primary sequence using a binary search algorithm, starting with a range of potential pIs from 4.0 to 12.0. Total charge was calculated at the midpoint of the range, summing the products for each amino acid of the charge in the ionized state, the number of residues, and the probability of ionization θ computed using the Henderson–Hasselbach equation,

$$pH = pK + \log[\theta/(1\text{-}\theta)]. \quad [4]$$

Convergence was defined when the range was smaller than 0.01 pH units. The code is available in sphereSim in the subroutine calculateIP included as Dataset S1. Bimodal fits to pI distributions were determined by nonlinear regression fitting of a sum of two Gaussians.

Molecular Simulations. Model proteins consisted of amino acids arranged in a spherical 3D cubic lattice, without sequence connectivity. The free energy of burial was calculated by summing the free energy of transfer with the product of the residue's probability of being ionized and the energetic penalty for burying a charge, computed from the change in p*K* with burial. Experimentally determined partition coefficients for water to *N*-methyl acetamide (32) were used for the free energies of burial. To compute the total free energy of folding, transfer energies summed over all residues and added to a constant factor approximately equivalent to the reduction in backbone entropy upon folding (31). Mutations were generated at random, with all amino acids except proline having an equal probability of occurrence. Mutations which improved the calculated energy of folding were accepted regardless of $\Delta G_{folding}$ (Eq. **1**). If $\Delta G_{folding}$ was larger than the selection threshold, the model accepted mutations with probability *a* determined by the Metropolis criterion (35, 36), with a system temperature $\beta = 1$,

$$a = \min[1, \exp(-\beta \cdot (E_i - E_{i-1}))], \quad [5]$$

where $E_i - E_{i-1}$ is the change in free energy of folding between the parent ($i - 1$) and current (i) sequence. Simulation output after 40,000 cycles included sequence, exposed and buried portions of the sequence, and the calculated pI. Ten thousand sequences were generated for each pH. Code for model construction and simulation is included in Dataset S1.

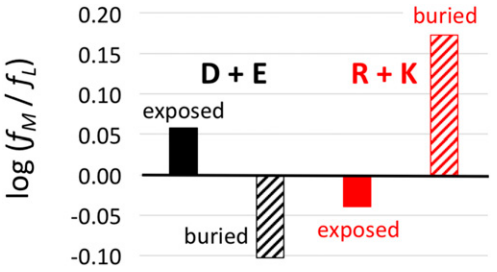

Fig. 7. Log odds ratios of titratable amino acid frequencies in mitochondria f_M versus lysosome f_L. Acidic residues are relatively underrepresented in mitochondrial protein cores, in contrast to basic residues which occur more frequently in mitochondrial cores. Counts and f_M and f_L values for individual environments are reported in *SI Appendix*, Table S1.

Protein Structure Analysis. Curated subsets of the PDB for the lysosome and mitochondria were taken from Chan et al. (4, 5). Amino acid solvent-accessible surface area (SASA) calculations were carried out using DSSP (29), and normalized to maximal SASA values (55) to identify buried versus exposed positions; a cutoff of 30% of maximal surface area was used to define buried positions.

ACKNOWLEDGMENTS. We thank Peter Lobel for useful discussions and access to lysosomal proteome datasets. This work was supported by National Institutes of Health Grant DP2 OD-006478 (to V.N.), National Aeronautics and Space Administration Grant 80NSSC18M0093 (to V.N.), the Rutgers University Bernstein Fellowship, Rutgers University Aresty Undergraduate Research Fellowship, and the CABM SURE program (K.L.).

1. Sillero A, Ribeiro JM (1989) Isoelectric points of proteins: Theoretical determination. *Anal Biochem* 179:319–325.
2. Russell ST, Warshel A (1985) Calculations of electrostatic energies in proteins. The energetics of ionized groups in bovine pancreatic trypsin inhibitor. *J Mol Biol* 185: 389–404.
3. Svensson H (1948) Preparative electrophoresis and ionophoresis. *Adv Protein Chem* 4: 251–295.
4. Chan P, Warwicker J (2009) Evidence for the adaptation of protein pH-dependence to subcellular pH. *BMC Biol* 7:69.
5. Chan P, Lovrić J, Warwicker J (2006) Subcellular pH and predicted pH-dependent features of proteins. *Proteomics* 6:3494–3501.
6. Schwartz R, Ting CS, King J (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res* 11:703–709.
7. Sun Q, Emanuelsson O, van Wijk KJ (2004) Analysis of curated and predicted plastid subproteomes of *Arabidopsis*. Subcellular compartmentalization leads to distinctive proteome properties. *Plant Physiol* 135:723–734.
8. Kiraga J, et al. (2007) The relationships between the isoelectric point and: Length of proteins, taxonomy and ecology of organisms. *BMC Genomics* 8:163.
9. Andrade MA, O'Donoghue SI, Rost B (1998) Adaptation of protein surfaces to subcellular location. *J Mol Biol* 276:517–525.
10. Brett CL, Donowitz M, Rao R (2006) Does the proteome encode organellar pH? *FEBS Lett* 580:717–719.
11. Khaldi N, Shields DC (2011) Shift in the isoelectric-point of milk proteins as a consequence of adaptive divergence between the milks of mammalian species. *Biol Direct* 6:40.
12. Garcia-Moreno B (2009) Adaptations of proteins to cellular and subcellular pH. *J Biol* 8:98.
13. Knight CG, Kassen R, Hebestreit H, Rainey PB (2004) Global analysis of predicted proteomes: Functional adaptation of physical properties. *Proc Natl Acad Sci USA* 101: 8390–8395.
14. Talley K, Alexov E (2010) On the pH-optimum of activity and stability of proteins. *Proteins* 78:2699–2706.
15. Alexov E (2004) Numerical calculations of the pH of maximal protein stability. The effect of the sequence composition and three-dimensional structure. *Eur J Biochem* 271:173–185.
16. Fersht AR (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman, New York).
17. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, London).
18. Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46: 105–109.
19. Shakhnovich EI, Gutin AM (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* 346:773–775.
20. Godoy-Ruiz R, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM (2004) Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. *J Mol Biol* 336:313–318.
21. Pakula AA, Young VB, Sauer RT (1986) Bacteriophage lambda cro mutations: Effects on activity and intracellular degradation. *Proc Natl Acad Sci USA* 83:8829–8833.
22. Matthews BW (1993) Structural and genetic analysis of protein stability. *Annu Rev Biochem* 62:139–160.
23. Yates A, et al. (2016) Ensembl 2016. *Nucleic Acids Res* 44:D710–D716.
24. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A (2007) UniProtKB/Swiss-Prot. *Methods Mol Biol* 406:89–112.
25. De Duve C (1971) Tissue fractionation. Past and present. *J Cell Biol* 50:20d–55d.
26. Jadot M, et al. (2017) Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Mol Cell Proteomics* 16:194–212.
27. Foster LJ, et al. (2006) A mammalian organelle map by protein correlation profiling. *Cell* 125:187–199.
28. Yang F, Moss LG, Phillips GN, Jr (1996) The molecular structure of green fluorescent protein. *Nat Biotechnol* 14:1246–1251.
29. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
30. Karp DA, et al. (2007) High apparent dielectric constant inside a protein reflects structural reorganization coupled to the ionization of an internal Asp. *Biophys J* 92: 2041–2053.
31. D'Aquino JA, et al. (1996) The magnitude of the backbone conformational entropy change in protein folding. *Proteins* 25:143–156.
32. Damodaran S, Song KB (1986) The role of solvent polarity in the free energy of transfer of amino acid side chains from water to organic solvents. *J Biol Chem* 261: 7220–7222.
33. Isom DG, Cannon BR, Castañeda CA, Robinson A, García-Moreno B (2008) High tolerance for ionizable residues in the hydrophobic interior of proteins. *Proc Natl Acad Sci USA* 105:17784–17788.
34. Kim J, Mao J, Gunner MR (2005) Are acidic and basic groups in buried proteins predicted to be ionized? *J Mol Biol* 348:1283–1298.
35. Li Z, Scheraga HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA* 84:6611–6615.
36. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092.
37. Weiller GF, Caraux G, Sylvester N (2004) The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution. *Proteomics* 4: 943–949.
38. Wu S, et al. (2006) Multi-modality of pI distribution in whole proteome. *Proteomics* 6: 449–455.
39. Isom DG, Castañeda CA, Cannon BR, García-Moreno B (2011) Large shifts in pKa values of lysine residues buried inside a protein. *Proc Natl Acad Sci USA* 108: 5260–5265.
40. Gunner MR, Saleh MA, Cross E, ud-Doula A, Wise M (2000) Backbone dipoles generate positive potentials in all proteins: Origins and implications of the effect. *Biophys J* 78: 1126–1144.
41. Baxa MC, Haddadian EJ, Jumper JM, Freed KF, Sosnick TR (2014) Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc Natl Acad Sci USA* 111:15396–15401.
42. Chan HS, Dill KA (1991) Sequence space soup of proteins and copolymers. *J Chem Phys* 95:3775–3787.
43. Sikosek T, Chan HS (2014) Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface* 11:20140419.
44. Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475:324–332.
45. Bloom JD, Raval A, Wilke CO (2007) Thermodynamics of neutral protein evolution. *Genetics* 175:255–266.
46. Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* 205: 581–598.
47. Taverna DM, Goldstein RA (2000) The distribution of structures in evolving protein populations. *Biopolymers* 53:1–8.
48. Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104: 16152–16157.
49. Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19:596–604.
50. Zhang Y, et al. (2015) Computational design and experimental characterization of peptides intended for pH-dependent membrane insertion and pore formation. *ACS Chem Biol* 10:1082–1093.
51. Lizatović R, et al. (2016) A de novo designed coiled-coil peptide with a reversible pH-induced oligomerization switch. *Structure* 24:946–955.
52. Guhaniyogi J, Sohar I, Das K, Stock AM, Lobel P (2009) Crystal structure and autoactivation pathway of the precursor form of human tripeptidyl-peptidase 1, the enzyme deficient in late infantile ceroid lipofuscinosis. *J Biol Chem* 284: 3985–3997.
53. Barrett RD, Hoekstra HE (2011) Molecular spandrels: Tests of adaptation at the genetic level. *Nat Rev Genet* 12:767–780.
54. DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat Rev Genet* 6:678–687.
55. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2007) Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics* 23:2231–2238.
56. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105–132.
57. Diamond R (1974) Real-space refinement of the structure of hen egg-white lysozyme. *J Mol Biol* 82:371–391.