Structural Correlates of Protein Melting Temperature

Eric A. Franzosa[1], Kevin J. Lynagh[2], and Yu Xia[1,*]

[1]Bioinformatics Program, Dept. of Chemistry, Dept. of Biomedical Engineering, Boston University, 24 Cummington Street, Boston, MA 02215, U.S.A.

[2]Reed College, MS 880, 3203 SE Woodstock Boulevard, Portland, Oregon 97202 – 8199, U.S.A.

E-Mail: *yuxia@bu.edu

Abstract

The stability of a protein's native state has important implications for its folding dynamics, function, and evolution. Here we report on a study investigating general relationships between sequence- and structure-based properties of a protein and its empirically determined stability (as measured by melting temperature experiments). Surprisingly, we find that contact density – a sequence-independent measure of protein compactness – is not significantly correlated with protein melting temperature; this property has been previously implicated as a correlate of protein stability in theoretical and evolutionary analyses. After incorporating residue type in the definition of residue-residue contacts, we find that increasing the fraction of hydrophobic contacts in a protein tends to raise melting temperature, consistent with a stabilizing effect, while increasing the fraction of repulsive charge contacts results in a marginally significant decrease in melting temperature, consistent with a destabilizing effect. Our work demonstrates that subtle sequence variation may be an important factor in fine-tuning the stability of a protein fold.

INTRODUCTION

A protein's stability can be thought of as its thermodynamic "preference" for achieving and maintaining the native (folded) state. This evolved property is critical for protein function and for preventing the accumulation of cytotoxic unfolded or misfolded protein forms. Stability can vary among proteins from a single species and among homologues of a given protein across species. A well-known example of the latter case is *Taq* polymerase: a bacterial DNA polymerase evolved for the high temperature environment of its thermophilic source organism, and also a key component in the development of the polymerase chain reaction (PCR).

A popular laboratory measure of protein stability is *melting temperature*. In melting temperature experiments, a solution of a protein in its native state is heated until the point of complete denaturation. The temperature at which the native and denatured states are equally populated at equilibrium is called the melting temperature, and it is inferred from a change in the optical properties of the solution. Melting temperature experiments are frequently carried out on wild type and mutant proteins to investigate the contributions of specific residues to protein stability. The ProTherm database contains a record of many such experiments [1].

Like all protein properties and functions, stability "information" must be encoded by the protein's structure, which is in turn specified by the protein's primary sequence. In principle it should therefore be possible to predict protein stability from sequence and structure data. Along these lines, many methods have been proposed for predicting *changes* in protein stability upon introduction of a mutation; see [2 – 7] for some recent examples. In this work, we aim to understand better which protein properties contribute to variation in experimental protein melting temperatures. Previous studies in this area have generally relied upon binning proteins to look for systematic differences between sequences and structures derived from organisms adapted to low, moderate, and high temperatures; see [8 – 14] for some examples. These studies revealed general and position-specific preferences for certain amino acids and amino acid pairs in different temperature regimes. Our approach has an advantage in that, instead of using "source organism" as a discrete proxy for protein temperature tolerance, we consider a direct, continuous measure of protein stability based on laboratory experiments. We hypothesize that uncovering the protein properties most correlated with melting temperature may provide new insight into the mechanisms underlying protein stability.

METHODS & RESULTS

We collected data for 72 single-domain, wild-type proteins with (i) known structures in the Protein Data Bank [15] and (ii) melting temperature data from at least one experiment reported in the ProTherm database [1]. We considered average melting temperature if more than one experimental value was given. Total surface area and volume were computed for

each protein structure using the program MSMS [16]. An *L*-by-*L contact matrix* was also constructed for each protein structure, where L is the protein's length. The (i, j) entry of such a matrix is 1 if the alpha carbons of residues i and j are separated by no more than 7 angstroms (representing contact) and 0 otherwise (representing no contact).

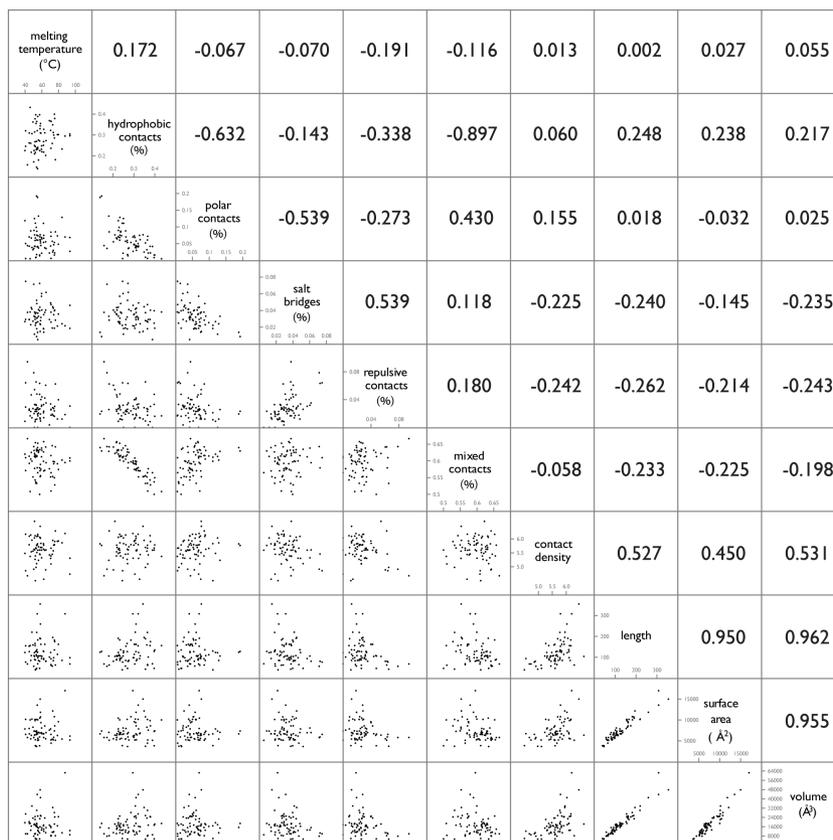|	hydrophobic contacts (%)	polar contacts (%)	salt bridges (%)	repulsive contacts (%)	mixed contacts (%)	contact density	length	surface area (Å²)	volume (Å³)
melting temperature (°C)	0.172	-0.067	-0.070	-0.191	-0.116	0.013	0.002	0.027	0.055
hydrophobic contacts (%)		-0.632	-0.143	-0.338	-0.897	0.060	0.248	0.238	0.217
polar contacts (%)			-0.539	-0.273	0.430	0.155	0.018	-0.032	0.025
salt bridges (%)				0.539	0.118	-0.225	-0.240	-0.145	-0.235
repulsive contacts (%)					0.180	-0.242	-0.262	-0.214	-0.243
mixed contacts (%)						-0.058	-0.233	-0.225	-0.198
contact density							0.527	0.450	0.531
length								0.950	0.962
surface area (Å²)									0.955

Figure 1. Relationships among protein melting temperature and protein structural properties. For each pair of properties we report the Pearson correlation coefficient (above the diagonal) and a scatter plot (below the diagonal) comparing data for the 72 proteins in our study.

The *contact density* of a protein is the average number of contacts per residue, and serves as a measure of protein "compactness". We also considered the maximum Eigen value of the contact matrix as a more sophisticated notion of contact density [17]; this measure generalizes a residue's degree of connectedness beyond its immediate neighbours. To incorporate sequence data, we then categorized each individual residue-residue contact as one of the following: hydrophobic (both residues are hydrophobic), polar (both residues are polar), mixed (one residue is polar, the other is hydrophobic), salt bridge (residues have opposite charge at physiological pH), and repulsive (residues have the same charge at physiological

pH). Correlations reported in the paper and the figure represent Pearson's correlation; one-tailed p-values were determined from 10,000 rounds of randomizing permutation. The Figure 1 also contains a scatter plot comparison of each property pair.

Basic protein geometric features (length, surface area, and volume) correlate extremely weakly with melting temperature. The correlation between melting temperature and contact density is also non-significant ($r = 0.013$; $p = 0.475$). This second fact is surprising given the theoretical importance of contact density: in simulated proteins, contact density has been demonstrated as an important determinant of fold designability [17], which is itself proposed to positively correlate with protein stability [18]. These ideas are supported by the observation that (across multiple species) contact density correlates positively with evolutionary rate [19, 20] – an observation that makes sense if proteins of high contact density are very stable and therefore robust against mutation. However, the observed lack of significant correlation between contact density and melting temperature suggests either that contact density is not directly connected to protein stability, or that the connection is too weak to detect using our small dataset (72 proteins). We also investigated a more sophisticated notion of contact density (the maximum Eigen value of the contact matrix), and reached similar conclusions.

One weakness of traditional contact density measurements is that they do not consider protein primary sequence information. Two proteins with the same backbone geometry will have the same contact density, even if their amino acid sequences are very different. To test the importance of amino acid sequence in determining melting temperature, we classified contacts in our protein dataset based on the biochemical nature of the involved pairs of amino acids. We identified a slight tendency for proteins with a larger fraction of hydrophobic contacts to have elevated melting temperature ($r = 0.172$; $p = 0.072$). Note that, although larger proteins tend to contain a larger fraction of hydrophobic contacts, the poor correlation between melting temperature and properties such as length, surface area, and volume precludes explanation based on a size effect. Maximizing hydrophobic contacts is a critical driving force in protein folding [21], and so it is reasonable to speculate that a larger fraction of such contacts might further stabilize a protein fold. Surprisingly, the correlation between melting temperature and fraction of salt bridges (contacts between oppositely charged residues) was weakly negative and non-significant ($r = -0.070$; $p = 0.275$). This seemingly counter-intuitive observation that salt bridge interactions do not contribute positively to protein stability can be explained by the large unfavourable free energy cost of desolvating these charged residues in the first place [22]. Finally, the fraction of repulsive contacts (interactions between like charged residues) showed a marginally significant, negative correlation with melting temperature ($r = -0.191$; $p = 0.053$), consistent with a destabilizing effect.

DISCUSSION

Although the significance of contact density in protein evolution and design has been theoretically and empirically demonstrated, it does not seem to play a significant role in determining protein melting temperature in our dataset. This observation can be rationalized by returning to the *Taq* polymerase example from the Introduction. *Taq* polymerase has a higher melting temperature than its counterpart in *Escherichia coli*. However, from a structural perspective, the polymerase domains of the two proteins are almost identical [23] and hence their computed contact densities will also be very similar. The same degree of similarity is not found at the sequence level. Indeed, a BLAST search reveals that the polymerase domains of the two proteins have fewer than 50% identically aligned residues [24]. This fraction is sufficient to produce highly similar folds as a result of the many-to-one nature of the sequence-structure relationship. At the same time, these numerous differences allow for great variation in the types of residue-residue contacts found in the two protein backbones, which may play a role in explaining the observed difference in their melting temperatures. We show here that a decrease in the fraction of repulsive contacts in a protein results is a marginally significant increase in melting temperature. An increase in the fraction of hydrophobic contacts was also associated with melting temperature increase. These findings provide support for the role of amino acid sequence in fine-tuning the stability of a protein fold.

The greatest limitation we have encountered thus far is the low availability of wild-type proteins of known structure and melting temperature. In the future we aim to expand our dataset through the addition of multi-domain proteins or by structural homology modelling. We are further limited by the general noisiness of protein melting temperature data and the fact that these data are derived from many independent experiments. Finally, the contact-based parameters presented here were based on alpha-carbon models of protein structures (by our own approximation). We will next expand the work to consider all-atom models, which may provide a more realistic picture of the contacts between residues. Nevertheless, our study here demonstrates that bioinformatics analysis of protein stability data can provide insights into the structural determinants of protein thermodynamics.

ACKNOWLEDGMENTS

REFERENCES

[1]   Kumar, M.D.*, et al.*(2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* **34**(Database issue): D 204 – 6.
        doi: http://dx.doi.org/10.1093/nar/gkj103.

[2]   Capriotti, E. *et al.* (2008) A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9**(Suppl 2):S 6.
        doi: http://dx.doi.org/10.1186/1471-2105-9-S2-S6.

[3]   Cheng, J., Randall, A. and Baldi, P. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **62**(4):1125 – 32.
        doi: http://dx.doi.org/10.1002/prot.20810.

[4]   Gromiha, M.M. (2007) Prediction of protein stability upon point mutations. *Biochem. Soc. Trans*. **35**(Pt 6):1569 – 73.
        doi: http://dx.doi.org/10.1042/BST0351569.

[5]   Huang, L.T. *et al.* (2007) Prediction of protein mutant stability using classification and regression tool. *Biophys. Chem.* **125**(2 – 3):462 – 70.
        doi: http://dx.doi.org/10.1016/j.bpc.2006.10.009.

[6]   Masso, M. and Vaisman, II (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics* **24**(18):2002 – 9.
        doi: http://dx.doi.org/10.1093/bioinformatics/btn353.

[7]   Parthiban, V. *et al.* (2007) Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins* **66**(1):41 – 52.
        doi: http://dx.doi.org/10.1002/prot.21115.

[8]   Cambillau, C. and Claverie, J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.* **275**(42):32383 – 6.
        doi: http://dx.doi.org/10.1074/jbc.C000497200.

[9]   Gianese, G., Bossa, F., and Pascarella, S. (2002) Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes. *Proteins* **47**(2):236 – 49.
        doi: http://dx.doi.org/10.1002/prot.10084.

[10]   Gromiha, M.M. (2001) Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophys. Chem.* **91**(1): 71 – 7.
        doi: http://dx.doi.org/10.1016/S0301-4622(01)00154-5.

[11]   Kannan, N. and Vishveshwara, S. (2000), Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Eng.* **13**(11):753 – 61.
doi: http://dx.doi.org/10.1093/protein/13.11.753.

[12]   Kumar, S., Tsai, C.J., and Nussinov, R. (2001) Thermodynamic differences among homologous thermophilic and mesophilic proteins. *Biochemistry* **40**(47):14152 – 65.
doi: http://dx.doi.org/10.1021/bi0106383.

[13]   Pack, S.P. and Yoo, Y.J. (2004) Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. *J. Biotechnol.* **111**(3):269 – 77.
doi: http://dx.doi.org/10.1016/j.jbiotec.2004.01.018.

[14]   Szilagyi, A. and Zavodszky, P. (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **8**(5):493 – 504.
doi: http://dx.doi.org/10.1016/S0969-2126(00)00133-7.

[15]   Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**(1):235 – 42.
doi: http://dx.doi.org/10.1093/nar/28.1.235.

[16]   Sanner, M.F., Olson, A.J., and Spehner, J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**(3):305 – 20.
doi: http://dx.doi.org/10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y.

[17]   England, J.L. and Shakhnovich, E.I. (2003) Structural determinant of protein designability. *Phys. Rev. Lett.* **90**(21):218101.
doi: http://dx.doi.org/10.1103/PhysRevLett.90.218101.

[18]   Bloom, J.D. *et al.* (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.* **103**(15):5869 – 74.
doi: http://dx.doi.org/10.1073/pnas.0510098103.

[19]   Zhou, T., Drummond, D.A. and Wilke, C.O. (2008) Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Evol.* **66**(4):395 – 404.
doi: http://dx.doi.org/10.1007/s00239-008-9094-4.

[20]   Bloom, J.D. *et al.* (2006) Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**(9):1751 – 61.
doi: http://dx.doi.org/10.1093/molbev/msl040.

[21]   Murphy, K.P. (2001) Stabilization of protein structure. *Methods Mol. Biol.* **168**:1 – 16.

[22]   Hendsch, Z.S. and Tidor, B. (1994) Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **3**(2):211 – 26.

[23]   Eom, S.H., Wang, J. and Steitz, T.A. (1996) Structure of Taq polymerase with DNA
        at the polymerase active site. *Nature* **382**(6588):278 – 81.
        doi: http://dx.doi.org/10.1038/382278a0.

[24]   Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of
        protein database search programs. *Nucleic Acids Res.* **25**(17):3389 – 402.
        doi: http://dx.doi.org/10.1093/nar/25.17.3389.