Protein structure and interaction under environmental stress: from quality control recognition to evolution of collective behavior

By

Kelly Paige Brock

S.B. Engineering Sciences, Harvard University 2011

Submitted to the Computational and Systems Biology Program in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the Massachusetts Institute of Technology

June 2016

© 2016 Massachusetts Institute of Technology. All rights reserved.

Signature of Author................. Signature redacted

Program in Computational and Systems Biology

May 20, 2016

Certified by .. Signature redacted ....

Jeremy England

Thomas D. and Virginia W. Cabot Career Development Professor of Physics

Thesis Supervisor

Accepted by .... Signature redacted ....

Chris Burge

Professor of Biology and Biological Engineering

Director, Computational and Systems Biology Ph.D. Program

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUN 14 2016
LIBRARIES
ARCHIVES

Protein structure and interaction under environmental stress: from quality control recognition to evolution of collective behavior

By

Kelly Paige Brock

Submitted in partial fulfillment of the requirements

of the degree of Doctor of Philosophy in Computational and Systems Biology

at the Massachusetts Institute of Technology

ABSTRACT

A protein's function in the cell depends on its structure, which in turn depends on the intracellular environment. Stress like heat shock or nutrient starvation can alter intracellular conditions, leading to protein misfolding – i.e. the inability of a protein to reach or maintain its native conformation. Since many proteins interact with each other, protein misfolding and cellular stress response must be examined both on the scale of individual protein conformational changes and on a more global level, where interaction patterns can reveal larger-scale protein responses to cellular stress.

On the individual scale, one example of a protein particularly susceptible to misfolding is the human von Hippel-Lindau (VHL) tumor suppressor. When expressed in the absence of its cofactors, VHL cannot fold correctly and is quickly degraded by the cell's quality control machinery. Here, I present a biophysical characterization of a VHL mutation that confers increased resistance to misfolding. Mathematical modeling provides an explanation for this mutant's increased stability in the cell by predicting how its cofactor and chaperone interaction sites are buried or exposed in the protein's predicted conformation.

On a more global level, a budding yeast cell undergoing glucose deprivation both acidifies its cytosol and exhibits widespread protein clustering. By employing a proteome-wide computational assay, I examine how this drop in pH could lead to the formation of higher order protein structures. This modeling framework also provides a rationale for why these two related phenotypes might be beneficial, since protein clustering can help regulate relevant metabolic pathways and provide protection from protein misfolding and/or degradation.

Thesis Supervisor: Jeremy England

Title: Thomas D. and Virginia W. Cabot Career Development Professor of Physics

Acknowledgements

I would first and foremost like to thank my amazing academic advisor, Jeremy England. Jeremy has been an incredible advisor, and is deeply invested in both the personal and professional well-being of his students. In terms of academic and research advising, he has helped me learn to formulate full research questions, propose experiments and controls to test them, and successfully convey the results. In short, I couldn't ask for more in a mentor, and I am very proud to be one of his students.

I would also like to thank my wonderful committee members, labmates and collaborators. My committee – Mark Bathe, Manolis Kellis, and for my defense, Debbie Marks – has given invaluable suggestions and has helped me to become a better scientist. Kolya Perunov was Jeremy's first student and the first of us to graduate – here's to lots of tea and learning random Russian words! I am so glad that I have gotten the chance to know and befriend Rotem Gura Sadovsky – here's to many, many far-reaching philosophical discussions, and thank you so much for your support both academic and personal. Thank you to my officemate Tal Kachman, for keeping your fingers and toes crossed for me and for lots of jokes and hugs. I would also like to thank Bobby Marsland, Pavel Chvykov, Sumantra Sarkar, and Jordan Horowitz – thank you so much for all your support over the years, and for all of our weirdly morbid breakfast discussions! I would also like to thank our collaborators at the Hebrew University of Jerusalem, Prof. Daniel Kaganovich and his students Triana Amen and Ayelet-Chen Abraham. This thesis wouldn't be possible without you.

I would also like to thank my classmates in the Computational and Systems Biology program – Sarah Spencer, Boyang Zhao, Thomas Willems, Max Wolf, and Asa Adadey. We have each other's backs, and I am so glad that I've gotten to know all of you during our time at MIT. I would also like to thank Jacqueline Blair Carota, our wonderful program admin, for always supporting us and going above and beyond. During my time at MIT I've also had the privilege to be both a teammate and a coach of the MIT Cheerleading Team – teamwork makes the dream work, and I love you all. I would also like to give a shout-out to an amazing group of women, the Graduate Women at MIT Exec Board 2015, for being such badass ladybosses. I would also like to thank the Kirkland House Tutor team for welcoming a new tutor couple in 2013 and supporting us ever since. And many, many hugs to my right brain Katie Barzee, Mandy Nagy soon-to-be Shepardson, Danyelle Thorpe, Jake Connors, and Meredith McGregor.

I wouldn't be here – either defending my doctoral thesis or, you know, alive – without my amazing family. To my dad Larry Brock and my grandma Elizabeth Street – I love you both so much, now and always, and you are always in my heart and I want y'all to be proud up there. To my Aunt Edna McClain – I miss you and love you. To my honorary aunt and uncle Judy and Bernie Rush, I love y'all. To my great-aunt Martha Timmerman, I love you a bushel and a peck and a hug around the neck and I get my smarts from you. And to my mom Linda Brock, who raised me, who loves me unconditionally, who is always there for me and who is forever and always the best mother ever and my champion. I love you so much, Mama.

And finally, to Tomo Lazovich, my fiancé. You are my best friend, my soulmate, and I still can't believe how lucky I am to have you in my life. Thank you so much for everything – for proofreading papers at 2 AM, for giving me pound cake, for dancing and crying with me, for my spot, for loving me back – I can't wait to spend the rest of my life with you. You are my certainty, and I love you with all my heart.

Table of Contents

Chapter 1

Protein folding and misfolding

Proteins, composed of chains of amino acids connected by peptide bonds, typically must undergo folding to reach a structure or set of structures that is required for its cellular function. A primary component in determining how a protein can fold is its amino acid sequence, which can determine inter-residue and extra-molecular interactions and is subject to energetic effects based on the hydrophobicity of each amino acid. The cellular interior, however, is a highly heterogeneous and densely crowded solution of proteins, which can interfere with how proteins reach and maintain these stable conformations (Luby-Phelps, 2000). A polypeptide chain must also be able to reach its native state on a biologically relevant timescale, despite the large number of available conformations of the protein. The protein folding problem, first articulated approximately fifty years ago, involves three primary questions: 1) how an amino acid chain gives rise to a unique native conformation, 2) how a protein can fold quickly enough to be useful to the cell, and 3) how we can predict computationally the 3D structure of a protein based solely on its sequence (Dill and MacCallum, 2012). In particular, solving the third component of this formulation – i.e., creating an algorithm to predict accurate structural information directly from the amino acid sequence – would be beneficial for a wide range of concepts, ranging from discovering new protein functions to designing better targeted drugs.

Current thought on how and why proteins fold can be traced back to Anfinsen's experiments in refolding denatured ribonuclease A. His finding that this protein could refold after denaturants were removed *in vitro* indicated that the information necessary to encode the protein's structure was contained in the amino acid sequence alone. This result then led to the thermodynamic hypothesis, which states that a protein's native conformation under physiological conditions is at its Gibbs free energy minimum (Anfinsen, 1973). The correlation between a protein's lowest energy structure and its native state appears to hold true for many proteins studied to date, and this dogma has helped guide our understanding of the protein folding process for over 50 years (Dill et al., 2008). Looking at protein folding through a thermodynamics lens can also give insight into the stability of a protein. Thermodynamic stability refers to the change in Gibbs free energy associated with the transition from the unfolded to the folded state, which reflects a balance between making enthalpically-favored inter-residue contacts, increasing the entropy of associated water molecules, and decreasing the entropy of the overall structure among other considerations. A Gibbs free energy value that is significantly lower for the folded state of the protein than for the unfolded state is said to be thermodynamically stable, with associated higher levels of the folded state present at equilibrium.

Several algorithms that aim to predict the 3D structure of a given protein rely on free energy minimization; however, the astronomically large number of conformations of an amino acid chain necessitates a better understanding of how proteins are able to traverse their energy landscapes quickly enough to find their native state on the timescale of cellular interactions. Overcoming this argument relies both on understanding the cellular milieu in which proteins fold and on how sections of proteins can fold locally to give rise to the global structure. Ab initio algorithms, which aim to determine fold based solely on sequence, tend to work based on small input protein sizes, vast computing power, and new approaches to creating starting points for energy minimization. The ROSETTA software suite, first published in the late 1990's as a relatively successful entry to the Critical Assessment of Protein Structure Prediction (CASP) contest, generates an ensemble of structures based on a simulated annealing program that builds up longer structures from short peptide chains found locally in the protein sequence. Discrimination among these 'decoy' models is then achieved by using a scoring function based on energy minimization (Simons et al., 1999). Currently, an expanded version of the software called Rosetta@Home (and its interactive version, FoldIt) aims to increase the available computing power by widely distributing folding problems to run on tens of thousands of privately-owned computers who donate idle processor time to the project.

Ab initio methods like Rosetta can provide structural information about small proteins (<100 residues), but for larger peptide chains homology modeling has given more accurate predictions at the cost of necessitating more experimental structure predictions. The query sequence is aligned with a database of other protein sequences whose structures have been experimentally determined, and the fold(s) of the highest scoring protein(s) are then mapped onto the target sequence. For a target sequence with >50% sequence identity to its template, the root mean square error can be within 1 Angstrom, while a sequence identity of less than 30% can have a completely incorrect fold (Baker and Sali, 2001). This type of modeling is limited by both the accuracy of sequence alignments and by the availability of experimentally determined structures.

Given sequence alignments of homologous proteins across species, recent work has also indicated that three-dimensional fold can be calculated based on evolutionary couplings of amino acids (Marks et al., 2011, 2012; Morcos et al., 2011). The underlying idea is that amino acids that are spatially correlated and form a residue-residue contact in the protein's native conformation are more likely to have undergone selective pressure to co-evolve, where a mutation in one will be more likely to necessitate a corresponding mutation in its interaction partner to maintain the contact. By increasing the sophistication of methods used to determine from multiple sequence alignments which evolutionary couplings authentically represent co-evolving residues, this technique has shown success in predicting three-dimensional folds of membrane proteins (Hopf et al., 2012) and inter-protein contacts in complexes (Hopf et al., 2014; Ovchinnikov et al., 2014).

Another consideration of the protein folding problem is examining the reaction rates associated with folding processes; even though a protein may be thermodynamically stable, it still may be unable to reach its free energy minimum on a biologically relevant timescale. The sample space of potential conformations is astronomically large, leading to Levinthal's paradox: if proteins randomly have to sample potential conformations, the protein would not be able to fold on a timescale compatible with life. Proteins also may be able to adopt nonfunctional, irreversible states distinct from the native state, leading to a depletion of the correctly folded protein. The irreversible states, including aggregated forms of different proteins, are separated from the native folded state by free energy barriers associated with intermediate transition states. If the free energy barrier to these nonfunctional states is high enough, the reaction rate of proteins achieving these terminal forms is low enough to be irrelevant on cellular time scales. Such proteins whose functional conformations are separated from their non-functional conformations by high energetic barriers are termed 'kinetically stable' (Sanchez-Ruiz, 2010). Several proteins like alpha-lytic protease have also been observed whose native state is not actually at a free energy minimum (and therefore fall outside of the thermodynamic hypothesis), which are able to stay in their native conformations due to kinetic trapping by experiencing large energetic barriers between the native state and a lower-energy conformation (Sohl et al., 1998). Low thermodynamic and/or kinetic stability can then lead to decreased amounts of correctly folded protein in cells, and are therefore potential contributors to cellular persistence – namely, the ability of a protein to maintain a fold that is resistant to cellular machinery that targets misfolded proteins.

Misfolding, where a protein either cannot reach or cannot maintain its native conformation in the cell, causes a loss of functionality associated with that protein; furthermore, these molecules can form potentially cytotoxic aggregates if left in the crowded intracellular environment (Chiti and Dobson, 2006, 2009). Mutations that lead to protein misfolding and/or aggregation have been implicated in proteopathies such as Huntington's and prion disease, emphasizing a need to better understand the cellular response to protein misfolding in the context of the physical driving forces that govern how an amino acid sequence can reach its native structure.

Protein quality control targets misfolded proteins

The protein quality control (PQC) machinery, found in different forms across all kingdoms of life, consists of the cellular pathways linked to protein folding and misfolding (Hartl et al., 2011). Chaperone proteins, a key component of PQC systems, can either assist a protein in folding or can target incorrect conformations for either refolding or destruction through the ubiquitin proteasome pathway or autophagy (Kim et al., 2013). Despite the prevalence of PQC machinery across organisms, many aspects of the system are not well understood. One current area of investigation involves how PQC systems recognize substrates that are misfolded. A recent discovery suggests that the ubiquitin ligase nuclear protein San1 can recognize exposed hydrophobicity and substrate protein insolubility and target these aberrant proteins for destruction in the nucleus of the cell (Rosenbaum et al., 2011). Subcellular localization has also been shown to play a role in how and when proteins become degraded, including differential shuttling of misfolded or aggregated proteins to juxtanuclear quality control compartments (JUNQs) or cytosolic insoluble protein deposits (IPODs) that uniquely determine protein fate (Kaganovich et al., 2008). Other hypotheses involve the lack of some specific tertiary structural element in misfolded proteins or the recognition of other structural motifs.

Complicating the question of how PQC systems can recognize misfolded substrates, proteins can contain regions (up to and including the entire sequence) that lack well-defined three-dimensional structures under physiological conditions. These intrinsically-disordered proteins (Uversky and Dunker, 2010) can often adopt a wide range of conformational states as part of their molecular function and are overrepresented in certain classes of proteins like those involved in cancer pathways (Iakoucheva et al., 2002). In eukaryotes, proteins with disordered regions >30 residues long comprise approximately half of the proteome (Dunker et al., 2008). Studying such proteins, whether the structural disorder is intrinsic or not, has proven difficult since these proteins do not respond well to experimental methods of determining structure like X-ray crystallography. These proteins often display a larger hydrodynamic radius than expected from their sequence length on non-denaturing PAGE assays, resulting from their inability to collapse completely into a well-folded state. Intrinsically-disordered proteins can also be recognized by their sensitivity to protease digestion – since increased flexibility is associated with more disordered regions, parts of the backbone become more exposed in solution and therefore more vulnerable to protease digestion (Fontana et al., 2004). This flexibility has also been shown to have functional advantages, allowing for a wider range of low-affinity interactions with multiple binding partners (Rosenbaum et al., 2011). Another method of examining intrinsic disorder in proteins is circular dichroism, which measures the difference in sample absorption of left and right circularly polarized light, with characteristic patterns observed for different secondary structural elements in far-UV circular dichroism and for tertiary elements in near-UV circular dichroism (Greenfield, 2006). Computational methods have also proven useful for studying IDPs. Artificial neural networks trained on disordered sequences form the basis of the PONDR (Prediction of Naturally Disordered Regions) family of algorithms, while biophysical predictors like FoldIndex rely on sequence characteristics like the relatively high ratio of mean net charge to mean hydropathy for IDPs to predict disordered regions (He et al., 2009). Experimental methods for studying IDPs, both in terms of classification and in terms of exploring more detailed structural and functional information, are an ongoing field of research.

Although NMR spectroscopy has provided some insight into examining structural fluctuations in proteins, obtaining an accurate snapshot of a disordered protein's allowed conformations remains challenging (Oldfield and Dunker, 2014). Recently, a phenomenological model developed by J.L. England to estimate the "burial trace," or the squared distance of each amino acid from the center of mass in the lowest-energy fold of a polypeptide chain, has shown promise as a method of computationally exploring the allowed conformational space of protein folds (England, 2011). The burial trace is computed by minimizing an energy function consisting of the hydropathies of each residue and the stretching between neighbor amino acids, subject to steric constraints. The calculation generally takes less than a second to run for short sequences, and adding noise to the parameters of the system can generate an ensemble of amino acid burial patterns for a given protein sequence that can be used to investigate the variability in structures that a protein can adopt.

Examining the benefits of using the model in this manner requires testing it on a well-defined system of structural disorder. The human von Hippel-Lindau protein (pVHL) is one such example that both contains intrinsically disordered regions and is particularly susceptible to incorrect folding. This model misfolding protein forms part of an E3 ubiquitin ligase complex that targets molecules like HIF-1$^{\alpha}$ for degradation, and has been catalogued in depth because hundreds of different mutations have been linked to cancer pathways in humans (Nordstrom-O'Brien et al., 2010). The first ~60 residues of the 213-residue protein remain disordered in the protein's native state, and it must undergo a distinct folding pathway *in vivo* including interacting with chaperones and binding with its cofactors elongin B and elongin C to achieve a state resistant to cellular degradation (McClellan et al., 2005; Schoenfeld et al., 2000). When folded correctly in complex with its cofactors, the non-disordered region adopts a well-defined tertiary structure; however, the protein adopts a molten globule state without its binding partners *in vitro* that consists of a partially collapsed state with some secondary structure but no tertiary structure (Sutovsky, 2004). This molten globule state, determined by gel filtration chromatography Stokes radius calculations, ease of low-concentration urea denaturation, and circular dichroism, indicates that pVHL has difficulty achieving its native state without interactions with other proteins. Perturbations to the system, including mutations to pVHL, are likely to lead to a misfolded or otherwise nonfunctional version of the protein *in vivo* (Hansen et al., 2002; Knauth et al., 2006). When VHL is introduced into non-native systems like *S. cerevisiae* or *E. coli*, where it does not exist naturally, the protein cannot achieve a biologically stable state and in yeast is quickly degraded by the cell (Melville et al., 2003; Sutovsky, 2004). In the context of understanding PQC in yeast, pVHL is a protein whose typical state is poised between adequate folding and being targeted for destruction through a misfolding recognition pathway.

To investigate the link between the underlying biophysics of protein folding and the role of the PQC machinery in recognizing and either repairing or clearing misfolded proteins, the England model was used to investigate the folding characteristics of the human von Hippel-Lindau tumor suppressor protein. Burial traces were calculated to predict exposed residues for the lowest energy conformations of different mutations of pVHL, 20 of which were created experimentally, to help elucidate the interplay between some of the physical forces that govern protein folding and the PQC response.

References

Anfinsen, C.B. (1973). Principles that Govern the Folding of Protein Chains. Science 181, 223–230.

Baker, D., and Sali, A. (2001). Protein Structure Prediction and Structural Genomics. Science 294, 93–96.

Chiti, F., and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. 75, 333–366.

Chiti, F., and Dobson, C.M. (2009). Amyloid formation by globular proteins under native conditions. Nat. Chem. Biol. 5, 15–22.

Dill, K.A., and MacCallum, J.L. (2012). The Protein-Folding Problem, 50 Years On. Science 338, 1042–1046.

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. (2008). The Protein Folding Problem. Annu. Rev. Biophys. 37, 289–316.

Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. 18, 756–764.

England, J.L. (2011). Allostery in protein domains reflects a balance of steric and hydrophobic effects. Struct. Lond. Engl. 1993 19, 967–975.

Fontana, A., de Laureto, P.P., Spolaore, B., Frare, E., Picotti, P., and Zambonin, M. (2004). Probing protein structure by limited proteolysis. Acta Biochim. Pol. 51, 299–321.

Greenfield, N.J. (2006). Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. 1, 2876–2890.

Hansen, W.J., Ohh, M., Moslehi, J., Kondo, K., Kaelin, W.G., and Welch, W.J. (2002). Diverse Effects of Mutations in Exon II of the von Hippel-Lindau (VHL) Tumor Suppressor Gene on the Interaction of pVHL with the Cytosolic Chaperonin and pVHL-Dependent Ubiquitin Ligase Activity. Mol. Cell. Biol. 22, 1947–1960.

Hartl, F.U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. Nature 475, 324–332.

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., and Dunker, A.K. (2009). Predicting intrinsic disorder in proteins: an overview. Cell Res. 19, 929–949.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell 149, 1607–1621.

Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradović, Z., and Dunker, A.K. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. J. Mol. Biol. 323, 573–584.

Kaganovich, D., Kopito, R., and Frydman, J. (2008). Misfolded proteins partition between two distinct quality control compartments. Nature 454, 1088–1095.

Kim, Y.E., Hipp, M.S., Bracher, A., Hayer-Hartl, M., and Ulrich Hartl, F. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis. Annu. Rev. Biochem. 82, 323–355.

Knauth, K., Bex, C., Jemth, P., and Buchberger, A. (2006). Renal cell carcinoma risk in type 2 von Hippel–Lindau disease correlates with defects in pVHL stability and HIF-1α interactions. Oncogene 25, 370–377.

Luby-Phelps, K. (2000). Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. Int. Rev. Cytol. 192, 189–221.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6, e28766.

Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nat. Biotechnol. 30, 1072–1080.

McClellan, A.J., Scott, M.D., and Frydman, J. (2005). Folding and Quality Control of the VHL Tumor Suppressor Proceed through Distinct Chaperone Pathways. Cell 121, 739–748.

Melville, M.W., McClellan, A.J., Meyer, A.S., Darveau, A., and Frydman, J. (2003). The Hsp70 and TRiC/CCT Chaperone Systems Cooperate In Vivo To Assemble the Von Hippel-Lindau Tumor Suppressor Complex. Mol. Cell. Biol. 23, 3141–3151.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. 108, E1293–E1301.

Nordstrom-O'Brien, M., van der Luijt, R.B., van Rooijen, E., van den Ouweland, A.M., Majoor-Krakauer, D.F., Lolkema, M.P., van Brussel, A., Voest, E.E., and Giles, R.H. (2010). Genetic analysis of von Hippel-Lindau disease. Hum. Mutat. 31, 521–537.

Oldfield, C.J., and Dunker, A.K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annu. Rev. Biochem. 83, null.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 3, e02030.

Rosenbaum, J.C., Fredrickson, E.K., Oeser, M.L., Garrett-Engele, C.M., Locke, M.N., Richardson, L.A., Nelson, Z.W., Hetrick, E.D., Milac, T.I., Gottschling, D.E., et al. (2011). Disorder Targets Misorder in Nuclear Quality Control Degradation: A Disordered Ubiquitin Ligase Directly Recognizes Its Misfolded Substrates. Mol. Cell 41, 93–106.

Sanchez-Ruiz, J.M. (2010). Protein kinetic stability. Biophys. Chem. 148, 1–15.

Schoenfeld, A.R., Davidowitz, E.J., and Burk, R.D. (2000). Elongin BC complex prevents degradation of von Hippel-Lindau tumor suppressor gene products. Proc. Natl. Acad. Sci. 97, 8507–8512.

Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3, 171–176.

Sohl, J.L., Jaswal, S.S., and Agard, D.A. (1998). Unfolded conformations of alpha-lytic protease are more stable than its native state. Nature 395, 817–819.

Sutovsky, H. (2004). The von Hippel-Lindau Tumor Suppressor Protein Is a Molten Globule under Native Conditions: IMPLICATIONS FOR ITS PHYSIOLOGICAL ACTIVITIES. J. Biol. Chem. 279, 17190–17196.

Uversky, V.N., and Dunker, A.K. (2010). Understanding protein non-folding. Biochim. Biophys. Acta BBA - Proteins Proteomics 1804, 1231–1264.

Chapter 2

Structural Basis for Modulation of Quality Control Fate in a Marginally Stable Protein

Originally published as:

Kelly P. Brock, Ayelet-chen Abraham, Triana Amen,Daniel Kaganovich, Jeremy L. England. "Structural basis for modulation of quality control fate in a marginally stable protein." *Structure*. 2015 Jul 7; 23 (7):1169-78. Epub 2015 May 28.

http://dx.doi.org/10.1016/j.str.2015.04.015

Abstract

The human von Hippel-Lindau (VHL) tumor suppressor is a marginally stable protein previously used as a model substrate of eukaryotic refolding and degradation pathways. When expressed in the absence of its cofactors, VHL cannot fold and is quickly degraded by the cell's quality control machinery. In this work, we combined computational methods with *in vivo* experiments to examine the basis of VHL's misfolding propensity. By expressing a set of randomly mutated VHL sequences in yeast, we discovered a mutant form stable against degradation. Subsequent modeling suggested the mutation had caused a conformational change affecting cofactor and chaperone interaction, and this hypothesis was then confirmed by additional knockout and overexpression experiments targeting a yeast cofactor homolog. These findings offer a detailed structural basis for the modulation of quality control fate in a model misfolded protein and highlight burial mode modeling as a rapid means to detect functionally important conformational changes in marginally stable globular domains.

Introduction

To function properly in the cell, a globular protein chain typically must remain folded into a specific conformation or set of conformations known as its native state. A primary determinant of how a globular protein folds is its amino acid sequence, which fixes the pattern of internal and external forces that act on the polypeptide chain in the aqueous environment. When a protein cannot reach or cannot maintain its native conformation in the cell it is considered to be in a misfolded state, which is often accompanied by a loss of cellular function. Misfolded proteins also exhibit a marked tendency to associate non-specifically, and sometimes form potentially cytotoxic aggregates if left to accumulate in the crowded intracellular environment (Amen and Kaganovich, 2015; Chiti and Dobson, 2006, 2009; Luby-Phelps, 2000). A large range of globular proteins have been shown to form amyloid fibrils under partially-denaturing conditions (Chiti and Dobson, 2009). Mutations or stochastic processes that lead to protein misfolding and/or aggregation have been implicated in proteinopathies such as Huntington's, Alzheimer's, Parkinson's, and the prion-based Creutzfeldt-Jakob disease, emphasizing a need to better understand the cellular response to protein misfolding in the context of the physical driving forces that govern how an amino acid sequence can reach its native structure (Mulligan and Chakrabartty, 2013).

The protein quality control (PQC) machinery consists of the cellular pathways linked to protein folding and misfolding. Chaperone proteins, a key component of PQC systems, can either assist a misfolded protein in refolding or can target incorrect conformations for destruction through the ubiquitin proteasome pathway or autophagy (Kim et al., 2013). Despite the prevalence of PQC machinery across organisms, many aspects of the system are not well understood. Misfolded proteins are able to reach a wide variety of different non-native states, and the PQC must be able to recognize this diverse group of conformations either to assist refolding or to target them for destruction before the misfolded polypeptide can detrimentally affect the cell. Hsp70, a pleiotropic heat-shock-induced chaperone conserved from bacteria to eukaryotes, has been shown to recognize exposed hydrophobic sites, particularly short (~5-7 residue) hydrophobic sequences flanked by positively-charged amino acids (Marcinowski et al., 2013; Rüdiger et al., 1997). The eukaryotic chaperonin TRiC, on the other hand, has eight distinct subunits that are each capable of recognizing distinct motifs in a variety of substrates, with mutations in different subunits leading to different cellular phenotypes (Amit et al., 2010; Spiess et al., 2006). Because the outcome of the PQC triage decision must ultimately depend on a protein's structure, investigating the role of small conformational perturbations in the sensitivity of a misfolded protein to the QC machinery can indicate how a PQC substrate's fate is modulated for typical substrates.

Marginally stable proteins, which can misfold easily and exist at a 'tipping point' between stable conformations and PQC-targeted misfolded variants, have been used as an experimental mechanism to explore PQC substrate recognition and subsequent refolding and degradation pathways. The human von Hippel-Lindau protein (VHL) is one such example that is particularly susceptible to incorrect folding. This model misfolding protein forms part of an E3 ubiquitin ligase complex that targets molecules like HIF-1$^{\alpha}$ for degradation, and has been catalogued in depth because hundreds of mutant forms have been linked to cancer pathways in humans (Nordstrom-O'Brien et al., 2010). The first ~60 residues of VHL remain disordered in the native state; however the 213-residue protein as a whole must traverse a distinct folding pathway *in vivo* including interacting with chaperones such as Hsp70 and TRiC and binding with its cofactors elongin B and elongin C to achieve a state resistant to cellular degradation

(McClellan et al., 2005; Schoenfeld et al., 2000). For TRiC, two short motifs in the VHL sequence – Box 1 and Box 2 – have been shown to be necessary and sufficient for TRiC binding to VHL in yeast (Feldman et al., 2003). When folded correctly in complex with its cofactors, the non-disordered region adopts a well-defined tertiary structure; however, the protein adopts a molten globule state without its binding partners *in vitro* that consists of a partially collapsed state with some secondary structure but no tertiary structure (Sutovsky, 2004). This molten globule state indicates that VHL has difficulty achieving its native state without interactions with other proteins. Perturbations to the system, including mutations to VHL, often lead to a misfolded or otherwise nonfunctional version of the protein *in vivo* (Hansen et al., 2002; Knauth et al., 2006). When VHL is introduced into non-native systems like *S. cerevisiae* or *E. coli*, where it does not exist naturally, the protein cannot achieve a biologically stable state and in yeast is quickly degraded by the cell (Kaganovich et al., 2008; Melville et al., 2003; Sutovsky, 2004; Weisberg et al., 2012). Since VHL is a protein whose typical state is poised between adequate folding and being targeted for destruction in yeast, it is ideal for use as a probe of how different folds (or misfolded variants) can lead to diverse outcomes through PQC pathways.

One of the persistent difficulties in understanding the physical mechanisms of protein misfolding and subsequent PQC interactions is that almost by definition, misfolded proteins are not amenable to conventional methods of structural characterization. Protein chains that adopt main different conformations cannot be crystallized easily, and aggregation-prone proteins are difficult to solubilize for *in vitro* characterization. Thus, in examining the effects of different mutations on a marginally stable protein like VHL, a computational model that could give insight into the resulting structural changes could offer a new and much needed perspective on the connection between sequence, structure, and recognition by PQC machinery for a large number of sequences. Recently, we developed a phenomenological model to predict tertiary structural information from sequence alone in globular proteins that has shown promise as a method of computationally exploring the allowed conformational space of fluctuating protein folds (England, 2011). The burial trace is computed by minimizing an energy function consisting of the hydropathies of each residue and the stretching between neighbor amino acids, subject to steric constraints. The calculation generally takes less than a second to run for short sequences, and adding noise to the parameters of the system can generate an ensemble of amino acid burial patterns for a given protein sequence that can be used to investigate the variability in structures that a protein can adopt. The rapidity of this model in determining structural information makes it an excellent candidate for probing large numbers of potential mutations of marginally stable proteins to understand PQC response to different conformations *in silico* and to guide *in vivo* experiments. This analysis could also shed light on a possible functional role for marginal stability, which may enable sensitive modulation of expression through qualitative transitions in conformational state.

To investigate the link between the underlying biophysics of protein folding and the PQC fate of a model misfolded protein, the burial mode model was used to investigate the folding characteristics of the human von Hippel-Lindau tumor suppressor protein. Burial traces were calculated to predict exposed residues for the lowest energy conformations of different mutations of VHL, 20 of which were generated experimentally and tested for their degradation properties *in vivo*. One of these mutations had markedly and consistently higher levels of VHL present at steady state. Through the use of burial mode analysis, the structural basis of its enhanced ability to persist in the cell was characterized. Our findings confirm that VHL sits on a structural tipping point in sequence space, where a single mutation can lead to a qualitative shift in folding stability that leads to an altered quality control outcome. Not only do these results highlight the power of a new computational model in gaining elusive information about the structure of intrinsically disordered proteins, they also raise the possibility that such proteins may generally be poised to exhibit strong sensitivity to mutation *in vivo*, where small perturbations can lead to large differences in the amount of folded protein that survives PQC supervision.

Results

Our initial goal was to see if we could use computational modeling of the VHL protein to design mutations that would alter its misfolding and degradation. Since VHL is a human protein, it lacks its elongin binding partners when expressed in yeast, and thus is quickly degraded. We began by using burial mode analysis to design mutations of VHL that would be expected to reduce the protein's misfolding propensity, and thus slow the degradation of the protein in the yeast cytosol.

Our previous work presented a phenomenological framework to estimate the "burial trace," or the distance of each amino acid from the center of mass in the lowest-energy fold of a globular polypeptide chain. This burial trace predictor assumes that each amino acid behaves like it is connected by a spring to its neighbor residues, simulating peptide bonds (Fig. 1). These hypothetical springs contribute to the overall energy for the system, along with the energetically unfavorable terms of putting hydrophilic residues far from the center of the protein and burying hydrophobic ones. This energy function can then be minimized through linear optimization under geometric constraints, specifically to ensure that the individual peptides are not clustered into a small volume that would lead to prohibitive steric clashing in a physical scenario. The calculation generally takes less than a second to run for short sequences, and results in a prediction of the distance of each residue from the center of the protein in its lowest energy state, called the burial trace.

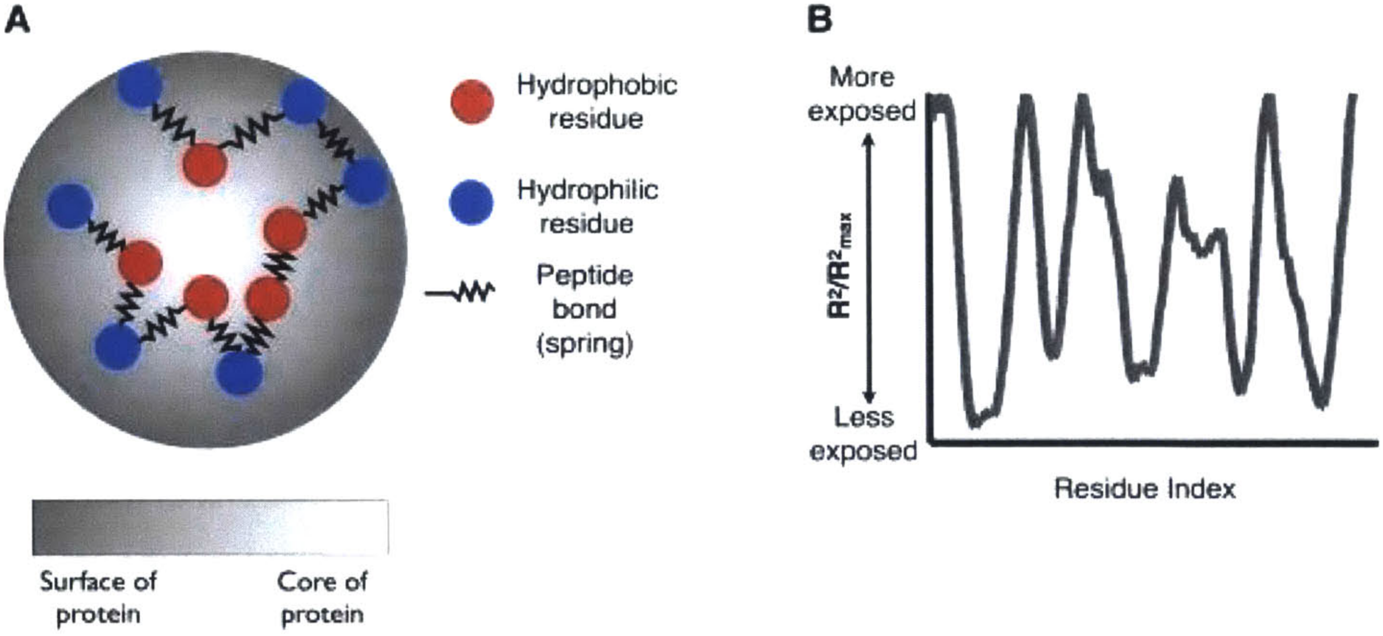


Fig. 1. **Overview of burial mode modeling** *(a). An amino acid chain can be modeled as points connected by springs representing peptide bonds. An energy function including costs for stretching neighbor residues apart and putting hydrophobic residues on the outside of the protein can be minimized under our desired steric constraints to predict low-energy burial patterns. (b) An example of a burial trace, plotting a measure of an amino acid's distance from the center of mass of the protein in its lowest-energy conformation versus that amino acid's index in the protein sequence.*

VHL's relatively small size (213 residues) and predominance of alpha helical structure in its non-disordered region make it a good candidate for use of the burial mode model. As a first attempt at modeling VHL stability, all $\binom{213}{2}$ possible VHL pair-swap mutations (where amino acids at two different locations in the sequence are switched, which preserves overall residue frequencies for the sequence) were ranked according to a parameter called structural variability. The hypothesis behind our procedure was that VHL mutants that were less capable of fluctuating between structurally different conformations at low energy (i.e. at energies comparable to the thermal energy scale kT above the energy of the optimal ("native") configuration) were less likely to be recognized as misfolded by the quality control system. Accordingly, for each pair-swap mutation, we first calculated the burial trace that was maximally different from the optimal lowest energy trace at fixed, low energy. Then, we ranked the mutations, and selected those ten for which this maximized difference was smallest. This procedure yielded ten "least structurally variable" mutants to test experimentally. We also selected a control group of ten mutations that were randomly chosen with uniform probability over all possible pair-swap mutations. The complete set of mutations is listed in Table 1.

Low Structural Variability Mutations			Random Mutations		
1.	K171	P61	11.	E173	K196
2.	R176	G14	12.	E70	G19
3.	V20	R3	13.	V155	G19
4.	P99	V74	14.	T124	K159
5.	V66	P40	15.	T157	V194
6.	K171	P5	16.	D28	A56
7.	V20	R4	17.	G93	V181
8.	R167	A11	18.	R69	S168
9.	C162	N7	19.	L201	E173
10.	Q203	N109	20.	S183	C162

Table 1. *List of mutants that were created experimentally, half of which were chosen based on a model parameter and half of which were chosen randomly.*

All 20 mutations were then created in an *S. cerevisiae* constitutive plasmid with an attached fluorescent Dendra2 tag and transfected into budding yeast cells, which contain neither native VHL protein nor the human elongin co-factors. If the VHL protein were able to adopt a fold that is resistant to PQC degradation, then the attached fluorescent tag would also be preserved and observable. Fluorescence was detected by using quantitative flow cytometry (Fig. 2a). A subset of these mutants were then tagged with GFP and integrated into the genome under the control of a galactose-operated promoter, and their expression level was quantified by fluorescence microscopy at steady-state levels (Fig. 2b).

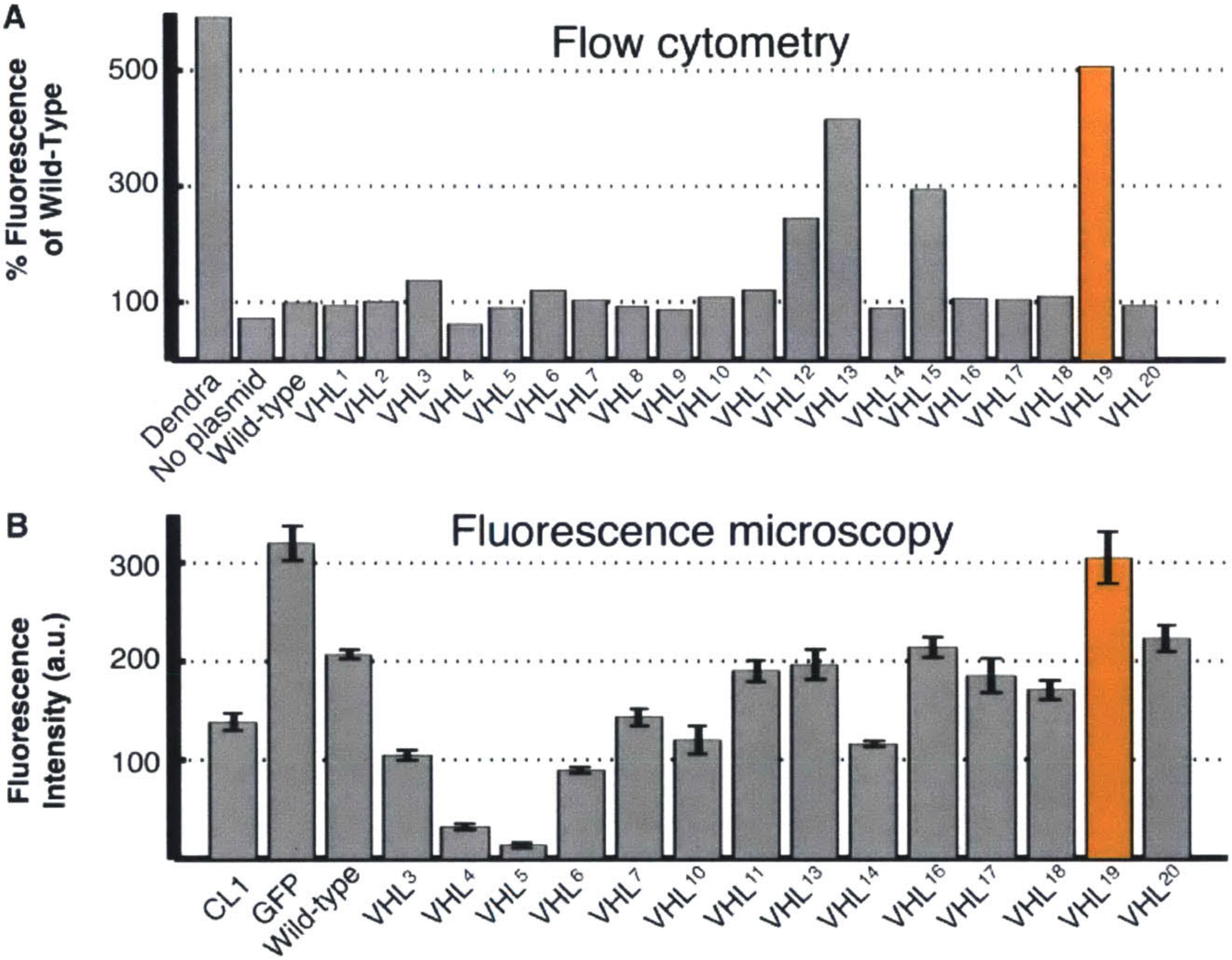


Fig. 2. Analysis of VHL mutant stability by flow cytometry and fluorescence microscopy

(a) Fluorescence measurements by quantitative flow cytometry for all experimentally-created VHL mutations and normalized to fluorescence of the wild type sequence. The orange entry indicates the most stable mutant, VHL^{19} (L201-E173) in both panels. (b) Selected mutants were also introduced to S. cerevisiae cells and expressed endogenously as a GFP fusion construct, and observed directly by fluorescence microscopy with results normalized to GFP levels without VHL. Error bars for all sections represent standard error. See also degradation curves in Figure S1.

Despite how they were chosen, the first ten non-randomly chosen mutations exhibited lower overall fluorescence than the ten randomly chosen control mutations. This result either could be interpreted to mean that the predictions of the burial mode model were wrong, or else that our chosen criterion for enhanced VHL folding (low predicted structural variability) was the wrong one. Had the random control set behaved no differently than the designed mutants, the trail would have gone cold at this point. Surprisingly, however, two mutations in the random control set achieved greater than a three-fold increase over the wild type sequence in the flow cytometry data. The mutant V155-G19 (henceforth referred to as VHL^{13}), which demonstrated 70% of the fluorescence of the Dendra2 tag alone, changed a hydrophobic residue in the Box 2 region (spanning residues 148-155) which mediates the chaperonin TRiC's interaction with VHL and replaced it with the smaller and more flexible glycine, which is normally located in the disordered N-terminal region. This mutation was distinct in our set because it was the only one that directly affected a known chaperonin binding site, which provided a natural rationale for its altered degradation.

The mutant VHL^{19}, L201-E173, however, exhibited the highest level of fluorescence with a five-fold increase in steady-state levels compared to the wild type sequence and ~85% of the baseline Dendra2 level, which corresponds to a five-fold increase in the protein's ability to escape PQC degradation. This mutation's high steady-state levels were further confirmed in the fluorescence microscopy. The biophysical characteristics of the residues swapped were dissimilar, with the hydrophobic leucine at residue 201 changing into a negatively charged glutamic acid (normally at residue 173) and vice-versa. Although both mutations occurred in the region that adopts a well-defined conformation during correct folding, neither was in a known binding site for cofactors or chaperones, raising the possibility that an allosteric response could explain the structural change necessary to evade PQC degradation systems.

Past work has established several interaction sites in the well-folded region of VHL, including two short motifs called Box 1 (residues 114-118) and Box 2 (residues 148-155) that are known to bind to chaperonins (Feldman et al., 2003) and a short binding motif for VHL's cofactors elongin B and elongin C. A BLAST search for VHL cofactors elongin B and elongin C also revealed that although elongin B does not have a counterpart in yeast, a homolog of elongin C is present. Furthermore, this homolog has been shown to bind to a VHL fragment containing residues from 157-171, which contains the known elongin B/C binding site, *in vitro* (Botuyan et al., 2001). Since elongin B and C stabilize VHL in human cells, one testable hypothesis is that burying the chaperonin interaction sites (and therefore possibly protecting VHL from chaperone-assisted degradation) and exposing the elongin interaction site could help explain why certain mutations caused the protein to resist degradation in the cell.

To explore how mutations could affect conformations, particularly in understanding exposure patterns of different interaction sites, the burial trace of each experimentally created mutant was computed using the burial model and compared to the actual burial trace calculated from the crystal structure of the well-folded region. Since the basic burial trace model works best on alpha-helical structures, each input sequence was truncated to residues 63-204 corresponding to the region that can be crystallized (PDB 1VCB) in an attempt to improve accuracy within the region that can adopt a well-defined folding state (Stebbins et al., 1999). Fig. 3 shows the predicted burial traces for the wild type and VHL^{19} sequences, along with maximal and minimal predicted burial values at each residue for all 20 experimentally created mutations. The experimental burial trace obtained directly from the crystal structure of wild type

VHL (PDB 1VCB chain C) is also shown. The crystal structure burial trace shows a distinctive burial of the Box 1 and Box 2 regions, suggesting that TRiC uses these regions to recognize the protein's failure to fold natively. The elongin binding site meanwhile is moderately exposed compared to Box 1 and Box 2 in the native conformation.

Out of all the experimentally tested mutants, the well-stabilized VHL[19] mutation (Fig. 3, orange line) is predicted to have the most buried Box 1 and Box 2 regions, similar to how TRiC is able to bury these motifs in achieving its native conformation in human cells. Furthermore, the elongin interaction region of the VHL[19] mutation was distinctive in that the burial mode model predicted that it was one of the most exposed regions in the entire sequence. This property of burying known chaperone interaction sites and exposing the cofactor binding site is unique among the other mutant and wild type sequences (Fig. 3, gray envelope and black line, respectively), and is more similar to the known native conformation of the protein derived from its crystal structure.

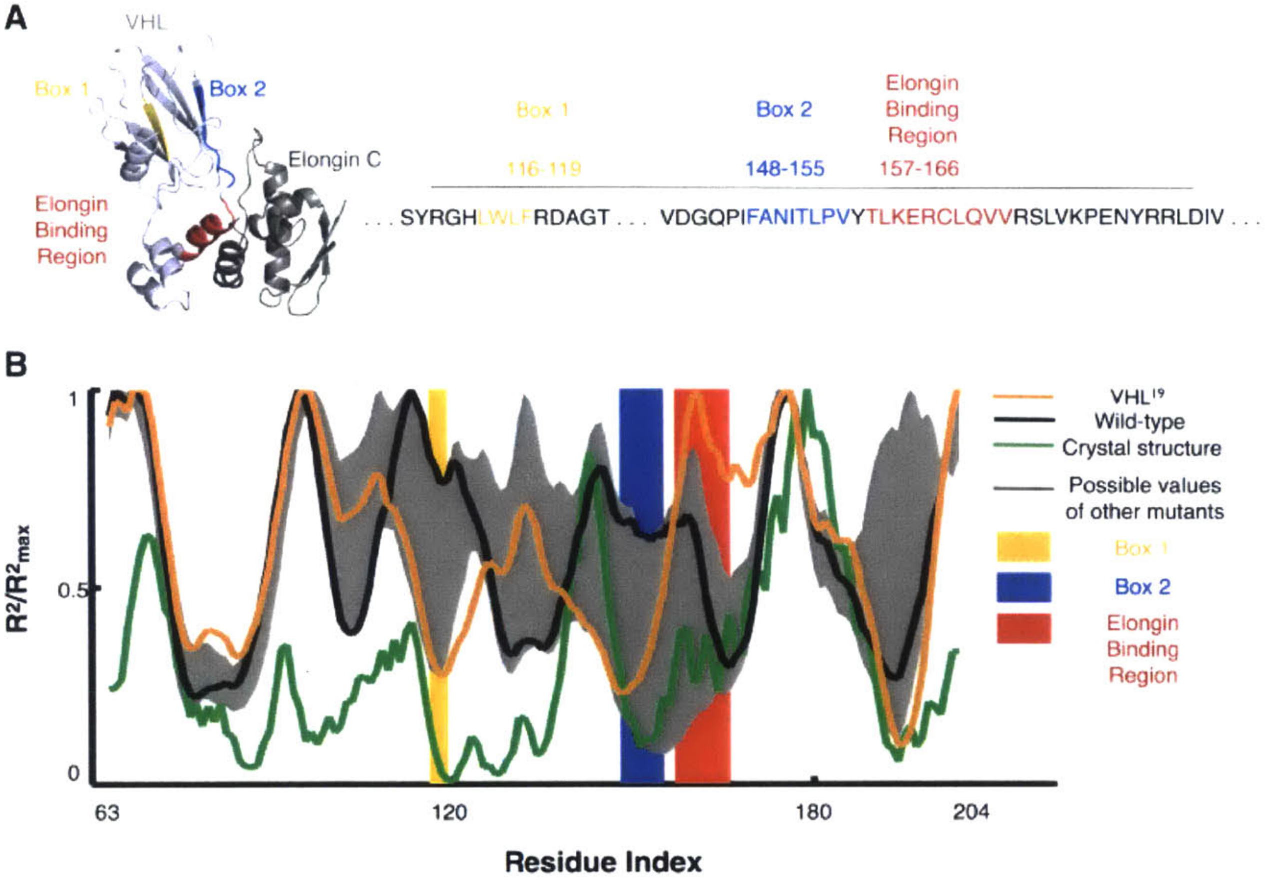


Figure 3. **Burial mode analysis of VHL mutants**

(a) The crystal structure of parts of VHL can be obtained when the protein is bound in complex to cofactors like elongin C, which binds to VHL between residues 157-166. VHL also is known to interact with chaperones in regions called Box 1 and Box 2. (b) The VHL[19] mutation (orange line) is predicted to bury its Box 1 and Box 2 regions, suggesting a conformation that is closer to the form the protein takes when stably bound as part of the VBC complex (green line). The elongin-interaction site of the mutation is also predicted to be highly exposed compared to both the wild type (black line) and the other experimental mutations (grey envelope). Location of specific mutations are shown in Figure S2.

To test the significance of this finding, we also calculated each possible pair swap mutation in the region of VHL that can be crystallized (residues 63-204) and compared their burial traces by Pearson correlation to the crystal structure's burial trace. Only 6.18% of possible mutations in this region had a higher correlation with the crystal structure than the VHL^{19} mutant (Fig 4a).

The two stabilizing mutants were unique in that they affected chaperone binding sites, either through direct mutation, as in the case of VHL^{13}, or through causing the burial of these sites to be more energetically favorable, as in VHL^{19}. The other distinctive component of the VHL^{19} burial trace, however, was its extreme exposure of its elongin cofactor binding site. We hypothesized that VHL^{19} evades degradation by combining reduced chaperonin interaction with increased stabilization through interaction with the elongin co-factor. Thus, we scored all possible pair swap mutations by ranking directly how well the Box 1 and Box 2 chaperonin interaction regions were buried and the elongin interaction site was exposed based on the burial model's predictions. This score can be obtained from a mutated sequence by calculating the sum of the predicted distances of the most-buried residues in the Box 1 and 2 regions from the center of mass of the protein, and adding this subscore to the predicted distance of the maximally-exposed residue in the elongin interaction region from the most exposed residue of the entire protein (Fig 4b). The smallest score would correspond to Box 1 and Box 2 being located at the center of the protein and the elongin interaction site being on the most external part of the protein's surface, which would in turn correspond to the hypothetically best way of evading degradation by achieving greater folding stability in the cell.

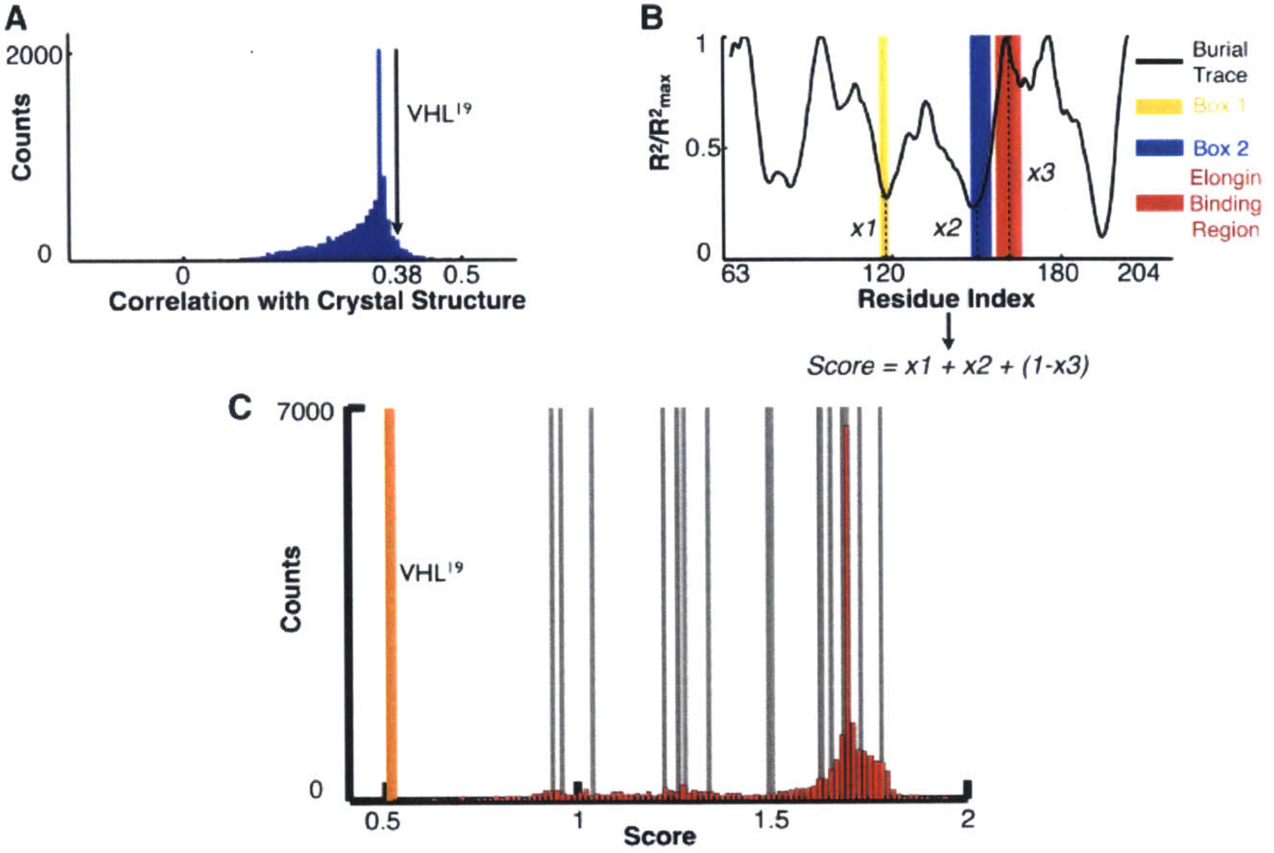


Figure 4. **VHL^{19}'s burial pattern is distinct from the other mutated sequences** *(a) Compared to all possible pair swap mutations in the region of VHL that is able to be crystallized (residues 63-204), VHL^{19} (black arrow) is predicted to have a higher burial pattern correlation with the crystal structure than 93.82% of all other possible mutations. (b) An illustration of how scores were generated for each possible mutant, where the smallest score would correspond to maximal burial of the Box 1 and 2 regions and maximal exposure of the elongin interaction site. (c) Each possible pair swap mutation in the region [63,204] was scored according to Fig. 4b. VHL^{19} has the lowest (best) score out of ~10,000 mutations. The locations of the other experimental mutants in the histogram are indicated by the gray lines, with VHL^{19} marked by the orange line.*

Amazingly, according to this scheme, VHL^{19} turned out to have the lowest (most optimal) score out of ~10,000 possible mutations (Fig. 4c). The location of the other experimentally determined mutants is also shown, indicating a large gap in scores between mutant VHL^{19} and the rest of the experimental sequences. The scores of the rest of the sequences, however, did not correlate with their fluorescence levels, indicating that this may be a binary effect only observed at the extreme achieved by VHL^{19}.

The homolog of elongin C, Elc1, is not an essential protein in *S. cerevisiae*, so to test our hypothesis that the VHL19 mutation helped stabilize the protein by decreasing chaperone interactions and/or interacting with Elc1, the mutated VHL sequences were transfected into an Elc1 knockout of yeast. Fluorescence microscopy was again used to measure persistence of the VHL mutant protein in the cell (Fig. 5). The stable expression of VHL^{19} in the knockout strain Δ*elc1* was decreased to that of the wild type, indicating that Elc1 could be helping to stabilize this mutation and lending credence to the findings of the burial model. A plasmid containing human elongin B and C, known to help stabilize VHL in humans, was then transfected to cells expressing the integrated VHL mutants and constitutively expressed (EBC OE strain). The presence of VHL's natural cofactors further stabilized the VHL^{19} mutant by ~70%, providing more evidence that the cofactor elongin homolog directly contributes to the exceptional extra stability of the VHL^{19} mutant.

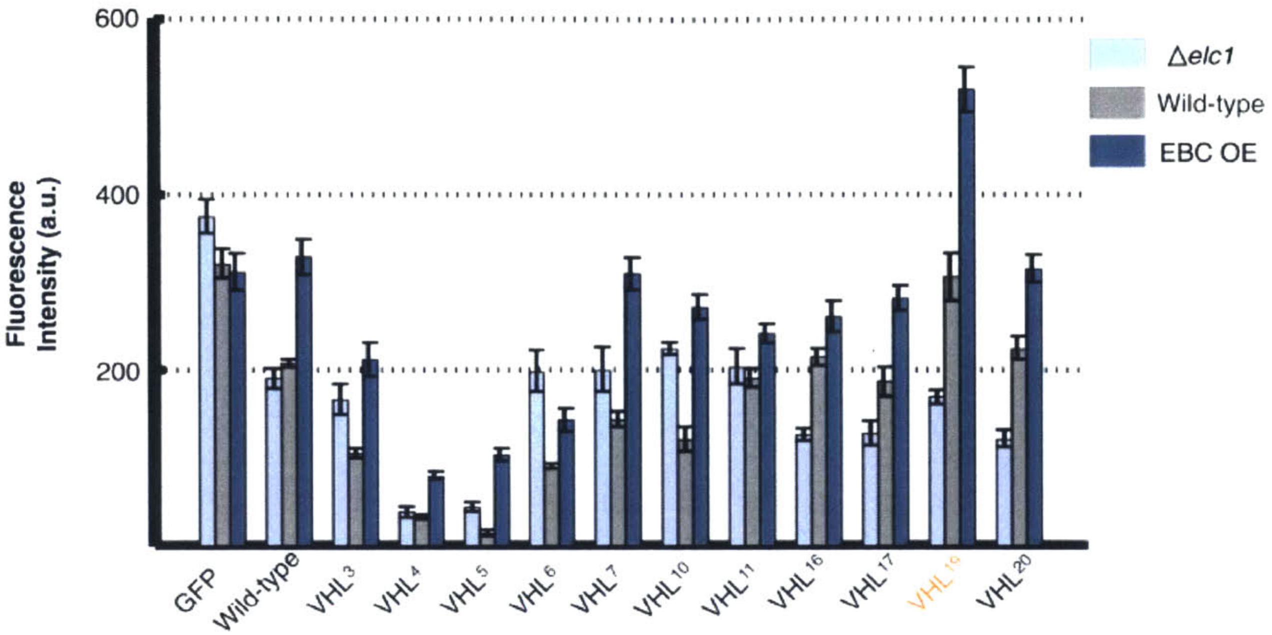


Fig. 5. Analysis of mutant VHL interaction with Elc1 through deletion and overexpression *GFP constructs with VHL were expressed in wild type S. cerevisiae (grey, center in grouping), a knockout strain for the yeast homolog of elongin C Δelc1 (light blue, leftmost in grouping), and in the EBC OE strain that introduced and overexpressed human versions of elongin B and C (dark blue, rightmost in grouping). Although expressing VHL^{19} in the* Elc1 *knockout strain decreased the mutant VHL's levels to that of wild type VHL, the presence of human elongin B and C markedly increased VHL^{19}levels. Error bars represent standard error in all sections. Degradation curves for the wild type sequence under both EBC OE and knockout conditions are also shown supplemental figure S1.*

Having quantified fluorescent levels, we then looked more closely at the cell biological phenotypes of the different mutants. Further observation of the behavior of different VHL mutants in the Δ*elc1* environment revealed clear puncta formation for most mutants, including VHL^{19}, in the absence of the yeast homolog of elongin C (Fig. 6). Puncta formation is linked to different aspects of the PQC system, including differential processing of terminally misfolded proteins and as a response to external and internal stress (Amen and Kaganovich, 2015; Narayanaswamy et al., 2009). Elongin C is itself part of the polyubiquitination process, functioning as a component of an E3 ubiquitin ligase complex in both yeast and mammals and therefore directly contributing to PQC systems through its participation in ubiquitin-mediated degradation pathways (Lisztwan et al., 1999; Ramsey et al., 2004). Intriguingly VHL^{13} (which had also exhibited high fluorescence in the original flow cytometry experiment) escaped this cellular phenotype, remaining diffuse throughout the cell. This VHL sequence is the only one tested that changed a residue in a known chaperonin binding site, indicating that this mutation might be able to escape PQC recognition and avoid being sequestered in puncta before degradation. This result is certainly consistent with our original hypothesis that the higher fluorescence of VHL^{13} was due to altered chaperonin interaction.

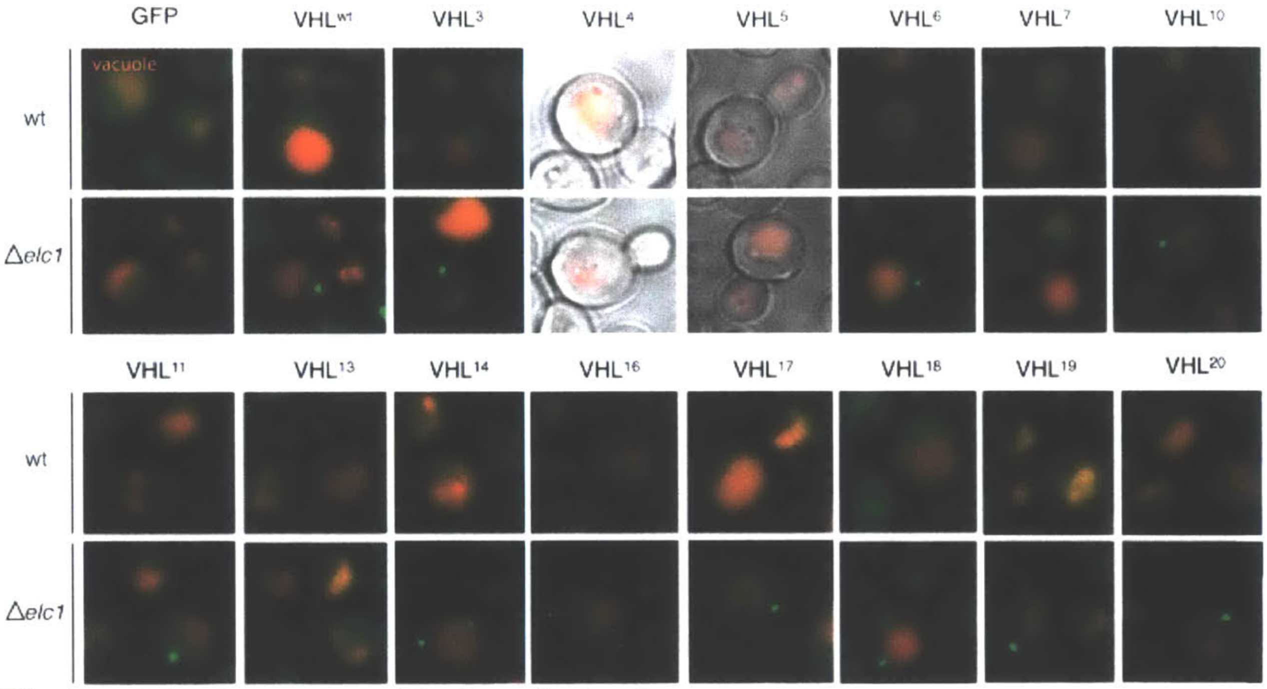


Fig. 6. Cellular phenotype resulting from Elc1 knockout

Fluorescently tagged VHL mutants display a distinct cellular phenotype when expressed in the Δelc1 strain compared to the wild type strain. Different panels show different mutant versions of VHL in both the wild type (top) and Δelc1 (bottom) strains, with green representing the fluorescent protein fusion construct and red the vacuole. VHL^{4} and VHL^{5} were unable to be detected in the cell at appreciable quantities, and only GFP alone and VHL^{13} escaped the puncta phenotype characterized by the other VHL forms.

Discussion

The ability of a cell to recognize aberrant proteins and target them for either refolding or destruction is crucial to maintaining protein homeostasis and preventing pathological aggregation. At the same time, allosteric interactions caused by relatively minor mutational changes can have a large impact on how a protein is categorized by the PQC system. Consequently, an improved understanding of how potential PQC substrates may be driven between qualitatively different conformational states by various perturbations is essential to a full understanding of how the cell maintains proteostasis.

A persistent challenge in studying many PQC substrates is that by nature they tend to be structurally disordered and are therefore not easily amenable to traditional methods of *in vitro* structural characterization (Oldfield and Dunker, 2014). As a result, there is an opportunity for computational methods to play a central role in providing valuable structural information that can be used in combination with experimental results to better understand how the PQC fate of misfolding-prone proteins may be affected by mutation or stress. This study presents a novel method of integrating computational biophysics techniques to gain an experimentally testable glimpse into how marginal stability affects protein interactions with the PQC machinery. This approach of using *in silico* results to guide *in vivo* experiments, and vice-versa, demonstrates that both approaches can be used in tandem to yield better explanations for these biophysical phenomena.

In particular, due to its marginal stability, VHL has been used as a model substrate for chaperone proteins involved in assisted folding pathways as well as for studying PQC systems. Hsp90, Hsp70, and TRiC have all been shown to interact with VHL in yeast, with the chaperonin TRiC potentially functioning as a 'holdase' for containing nascent VHL until it can bind to its elongin cofactors (Feldman et al., 1999). Although TRiC hasn't been directly implicated in the degradation processes, some mutant forms of VHL have also been shown to be able to refold after denaturing events in a chaperonin-independent way *in vitro* but not *in vivo* (Feldman et al., 2003; Yang et al., 2013).

In this study we present 20 novel pair swap mutations of VHL that switch the position of two residues in the sequence and examine the protein products in yeast for a measure of their ability to escape rapid degradation by the cell. In particular, two mutations caused VHL to be detected in much larger amounts than that of transfected wild type VHL: swapping the valine and glycine residues at positions 155 and 19 respectively (VHL^{13}) and swapping the leucine and glutamic acid residues at positions 201 and 173 (VHL^{19}). The first mutant directly affected a chaperone binding site, while the second did not mutate residues in any known chaperone or cofactor interaction motifs.

Initially, these results were surprising because both stabilizing mutations were chosen randomly as part of the control group, while the test group of ten mutants designed to exhibit more stability actually produced less-stable behavior on average. The underlying rationale behind choosing this latter group of mutants focused on predicting how easily a protein structure could access different structures at a low energy, with the idea that sequences displaying the least variability in predicted structure also might indicate increased thermodynamic stability and decreased degradation by the PQC system. The behavior of these mutants *in vivo* illustrates the difficulty in teasing apart the mechanisms of PQC response for marginally stable proteins. In the case of the ten designed mutants, the lower-than-expected observed amounts present in the cell may have been indicative that the mutated sequences were 'kinetically trapped' in one configuration that was more easily identified by the PQC system. An estimate between the random and non-random sets bounded the probability of observing no highly fluorescent hits in the non-random set while obtaining two or more hits in the random control set at a maximum value slightly below 0.1. It is possible that the structural variability calculations actually predicted the opposite effect than originally thought – namely, lower ability to persist in the cell rather than higher thermodynamic stability. In this case, our designed parameter to maintain protein structure despite small energetic fluctuations may have made these mutants more susceptible to degradation, underscoring the importance of examining how protein conformational changes can affect the outcome of PQC interactions. Future sequence design protocols along the lines of our initial attempt should therefore be modified to include not only the effect of structural variability, but also the importance of maintaining resemblance to a native fold that exposes the right parts of the chain on the surface.

To explain the two unexpectedly stable hits in the random control sample, the biophysical burial mode model was used to examine each of the twenty mutated sequences. These structural predictions were also compared to the wild type sequence and its experimentally determined structure. Strikingly, the burial mode model predicted that the mutated sequence of VHL^{19} would fold such that it would bury known hydrophobic chaperone interaction sites while exposing a motif known to be necessary and sufficient for binding to a homolog of its human cofactor elongin C in yeast (Figure 3). The two chaperone interaction sites have been identified as binding to TRiC, with the first motif also similar to the known Hsp70 recognition motif of a short hydrophobic sequence flanked by charged residues (Kim et al., 2013; Rüdiger et al., 1997). Thus, one possible effect of the VHL^{19} mutation could be to suppress interaction between the VHL protein and the PQC machinery. Even more intriguingly, the burial mode modeling also points to the possibility that this mutant form of VHL is stabilized because of interactions with a cofactor homolog that does not stabilize the wild type sequence. The cofactor elongin C exists in yeast despite the absence of other VHL cofactors, but previous studies had indicated that it does not increase the half-life of wild type VHL (McClellan et al., 2005). However, when we introduced VHL^{19} into a yeast strain lacking the elongin C homolog, levels of mutant VHL were decreased to that of the wild type protein, providing compelling experimental evidence in support of the hypothesis that the yeast homolog of elongin C does substitute for the human co-factor's effect as predicted and stabilizes the VHL^{19} mutant. This hypothesized interaction between the elongin cofactor homolog Elc1 and VHL in yeast was further supported when the different mutants were examined by microscopy (Figure 6), since the absence of Elc1 gives rise to a new quality control phenotype in wild type and most mutated forms of VHL.

Using VHL as a model protein, the burial model has been shown to be capable of providing explanations of the relative likelihood of a mutated protein being degraded based on its ability to adopt different conformational states. Furthermore, this predictive capability was able to scan a large set of sequences in a relatively short time, providing rapid structural information about all possible mutations to narrow and refine experimental studies. The effectiveness of the burial mode model in explaining structural changes in mutant forms of VHL has significant implications for the underlying physical driving forces behind VHL stability. The model predicts structural change by calculating the trade-off between hydrophobic and steric effects; thus, our findings point to the possibility that VHL's marginal stability originates in a fluctuating balance struck among three distinct stretches of moderately hydrophobic sequence (Box 1, Box2, and the ELC-binding region) that compete for limited space in the protein's packed hydrophobic core.

Meanwhile, the elevated levels of VHL[13], V155-G19, that we observed are more likely to have been the result of altered interaction with chaperones that directly followed from changes to known chaperone binding sites. The fact that this mutant was alone in escaping the puncta phenotype observed by microscopy in other VHL variants in the Δ*elc1* strain further illustrates the complex interactions that characterize cellular PQC systems involving both segregation and degradation of terminally misfolded proteins.

Stabilization both through modulation of chaperone interactions and through cooperative interactions with binding partners is implicated for VHL in determining how the PQC responds to conformational variants. Past computational work has underlined the importance of allosteric conformational change to VHL structure and function (Liu and Nussinov, 2008). Our findings confirm the importance of these effects specifically in connection to cofactor binding. Furthermore, this analysis also raises the question of whether or not a marginally stable protein's susceptibility to adopting different structural forms might be advantageous to the cell, since this differential folding pattern based on small perturbations is directly coupled to PQC response. For example, an enhanced sensitivity to mutation can allow an additional level of cellular modulation of a protein's functional fold, or can contribute to a protein's enhanced ability to rapidly traverse an evolutionary landscape to find potential new folds.

Expanding our computational and experimental analysis to other proteins besides VHL may also prove informative in how the PQC responds to other marginally stable proteins, and in elucidating the importance of such a protein's structural flexibility in a cellular context. The well-characterized p53 protein is similar to VHL in that it is small enough for burial trace analysis (393 amino acids), includes an N-terminal intrinsically disordered region, shows low thermodynamic and kinetic stability, and is characterized by a large number (>1000) of cancer-related mutations found in humans (Joerger and Fersht, 2007). Additionally, a recent study has indicated that p53 interacts with the same chaperonin, TRiC, that is involved in VHL folding (Trinidad et al., 2013). This protein's similarity to VHL, both in terms of its functional properties and its clinical importance, makes it an enticing target for burial analysis. Further exploration of the structural basis of misfolded protein recognition, including of client tumor suppressors like VHL and p53, can give new insights into how cells maintain proteostasis and what structural mechanisms can drive qualitative shifts in the PQC fates of marginally stable proteins on the brink of structural disorder.

Experimental Procedures

Yeast Strains, growth conditions and materials

Yeast growth, media preparation and manipulations were performed according to standard protocols (Adams et al., 1997). The strains used are listed in Table 2.

Strain	Genotype	Reference
BY4741	BY4741 *Mata his3del0 leu2del0 met15del0 ura3del0*	(Brachmann et al., 1998)
BY4741 Δ*elc1*	BY4741 *Mata his3del0 leu2del0 met15del0 ura3del0 elc1::KANMX*	(Winzeler et al., 1999)
BY4741 *GFP*	BY4741 *his3::GALp-GFP -HIS3*	This study
BY4741 *GFP-VHL* (wt or mutants)	BY4741 *his3::GALp-GFP-VHL*$^{(wt\ or\ mutants)}$*-HIS3*	This study

Plasmids used in this study are summarized in Table 3. VHL gene was fused to GFP or DDR2 and was expressed under the control of a galactose-regulated promoter (Gal1p), nls-TFP was used as a nuclear marker and expressed under GAL10p.

Plasmid	Description	Reference

pDK399	pESC-*LEU-ELONGIN B/C*	This study
pDK2-22	pRS303-*GAL1p GFP GAL10p nls-TFP*	This study
pDK2-(2-21)	pRS303 *GAL1p GFP-VHL*$^{wt\ or\ mutants}$ *GAL10p nls-TFP*	This study
pDK5 (1-21)	pESC-*URA GAL1p GFP-VHL*$^{wt\ or\ mutants}$ *GAL10p nls-TFP*	This study
pDK437	pRS316 GPDp *DDR*2	This study
pDK437 (2-21)	pRS316 GPDp *DDR2 -VHL*$^{wt\ or\ mutants}$	This study

Microscopy

For imaging yeast cells were grown on galactose containing media to middle log phase and seeded on concanavalin A (Sigma) coated 4-well microscope plates (IBIDI). Confocal 3D images were acquired using a dual point-scanning Nikon A1R-si microscope equipped with a PInano Piezo stage (MCL), using a 60x PlanApo VC oil objective NA 1.40. Calculations of the fluorescence intensity and image processing were performed using NIS-Elements software.

Single cell degradation assay

Cells were grown as described earlier. The expression of GFP-VHL variants was induced by switching to galactose-containing media for 6 hours. The degradation was calculated as a decrease in the green fluorescence intensity after cells were transferred to glucose media (time 0). The fluorescence intensity of single cells was calculated using NIS-software. All data points were normalized to GFP fluorescence decay under the same conditions.

FACS analysis

Yeast strains were grown on selection media for 2 days, diluted twice a day to log phase, and at the 3 day the fluorescence intensity of the DDR2 tag fused to VHL were analyzed by FACS with 488 laser using BD FACSDiva (BD Biosciences, San Jose, CA) software.

Mutagenesis

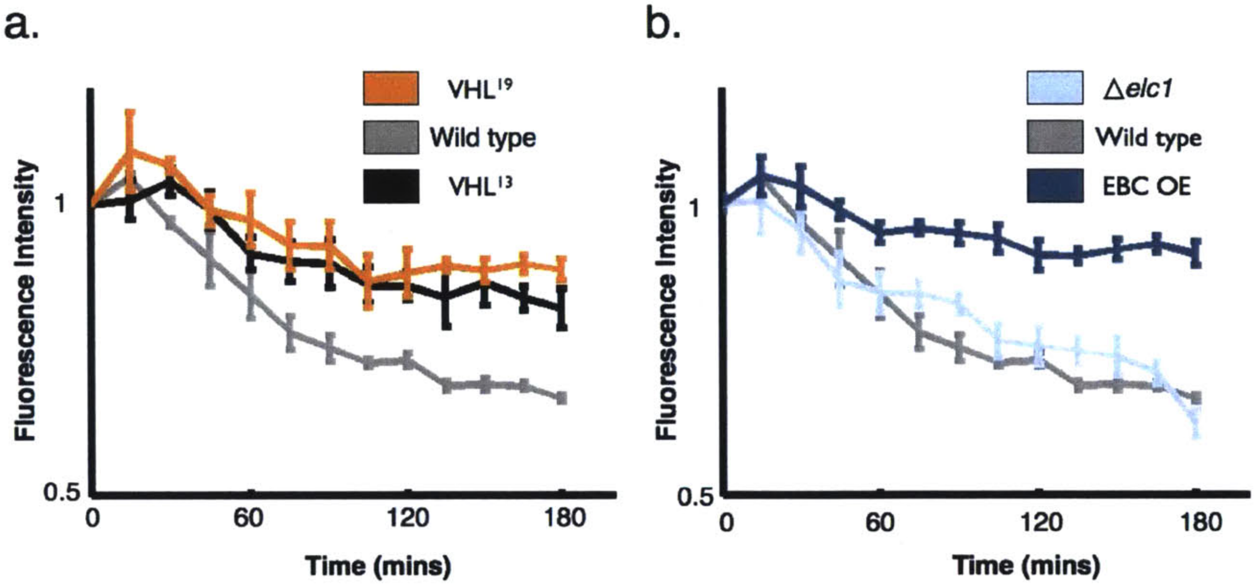
Different VHL mutants were cloned using restriction free cloning protocol as described (van den Ent and Löwe, 2006). Briefly, primers were designed according to desired mutation and used for first PCR reaction to amplify a mega primer caring two mutations on the VHL sequence. Mega primers were used for a second PCR reaction to amplify the whole pDK5 plasmid with the desired two mutations using high fidelity DNA polymerase (KAPA KR0370). Dpn1 digestion were used after second PCR to get rid of the old methylated plasmids and left only with the new mutated plasmids. VHL mutants were then subcloned to pRS plasmids for FACS analysis.

Statistical Analysis

For comparing the designed mutants to the control group, we estimated our probability of the result arising from chance by calculating the number of "hits" (fluorescence levels above a threshold T) in both the designed and control group:

Max(P(no hits > T in designed | f) * P(2 or more hits > T in control | f))

where each probability is calculated as a binomial distribution arising from f, the underlying fraction of hits expected to be seen above the threshold. For the calculation in this paper, the threshold was set at 60%.



Supplemental Figures

Figure S1, related to Figure 2. Degradation curves for VHL.

(a) Degradation curves are shown for the two most stable mutations, VHL^{19} (orange) and VHL^{13} (black), and wildtype sequence (grey). Both mutated sequences exhibit slower degradation compared to the wildtype. Curves are normalized to GFP degradation as well as initial values. (b) The degradation curves for the wildtype sequence under normal (grey), elc1 knockout (light blue), and human elongin BC overexpression (dark blue) are shown, using the same normalization scheme as in (a). Error bars for (a) and (b) are standard error over a triplicate experiment.

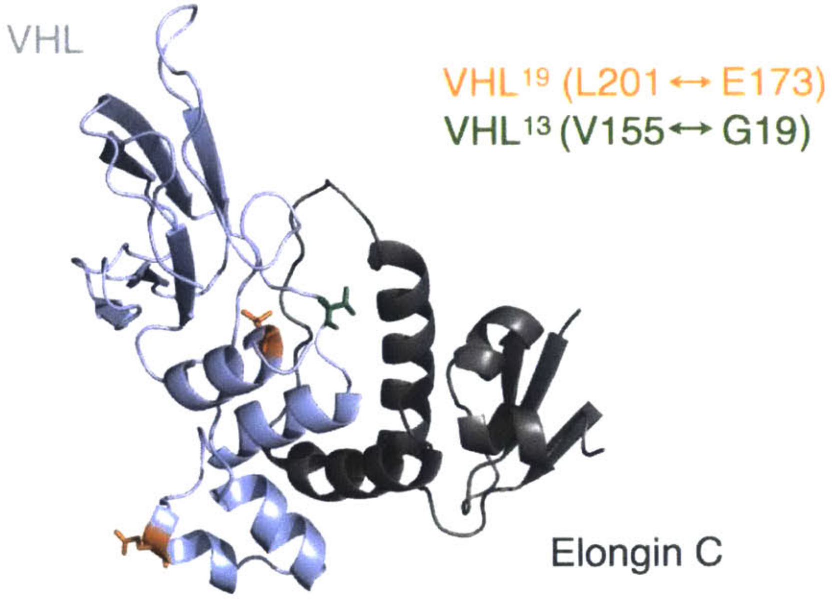


Figure S2, related to Figure 3. Positions of relevant mutations of VHL.

The crystal structure of VHL (light grey) with its binding partner elongin C (dark grey) is shown, with the mutated residues for VHL19 (orange) and VHL13 (green). The second mutation in VHL13, at position 19, falls in a disordered region that has not been crystallized.

Author Contributions

K.B. and J.E. conducted all computational modeling and wrote the manuscript. A.A. created Dendra-tagged mutants and performed flow cytometry experiments, and T.A. created GFP-tagged constructs and performed fluorescence miscroscopy including knockout studies. J.E. and D.K. oversaw all studies and helped formulate experimental design. All authors helped edit the manuscript.

Acknowledgements

We thank members of the Kaganovich lab for discussion and comments on the manuscript. This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC-StG2013 337713 DarkSide starting grant, an Israel Science Foundation Grant ISF 843/11, an Abisch-Frenkel Foundation for the Promotion of Life Sciences grant, a German Israel Foundation Grant GIFI-1201-242.13/2012; an Israel Health Ministry grant under the framework of E-Rare-2, a Niedersachsen-Israel Research Program grant, and a grant from the American Federation for Aging Research. K.B. thanks MIT for a research fellowship supporting this work.

References

Adams, A., Gottschling, D.E., and Kaiser, C. (1997). Methods in Yeast Genetics: A Laboratory Course Manual (Plainview, N.Y: Cold Spring Harbor Laboratory Press).

Amen, T., and Kaganovich, D. (2015). Dynamic droplets: the role of cytoplasmic inclusions in stress, function, and disease. Cell. Mol. Life Sci. CMLS 72, 401–415.

Amit, M., Weisberg, S.J., Nadler-Holly, M., McCormack, E.A., Feldmesser, E., Kaganovich, D., Willison, K.R., and Horovitz, A. (2010). Equivalent Mutations in the Eight Subunits of the Chaperonin CCT Produce Dramatically Different Cellular and Gene Expression Phenotypes. J. Mol. Biol. 401, 532–543.

Botuyan, M.V., Mer, G., Yi, G.-S., Koth, C.M., Case, D.A., Edwards, A.M., Chazin, W.J., and Arrowsmith, C.H. (2001). Solution structure and dynamics of yeast elongin C in complex with a von hippel-lindau peptide. J. Mol. Biol. 312, 177–186.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P., and Boeke, J.D. (1998). Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast Chichester Engl. 14, 115–132.

Chiti, F., and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. 75, 333–366.

Chiti, F., and Dobson, C.M. (2009). Amyloid formation by globular proteins under native conditions. Nat. Chem. Biol. 5, 15–22.

England, J.L. (2011). Allostery in protein domains reflects a balance of steric and hydrophobic effects. Struct. Lond. Engl. 1993 19, 967–975.

Van den Ent, F., and Löwe, J. (2006). RF cloning: a restriction-free method for inserting target genes into plasmids. J. Biochem. Biophys. Methods 67, 67–74.

Feldman, D.E., Thulasiraman, V., Ferreyra, R.G., and Frydman, J. (1999). Formation of the VHL-elongin BC tumor suppressor complex is mediated by the chaperonin TRiC. Mol. Cell 4, 1051–1061.

Feldman, D.E., Spiess, C., Howard, D.E., and Frydman, J. (2003). Tumorigenic mutations in VHL disrupt folding in vivo by interfering with chaperonin binding. Mol. Cell 12, 1213–1224.

Hansen, W.J., Ohh, M., Moslehi, J., Kondo, K., Kaelin, W.G., and Welch, W.J. (2002). Diverse Effects of Mutations in Exon II of the von Hippel-Lindau (VHL) Tumor Suppressor Gene on the Interaction of pVHL with the Cytosolic Chaperonin and pVHL-Dependent Ubiquitin Ligase Activity. Mol. Cell. Biol. 22, 1947–1960.

Joerger, A.C., and Fersht, A.R. (2007). Structure–function–rescue: the diverse nature of common p53 cancer mutants. Oncogene 26, 2226–2242.

Kaganovich, D., Kopito, R., and Frydman, J. (2008). Misfolded proteins partition between two distinct quality control compartments. Nature 454, 1088–1095.

Kim, Y.E., Hipp, M.S., Bracher, A., Hayer-Hartl, M., and Ulrich Hartl, F. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis. Annu. Rev. Biochem. 82, 323–355.

Knauth, K., Bex, C., Jemth, P., and Buchberger, A. (2006). Renal cell carcinoma risk in type 2 von Hippel–Lindau disease correlates with defects in pVHL stability and HIF-1α interactions. Oncogene 25, 370–377.

Lisztwan, J., Imbert, G., Wirbelauer, C., Gstaiger, M., and Krek, W. (1999). The von Hippel–Lindau tumor suppressor protein is a component of an E3 ubiquitin–protein ligase activity. Genes Dev. 13, 1822.

Liu, J., and Nussinov, R. (2008). Allosteric effects in the marginally stable von Hippel–Lindau tumor suppressor protein and allostery-based rescue mutant design. Proc. Natl. Acad. Sci. 105, 901–906.

Luby-Phelps, K. (2000). Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. Int. Rev. Cytol. 192, 189–221.

Marcinowski, M., Rosam, M., Seitz, C., Elferich, J., Behnke, J., Bello, C., Feige, M.J., Becker, C.F.W., Antes, I., and Buchner, J. (2013). Conformational Selection in Substrate Recognition by Hsp70 Chaperones. J. Mol. Biol. 425, 466–474.

McClellan, A.J., Scott, M.D., and Frydman, J. (2005). Folding and Quality Control of the VHL Tumor Suppressor Proceed through Distinct Chaperone Pathways. Cell 121, 739–748.

Melville, M.W., McClellan, A.J., Meyer, A.S., Darveau, A., and Frydman, J. (2003). The Hsp70 and TRiC/CCT Chaperone Systems Cooperate In Vivo To Assemble the Von Hippel-Lindau Tumor Suppressor Complex. Mol. Cell. Biol. 23, 3141–3151.

Mulligan, V.K., and Chakrabartty, A. (2013). Protein misfolding in the late-onset neurodegenerative diseases: Common themes and the unique case of amyotrophic lateral sclerosis. Proteins Struct. Funct. Bioinforma. 81, 1285–1303.

Narayanaswamy, R., Levy, M., Tsechansky, M., Stovall, G.M., O'Connell, J.D., Mirrielees, J., Ellington, A.D., and Marcotte, E.M. (2009). Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. Proc. Natl. Acad. Sci. U. S. A. 106, 10147–10152.

Nordstrom-O'Brien, M., van der Luijt, R.B., van Rooijen, E., van den Ouweland, A.M., Majoor-Krakauer, D.F., Lolkema, M.P., van Brussel, A., Voest, E.E., and Giles, R.H. (2010). Genetic analysis of von Hippel-Lindau disease. Hum. Mutat. 31, 521–537.

Oldfield, C.J., and Dunker, A.K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annu. Rev. Biochem. 83, null.

Ramsey, K.L., Smith, J.J., Dasgupta, A., Maqani, N., Grant, P., and Auble, D.T. (2004). The NEF4 Complex Regulates Rad4 Levels and Utilizes Snf2/Swi2-Related ATPase Activity for Nucleotide Excision Repair. Mol. Cell. Biol. 24, 6362–6378.

Rüdiger, S., Buchberger, A., and Bukau, B. (1997). Interaction of Hsp70 chaperones with substrates. Nat. Struct. Mol. Biol. 4, 342–349.

Schoenfeld, A.R., Davidowitz, E.J., and Burk, R.D. (2000). Elongin BC complex prevents degradation of von Hippel-Lindau tumor suppressor gene products. Proc. Natl. Acad. Sci. 97, 8507–8512.

Spiess, C., Miller, E.J., McClellan, A.J., and Frydman, J. (2006). Identification of the TRiC/CCT Substrate Binding Sites Uncovers the Function of Subunit Diversity in Eukaryotic Chaperonins. Mol. Cell 24, 25–37.

Stebbins, C.E., Kaelin, W.G., Jr, and Pavletich, N.P. (1999). Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function. Science 284, 455–461.

Sutovsky, H. (2004). The von Hippel-Lindau Tumor Suppressor Protein Is a Molten Globule under Native Conditions: IMPLICATIONS FOR ITS PHYSIOLOGICAL ACTIVITIES. J. Biol. Chem. 279, 17190–17196.

Trinidad, A.G., Muller, P.A.J., Cuellar, J., Klejnot, M., Nobis, M., Valpuesta, J.M., and Vousden, K.H. (2013). Interaction of p53 with the CCT Complex Promotes Protein Folding and Wild-Type p53 Activity. Mol. Cell 50, 805–817.

Weisberg, S.J., Lyakhovetsky, R., Werdiger, A., Gitler, A.D., Soen, Y., and Kaganovich, D. (2012). Compartmentalization of superoxide dismutase 1 (SOD1G93A) aggregates determines their toxicity. Proc. Natl. Acad. Sci. 109, 15811–15816.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285, 901–906.

Yang, C., Huntoon, K., Ksendzovsky, A., Zhuang, Z., and Lonser, R.R. (2013). Proteostasis Modulators Prolong Missense VHL Protein Activity and Halt Tumor Progression. Cell Rep. 3, 52–59.

Chapter 3

Understanding protein sequence evolutionary constraints

The previous chapter illustrated that a combination of computational and experimental techniques can be used to understand how an individual misfolded protein can be targeted for degradation by the cell. Under a fluctuating environment, however, many different pathways can be affected by changing external conditions. Individual protein misfolding is also only one potential outcome of external stresses that cause change in the intracellular environment. Therefore, environmental perturbations can represent a driving force that acts to select for specific protein biophysical characteristics. Since these stress responses can change intracellular conditions, one idea is that cytosolic proteins can act collectively to improve overall cellular survival. In this context, a large, coordinated response across both proteins and pathways could help explain how yeast cells are able to exhibit resistance to glucose starvation by entering a quiescent state where metabolism, translation, and division are downregulated. By examining how biophysical characteristics like electrostatic attraction change as a result of starvation stress, we can explore the evolution of collective cytosolic properties.

Previous work has helped elucidate the general evolutionary constraints that act on individual protein sequences, with recent work also providing insight into the constraints on the network topology of protein-protein interaction networks. Epistasis, the phenomenon where genotypic changes can contribute non-linearly to larger phenotypic changes, also represents a promising avenue for exploring how proteins can exhibit both collective behavior and collective evolution. However, understanding selective pressures using this framework has been challenging because it requires identifying relatively weak interactions within a large potential sequence space. In this chapter, I include a discussion of the general constraints that act on individual protein sequences, and will then connect these constraints to selective pressures that act on protein interactions. In the context of environmental perturbations, we can then examine how protein collective behavior evolves to enhance cellular survival under adverse conditions.

Early observations of how proteins differ among species came from examining the sequence compositions of different proteomes, particularly between microorganisms (Freeland and Gale, 1947; Stokes and Gunness, 1946). The line of reasoning that species might have different amino acid frequencies has helped researchers identify key variations; by using the mathematical techniques of principal components analysis, full-proteome composition signatures between archaea, eubacteria, and prokaryota were found to be distinguishable (Pe'er et al., 2004). Disordered proteins, which are more commonly found in multicellular eukaryotes, are also more correlated with an increase in the residues R, K, E, P, and S (Dunker and Obradovic, 2001; Romero et al., 2001). These proteome characteristics have also been used to postulate evolutionary advantages. One study reported that the negatively-charged glutamic acid and the positively-charged lysine and arginine residues have a correlated enrichment in hyperthermophilic species (Tekaia et al., 2002). Multiple evolutionary explanations have been proposed to explain this difference in charged residue frequencies between thermophiles and mesophiles, including increased stability through ionic interactions (Tekaia et al., 2002) and a decreased tendency to participate in hydrophobic-based aggregation (Greaves and Warwicker, 2007). However, these studies have been limited in their ability to examine emergent collective protein behavior that forms a direct response to environmental stresses.

Another evolutionary constraint that has been observed is selection for the thermodynamic and kinetic properties of individual proteins. Protein misfolding can lead to the formation of irreversible aggregates, which are toxic to the cell and require dedicated quality control systems that can be overwhelmed (Chiti and Dobson, 2006; Kaganovich et al., 2008; Mulligan and Chakrabartty, 2013). Furthermore, many globular proteins have the ability to form amyloid-like structures under native conditions based on local unfolding, indicating the degree to which cells must be able to handle potential misfolding events (Chiti and Dobson, 2009). Therefore, the dynamics of protein folding and misfolding represent a potential biophysical evolutionary constraint.

One selective pressure that has been demonstrated is a heightened tendency towards fast (but not the fastest possible) folding kinetics of a group of protein sequences, which can help avoid aggregation of long-lived intermediate states (Mirny and Shakhnovich, 1999). An even stronger evolutionary pressure has also been demonstrated in the SH3 domain family, specifically a strong tendency to have slower unfolding rates (Di Nardo et al., 2003; Sikosek and Chan, 2014). Highly abundant protein species also evolve more slowly, which has been connected to the correlation between protein abundance and aggregation (Ciryam et al., 2013; Serohijos et al., 2012). This evolutionary constraint pushing proteins towards being less likely to misfold is a global trend, but tends to act at the level of individual sequences.

An additional factor to consider in the evolution of protein folding kinetics is the presence of chaperone proteins. This class of proteins, discussed in the previous chapters, helps proteins to fold and additionally can target misfolded proteins for either refolding or degradation. The thermodynamic stability of proteins presents a barrier to the evolution of individual proteins; namely, many mutations tend to be destabilizing and are not well-tolerated by the cell due to an increase in misfolding (Tokuriki et al., 2008). Chaperones can help mitigate this limitation on mutational exploration, specifically by allowing proteins that might have been selected against otherwise on the basis of fold stability or folding kinetics to explore new regions of the fold's fitness landscape (Tokuriki et al., 2004; Wyganowski et al., 2013). Therefore, a protein's evolvability – the ability of a sequence to try new mutations and therefore new functions and/or folds – can be tied to the presence and co-evolution of proteostasis networks.

This connection between protein quality control systems and their client proteins also leads to the question of how molecular chaperones have evolved and what selective pressures act on them. From a biomedical standpoint, using a directed evolution approach in the lab can be beneficial to engineering chaperone systems that are better able to target the cytotoxic aggregates found in neurodegenerative diseases like Huntington's, Parkinson's, and Alzheimer's. One rationale for this methodology is that these diseases often onset after reproductive age and may not provide strong deterrents to maintaining full proteostasis activity against "late-expressing" mutations (Balch et al., 2008; Gavrilov and Gavrilova, 2002; Mack and Shorter, 2016).

Recent forays into this field have also provided examples of the tradeoffs involved in increasing efficiency of specific chaperone proteins. An early study attempted to improve the efficiency of the bacterial chaperonin system GroEL/ES and its involvement in the folding process of the green fluorescent protein GFP (Wang et al., 2002). Although they were able to increase the folding efficiency of GFP significantly, their engineered GroEL/ES system subsequently displayed increased specificity as well. The variants were generally unable to assist their normal clients with folding, which in E. coli amounts to roughly 10% of cytosolic proteins (Clare and Saibil, 2013). These results indicate that there exists an evolutionary tradeoff between being general enough to interact with multiple client substrates and overall functional efficiency for molecular chaperones.

Selective pressure that acts on interaction sites also forms an evolutionary constraint on intrinsically disordered proteins (IDPs) as well. As discussed in Chapter 1, these proteins are characterized by a much more diverse and flexible ensemble of structures that define their native conformational state(s). Due to this plasticity, IDPs are not necessarily subject to the same overall constraints that characterize globular proteins; however, their sequences are not random and they are also susceptible to biophysical limitations (Sikosek and Chan, 2014). IDPs tend to be the hubs in protein-protein interaction networks, and are capable of binding to multiple interaction partners (Dunker et al., 2005; Oldfield et al., 2008). IDPs can undergo partial or almost complete folding upon binding (Dyson and Wright, 2005), so mutations that affect binding sites are subject to increased evolutionary pressures that are still being explored (Brown et al., 2010).

Experimental and evolutionary characterization of protein interaction networks

Client substrate recognition as an evolutionary constraint extends beyond chaperones and IDPs as well. The expansion of high-throughput proteomic technologies over the past few decades has enabled researchers to elucidate protein-protein interaction networks for different species, including *S. cerevisiae*. One particular methodology is yeast two-hybrid (Y2H) screening, a high-throughput experiment that is able to measure the presence of an interaction between a bait and prey protein through the expression of a reporter gene; recent advances in experimental design have allowed for increasingly accurate binary interaction networks to be found (Brückner et al., 2009). In combination with massive libraries, this technique has enabled probing of ~70% of all possible *S. cerevisiae* binary interactions within the yeast proteome (Yu et al., 2008). Mass spectrometry has also emerged as a powerful new tool capable of handling interactome-level questions. Based on how ions derived from peptide sequences can be characterized by their mass-to-charge ratio, mass spectrometry can detail with high accuracy which proteins were present in a complex that had previously been purified (Brückner et al., 2009). Simultaneously, the frameworks of experimental and computational systems biology and network topology have contributed to a better understanding of how protein interactions function in the cellular environment.

The evolution of experimental techniques to identify the components of protein interaction networks has also led to an increased ability to probe the evolution of the protein interaction networks themselves. The expansion of available genomic information has enabled researchers to use multiple sequence alignments and phylogenetic trees to identify co-evolving amino acids within and between proteins which share an aspect of the protein's evolutionary history and are likely to be highly spatially correlated (Göbel et al., 1994; Lockless and Ranganathan, 1999; Pazos and Valencia, 2001; Pazos et al., 2005). This approach has been used to predict protein-protein interactions and to infer three-dimensional structural information from genetic data alone (Hopf et al., 2014; Ovchinnikov et al., 2014; Skerker et al., 2008; Weigt et al., 2009). Beyond general biophysical constraints imposed on a protein's stability and folding kinetics, these results indicate that understanding how evolutionary pressures act at a systems level can provide information about how cells have evolved to survive and thrive under different environmental conditions that can affect many protein pathways at once.

One evolutionary constraint that acts on the entire protein interaction network is the degree of binding partners for each protein within the interactome. Individual proteins can act as 'hubs' that have many potential interactions; work from 2002 showed that these proteins evolve more slowly than less-connected sequences in the network (Fraser et al., 2002). Furthermore, the same study also showed that proteins that interact have similar evolutionary rates, pointing to inter-protein coevolution. In an artificial minimum proteome, a correlation between protein copy number, charge, and mass were also found to be optimized for protein-protein interaction networks (Xu et al., 2013). Protein abundance has also been shown to be negatively correlated with evolutionary rate, likely as a result of a selective pressure to avoid misinteraction between proteins (Pál et al., 2001; Yang et al., 2012).

The concept of promiscuity as a central feature of proteins within an interaction network can also provide information on how specificity evolves. A toxin-antitoxin system was used in a recent study to show that interaction specificity – i.e. the ability of one protein to exclude most strong binding except to its interaction partner – likely evolves through a promiscuous intermediate state (Aakre et al., 2015). For signal transduction systems like histidine kinases and their response regulators in bacteria, specificity allows for cells to accurately interpret environmental information, with crosstalk between different signaling components having the capacity for deleterious effects. Therefore, residues in multiple proteins have been shown to coevolve to target different substrates and maintain a degree of pathway specificity (Skerker et al., 2008).

In the crowded cellular environment, minimization of nonfunctional interactions can also serve as a topological constraint on protein-protein interaction networks. Work based on assuming that nonfunctional interactions are a waste of protein resources predicted that yeast have close to the theoretical maximum number of allowable proteins (Zhang et al., 2008). Later work also showed that a selective pressure against potentially disease-causing nonfunctional binding can cause protein interaction networks in multicellular organisms both to limit the size of their proteomes and to favor a scale-free topology with a few hub proteins and many proteins with few interaction partners (Johnson and Hummer, 2011). This tendency towards decreasing nonfunctional interactions is also echoed in cell's expression profile, where hub proteins and sequences with increased ability to interact nonspecifically tend to have decreased abundances (Heo et al., 2011; Levy et al., 2012). Another constraint imposed on hub proteins as part of their interaction network topology focuses on how they contact their binding partners – hub proteins that tend to be more disordered with a single promiscuous binding interface also tend to interact more through hydrophobic contacts than electrostatic contacts, which impacts the overall network formation (Dosztányi et al., 2006; Peleg et al., 2014).

Additionally, protein quality control systems that impact the evolvability of individual proteins can also impact how protein-protein interactions evolve. Active chaperones that directly catalyze folding have been shown to allow a fuller exploration of sequence space to strengthen functional interactions and to decrease nonfunctional interactions by lessening the selective pressure for individual protein stability (Çetinbaş and Shakhnovich, 2013). These results indicate the degree to which protein quality control and stress response systems can impact the evolutionary trajectory of both individual proteins and their interactions.

Stress response as a proteomic-level selective pressure

Constraints on the evolution of individual protein sequences, like selection for stability and slower unfolding, are examples of global pressures that can also affect how protein interaction networks form. Mutational effects, like the relative advantages of promiscuity versus specificity in network topologies, also might only become apparent when viewed from a systems level. However, these evolutionary trajectories of protein groups have only been studied extensively in the context of protein-protein interaction networks, and further work is needed to understand fully the degree to which co-adaptation of complexed proteins underlies observed patterns of amino acid coevolution (Pazos and Valencia, 2008).

Epistasis, which refers to multiple mutations in a gene or set of genes having compounding instead of additive effects, has been identified as a potential framework for understanding how proteins can evolve collectively (Weinreich et al., 2013). In this context, mutations spread out across many genes or proteins might have a disproportionate effect on a cell's phenotype. In particular, epistasis has been implicated as a driving force behind the evolutionary trajectory of a transcription factor and its DNA binding sites (Anderson et al., 2015). However, detecting these signals can be computationally challenging, since it requires detecting relatively weak yet potentially related genotypic changes across large groups of proteins. Epistatic interactions also remains elusive to quantify, with different studies reporting either minor or major effects on protein evolution (Papp et al., 2011; Starr and Thornton, 2016). Furthermore, determination of the degree to which epistasis acts between proteins has been identified as an appropriate framework to better understand evolutionary processes (Poelwijk et al., 2015; Starr and Thornton, 2016).

Previously, I have focused on understanding protein quality control systems in the context of protein misfolding and its impact on protein evolvability. However, these systems also form part of a cell-wide response to different types of intracellular and extracellular stress. Environmental perturbations like nutrient deprivation or heat shock result in large phenotypic changes in cellular behavior and affect the cell's ability to maintain homeostasis. The degree to which nutrient starvation, for example, can directly impact the local environment of many cytosolic proteins in *S. cerevisiae* provides a lens with which to view protein evolution in the context of enhancing survival given selection on the collective behavior of cytosolic proteins.

Particularly, glucose deprivation in *S. cerevisiae* typically causes a drop in cytosolic pH, which impacts the local environment of cytosolic proteins (Orij et al., 2009). Changes in the local environment have also been linked to collective protein behavior like higher-order assembly formation in this yeast (Petrovska et al., 2014). The large-scale nature of this environmental stress and emerging evidence of its correlated widespread protein response make it an ideal system for examining how epistasis can shape proteome-level selection pressure as a response to nutrient deprivation. Evidence for the degree to which yeast proteins have evolved to exhibit this collective behavior can be found in the next chapter.

References

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. Cell 163, 594–606.

Anderson, D.W., McKeown, A.N., and Thornton, J.W. (2015). Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. eLife 4, e07864.

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

Brown, C.J., Johnson, A.K., and Daughdrill, G.W. (2010). Comparing models of evolution for ordered and disordered proteins. Mol. Biol. Evol. 27, 609–621.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast Two-Hybrid, a Powerful Tool for Systems Biology. Int. J. Mol. Sci. 10, 2763–2788.

Çetinbaş, M., and Shakhnovich, E.I. (2013). Catalysis of Protein Folding by Chaperones Accelerates Evolutionary Dynamics in Adapting Cell Populations. PLOS Comput Biol 9, e1003269.

Chiti, F., and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. 75, 333–366.

Chiti, F., and Dobson, C.M. (2009). Amyloid formation by globular proteins under native conditions. Nat. Chem. Biol. 5, 15–22.

Ciryam, P., Tartaglia, G.G., Morimoto, R.I., Dobson, C.M., and Vendruscolo, M. (2013). Widespread Aggregation and Neurodegenerative Diseases Are Associated with Supersaturated Proteins. Cell Rep. 5, 781–790.

Clare, D.K., and Saibil, H.R. (2013). ATP-driven molecular chaperone machines. Biopolymers 99, 846–859.

Di Nardo, A.A., Larson, S.M., and Davidson, A.R. (2003). The Relationship Between Conservation, Thermodynamic Stability, and Function in the SH3 Domain Hydrophobic Core. J. Mol. Biol. 333, 641–655.

Dosztányi, Z., Chen, J., Dunker, A.K., Simon, I., and Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res. 5, 2985–2995.

Dunker, A.K., and Obradovic, Z. (2001). The protein trinity—linking function and disorder. Nat. Biotechnol. 19, 805–806.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 272, 5129–5148.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. 6, 197–208.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science 296, 750–752.

Freeland, J.C., and Gale, E.F. (1947). The amino-acid composition of certain bacteria and yeasts. Biochem. J. 41, 135–138.

Gavrilov, L.A., and Gavrilova, N.S. (2002). Evolutionary theories of aging and longevity. ScientificWorldJournal 2, 339–356.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins Struct. Funct. Bioinforma. 18, 309–317.

Greaves, R.B., and Warwicker, J. (2007). Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. BMC Struct. Biol. 7, 18.

Heo, M., Maslov, S., and Shakhnovich, E. (2011). Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc. Natl. Acad. Sci. U. S. A. 108, 4258–4263.

Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3.

Johnson, M.E., and Hummer, G. (2011). Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. Proc. Natl. Acad. Sci. U. S. A. 108, 603–608.

Kaganovich, D., Kopito, R., and Frydman, J. (2008). Misfolded proteins partition between two distinct quality control compartments. Nature 454, 1088–1095.

Levy, E.D., De, S., and Teichmann, S.A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc. Natl. Acad. Sci. U. S. A. 109, 20461–20466.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science 286, 295–299.

Mack, K.L., and Shorter, J. (2016). Engineering and Evolution of Molecular Chaperones and Protein Disaggregases with Enhanced Activity. Front. Mol. Biosci. 3.

Mirny, L.A., and Shakhnovich, E.I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1. J. Mol. Biol. 291, 177–196.

Mulligan, V.K., and Chakrabartty, A. (2013). Protein misfolding in the late-onset neurodegenerative diseases: Common themes and the unique case of amyotrophic lateral sclerosis. Proteins Struct. Funct. Bioinforma. 81, 1285–1303.

Oldfield, C.J., Meng, J., Yang, J.Y., Yang, M.Q., Uversky, V.N., and Dunker, A.K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 9 Suppl 1, S1.

Orij, R., Postmus, J., Ter Beek, A., Brul, S., and Smits, G.J. (2009). In vivo measurement of cytosolic and mitochondrial pH using a pH-sensitive GFP derivative in Saccharomyces cerevisiae reveals a relation between intracellular pH and growth. Microbiology 155, 268–278.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 3, e02030.

Pál, C., Papp, B., and Hurst, L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931.

Papp, B., Notebaart, R.A., and Pál, C. (2011). Systems-biology approaches for predicting genomic evolution. Nat. Rev. Genet. 12, 591–602.

Pazos, F., and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14, 609–614.

Pazos, F., and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. EMBO J. 27, 2648–2655.

Pazos, F., Ranea, J.A.G., Juan, D., and Sternberg, M.J.E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 352, 1002–1015.

Pe'er, I., Felder, C.E., Man, O., Silman, I., Sussman, J.L., and Beckmann, J.S. (2004). Proteomic signatures: Amino acid and oligopeptide compositions differentiate among phyla. Proteins Struct. Funct. Bioinforma. 54, 20–40.

Peleg, O., Choi, J.-M., and Shakhnovich, E.I. (2014). Evolution of specificity in protein-protein interactions. Biophys. J. 107, 1686–1696.

Petrovska, I., Nüske, E., Munder, M.C., Kulasegaran, G., Malinovska, L., Kroschwald, S., Richter, D., Fahmy, K., Gibson, K., Verbavatz, J.-M., et al. (2014). Filament formation by metabolic enzymes is a specific adaptation to an advanced state of cellular starvation. eLife 3.

Poelwijk, F.J., Krishna, V., and Ranganathan, R. (2015). The context-dependence of mutations: a linkage of formalisms. ArXiv150200726 Q-Bio.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. Proteins 42, 38–48.

Serohijos, A.W.R., Rimas, Z., and Shakhnovich, E.I. (2012). Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2, 249–256.

Sikosek, T., and Chan, H.S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. J. R. Soc. Interface R. Soc. 11, 20140419.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell 133, 1043–1054.

Starr, T.N., and Thornton, J.W. (2016). Epistasis in protein evolution. Protein Sci. Publ. Protein Soc.

Stokes, J.L., and Gunness, M. (1946). The Amino Acid Composition of Microorganisms. J. Bacteriol. 52, 195–207.

Tekaia, F., Yeramian, E., and Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297, 51–60.

Tokuriki, N., Kinjo, M., Negi, S., Hoshino, M., Goto, Y., Urabe, I., and Yomo, T. (2004). Protein folding by the effects of macromolecular crowding. Protein Sci. Publ. Protein Soc. 13, 125–133.

Tokuriki, N., Stricher, F., Serrano, L., and Tawfik, D.S. (2008). How Protein Stability and New Functions Trade Off. PLOS Comput Biol 4, e1000002.

Wang, J.D., Herman, C., Tipton, K.A., Gross, C.A., and Weissman, J.S. (2002). Directed evolution of substrate-optimized GroEL/S chaperonins. Cell 111, 1027–1039.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl. Acad. Sci. 106, 67–72.

Weinreich, D.M., Lan, Y., Wylie, C.S., and Heckendorn, R.B. (2013). Should evolutionary geneticists worry about higher-order epistasis? Curr. Opin. Genet. Dev. 23, 700–707.

Wyganowski, K.T., Kaltenbach, M., and Tokuriki, N. (2013). GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates. J. Mol. Biol. 425, 3403–3414.

Xu, Y., Wang, H., Nussinov, R., and Ma, B. (2013). Protein charge and mass contribute to the spatio-temporal dynamics of protein-protein interactions in a minimal proteome. Proteomics 13, 1339–1351.

Yang, J.-R., Liao, B.-Y., Zhuang, S.-M., and Zhang, J. (2012). Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc. Natl. Acad. Sci. U. S. A. 109, E831–E840.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110.

Zhang, J., Maslov, S., and Shakhnovich, E.I. (2008). Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. Mol. Syst. Biol. 4, 210.

Chapter 4

Sequences of yeast cytosolic proteins reveal selective pressure on a collective behavior

Kelly P. Brock, Jeremy L. England

Abstract

Under rapidly fluctuating environments, cells that are able to respond appropriately to harsh conditions are more likely to survive and reproduce. In the case of nutrient starvation, many proteins must act in concert to keep a unicellular organism alive, pointing to a concerted selective pressure that acts on many proteins simultaneously. However, understanding how the collective behavior of cytosolic proteins evolves can be difficult since this process involves looking for weak signals spread across many protein sequences. Here, we present the budding yeast glucose starvation response as an example of proteome-level selection pressure. In *S. cerevisiae*, starvation conditions – specifically a lack of glucose – lead to a rapid pH drop in the cytosol. This deficit in nutrients is also linked to an enhanced ability of proteins to form higher-order structures. In this study, we present a proteome-wide computational analysis to examine how protein interactions might be affected by a drop in pH, and present evidence that proteins that are normally associated under physiological conditions may actually be more electrostatically attractive at lower pH values. This tighter association of different proteins could help the cell control large-scale metabolic pathways and store and protect individual proteins from terminal misfolding and aggregation. Studying this phenomenon allows for exploration of proteomic-scale evolution, something that can be difficult to characterize by current experimental methods.

Introduction

Studying the evolutionary trajectories of protein sequences has given a greater understanding of how proteins fold and behave in the crowded intracellular environment. Environmental fluctuations can occur rapidly as well, exposing cells to a wide variety of extracellular stimuli that can have profound effects on their intracellular states. Since external stress can affect many functions throughout the cell, it also represents a selective pressure that acts on many proteins at once. Since many cytosolic proteins are either involved or affected by environmental fluctuations like nutrient depletion, one possibility is that an emergent collective behavior of these proteins can confer increased resilience to suboptimal conditions. The large-scale responses that occur as a response to environmental stress therefore provide a way to study the evolutionary forces that underlie coordinated protein behavior.

Prior evolutionary studies have tended to focus on individual protein sequences. For example, global evolutionary constraints like those imposed on protein minimal stability and unfolding kinetics have been identified as common selective pressures (Di Nardo et al., 2003; Mirny and Shakhnovich, 1999; Sikosek and Chan, 2014). However, recent work has been able to demonstrate that the coevolution of amino acids between proteins is sufficient to make predictions about three-dimensional structures of protein complexes (Hopf et al., 2014; Ovchinnikov et al., 2014). At the level of proteomes, amino acid sequence composition also has been shown to be capable of distinguishing among both kingdoms and individual species (Freeland and Gale, 1947; Pe'er et al., 2004). Evolutionary explanations like increased protein stability or decreased tendency to aggregate have been suggested as possible explanations for differences in charged residues for thermophilic species when compared to mesophiles (Greaves and Warwicker, 2007; Tekaia et al., 2002). However, due to the nature of tracking relatively small, widespread signals across many sequences, examining how the proteome as a unit evolves to survive stress – and how proteins might coordinate emergent behavior as a result - remains difficult to quantify.

One large-scale perturbation that imposes severe constraints on a population of cells is an absence of necessary nutrients. These starved cells, either individually or communally, must be able to respond appropriately to this environmental cue to mitigate potential damage or to avoid premature apoptosis. Such large-scale cellular insults also have the capacity to alter how each individual protein behaves and how these proteins form interactions with each other. Many species have dedicated protein quality control (PQC) systems that can attempt to rescue misfolded protein variants that become more abundant under stress conditions or target them for degradation (Hartl et al., 2011; Kim et al., 2013). Although these systems are complex and well-suited for a variety of responses, they rely on ATP and other energy carriers being readily available to function - which can be depleted under nutrient deprivation. Since multiple pathways are affected by extracellular conditions, having a signal capable of quickly initiating stress response tactics for many proteins at once can be advantageous for cellular survival.

One potential example of both a rapid and large-scale phenotypic change that occurs during nutrient depletion can be found in intracellular pH shifts. In the budding yeast *Saccharomyces cerevisiae*, starvation conditions - specifically a lack of glucose – can lead to a rapid pH drop in the cytosol. The underlying mechanism is thought to be related to the inability of vacuolar and plasma membrane ATPases to properly maintain intracellular pH in yeast's preferred acidic environment (Diakov et al., 2013; Martínez-Muñoz and Kane, 2008; Orij et al., 2009, 2011). This acidification process occurs on the order of seconds to minutes, and upon re-addition of glucose the pH quickly returns to a physiological level. However, the effects of this pH drop on cytosolic proteins and their interactions is an open question.

Because this drop in pH subsequently changes the local environment of cytosolic proteins, it also represents a chance to examine how collective protein behavior might have evolved under this different intracellular context. Acidification can sometimes be an essential part of a molecular process, as is the case with spider silk formation (Askarieh et al., 2010) and brine shrimp's entry into a dormant state (Busa and Crowe, 1983). A shift in the pH of the cellular cytosol could also have a profound effect on the proteome as a whole, since shifts in pH cause a change in the charge profile of individual proteins. Recent work has shown that the pH shift associated with yeast starvation causes the protein glutamine synthetase to form higher-order structures, which could represent an emergent collective behavior than can be examined in the context of proteome-scale evolution (Petrovska et al., 2014). Furthermore, starvation-induced changes have been linked to protein assemblies like puncta structures in a wide range of other yeast cytosolic proteins, characterized by a protein transitioning from being diffuse in the cytosol to forming tightly colocalized clusters that can be detected as a small bright spot using fluorescence microscopy and are characterized by their quick reversibility upon glucose readdition (Narayanaswamy et al., 2009; O'Connell et al., 2014; Peters et al., 2013; Petrovska et al., 2014; Sagot et al., 2006; Suresh et al., 2015). This formation might help protect proteins from adverse structural changes during quiescence, which is the non-growing state cells can enter during suboptimal conditions (Laporte et al., 2011; Peters et al., 2013).

A link between glucose starvation, cytosolic acidification, and higher-order protein structure formation in budding yeast has been implicated in several studies, particularly filament formation of glutamine synthetase (Petrovska et al., 2014) and the clustering of proteasomes into 'proteasome storage granules' under low pH conditions (Peters et al., 2013). Since acidification can change the protonation state of multiple amino acids within the relevant pH range, one possible hypothesis is that acidification can cause proteins to become more electrostatically attractive to each other. In other words, the pH shift – traditionally thought to be only a by-product of the decreased ATP availability necessary for maintaining intracellular pH at homeostatic levels – may also act as a modulator that can regulate collective protein behavior on a global scale as a direct response to starvation stress.

However, this observation leads to another question: can collective protein behavior be beneficial to cellular survival? Under starvation conditions, cells must optimize their overall energy usage yet be available to re-enter the cell cycle quickly when conditions improve. These constraints necessitate shutting down flux through many metabolic pathways; when the cell is in a quiescent state, cellular growth and division are no longer prioritized. Alternatively, colocalizing proteins that mediate pathways critical for cellular survival (but not growth) could yield more efficient pathway organization (Castellana et al., 2014). Since harsh conditions can also cause proteins to unfold, which leads to cytotoxic and irreversible aggregates, protecting proteins by temporarily storing them in reversible puncta might also confer a selective advantage. Additionally, this quick puncta dissolution could also allow cells to exit quiescence rapidly and prepare them to resume division, as seen in the case of actin bodies that act as a reserve of actin molecules during quiescence and enable rapid cell-cycle re-entry (Sagot et al., 2006).

Analysis of the functional benefits of this putative large-scale protein coordination also provides a framework for examining a broader question: proteomic-scale evolution. If these puncta or other higher-order assemblies actually enable cells to better survive starvation conditions, then this "by-product" pH shift might have acted as a selection pressure over an evolutionary timescale that enabled budding yeast to take advantage of an extracellular-mediated process. Puncta formation is also associated with metabolic proteins, opening the idea that we may be observing a selection pressure that acts on the entire protein interaction network instead of targeting one particular protein or pathway subset. Studying how different species have evolved to survive and thrive in adverse conditions requires looking for not only individual protein evolutionary trajectories but also relatively obscure signals hidden across many proteins – something that is difficult to test experimentally but can be an ideal candidate for computational approaches. Here, we establish a mechanism by which pH can act as a tunable and readily reversible knob in the cell that can modulate protein interactions like puncta formation on the scale of the entire proteome. Furthermore, we present these results as an example of selection pressures acting on the proteome as a unit instead of at the level of an individual protein, illustrating how the yeast protein interaction network has evolved to enable cells to survive prolonged starvation.

Results

When the pH of a protein's environment changes, amino acid side chains have the potential to change protonation state and thus change the overall charge of the protein. Since the acidification associated with glucose starvation occurs on a short timescale, many yeast cytosolic proteins have the potential to change their net charge as a result. From an electrostatics perspective, protein interactions also have the potential to be affected; if one protein remains negatively charged under acidic conditions while an interaction partner goes from negatively to positively charged, these two proteins might display an increased mutual attraction and therefore a different interaction pattern under starvation stress. This large-scale variability in protein characteristics in response to changing cytosolic pH provides an opportunity to examine how the full set of protein interactions can be modulated under stress conditions.

Mathematical model of structure formation using electrostatic attraction

To explore the idea that the pH drop can cause proteins to co-assemble based on electrostatic attraction, the predicted charge of each cytosolic yeast protein can be calculated given amino acid side chain pKa values. These predictions have been used previously to understand distributions of isoelectric points (the pH at which a protein exhibits no net charge) in various organisms (Knight et al., 2004). Based on the Henderson-Hasselbalch relation, the net charge of a given protein from its residues can be given by

$$q_{protein} = \sum_{residue=\{R,H,K,N-terminus\}}^{Sequence} \frac{1}{1+10^{pH-pKa_{residue}}} - \sum_{residue=\{Y,C,D,E,C-terminus\}}^{Sequence} \frac{10^{pH-pKa_{residue}}}{1+10^{pH-pKa_{residue}}}$$

where the first summation on the right side of the equation gives a partial charge from the positively-charged arginines, histidines, lysines, and the N-terminus while the second summation gives the partial charge contributed by the negative tyrosines, cysteines, aspartic acids, glutamic acids, and the C-terminus given the pKa values of each residue and the pH of the environment (Cameselle et al., 1986).

Considering just electrostatic interactions, proteins may be induced to come closer together based on two overlapping ideas: decreasing electrostatic repulsion of a protein pair and increasing electrostatic attraction. For example, two proteins that interact under normal pH conditions might become more electrostatically attractive to each other as the pH drops and therefore associate more closely (Fig. 1a). Protein pairs that are already colocalized, like those known to interact with each other, provides a way to test whether proteins that are already spatially correlated with each other preferentially undergo an increase in electrostatic attraction to form higher-order structures.

To test this idea that pH can cause interacting protein pairs to form higher-order structures, we also need a metric that can stratify interactions based on their ability to become more attractive at a lower pH. To keep from introducing bias based on longer sequences having a propensity for larger charge values, protein charges are first normalized by the number of amino acids in their sequences. Multiplying the charge (*q*) per number of amino acids (*l*) of the *i*th protein with the *j*th protein gives the following metric:

$$E_{\{i,j\},pH} = \frac{q_{i,pH}q_{j,pH}}{l_i l_j}$$

This heuristic will be more negative when two proteins are predicted to be more electrostatically attractive at a given pH, and more positive when the two proteins are more likely to be repulsive (Fig. 1b). This model is simplistic in that it doesn't account for localized charge density, ion concentration in the cytosol, or other environmental considerations beyond predicted net charge of the proteins themselves. However, looking at protein interaction pairs through this phenomenological framework allows us to examine many protein pairings quickly while helping us understand the degree to which simple charge differences can mediate larger-scale changes in protein interaction networks.

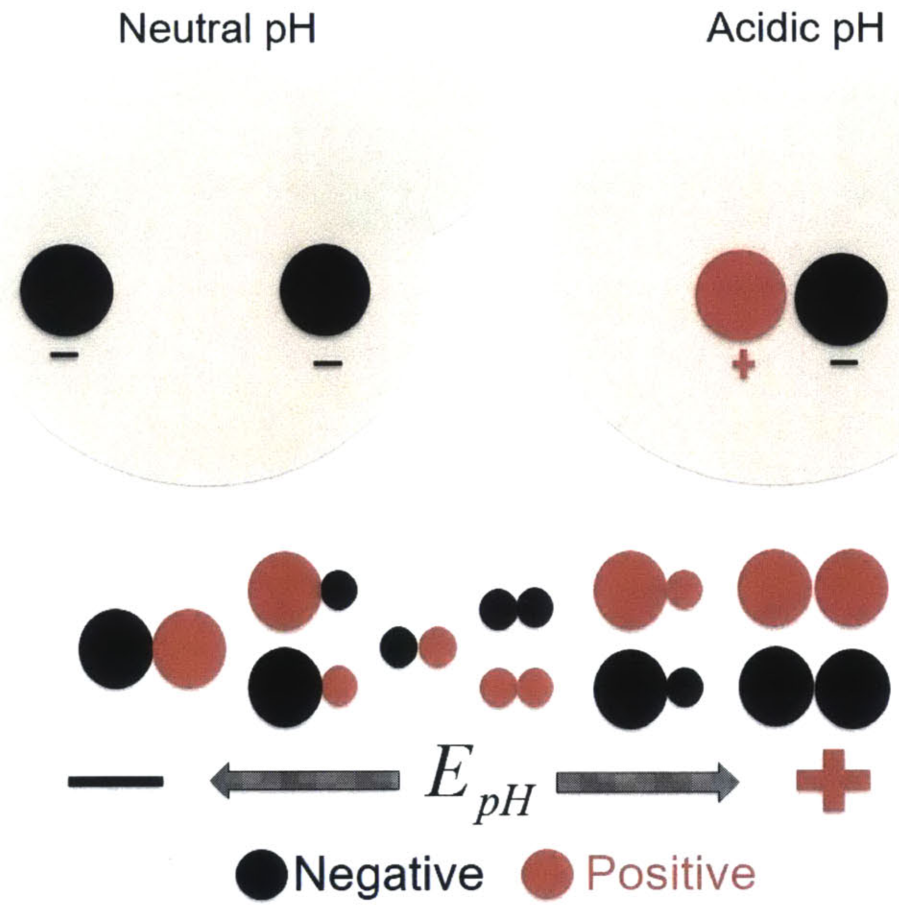


Figure 1. Changes in electrostatic charge mediate protein interactions

A change in pH can cause several amino acids either to gain or lose protons, which can cause a change in the net charge of the protein. If two proteins undergo charge shifts such that their electrostatic attractiveness increases with decreasing pH, then they could form higher-order structures like puncta. By using the heuristic described in Equation 2, we can stratify interactions based on their electrostatic attraction. Two proteins that are oppositely charged with large charge magnitudes would give a very negative E_{pH} score; alternatively, two proteins with the same charge at a given pH will display a positive E_{pH}.

Y2H interactions distinguishable using model

Protein pairs that interact under normal pH conditions may swing through shifts in charge that make them more likely to associate more closely, forming the basis for how puncta form under starvation conditions. To test this idea, we took the yeast interactome determined from the union of multiple yeast two-hybrid (Y2H) screenings that together contained ~500 cytosolic binary protein interactions (Yu et al., 2008). For each pair, we computed $E_{\{i,j\},pH}$ for pH 7 (approximately physiological) and for pH 5 (estimated cytosolic pH under starvation conditions). Summing these scores gives one number, $\sum E_{\{i,j\},pH}$, that will be negative if the interactome as a whole is predicted to have strengthened interactions on average.

If the set of yeast cytosolic proteins has undergone a collective selective pressure to avoid cellular death during starvation based on modulating charged interactions, then the group of actual interactions is likely to be distinct from random pairings of proteins that do not necessarily interact *in vivo*. 10,000 random interactomes were created, each of which drew from the same group of proteins as the Y2H set and included the same number of protein pairs. Importantly, restricting the random interaction sets to actual yeast sequence information allows us to examine the collective behavior of real protein interactions instead of identifying generalized bias in sequence composition. The comparison of this distribution of random sets to the real interactome is shown in Fig. 2. Under neutral pH, the real interactome is more negative than 99.4% of the random interactomes, indicating that under normal conditions electrostatic attraction can help mediate protein-protein interactions. This expected result gives us more confidence that our simple heuristic can isolate biophysical characteristics from large protein datasets. When the pH drops to 5, the distribution of random interactomes becomes less attractive on average; however, the real set of Y2H protein pairings remains at the same level.

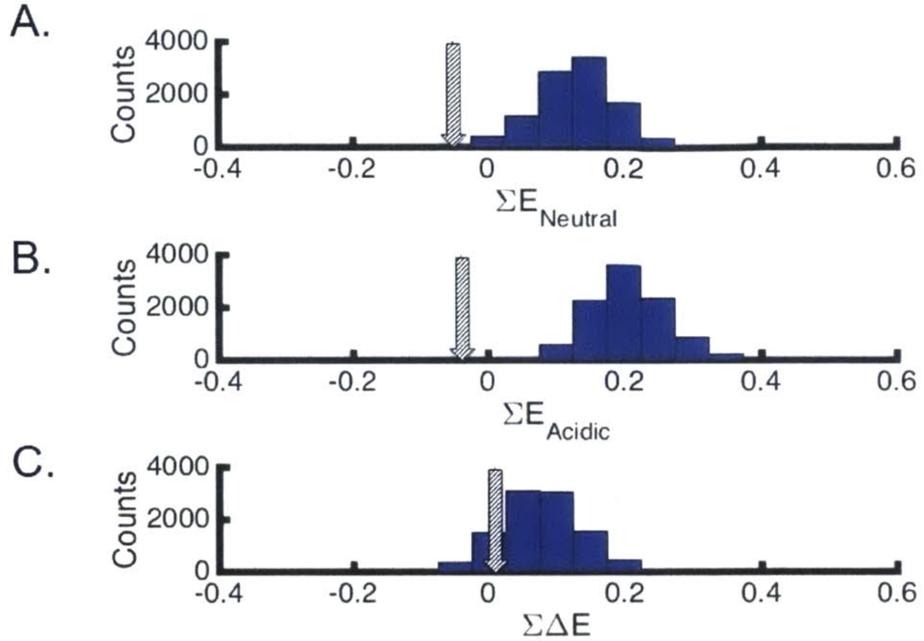


Figure 2. Real interactions predicted to be more attractive than random interactions under both neutral and acidic conditions.

In all panels, the black striped arrow indicates the score for the real set of Y2H interactions. A) Histogram of the scores of randomized interaction sets at neutral pH. Real interactions are more attractive on average than random ones. B) Randomized scores at acidic pH (pH = 5). Most random interactions become less attractive on average, but the real set of interactions remains roughly the same as at pH 7. C) Histogram of the score at pH 5 minus the score at pH 7 for randomized interactions.

This effect, where real interactions appear to be 'buffered' against changes in pH, was surprising – our initial prediction was that the real interactions would have a more negative score under acidic conditions, corresponding to becoming tightly colocalized enough to form higher-order structures. While the real interaction set is more attractive than all 10,000 random interaction sets tested under acidic conditions, it does not become more attractive under the shift in pH. Additionally, these results were unchanged when the underlying topology – namely, the degree of each protein node in the interaction network – was kept consistent between the real and random interactomes (results not shown). These findings suggest that the budding yeast interaction network displays a non-random pattern under pH stress, where an average real interaction is more likely to stay at the same level of electrostatic attraction under cytosolic acidification compared to random pairs of proteins.

In the context of understanding puncta formation, one explanation might be that proteins with different roles in the cell might also have different abilities to form higher-order structures. Sequestering proteins in puncta and therefore decreasing their function might prove advantageous under starvation conditions, particularly if those proteins are involved in cell cycle and division processes that are detrimental when energy sources are scarce. Therefore, determining whether S. cerevisiae sequences are distinct from other species can help distinguish whether the interaction network has undergone selective pressure to exhibit collective behavior.

Large-scale functional control represents a selective pressure

If this pH drop has provided a selection pressure on proteins in yeast to display the experimentally observed clustering behavior, then we can examine similar, homologous protein sequences in different organisms that do not display pH-shifting behavior. If their protein interaction pairs are less likely to be electrostatically attractive at lower pH, then this finding would provide evidence that the pH shift has acted as a selection pressure on the yeast proteome.

We identified homologous sequences between the yeast Y2H set of interacting protein pairs and the proteomes of different organisms, ranging from close unicellular relatives like *S. pombe* to more diverged species like humans using the Ensembl Biomart database (Cunningham et al., 2015). An interaction was considered to be homologous if an ortholog for both partners in each Y2H interacting set were also found in the cross species, and an interaction score averaged over number of homologous interactions was calculated for each cross species. Fig. 3a shows the average interaction score for the subset of yeast interactions that exist in each organism. Only two organisms, *S. cerevisiae* and the blind cave fish *A. mexicanus*, are predicted to have attractive interactions under both neutral and acidic conditions. Since the blind cave fish lives under extreme conditions anyway – typically cold and dark caves – we believe that this organism has unusual pH-dependent properties at baseline.

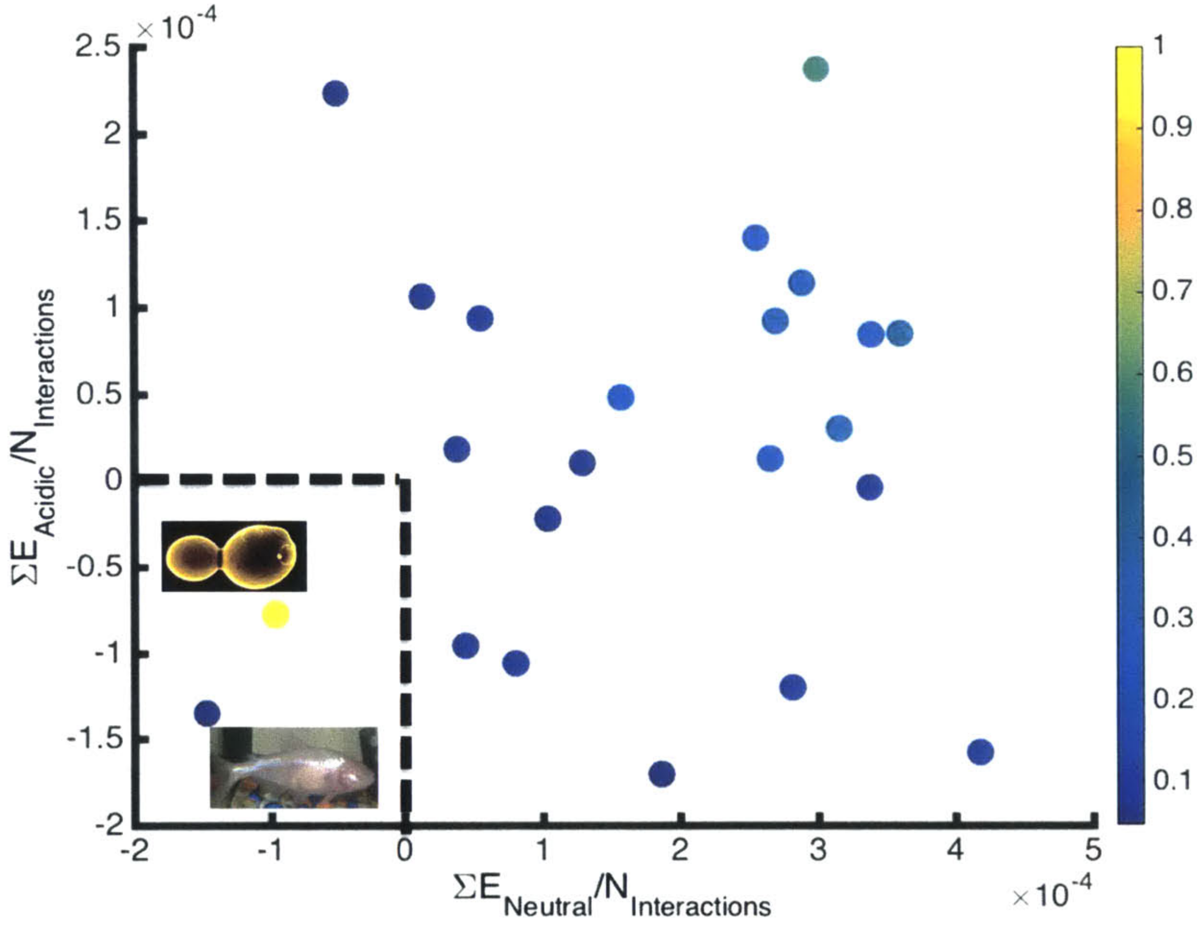


Figure 3a. S. cerevisiae proteins distinguishable from other organisms by their predicted response to different pH values.

Each data point represents the set of budding yeast sequences that have direct homologs in a different species – the list is given in the Methods section. The colorbar represents the fraction of the budding yeast sequences that have homologs for that organism. Dotted lines demarcate attractive versus non-attractive for interactions at neutral and acidic pH. The only organisms that fall within this attractive region for both pH values are S. cerevisiae and A. mexicanus, the blind Mexican cave fish. (Photo of S. cerevisiae: Ppdictionary.com/mycology. Photo of A. mexicanus: CC License JohnstonDJ)

Each average interaction score was computed for sequences directly from the comparison organism as well, as shown in Fig. 3b. Only baker's yeast sequences display attraction, characterized by having an interaction score that is negative, under both neutral and acidic regimes. These results indicate that S. cerevisiae sequences have preferentially evolved to display properties that are characteristic of large scale puncta formation under low pH. This finding becomes more apparent when we directly compare yeast sequences to those of other species, as shown in Fig. 3c. For all organisms studied, yeast sequences are either more attractive at acidic pH or display very little difference when compared to their cross-species homologous sequences. Together, these results suggest that the S. cerevisiae protein interaction network has evolved to take advantage of the large-scale acidification that occurs during glucose starvation.

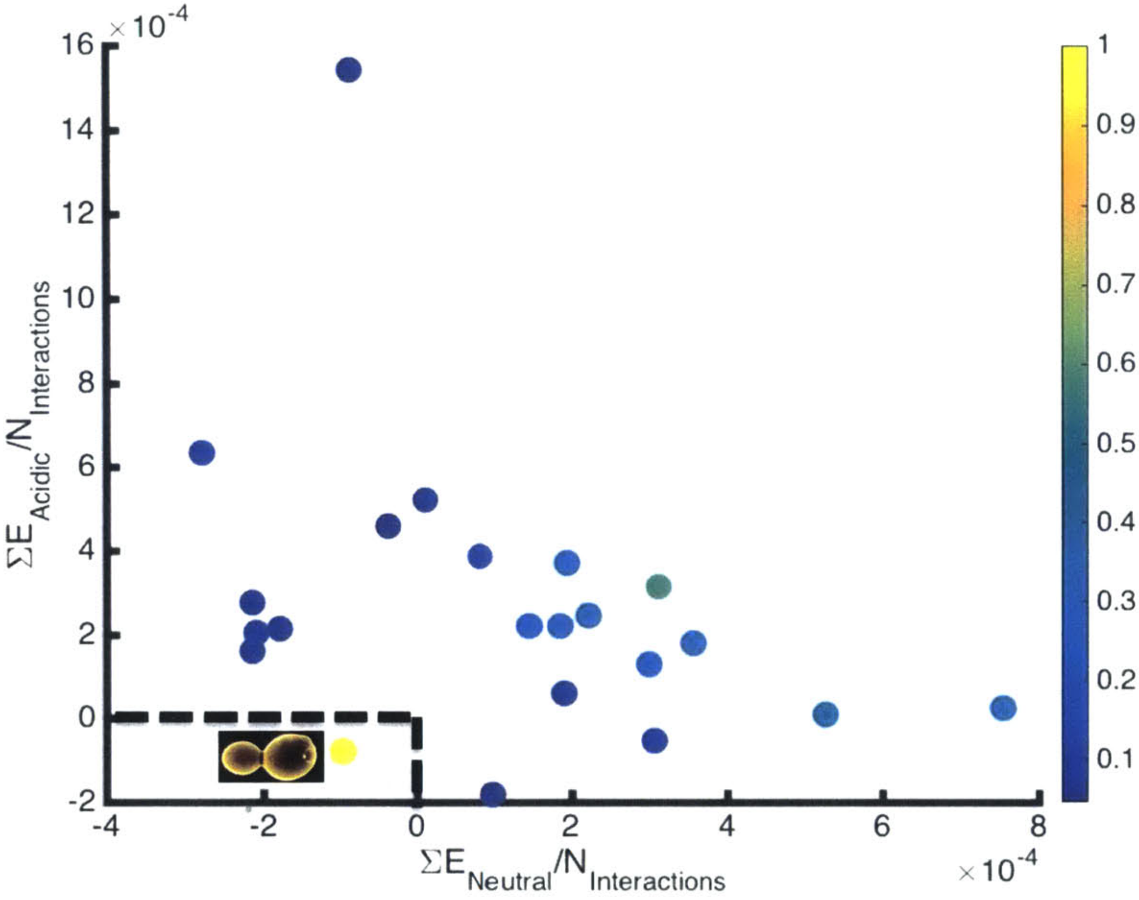


Figure 3b. S. cerevisiae proteins distinguishable from other organisms by their predicted response to different pH values

Colors, dotted lines, and axes are the same as in Fig. 3a. However, scores were calculated using the homologous sequences found in the cross-species directly. S. cerevisiae is the only organism whose protein sequences display on average a pattern of being electrostatically attractive under both neutral and acidic pH conditions.

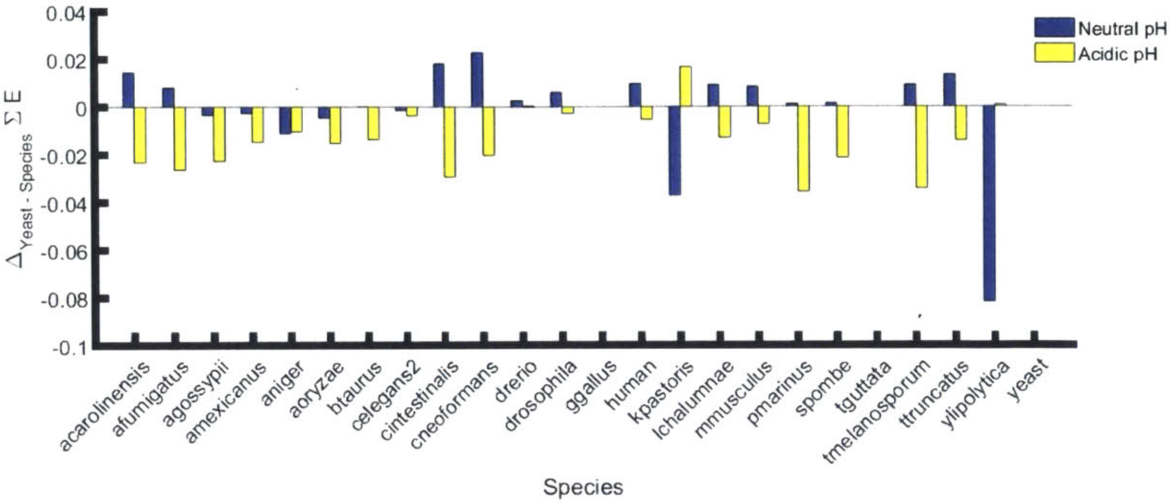


Figure 3c. S. cerevisiae proteins distinguishable from other organisms by their predicted response to different pH values

Each bar represents the difference in average interaction score between budding yeast sequences and the sequences of other species, listed on the independent axis, with calculations done for both neutral (blue) and acidic (yellow) conditions. Although the pattern varies at neutral pH, for almost all organisms budding yeast sequences tend to display more attractiveness at acidic pH than their cross-species sequence counterparts, shown by most yellow bars being negative.

To provide evidence that this difference in the yeast interaction network has the functional advantage required for selection, we then decided to examine which protein interaction pairs in S. cerevisiae were more likely to become more attractive at low pH. One such candidate group is the set of essential proteins, which are characterized by their knockouts being unable to form colonies when plated. Because they are crucial for ensuring yeast division, essential proteins might therefore be expected to form non-functional structures preferentially under quiescence when compared to non-essential proteins.

To better understand which proteins might cluster to decrease their functional activity and which might rely on continuing normal operations, the group of Y2H interactions was split into three groups: interactions where both proteins are considered essential, interactions where only one protein is considered essential, and interactions where neither protein is essential. $E_{\{i,j\},pH}$ was calculated for all interactions, and histograms of the three different classes were made for neutral and acidic pH values as shown in Fig. 4. When the cytosolic pH is roughly physiological, interactions containing essential proteins were virtually indistinguishable from those containing two non-essential proteins (p = 0.4657, Kolmogorov-Smirnov test). However, under acidic conditions interactions containing either one or two essential proteins shift from having a positive (less attractive) to negative (more attractive) median score when compared to physiological pH. In comparison, the median of the class of interactions containing no essential proteins remains greater than zero, indicating that these interactions are not as likely to become more attractive under our prediction scheme (p = 0.009, Kolmogorov-Smirnov test between two-essential and no-essential histograms). Furthermore, more interactions with either one or both proteins being essential become more attractive when they go from a neutral to an acidic environment compared to interactions where neither protein is essential; 58.8% of interactions with 2 essential proteins, 54.2% of interactions with 1 essential protein, and 37.7% of interactions where neither protein is essential have lower scores at pH 5 than pH 7.

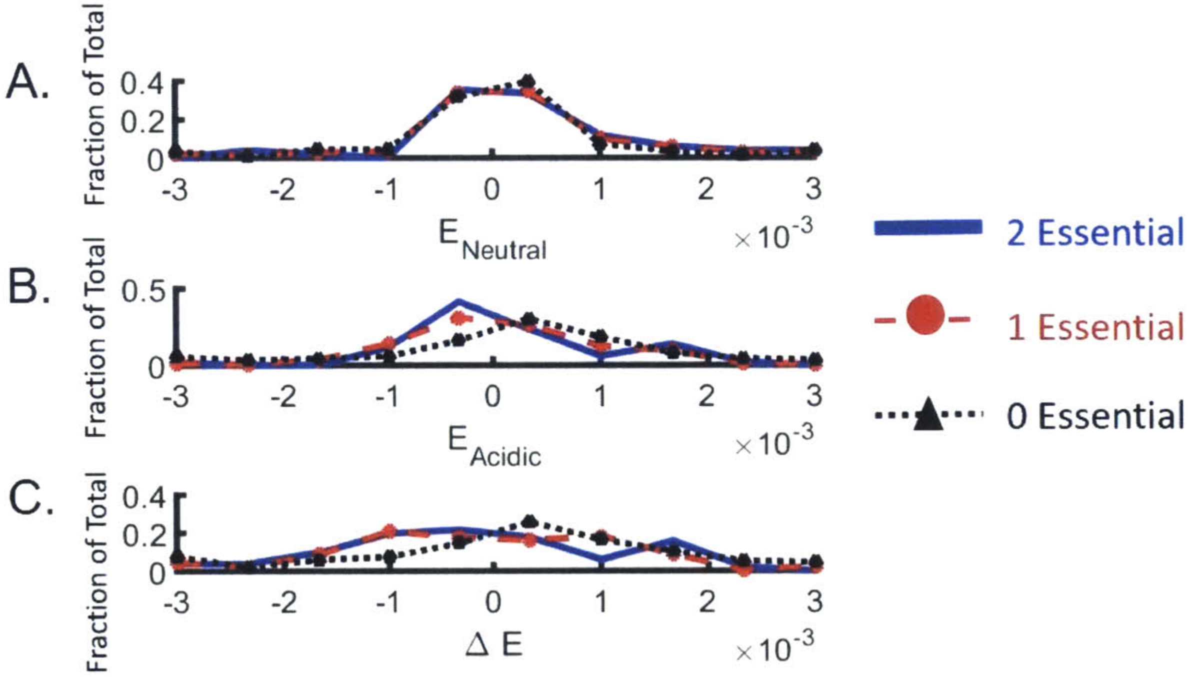


Figure 4. Essential proteins more likely to strengthen Y2H interactions at low pH. All Y2H interactions were classified based on whether they consisted of two essential proteins (blue line), exactly one essential and one non-essential protein (red dotted line), or two non-essential proteins (black triangles). Histograms for each classification are shown for: A) neutral pH, B) acidic pH, and C) the change in score from pH 7 to pH 5. Interactions that contain at least one essential protein are more likely to become more electrostatically attractive under acidic conditions.

Functional relevance of high-scoring proteins

These results indicate that essential proteins are more likely to tightly colocalize on an electrostatic basis than non-essential proteins, which provides evidence that puncta formation can shut off different processes dependent on their functional role in the cell. To further explore the relationship between function and puncta formation, we enumerated all possible triplets of cytosolic proteins included in the Y2H data set. Instead of filtering through the lens of pre-determined interactions, each protein was assumed to interact with the remaining two proteins in the triplet. Scores for all three interacting pairs within a triplet were calculated and summed for pH 5 and 7, and the groupings were ranked based on going through the largest change to become more attractive under the acidic pH. The top scoring group is shown in Table 1.

Protein Name	Description
YMR227C	TFIID subunit; involved in RNA polymerase II transcription initiation
YLR435W	Protein with a potential role in pre-rRNA processing
YEL012W	Ubiquitin-conjugating enzyme that regulates gluconeogenesis

Together, YMR227C and YLR435W appear as a pair in eight of the top-scoring triplets. Functionally, these two proteins are relevant in the context of known phenotypic changes under starvation conditions. When budding yeast cells are deprived of glucose, gene expression is altered – although some proteins like those involved in aspects of stress response become more abundant, both transcription and protein synthesis decrease overall (Ashe et al., 2000; Gasch et al., 2000; Gray et al., 2004; Klosinska et al., 2011). One class of proteins that is heavily downregulated during carbon starvation is the group of proteins associated with ribosome biogenenesis, ribosome processing, and translation, with transcription of these genes increasing within minutes of glucose readdition (Kresnowati et al., 2006). YMR227C, the yeast gene TAF7, is part of the TFIID complex that is thought to function as a regulatory checkpoint responsible for repressing premature transcription (Gegonne et al., 2006). It is also required for transcription of up to 24% of the *S. cerevisiae* proteome (Shen, 2003). Remarkably, translation-associated genes are statistically overrepresented in the set of genes that YMR227C modulates (p < 10^-4 using PANTHER), indicating that abrogating its function would directly lead to downregulation of a large group of proteins involved in translation (Mi et al., 2013). Together with YLR435W, a protein involved in ribosomal biogenesis through 20S pre-rRNA processing (Peng et al., 2003), the increased predicted ability of them to cluster electrostatically could help explain part of the halt in transcription and ribosomal biogenesis associated with glucose deprivation.

This example, where higher order protein structure formation can control an array of cellular functions, indicates that targeting one protein through sequestration in puncta can also affect transcriptional control that occurs during glucose starvation. Pathways like the PKA and Snf1 systems become activated under starvation conditions and help to regulate these differential gene expression patterns (Broach, 2012). Recent work has suggested a link between the Snf1 pathway and higher order protein structure formation, specifically the protein Kog1 (YHR186C) that forms part of the target of rapamycin complex 1 (TORC1) complex in yeast (Hallett et al., 2015). In the absence of glucose, a Snfl protein phosphorylates Kog1, which causes it to break away from the TORC1 complex and form an inclusion – a puncta structure – that reverses upon glucose addition allowing Kog1 to rejoin the complex. This reversible, glucose-dependent assembly formation of Kog1 therefore helps regulate downstream targets of the TORC1 complex, which includes a variety of protein and nucleotide biosynthesis whose control is associated with yeast entering a quiescent state (Barbet et al., 1996). These experimental results, together with our observations of the proteins most likely to undergo a shift in attractiveness under starvation conditions, point to a larger role for pH in modulating protein interactions that could affect phenotypic gene expression levels in the absence of glucose.

The third protein in the triplet, YEL012W, exhibits functional relevance as well. This protein forms part of an ubiquitin-conjugating enzyme that downregulates the gluconeogenesis pathway in yeast (Schüle et al., 2000). Specifically, gluconeogenesis is a pathway that converts malate into glucose while S. cerevisiae are not in the presence of the fermentable carbon source glucose. Since this pathway is only activated when cells are either carbon-starved or only have access to non-fermentable carbon sources, sequestering downregulators byby electrostatic colocalization may allow the pathway to proceed during glucose starvation, which would constitute a direct example of a metabolic pathway that could be regulated by cytosolic pH state as a messenger for glucose presence.

To further explore this idea, we looked at the eight downregulators and two upregulators of gluconeogenesis in *S. cerevisiae*. If the downregulators are confined to puncta while the upregulators remain diffuse at an acidic pH, then this higher order protein structure formation could help explain how starvation-associated acidification can regulate an individual metabolic pathway. We generated random groups of eight cytosolic proteins, and calculated the sum of interaction scores at each possible pairing within this group at pH 7 and 5 for 10000 trials. We then compared those scores to that of the gluconeogenesis downregulators, and did the same for the pair of upregulators using randomized cytosolic protein pairs. Results are shown in Fig. 5. At neutral pH, gluconeogenesis downregulators are significantly less attractive than an average grouping of cytosolic proteins. However, when the pH drops, gluconeogenesis negative regulators are predicted to become much more likely to form electrostatically-mediated higher order structures according to our criteria, with only 0.23% of random groupings having a more attractive score. This group also undergoes one of the largest changes in interaction score compared to the control distribution. On the other hand, the upregulators retain an average score under both neutral and acidic conditions, indicating that their collective charge profile is much less affected by glucose-mediated acidification than the downregulators.

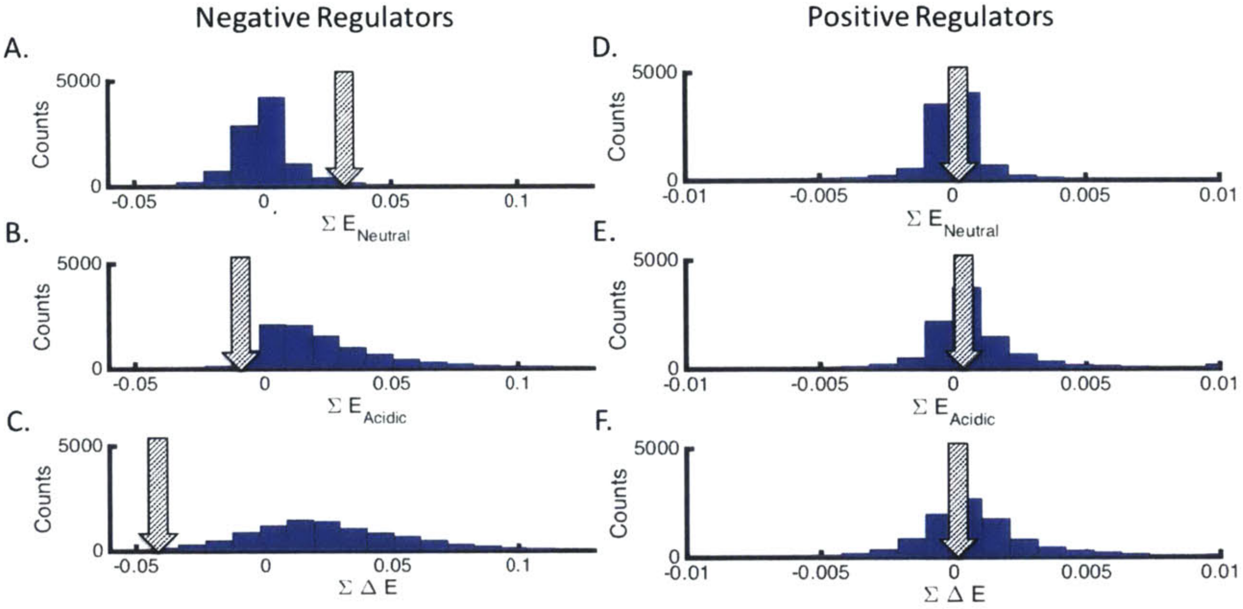


Figure 5. Down regulators of gluoconeogenesis predicted to form puncta under acidic conditions, while positive regulators are unaffected by pH.

Random groups of cytosolic proteins, chosen to match the number of down or up regulators of gluconeogenesis, were scored and the histograms for trials under different conditions are shown. In all cases, the black striped arrow shows the location of the real group of down or up regulators. Conditions are as follows: A) Neutral pH for real down regulators and random cytosolic protein groupings, B) acidic pH for real down regulators and random cytosolic proteins groupings, C) change in score from pH 7 to pH 5 for down regulators and randomized groupings, D) neutral pH for real up regulators and random groupings, E) acidic pH for real up regulators and randomized groupings, and F) the change in scores between neutral and acidic pH for up regulators and randomized groupings. While down regulators show a pattern of being more likely to become more attractive at low pH, the up regulators do not.

Entry into quiescence due to glucose starvation also affects many other pathways beyond transcriptional control of translational products and activation of gluconeogenesis pathways. Cell cycle control, translation, and large-scale metabolic responses all undergo phenotypic changes in response to starvation. To tease out how starvation-induced protein collective behavior can control different pathways on a larger scale, we next used Metropolis-Hastings Monte Carlo sampling to generate a group of 30 cytosolic proteins that undergo the largest change in group interaction score that involves an increase in attractiveness at low pH compared to neutral pH. Each protein was then classified according to its Gene Ontology GO Slim term, and was grouped into one of four categories: cell cycle, translation and RNA control, carbohydrate and energy metabolism, and other. The first three titles are functional classifications that tend to be suppressed during quiescence, which we predict are more likely to become co-insoluble as a means of controlling cellular processes on a larger scale (Gray et al., 2004). A comparison of our optimized group of proteins to the full cytosolic set is shown in Fig. 6. Together, cell cycle, translation, and metabolic control account for 69% of the selected group of proteins. In contrast, these same groups only account for 29% of the full group of cytosolic proteins.

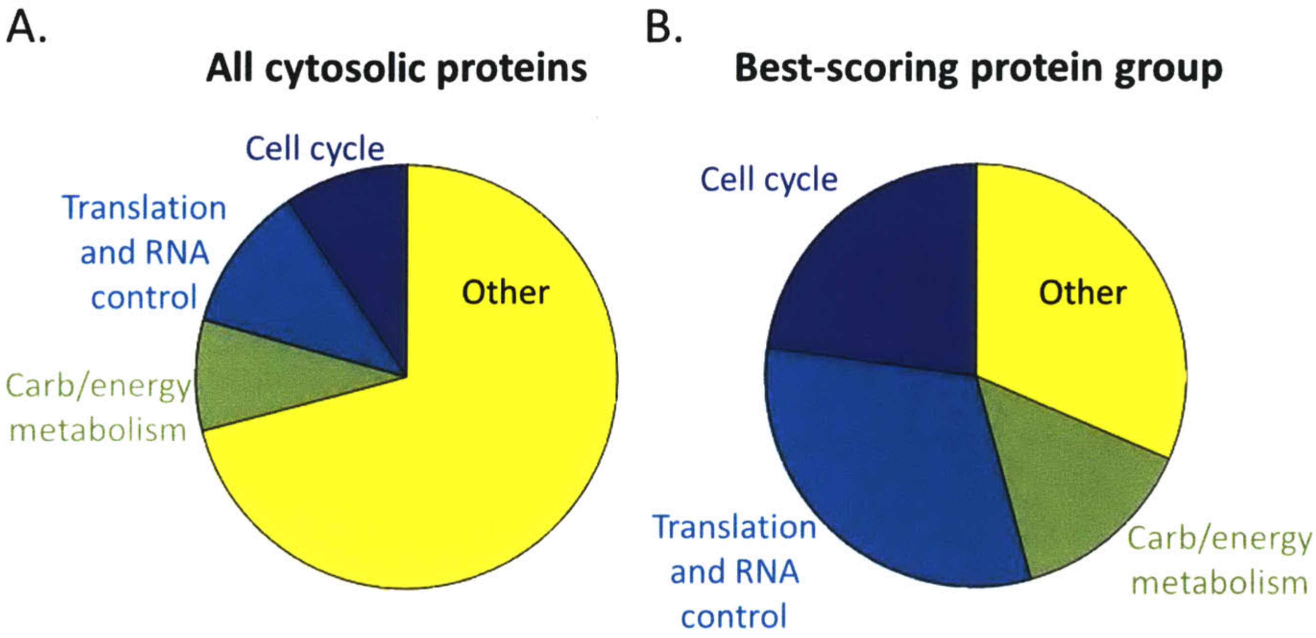


Figure 6. Cell cycle, translation, and metabolic proteins dominate group selected for undergoing the largest downwards shift in total interaction score.

A) The fraction of all cytosolic proteins that can be classified as being involved in cell cycle (purple), translation and RNA control (light blue), carbohydrate and energy metabolism (green), and none of the above (yellow). The first three categories are typically decreased during glucose starvation in S. cerevisiae. B) Fraction of proteins that fall into the same categories for the optimized group of 30 proteins that undergo the largest shift to become more attractive at acidic pH.

These potential functional benefits from controlling metabolic pathways, including gluconeogenesis and ribosome biogenesis, on a large scale indicate that the yeast proteome evolved to take advantage of the glucose deprivation pH drop. By clustering together due to increased electrostatic attraction, these proteins could be protected from detrimental non-reversible aggregation and premature degradation. The reversible nature of the clustering also means that when the cellular pH returns to baseline levels upon nutrient availability, cytosolic proteins can reactivate quickly and help bring the cell out of a quiescent state by returning to normal function.

Discussion

This work highlights the importance of understanding intracellular spatial heterogeneity in response to changing environmental conditions, and further suggests that understanding how yeast form puncta structures in response to starvation can help us understand the physical forces that drive protein colocalization in cells. Previous studies have shown that maximum metabolic efficiency is linked to protein cytosolic distribution, with different enzymes in a pathway colocalized in a common volume to improve downstream processing (Castellana et al., 2014). These metabolic deposits might lead to different collective behavior as interactions between different colocalized proteins become stronger and repulsion between like protein sequences becomes weaker, which corresponds in our electrostatic framework to protein charge becoming closer to neutral as the pH drops. From a functional standpoint, metabolic processes like ribosome biosynthesis might become less crucial for a cell entering a starvation-induced quiescent state where protein transcription rate decreases. The same may also be true for production of nucleotides; prior research indicates that flux through the *de novo* nucleotide biosynthetic pathway sharply decreases upon removal of carbon-source nutrients in yeast (Brauer et al., 2006). Therefore, proteins that are part of this pathway may form structures like puncta to halt unnecessary metabolite production.

Here, we have examined whether the drop in pH associated with glucose depletion in yeast can act as a large-scale modulator that controls the collective behavior of cytosolic proteins. Our heuristic for determining degree of attraction between two proteins has been able to distinguish real, Y2H-confirmed interactions from random pairings of proteins. Furthermore, we have determined that the proteins that are most likely to assemble according to our criteria correspond to pathways that control growth, cell cycle and division, translation, and metabolism, which also represent functions that tend to be downregulated during starvation conditions. Two proteins that we predict are likely to become more "sticky" and help mediate puncta formation are also associated with transcriptional and post-transcriptional control of translation and ribosome biogenesis, which have been experimentally determined to decrease substantially in the absence of glucose in *S. cerevisiae*.

In addition to sequestering proteins that would then shut off different functional processes, higher-order structure formation might also form protective storage depots of cytosolic proteins. The cell expends energy during translation to produce these proteins, so storing them reversibly in anticipation of improved glucose availability can help the cell survive and re-enter the cell cycle. In addition to keeping the proteins from being degraded, puncta formation can also help protect them from misfolding events which, on a large scale, can be lethal to the cell. Storage depots have been suggested in the context of proteasome storage granules, which are also reversible assemblies composed of the large complex of the proteasome degradation system in yeast (Peters et al., 2013). In this case, the proteasome is a large structure that would be costly to degrade and is crucial for cellular survival under growth conditions.

Similarly, the GO group that underwent the largest shift to become more attractive at pH 5 on average was categorized by "proteolysis involved in cellular protein catabolic process," which includes units of the yeast proteasome.

These functional benefits – large-scale and rapid pathway control and reversible storage – point to the presence of a selective pressure that has acted on a large subset of cytosolic proteins. *S. cerevisiae* are distinct in their ability to survive prolonged periods of starvation, and the associated cytosolic pH shift and puncta formation across large subsets of the proteome have only been observed in this species and, conditionally, *D. discoideum* (Petrovska et al., 2014). This situation provides an example of putative evolutionary selection that has shaped an entire protein interaction network. Traditionally, observing such widespread selection has been challenging since it necessitates looking for a small signal spread across a large set of proteins; instead, prior work has tended to focus on understanding the evolution of individual protein sequences or the general topology of protein-protein interaction networks. This computational study could provide evidence that the drop in pH has acted as a selection pressure on the yeast proteome as a whole, with the significance of small-scale sequence alterations only becoming apparent when examining large-scale, cellular-level processes. This notion that evolution can act on the level of the proteome to modulate a collective property of the cytosol, in addition to the selective pressures that target individual genes and proteins, can provide new insight into evolutionary processes.

Materials and Methods

Calculating pH-based protein charges

Yeast cytosolic proteins were identified by comparison against the YeastGFP Fusion Localization database (Ghaemmaghami et al., 2003; Howson et al., 2005; Huh et al., 2003). Proteins were included in the set if they were categorized as 'cytosolic', including proteins that were tagged in multiple subcellular localizations. Protein charges were calculated according to the Henderson-Hasselbalch equation as described in Eq. 1 in the Results section. Amino acid pKa values used are as listed: D (3.9), E (4.3), R (12.0), K (10.5), H (6.08), C (8.28), Y (10.1) (Gunner et al., 2006). All calculations were done using Matlab.

Calculating Y2H interactions

Yeast two-hybrid data were derived from a collective dataset (Yu et al., 2008). Entries were screened to remove self-self entries, leaving 535 interactions composed of 539 unique proteins. For randomized interaction sets, 535 pairs were drawn with replacement from the set of 539 pairs were used for each trial. The figures show histograms with 10,000 trials. For comparison of interactions with cross-species sequences, homologs of each yeast cytosolic protein were found using BioMart with the Ensembl database release 77 (Cunningham et al., 2015; Herrero et al., 2016). The list of organisms used for the cross-species comparisons are listed below, along with their percentage of homologous proteins present:

A. carolinensis	0.204082
A. fumigatus	0.506494
A. gossypii	0.71243

A. mexicanus	0.170686
A. niger	0.486085
A. oryzae	0.458256
B. taurus	0.19295
C. elegans	0.233766
C. intestinalis	0.246753
C. neoformans	0.458256
D. rerio	0.187384
D. melanogaster	0.2282
G. gallus	0.198516
H. sapiens	0.204082
K. pastoris	0.61039
L. chalumnae	0.200371
M. musculus	0.200371
P. marinus	0.152134
S. pombe	0.445269
T. guttata	0.183673
T. melanosporum	0.486085
T. truncatus	0.220779
Y. lipolytica	0.55102
Yeast	1

Monte Carlo simulations

To generate the set of 30 proteins used for functional classification, Monte Carlo sampling was implemented with a probability of accepting a randomly substituted protein into the group proportional to a Boltzmann distribution. Each interaction score was calculated as the sum of all possible pairwise interactions within the current group at both pH 5 and 7, and the energy analog in the Boltzmann distribution was taken to be $\Sigma_{i,j}\Delta_{pH\ 5-pH\ 7}E_{i,j}$ for the ith and jth proteins within the current set of 30. Trials were run for >100000 generations and checked for reasonable convergence of the difference in summed interaction scores at pH 5 and pH 7.

Classifying protein functional classes

For functional classifications, proteins were first categorized by their Gene Ontology (GO) Slim term (Ashburner et al., 2000; Gene Ontology Consortium, 2015). Proteins that were associated with the following labels were classified under the Cell Cycle and Division category: 'mitotic cell cycle', 'meiotic cell cycle', and 'regulation of cell cycle'. The grouping 'ribosomal small subunit biogenesis', 'cytoplasmic translation', and 'RNA modification' was categorized under translational control. Finally, 'carbohydrate metabolic process' and 'generation of precursor metabolites and energy' protein labels were grouped under the heading of carbohydrate and energy metabolism.

References

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. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Ashe, M.P., Long, S.K.D., and Sachs, A.B. (2000). Glucose Depletion Rapidly Inhibits Translation Initiation in Yeast. Mol. Biol. Cell 11, 833–848.

Askarieh, G., Hedhammar, M., Nordling, K., Saenz, A., Casals, C., Rising, A., Johansson, J., and Knight, S.D. (2010). Self-assembly of spider silk proteins is controlled by a pH-sensitive relay. Nature 465, 236–238.

Barbet, N.C., Schneider, U., Helliwell, S.B., Stansfield, I., Tuite, M.F., and Hall, M.N. (1996). TOR controls translation initiation and early G1 progression in yeast. Mol. Biol. Cell 7, 25–42.

Brauer, M.J., Yuan, J., Bennett, B.D., Lu, W., Kimball, E., Botstein, D., and Rabinowitz, J.D. (2006). Conservation of the metabolomic response to starvation across two divergent microbes. Proc. Natl. Acad. Sci. U. S. A. 103, 19302–19307.

Broach, J.R. (2012). Nutritional Control of Growth and Development in Yeast. Genetics 192, 73–105.

Busa, W.B., and Crowe, J.H. (1983). Intracellular pH Regulates Transitions Between Dormancy and Development of Brine Shrimp (Artemia salina) Embryos. Science 221, 366–368.

Cameselle, J.C., Ribeiro, J.M., and Sillero, A. (1986). Derivation and use of a formula to calculate the net charge of acid-base compounds. Its application to amino acids, proteins and nucleotides. Biochem. Educ. 14, 131–136.

Castellana, M., Wilson, M.Z., Xu, Y., Joshi, P., Cristea, I.M., Rabinowitz, J.D., Gitai, Z., and Wingreen, N.S. (2014). Enzyme clustering accelerates processing of intermediates through metabolic channeling. Nat. Biotechnol. 32, 1011–1018.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. 43, D662–D669.

Diakov, T.T., Tarsio, M., and Kane, P.M. (2013). Measurement of Vacuolar and Cytosolic pH In Vivo in Yeast Cell Suspensions. J. Vis. Exp. JoVE.

Di Nardo, A.A., Larson, S.M., and Davidson, A.R. (2003). The Relationship Between Conservation, Thermodynamic Stability, and Function in the SH3 Domain Hydrophobic Core. J. Mol. Biol. 333, 641–655.

Freeland, J.C., and Gale, E.F. (1947). The amino-acid composition of certain bacteria and yeasts. Biochem. J. 41, 135–138.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol. Biol. Cell 11, 4241–4257.

Gegonne, A., Weissman, J.D., Zhou, M., Brady, J.N., and Singer, D.S. (2006). TAF7: A possible transcription initiation check-point regulator. Proc. Natl. Acad. Sci. U. S. A. 103, 602–607.

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature 425, 737–741.

Gray, J.V., Petsko, G.A., Johnston, G.C., Ringe, D., Singer, R.A., and Werner-Washburne, M. (2004). "Sleeping Beauty": Quiescence in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. 68, 187–206.

Greaves, R.B., and Warwicker, J. (2007). Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. BMC Struct. Biol. 7, 18.

Gunner, M.R., Mao, J., Song, Y., and Kim, J. (2006). Factors influencing the energetics of electron and proton transfers in proteins. What can be learned from calculations. Biochim. Biophys. Acta BBA - Bioenerg. 1757, 942–968.

Hallett, J.E.H., Luo, X., and Capaldi, A.P. (2015). Snf1/AMPK promotes the formation of Kog1/Raptor-bodies to increase the activation threshold of TORC1 in budding yeast. eLife 4, e09181.

Hartl, F.U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. Nature 475, 324–332.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database 2016, bav096.

Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3.

Howson, R., Huh, W.-K., Ghaemmaghami, S., Falvo, J.V., Bower, K., Belle, A., Dephoure, N., Wykoff, D.D., Weissman, J.S., and O'Shea, E.K. (2005). Construction, verification and experimental use of two epitope-tagged collections of budding yeast strains. Comp. Funct. Genomics 6, 2–16.

Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature 425, 686–691.

Kim, Y.E., Hipp, M.S., Bracher, A., Hayer-Hartl, M., and Ulrich Hartl, F. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis. Annu. Rev. Biochem. 82, 323–355.

Klosinska, M.M., Crutchfield, C.A., Bradley, P.H., Rabinowitz, J.D., and Broach, J.R. (2011). Yeast cells can access distinct quiescent states. Genes Dev. 25, 336–349.

Knight, C.G., Kassen, R., Hebestreit, H., and Rainey, P.B. (2004). Global analysis of predicted proteomes: functional adaptation of physical properties. Proc. Natl. Acad. Sci. U. S. A. 101, 8390–8395.

Kresnowati, M.T.A.P., van Winden, W.A., Almering, M.J.H., ten Pierick, A., Ras, C., Knijnenburg, T.A., Daran-Lapujade, P., Pronk, J.T., Heijnen, J.J., and Daran, J.M. (2006). When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. Mol. Syst. Biol. 2.

Laporte, D., Lebaudy, A., Sahin, A., Pinson, B., Ceschin, J., Daignan-Fornier, B., and Sagot, I. (2011). Metabolic status rather than cell cycle signals control quiescence entry and exit. J. Cell Biol. 192, 949–957.

Martínez-Muñoz, G.A., and Kane, P. (2008). Vacuolar and Plasma Membrane Proton Pumps Collaborate to Achieve Cytosolic pH Homeostasis in Yeast. J. Biol. Chem. 283, 20309–20319.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 8, 1551–1566.

Mirny, L.A., and Shakhnovich, E.I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1. J. Mol. Biol. 291, 177–196.

Narayanaswamy, R., Levy, M., Tsechansky, M., Stovall, G.M., O'Connell, J.D., Mirrielees, J., Ellington, A.D., and Marcotte, E.M. (2009). Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. Proc. Natl. Acad. Sci. U. S. A. 106, 10147–10152.

O'Connell, J.D., Tsechansky, M., Royall, A., Boutz, D.R., Ellington, A.D., and Marcotte, E.M. (2014). A proteomic survey of widespread protein aggregation in yeast. Mol. Biosyst. 10, 851–861.

Orij, R., Postmus, J., Ter Beek, A., Brul, S., and Smits, G.J. (2009). In vivo measurement of cytosolic and mitochondrial pH using a pH-sensitive GFP derivative in Saccharomyces cerevisiae reveals a relation between intracellular pH and growth. Microbiology 155, 268–278.

Orij, R., Brul, S., and Smits, G.J. (2011). Intracellular pH is a tightly controlled signal in yeast. Biochim. Biophys. Acta BBA - Gen. Subj. 1810, 933–944.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 3, e02030.

Pe'er, I., Felder, C.E., Man, O., Silman, I., Sussman, J.L., and Beckmann, J.S. (2004). Proteomic signatures: Amino acid and oligopeptide compositions differentiate among phyla. Proteins Struct. Funct. Bioinforma. 54, 20–40.

Peng, W.-T., Robinson, M.D., Mnaimneh, S., Krogan, N.J., Cagney, G., Morris, Q., Davierwala, A.P., Grigull, J., Yang, X., Zhang, W., et al. (2003). A Panoramic View of Yeast Noncoding RNA Processing. Cell 113, 919–933.

Peters, L.Z., Hazan, R., Breker, M., Schuldiner, M., and Ben-Aroya, S. (2013). Formation and dissociation of proteasome storage granules are regulated by cytosolic pH. J. Cell Biol. 201, 663–671.

Petrovska, I., Nüske, E., Munder, M.C., Kulasegaran, G., Malinovska, L., Kroschwald, S., Richter, D., Fahmy, K., Gibson, K., Verbavatz, J.-M., et al. (2014). Filament formation by metabolic enzymes is a specific adaptation to an advanced state of cellular starvation. eLife 3.

Sagot, I., Pinson, B., Salin, B., and Daignan-Fornier, B. (2006). Actin bodies in yeast quiescent cells: an immediately available actin reserve? Mol. Biol. Cell 17, 4645–4655.

Schüle, T., Rose, M., Entian, K.D., Thumm, M., and Wolf, D.H. (2000). Ubc8p functions in catabolite degradation of fructose-1, 6-bisphosphatase in yeast. EMBO J. 19, 2161–2167.

Shen, W.-C. (2003). Systematic analysis of essential yeast TAFs in genome-wide transcription and preinitiation complex assembly. EMBO J. 22, 3395–3402.

Sikosek, T., and Chan, H.S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. J. R. Soc. Interface R. Soc. 11, 20140419.

Suresh, H.G., da Silveira Dos Santos, A.X., Kukulski, W., Tyedmers, J., Riezman, H., Bukau, B., and Mogk, A. (2015). Prolonged starvation drives reversible sequestration of lipid biosynthetic enzymes and organelle reorganization in Saccharomyces cerevisiae. Mol. Biol. Cell 26, 1601–1615.

Tekaia, F., Yeramian, E., and Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297, 51–60.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110.

Chapter 5

Discussion and epilogue

The field of protein biophysics has expanded dramatically over the past decade. Protein folding has emerged as a new horizon, fueled by the complexity involved in understanding the ensemble of structures that is more likely to characterize a sequence than one rigid conformation. New algorithms designed to tackle the protein folding problem have shown great promise in helping to elucidate both the biophysical and evolutionary parameters that explain how proteins function in the intracellular environment.

Understanding protein folding has become critically important in light of the wide variety of human diseases that have been determined to have their origin in protein misfolding including Alzheimer's disease, Parkinson's, Huntington's, amyotrophic lateral sclerosis, and Type II diabetes mellitus (Mukherjee et al., 2015; Mulligan and Chakrabartty, 2013, 2013). Prion diseases like the transmissible spongiform encephalopathies also share the hallmarks of misfolding; their infectiousness stems from the ability of a prionic, misfolded variant to induce native copies of the protein to adopt the prionic form and aggregate (Jackson and Krost, 2014). However, functional roles for prion proteins in yeast have been identified in connection with their ability to be inherited in their prionic form across generations (Halfmann et al., 2012; True and Lindquist, 2000). These findings illustrate the need for a better understanding of protein misfolding, aggregation, and sequence evolution.

Two original projects have been presented in this dissertation, both of which have been focused towards understanding how protein characteristics can change in the context of cellular stress. In the first paper, published in Structure in 2015 (Chapter 2), we explored how a marginally stable protein is targeted for degradation in the cell (Brock et al., 2015). The von Hippel-Lindau protein pVHL, used as a model misfolded protein, was mutated and transfected into a unicellular environment that lacks at least one crucial cofactor. One mutation was able to resist cellular degradation in this environment. Through a combination of computational and experimental techniques, we saw that the mutant form of pVHL was predicted to shift into an alternate conformation that was better able to bind to the homologous cofactor that was still present in yeast. This enhanced binding abrogated the necessity of the secondary cofactor that is only found in multicellular organisms. The conformational change also predicted that the mutant pVHL could also hide its chaperone target sites better than the wild-type, providing another method in which it could escape protein quality control recognition.

Together, these results provided insight into the factors that can modulate quality control fate for an example marginally stable protein. This paper also highlights the strength of adding computational analyses to experimental data, especially for proteins like pVHL that contain large disordered regions that are difficult to characterize using solely experimental methods. Since pVHL is a tumor suppressor protein, future directions could include expanding this analysis to include other proteins like p53 that are also involved in cancer, exhibit significant thermodynamic and kinetic instability, and contain large disordered regions. In particular, using the England burial trace model could help find new ways to exploit protein-protein interactions to stabilize protein variants that are connected with disease states in humans. Many point mutations tend to be destabilizing, and the frequency of mutations in tumor suppressor proteins in inherited cancers suggests that introducing a variant protein form that could help stabilize the native tumor suppressor could provide a viable avenue for therapeutics research (Sikosek and Chan, 2014).

The interaction of pVHL with its cofactors and chaperones also points to the necessity of understanding protein folding at a systems level. Since many functional forms of proteins bind either transiently or long-term, examining protein interactions can allow us to better understand functional pathways in the cell. Complex collective behavior of proteins has also been observed in the context of stresses like nutrient starvation (An et al., 2008; Narayanaswamy et al., 2009; Peters et al., 2013). The exceptionally crowded environment of the cytosol also provides another facet to understanding how proteins form both transient and longer-lived interactions in both functional and nonfunctional contexts (Luby-Phelps, 2000).

Understanding how proteins interact differently under both normal and abnormal conditions also provides a framework to probe protein coevolution. Evolutionary constraints on individual sequences to maintain thermodynamic properties, sequence composition, charge profiles, or abundances have been observed, along with constraints on the overall topology of interaction networks (Sikosek and Chan, 2014). The rapid advancements in sequencing technology have also fueled the expansion of the number of sequenced genomes. This availability in data has enabled us to probe the evolutionary history of many protein sequences. However, understanding how proteins evolve collectively through the cumulative effect of amino acid variants spread throughout the proteome remains an exciting challenge.

In Chapter 4, we attempt to answer how a proteome can evolve to exhibit collective cytosolic behavior as a response to nutrient starvation. In *S. cerevisiae*, the cytosol tends to acidify under glucose starvation, and this condition has also been connected to higher order protein assembly formation (Petrovska et al., 2014). We propose that this pH shift can cause many proteins to be inactivated simultaneously, which would regulate metabolism generally and could also provide a protective benefit. Modeling results show that protein-protein interactions retain their electrostatic attractiveness even when the pH drops to levels found in glucose starvation, while randomized interaction sets do not display this behavior. Furthermore, this behavior only seems to be found in *S. cerevisiae* when compared to other organisms that do not display the same response to nutrient depletion, indicating that a selective pressure could explain this collective response. Proteins that are more likely to tightly colocalize and form structures under this condition are also more likely to be involved in functional roles that are attenuated under starvation condition, including gluconeogenesis downregulators and metabolic and cell cycle inducing pathways. These results indicate that a collective property of the cytosol – namely, tightly colocalizing protein groups together – can serve to modulate metabolic response in general and represents a selective pressure that has affected the systems behavior of cytosolic proteins as a group.

These results exemplify a novel approach to understanding how evolution can act on a level that has been difficult to characterize previously. By looking at the collective behavior of cytosolic proteins under stress conditions, we were able to tease out a selective pressure that has acted in concert across a large swathe of the *S. cerevisiae* proteome. This new framework could be applied to look at broader contexts of stress response; however, the implications extend beyond the specific model and system used in this paper. By demonstrating that selective pressures can act at a systems level to shape collective behavior, these results highlight the complexities involved in protein evolution and the necessity of examining proteins from a systems perspective to gain new insights.

References

An, S., Kumar, R., Sheets, E.D., and Benkovic, S.J. (2008). Reversible Compartmentalization of de Novo Purine Biosynthetic Complexes in Living Cells. Science 320, 103–106.

Brock, K.P., Abraham, A., Amen, T., Kaganovich, D., and England, J.L. (2015). Structural Basis for Modulation of Quality Control Fate in a Marginally Stable Protein. Struct. Lond. Engl. 1993 23, 1169–1178.

Halfmann, R., Jarosz, D.F., Jones, S.K., Chang, A., Lancaster, A.K., and Lindquist, S. (2012). Prions are a common mechanism for phenotypic inheritance in wild yeasts. Nature 482, 363–368.

Jackson, W.S., and Krost, C. (2014). Peculiarities of Prion Diseases. PLOS Pathog 10, e1004451.

Luby-Phelps, K. (2000). Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. Int. Rev. Cytol. 192, 189–221.

Mukherjee, A., Morales-Scheihing, D., Butler, P.C., and Soto, C. (2015). Type 2 diabetes as a protein misfolding disease. Trends Mol. Med. 21, 439–449.

Mulligan, V.K., and Chakrabartty, A. (2013). Protein misfolding in the late-onset neurodegenerative diseases: Common themes and the unique case of amyotrophic lateral sclerosis. Proteins Struct. Funct. Bioinforma. 81, 1285–1303.

Narayanaswamy, R., Levy, M., Tsechansky, M., Stovall, G.M., O'Connell, J.D., Mirrielees, J., Ellington, A.D., and Marcotte, E.M. (2009). Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. Proc. Natl. Acad. Sci. U. S. A. 106, 10147–10152.

Peters, L.Z., Hazan, R., Breker, M., Schuldiner, M., and Ben-Aroya, S. (2013). Formation and dissociation of proteasome storage granules are regulated by cytosolic pH. J. Cell Biol. 201, 663–671.

Petrovska, I., Nüske, E., Munder, M.C., Kulasegaran, G., Malinovska, L., Kroschwald, S., Richter, D., Fahmy, K., Gibson, K., Verbavatz, J.-M., et al. (2014). Filament formation by metabolic enzymes is a specific adaptation to an advanced state of cellular starvation. eLife 3.

Sikosek, T., and Chan, H.S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. J. R. Soc. Interface R. Soc. 11, 20140419.

True, H.L., and Lindquist, S.L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. Nature 407, 477–483.

Appendix A

Alternative models to examine VHL allosteric response

In the chapter dedicated to understanding how a misfolded variant of VHL can escape quality control degradation, we used a previously developed biophysical model that had shown success in predicting allosteric response (Brock et al., 2015; England, 2011). This model, which uses sequence information to calculate a burial trace based on amino acid hydrophobicity and sequential order, is particularly well suited to this type of analysis. Each calculation typically takes less than a second to run, which allows us to screen many possible mutations on a feasible timescale. In contrast, using typical force-field folding algorithms like Rosetta can be prohibitively slow, especially for VHL's crystallizable region that contains > 100 residues.

Previous work has shown that using experimentally characterized hydropathies like the Kyte-Doolittle values are already optimal for correlating predicted burial trace with known crystal structure values (Perunov and England, 2014). These results confirm that hydrophobicity is a dominant factor in protein folding, at least in globular domains, and that these parameters in the model are already at near-optimal levels. Therefore, one alternative model might examine averaged hydrophobicity values taken over the sequence of VHL that can be crystallized. These results would provide information about whether the predicted structural change caused by the mutation can be explained more simply in terms of hydropathies, without relying on the steric constraints that the burial trace model provides.

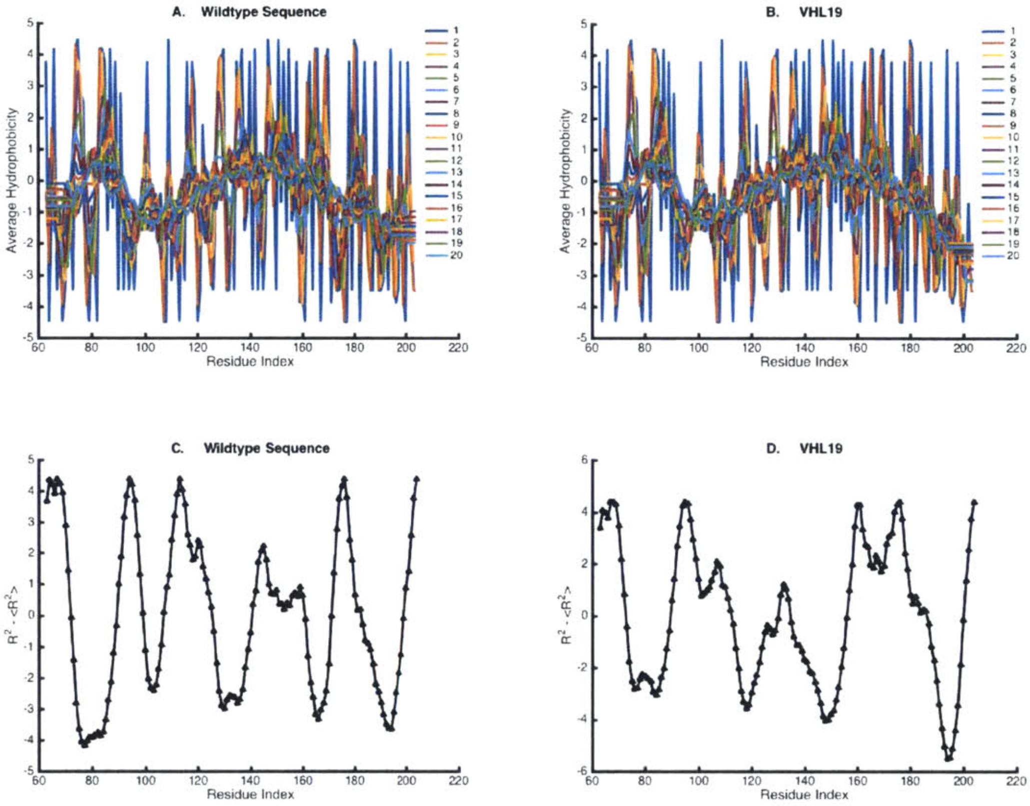


Figure 1. Average hydropathies of residue windows compared to burial trace model. Hydrophobicity was averaged along a window size denoted by color for (A) the wildtype sequence and (B) VHL19, the stabilizing mutation. For comparison, the burial trace for (C) wildtype and (D) VHL19 are shown below as well.

The averaged hydrophobicity at each residue was calculated by finding the mean hydrophobicity using the Kyte-Doolittle scale of the immediately preceding and following amino acids, and results are shown in Figure 1. The total number of amino acids averaged is referred to as the window size; even window sizes calculated an average hydrophobicity for the first of the two residues in the middle of the stretch, while odd window sizes were averaged around and reported on the true median. Importantly, more hydrophobic residues are more positive on this scale. To further examine whether average hydrophobicity is sufficient to explain the burial trace products by itself, I also plotted correlation of the average hydrophobicity with the normalized burial trace for both the wildtype and mutated sequences. In this case, the degree of anticorrelation corresponds to how well the averaged hydrophobicity model is able to recapitulate the predicted burial traces.

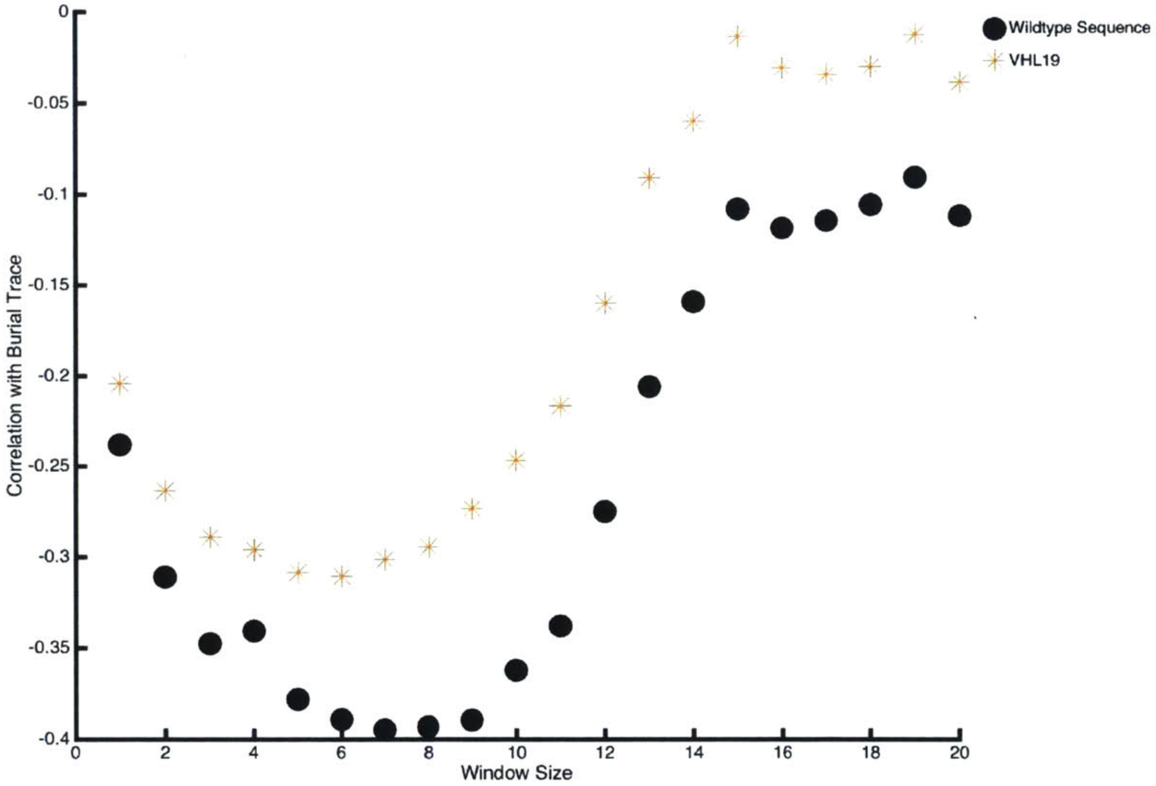


Figure 2. Correlation between burial trace model and average hydrophobicity for the mutated sequence VHL19 (orange stars) and the wildtype sequence (black circles).

Here, we see that the optimal window size is 7 for the wildtype and 6 for the mutant sequence, with a correlation of roughly -0.4 and -0.3 respectively. Therefore, although average hydrophobicity contributes to the results of the burial trace model, the level of correlation is not sufficient to reproduce the results even at the optimal window sizes. This finding is particularly true when discussing allostery, as is the case for the VHL19 mutation. The burial trace model optimizes overall exposed hydrophobicity against steric constraints and energetic costs associated with stretching peptide bonds. The average hydrophobicity model, on the other hand, is much more localized – for our optimal window sizes, changes in a distal region do not have an impact on local sequence. Therefore, this model would have been insufficient to view allosteric contributions from mutations in this context, which can be seen by comparing the lessened anticorrelation of the mutant and its burial trace when compared to the wildtype.

By using simple biophysical characterizations, this technique can provide an understanding of when hydrophobicity and steric constraints alone are enough to explain protein structural changes due to mutations. For example, VHL19's differential burial pattern in our predictions illustrates that the mutant is able to alter the exposure of its interaction motifs directly from these sequence characteristics, as opposed to necessitating prior interactions to change the conformational pattern of the cofactor and chaperone motifs. Furthermore, this insight is distinct to the burial mode model and it is particularly helpful in capturing complex, nonlinear responses to mutational change.

Another useful exercise is to study VHL in alternate models of allostery to see how our model's results compare to other predictive methods. One technique that has shown promise in understanding large-scale protein motion is normal mode analysis, which uses eigendecomposition of a vibrational system to describe residue fluctuations (Grant et al., 2006; Skjaerven et al., 2009; Skjærven et al., 2014). For example, predicted fluctuations of the amino acids in VHL derived from its crystal structure in PDB 1VCB are shown below.

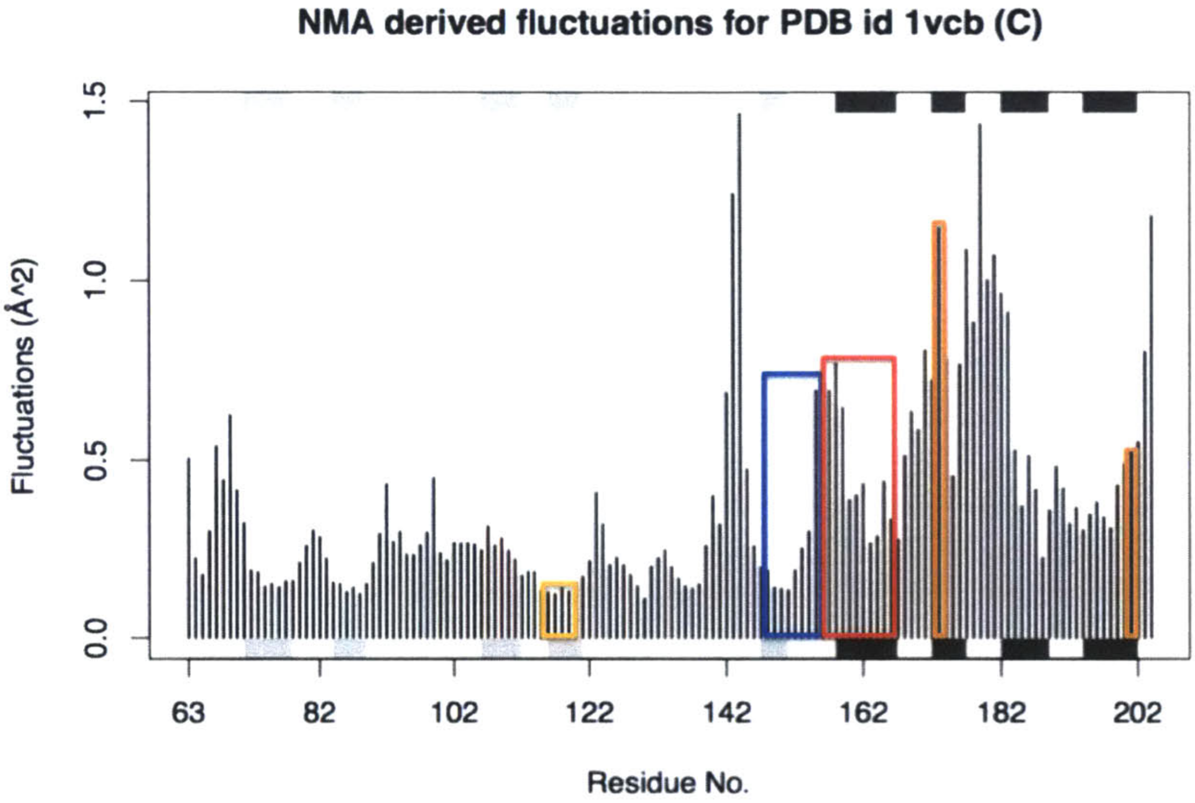


Figure 3. Fluctuations of each residue in VHL (derived from the crystal structure, PDB ID 1VCB) derived from normal modes analysis using Bio3D (Grant et al., 2006; Skjærven et al., 2014). The chaperonin interaction sites Box 1 and Box 2 are highlighted by the yellow and blue boxes, the elongin interaction region is highlighted by the red box, and the orange boxes at 173 and 201 show the locations of the residues involved in the VHL19 pair swap mutation.

Here, we see that Box 1 experiences relatively little fluctuation, while the other interaction sites exhibit a mix between low and high levels of local fluctuations. Furthermore, residue 173 – one of the two amino acids mutated in VHL19 – has a higher predicted ability to fluctuate than its counterpart 201.

Normal mode analysis can also be combined with other techniques to study allosteric response directly. One example, which also incorporates machine learning algorithms, is

AlloPred (Greener and Sternberg, 2015). This tool takes in a PDB crystal structure and the active site, and uses this information to predict protein allosteric pockets that can influence the active site. To compare to our burial trace prediction that VHL19 can alter the location of the VHL19 mutated residues, L201-E173, I defined the 'active sites' to be either the elongin interaction region or the elongin interaction region + chaperonin interaction sites, since these motifs were the places that our model predicted a large-scale structural change. In both cases, the top-scoring predicted pocket contained residue 201, as shown in the figure below.

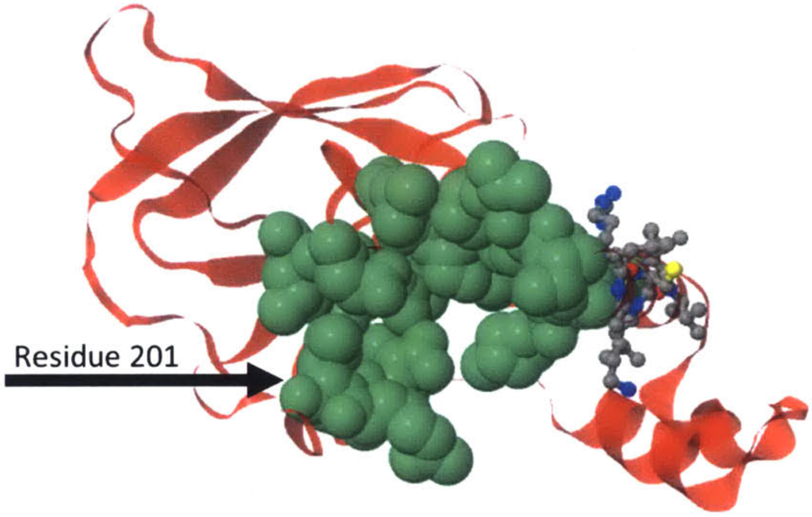


Figure 4. Structure of VHL, with the AlloPred predicted allosteric pocket shown with green balls and the elongin interaction region shown using grey sticks. The rest of the structure is shown with red ribbon cartoons. The black arrow denotes the location of residue 201.

Although the predicted binding pocket contains a relatively large number of residues, the inclusion of residue 201 agrees with the outcome of our burial trace model and its experimental validation. In the plain normal mode analysis calculation, residue 201 was predicted to be less able to experience large fluctuations. Therefore, one possible suggestion is that replacing the leucine at 201 with the charged residue glutamic acid might cause that chain to exhibit a different flexibility pattern, which could then affect other regions of the protein. From the burial trace model, we predict that swapping residues 201 and 173 changes the optimization of hydrophobic and steric constraints such that the interaction motifs can either be pushed into or out of the protein core relative to the wildtype sequence without its binding partners. Further exploration of the physical considerations underlying exactly how this pair-swap mutation causes this differential pattern could be an interesting avenue for future study.

References

Brock, K.P., Abraham, A., Amen, T., Kaganovich, D., and England, J.L. (2015). Structural Basis for Modulation of Quality Control Fate in a Marginally Stable Protein. Struct. Lond. Engl. 1993 23, 1169–1178.

England, J.L. (2011). Allostery in protein domains reflects a balance of steric and hydrophobic effects. Struct. Lond. Engl. 1993 19, 967–975.

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., and Caves, L.S.D. (2006). Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22, 2695–2696.

Greener, J.G., and Sternberg, M.J. (2015). AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. BMC Bioinformatics 16, 335.

Perunov, N., and England, J.L. (2014). Quantitative theory of hydrophobic effect as a driving force of protein structure. Protein Sci. 23, 387–399.

Skjaerven, L., Hollup, S.M., and Reuter, N. (2009). Normal mode analysis for proteins. J. Mol. Struct. THEOCHEM 898, 42–48.

Skjærven, L., Yao, X.-Q., Scarabelli, G., and Grant, B.J. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics 15, 399.

Appendix B

Additional aspects of collective behavior under starvation conditions

Sequence composition for S. cerevisiae

Sequence composition differs between species, and characteristics of sequence composition have been demonstrated to be under potential selective pressures as discussed in Chapters 3 and 4. In this study, we provided evidence that the cumulative effect of changes in charge across protein interaction partners and functional groups can affect overall cytosolic collective behavior under glucose deprivation. To tease out the degree to which these correlations between proteins might be important, we also examined S. cerevisiae sequence composition to determine whether the results that we saw were explainable purely in terms of overall sequence composition. For comparison purposes, the sequence compositions of *S. cerevisiae* and *H. sapiens* are shown below.

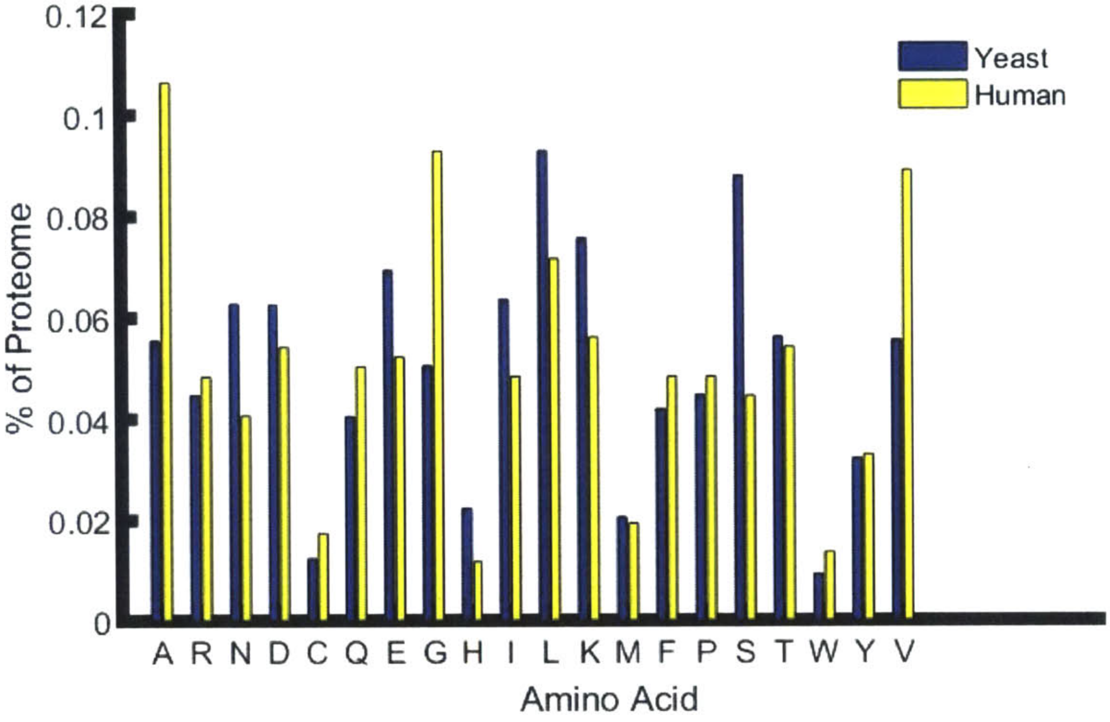


Figure 1. Sequence composition for S. cerevisiae (denoted as yeast, blue bars) and H. sapiens (human, yellow bars) proteomes.

Several amino acids, like glycine, valine, and alanine, are overrepresented in humans when compared to baker's yeast. Histidine also exhibits an increased frequency in *S. cerevisiae*, which might help contribute to this species' collective protein behavior under low pH conditions since histidine's side chain isoelectric point falls within the pH range that the yeast's cytosol experiences during glucose starvation. To further explore differences in overall sequence composition between S. cerevisiae and other organisms, we took a group of homologous proteins between *S. cerevisiae*, *H. sapiens*, and *A. gossypii* (a fungal species closely related to yeast) and compared their sequence compositions directly.

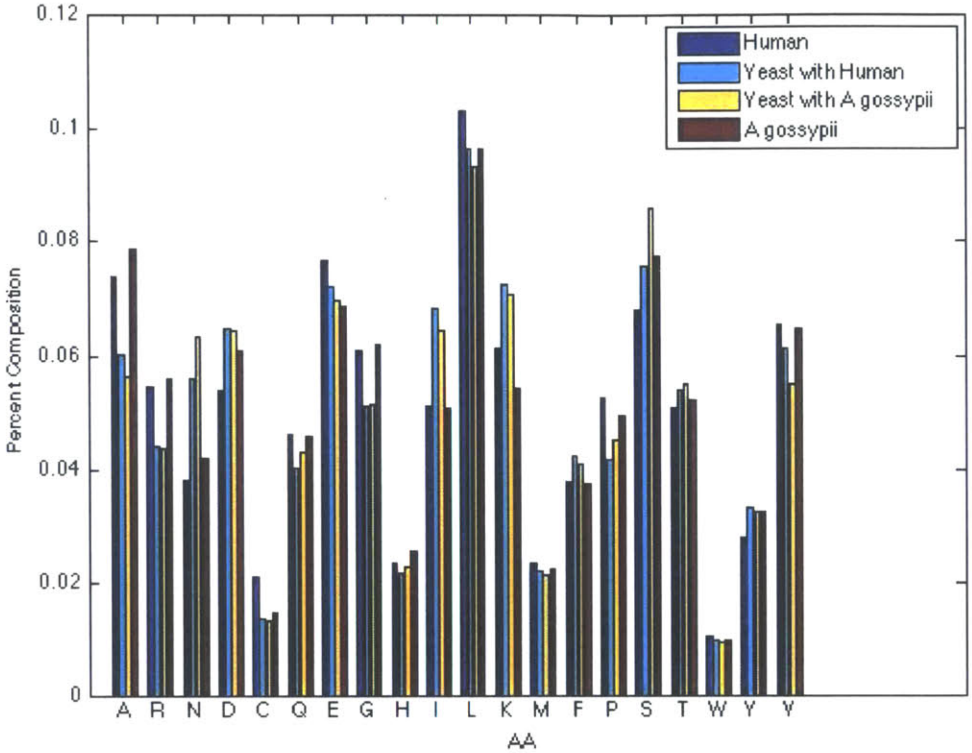


Figure 2. Sequence composition for human sequences that have homologues in S. cerevisiae (dark blue), S. cerevisiae sequences that have homologs in human (light blue), S. cerevisiae sequences that have homologs in A. gossypi (yellow), and A. gossypi sequences that have homologs in S. cerevisiae.

Although relatively minor differences exist within this smaller grouping of proteins, overall frequency of charged residues is relatively consistent between the three species. However, this grouping of homologous proteins was still predicted to have differential interaction patterns based on the electrostatic attractiveness of putative interactions (Chapter 4 Figure 3), indicating that sequence composition is not sufficient to explain our results. For further clarification, the overall sequence composition for charged groupings of residues is shown below.

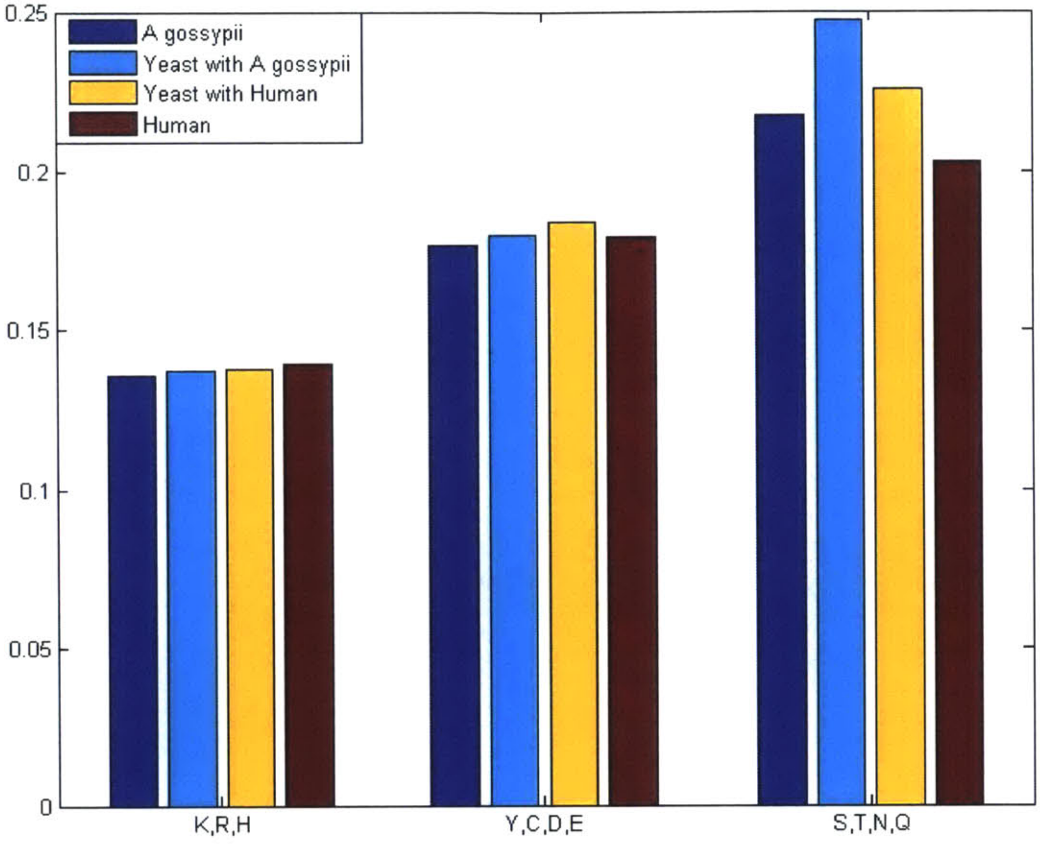


Figure 3. Sequence composition for human sequences that have homologs in S. cerevisiae (red), S. cerevisiae sequences that have homologs in human (yellow), S. cerevisiae sequences that have homologs in A. gossypi (light blue), and A. gossypi sequences that have homologs in S. cerevisiae (dark blue).

The two sets of yeast sequences, one with homologs in humans and one with homologs in *A. gossypii*, have very similar sequence compositions. For positively charged (K, R, and H) and negatively charged (Y, C, D, and E) residues, homologous sequences in yeast are not distinct in their sequence composition when compared to two other species. Since we still saw a differential response in predicted electrostatics when comparing *S. cerevisiae* interactions to their homologous protein partners in other species, these results indicate that a pattern of charge spread across an interaction network is more likely to form the basis of our predicted behavior as opposed to being purely an effect of overall amino acid frequencies.

Influence of abundance on charged proteins

Another important aspect in understanding protein collective behavior is by filtering through the lens of protein abundance. The cytosol is very crowded, and protein abundance levels can differ by several orders of magnitude. Therefore, connecting relative protein abundance to our earlier results on collective behavior under glucose starvation could provide a powerful future research direction.

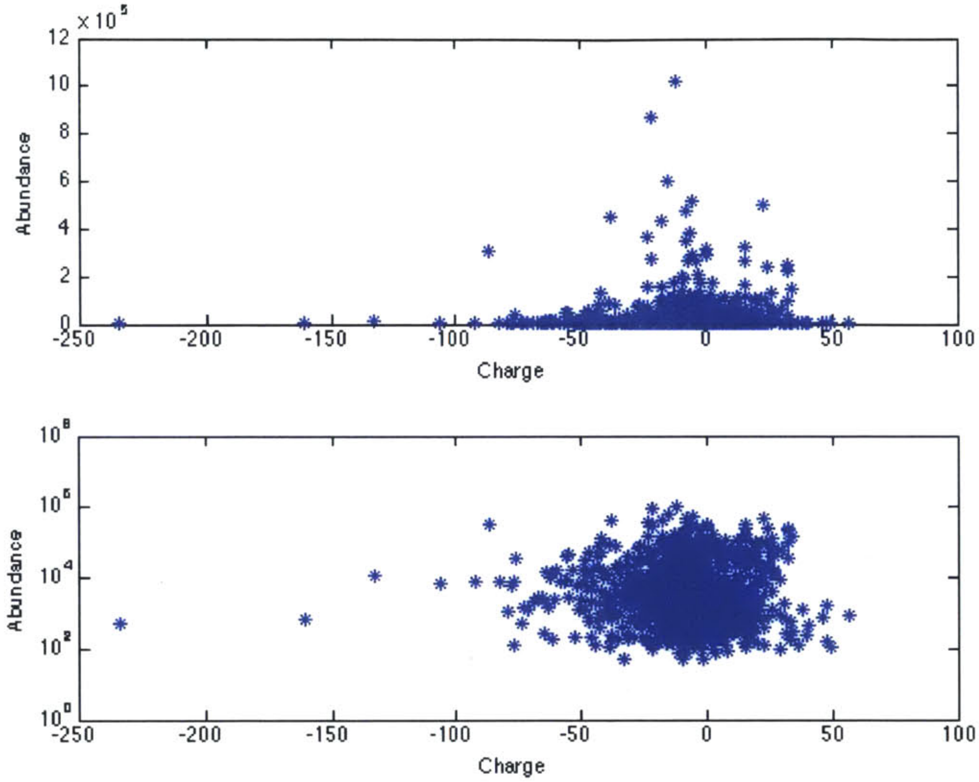


Figure 4. S. cerevisiae cytosolic proteins plotted with their abundance versus charge at neutral pH on a linear (top) and log scale (bottom).

In Figure 4, proteins that have a significant degree of charge tend to be less abundant, while proteins that are highly abundant tend to have a relatively lower charge magnitude. However, this pattern is more consistent for outliers – proteins that are distinguished in terms of either characteristic – than for the majority of proteins. A plot showing the distribution of isolectric points for proteins that display either very high or very low abundance is shown in Figure 5.

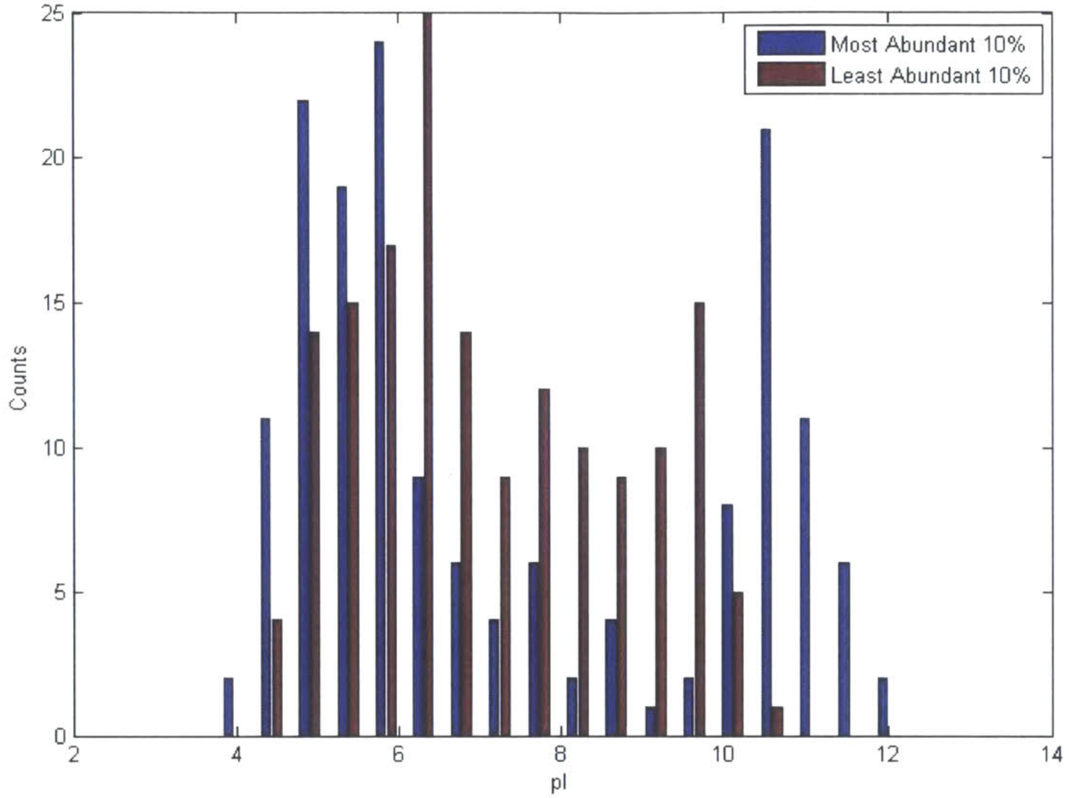


Figure 5. The distribution of isoelectric points (pI) for S. cerevisiae cytosolic proteins that fall into either the top 10% of highly abundant proteins (blue) or the bottom 10% of low-abundance proteins (red).

Interestingly, the high-abundance group only has 18% of its proteins in the isoelectric point range between pH 6 and 8, while the low-abundance proteins have 38% in this range. Future directions for this work could involve better understanding this relationship in the context of protein collective behavior; one example could be probing whether highly abundant proteins that form higher-order structures are more likely to do so through self-interactions or whether the change in charge makes them more promiscuous interactors.

Additional controls for Y2H interactions

In Chapter 4, we presented results that Y2H interactions have lower interaction scores under both neutral and acidic conditions when compared to random interaction sets drawn from the same group of proteins. One possible cause of this behavior at acidic pH is that interactions that are already attractive at neutral pH – which real, Y2H-determined interactions are – might be more likely to remain attractive at acidic pH. The distributions of the full group of 10,000 random interactions (blue stars) and the subset of that group that has interaction scores < 0 at neutral pH (red stars) at both neutral (top) and acidic (bottom) pH are shown below.

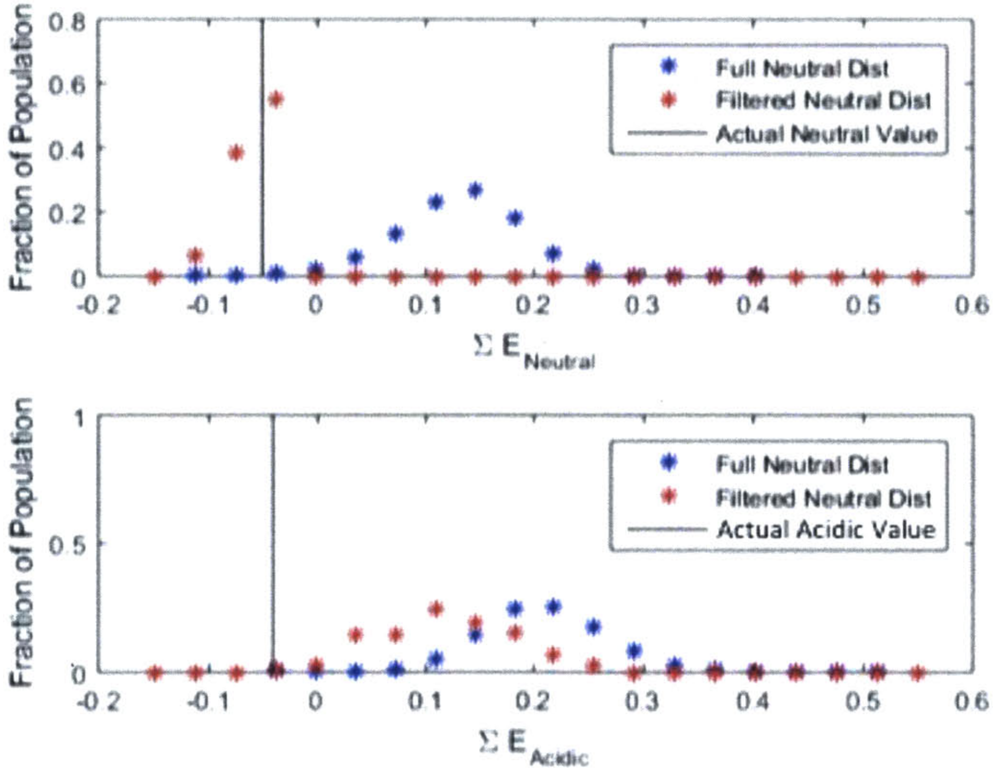


Figure 6. Normalized histograms showing interaction scores at neutral pH (top) and acidic pH (bottom) for 10000 random protein pairings (blue stars) and 78 random interactions that have very attractive interaction scores at neutral pH (blue stars). The real set of Y2H interactions are shown by the black vertical lines.

The subset of random interaction sets that is more attractive is pulled more towards being attractive at acidic pH as well, as shown in the bottom plot of Figure 6. However, these results are not sufficient to pull the distribution to the level of the real interactions, indicating that this effect alone does not explain our real results.

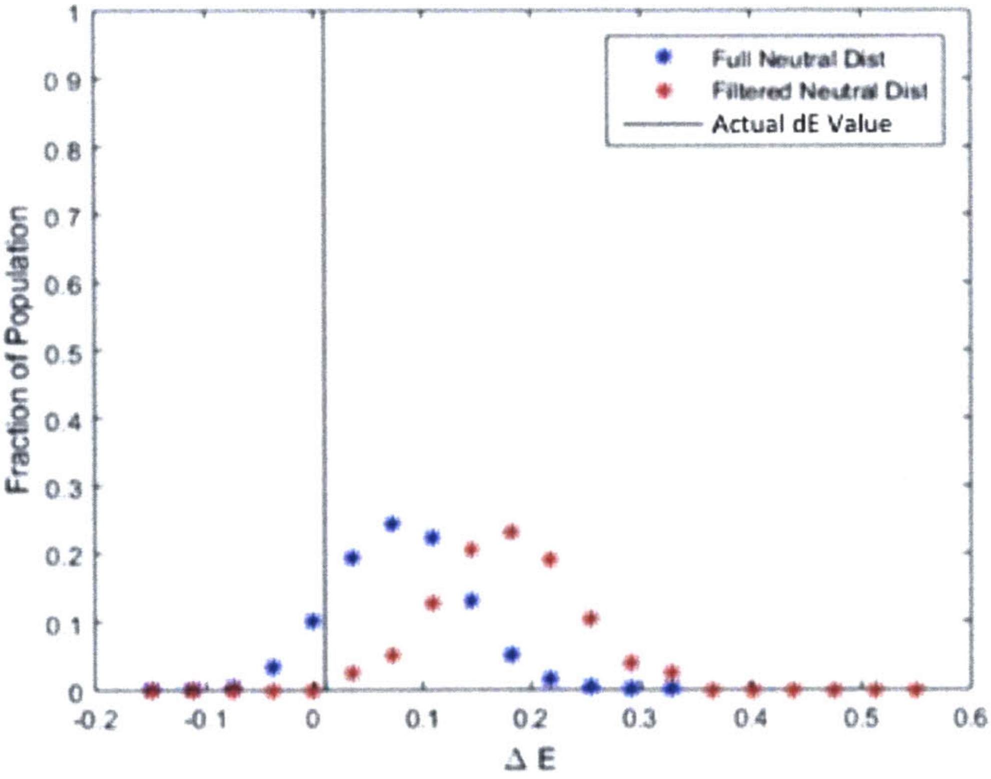


Figure 7. Distribution of change in interaction score (score at pH 5 – pH 7) for the full group of random interactions (blue stars), the subset that are very attractive at pH 5 (red stars), and the interaction score for the real set of Y2H interactions (black line).

This distinction can also be seen when we look at the change in interaction score as the pH is lowered from pH 7 to pH 5. Using the same set of filtered random interactions that have behavior similar to the Y2H interaction set at neutral pH, we compared the change in interaction score from pH 7 to pH 5 for this filtered random set to the full set of random interaction trials. However, this filtered set is more likely to become less attractive at acidic pH than both the full random set and the real Y2H set, as shown in Figure 7.

Another possible control is to see if the distribution of charges for real interactions is somehow distinct from those of random pairings, which might indicate that this effect can be explained at the individual sequence level as opposed to a collective effect arising from the network of protein interactions. We can examine this idea qualitatively in the plot below.

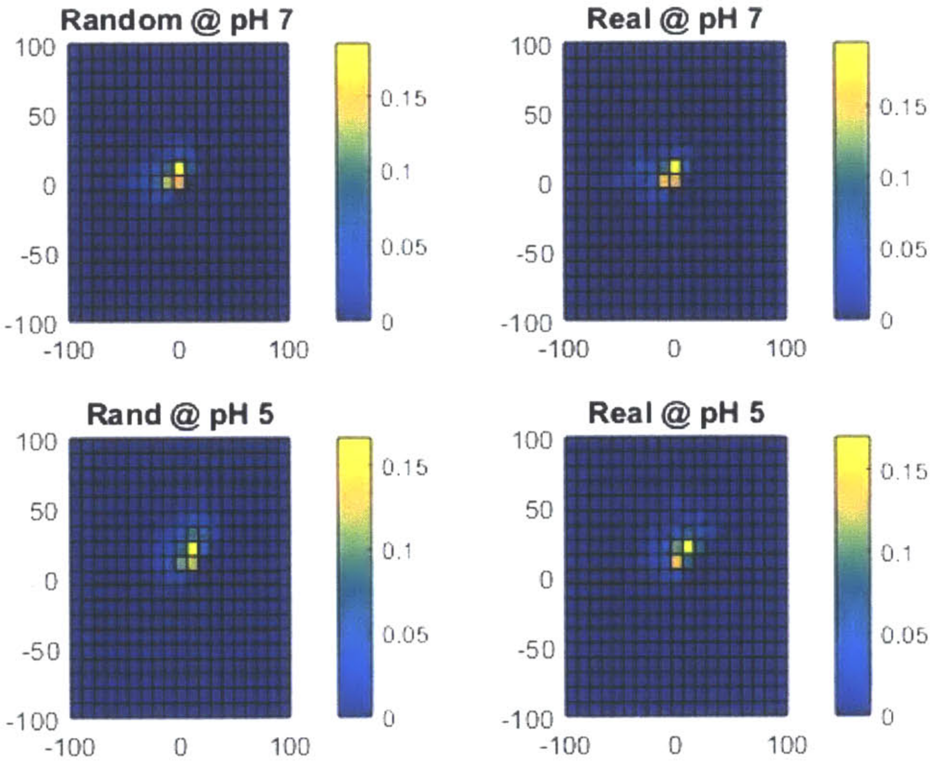


Figure 8. Distribution of charges for interacting protein pairs. The color bars represent the fraction of protein interacting pairs that fall within a bin defined by charges of proteins on the x and y axes of each plot. Random interactions at pH 7 and pH 5 are shown on the top and bottom left respectively, while real interactions are shown at pH 7 and pH 5 on the top right and bottom right.

As expected, the higher sampling in the random set provides a smoother overall shape when compared to the subset of real interactions. However, comparing real versus random charges at the same pH does not reveal any striking differences, particularly in the region where the proteins become oppositely charged at acidic pH.

Another potential test is to be more stringent in how we define random interactions. In the results reported in Chapter 4, we presented randomized interactions that were created by choosing two proteins randomly from the set of Y2H-involved proteins. Alternatively, we could also place a constraint on how promiscuous different proteins are – in other words, we could retain the degrees in the network topology by ensuring that the number of proteins that have N interaction partners is constant between the real Y2H data set and the random samples. If the results differed from our earlier plots, this finding would indicate that the Y2H results might be a function of network topology as opposed to being a consequence of how specific proteins interact with each other.

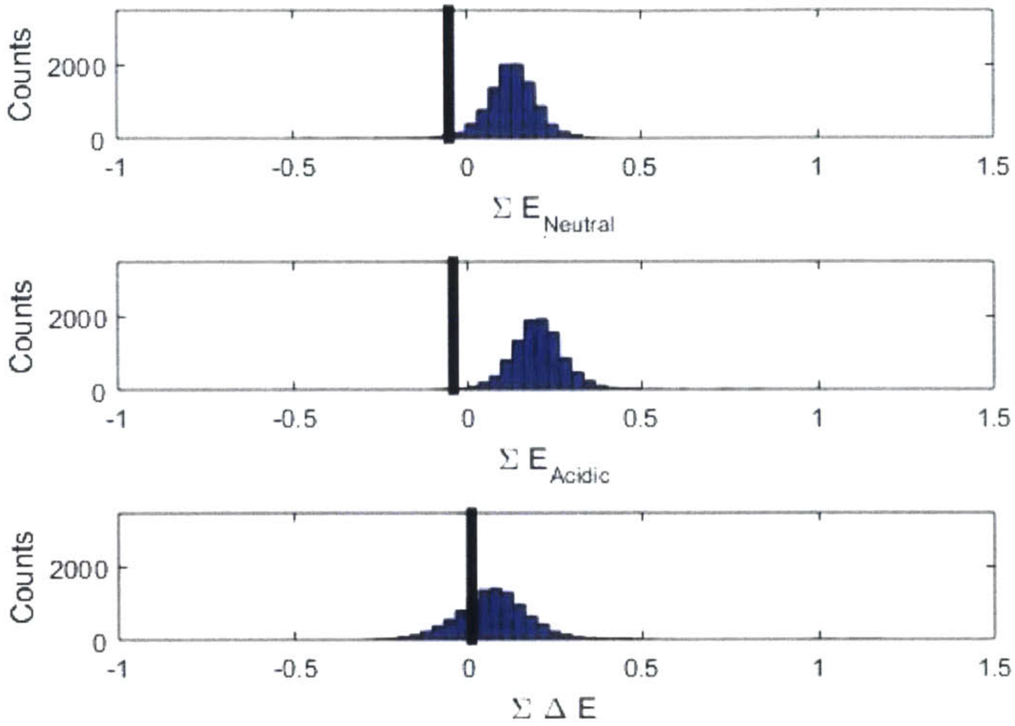


Figure 9. Random distributions with consistent overall topology for comparison to the Y2H data set. Distribution of random interaction networks are shown at neutral pH (top), acidic pH (middle), and for the change in score between the two values (bottom). In all cases, the black vertical line shows where the real Y2H interactions are located in the distribution.

These results are very similar to those obtained from naïve random pairings, with the Y2H interaction set being distinguishable in its level of attractiveness at both neutral and acidic pH.

An alternate way to examine changing protein interactions, that incorporates changes in charge but does not use the interaction score defined in Chapter 4, is to test whether a protein pair goes from having the same charge at pH 7 to having opposite charges at pH 5. We can calculate the fraction of real Y2H interactions that exhibit this behavior out of the complete Y2H set, and compare this to a random distribution of pairings drawn from the same group of proteins.

This idea is analogous to computing a conditional probability of an interaction turning 'on' (becoming more electrostatically attractive) at pH 5 given that it was 'off' (repulsive) at pH 5. The opposite effect, where a protein pair goes from being attractive at pH 7 to being repulsive at pH 5 (or the conditional probability of a protein pair turning 'off' at pH 5 given that it was 'on' at pH 7), can be calculated as well as shown in Figure 10.

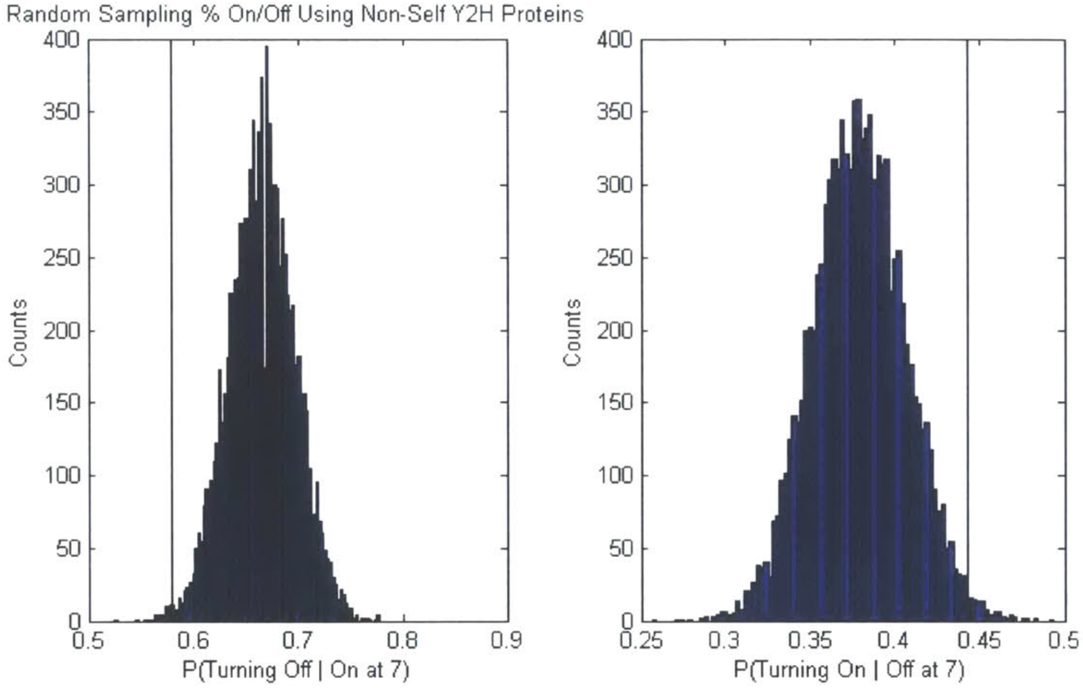


Figure 10. The distribution of the fractions of protein pairs in randomized trials that go from being attractive at pH 7 to repulsive at pH 5 (left) and that go from being repulsive at pH 7 to attractive at pH 5 (right). In both cases, the black vertical line shows the location of the real set of Y2H interactions.

For 10,000 random groups of proteins pairs, the real set of Y2H interactions is defined by its enhanced ability to have a higher fraction of interactions be more attractive at pH 5 – either by remaining attractive from their values at pH 7, shown by the plot on the left where 99.53% of random trials have a higher conditional probability of turning 'off' and having the same charge sign at pH 5 compared to the real interaction set, or by going from repulsive to attractive as the pH shifts from pH 7 to pH 5, as shown by the plot on the right where 98.97% of trials fall below our actual fraction of interactions that become oppositely charged when going from neutral to acidic pH. Therefore, these results still show the same overall charge-based collective effect, but they do not rely on our formulation of interaction score.

Charge-based analysis of Gene Ontology (GO) groups

In Chapter 4, we provided several computational results that explore the idea that pH-based increases in electrostatic attractiveness can cause proteins to cluster together to inactivate functional responses like metabolic pathways. This analysis was first motivated by several experimental studies that showed general clustering and inactivation of metabolic and functionally-related proteins, as outlined in Chapters 3 and 4. An additional test that we performed was to look at groups of proteins that share a similar function, as defined by their Gene Ontology (GO) characterization (Ashburner et al., 2000; Gene Ontology Consortium, 2015). We took our group of cytosolic proteins and classified them based on 100 GO Slim categories. We then modeled each protein as interacting with every other protein within its group, and calculated the summed group interaction score. For each GO group, we performed many trials that either chose the same number of a random grouping of cytosolic proteins or generated the same number of random sequences that overall shared the same amino acid frequencies with the original GO group. We then compared the real group to both sets of controls and, to see what their position was on that randomized distribution, we calculated a Z-score (real value minus the random mean divided by the standard deviation of the random distribution). Histograms of Z-scores for both control types are shown below.

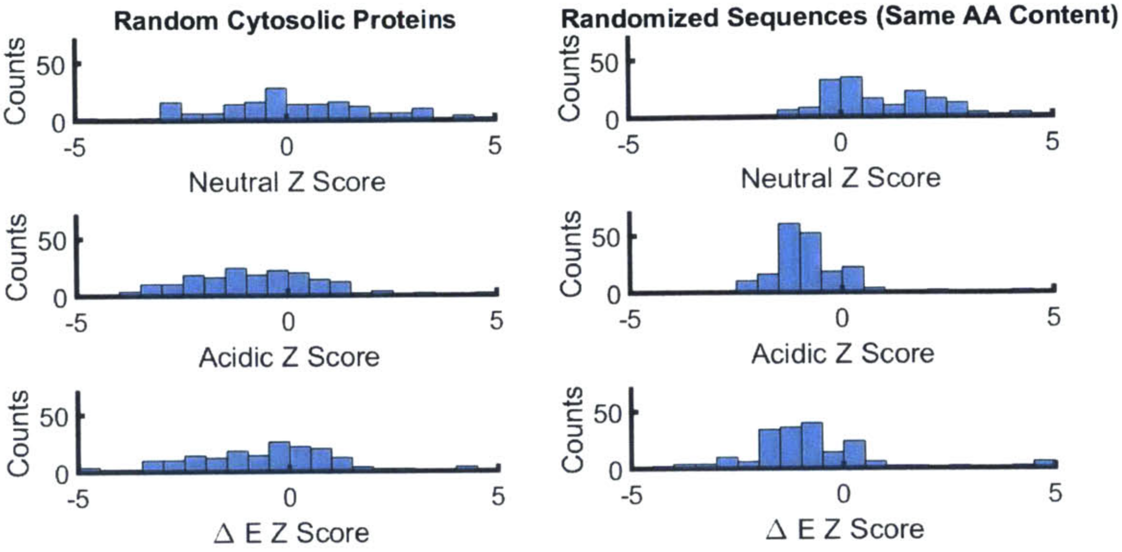


Figure 11. Distribution of GO group Z-scores compared to random cytosolic protein groupings (left) and randomized sequences that retain the same amino acid content (right). Distributions at neutral pH (top), acidic pH (middle), and the change in scores between pH values (bottom) are shown.

For random cytosolic proteins, the fraction of GO groups that have Z-scores less than -1, indicating that the real group of proteins is further away than one standard deviation below the mean, jumps from 24 to 42. The results become more pronounced when we look at randomized sequences – at neutral pH, only 3 GO groups have a summed interaction score that is at least a standard deviation less than the random control distribution, while that number jumps to 43 at acidic pH. The results indicate that GO groups may become more attractive within their functional classes as the pH drops, lending evidence that functional relationships can underlie which proteins are more likely to form higher order structures.

References

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. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056.