Eur Biophys J (2009) 38:577–587
DOI 10.1007/s00249-009-0409-1

ORIGINAL PAPER

A new computational model to study mass inhomogeneity and hydrophobicity inhomogeneity in proteins

Anirban Banerji · Indira Ghosh

Received: 29 August 2008 / Revised: 5 January 2009 / Accepted: 14 January 2009 / Published online: 13 February 2009

Abstract We propose a simple yet reliable computational framework that characterizes the differential mass and hydrophobicity distribution within structural classes of proteins. Radial partitioning of protein interior that could successfully distinguish the mass and hydrophobicity distribution patterns in extremophilic proteins from that in their structurally aligned mesophilic counterparts. Distance-dependent mass and hydrophobicity magnitudes could retrieve vital structural insights; needed to probe the hidden connections between packing, folding and stability within different structural classes of proteins, with causality. New computational markers; one, to represent the total mass content; other, related to hydrophobic centrality of proteins, are proposed as well. Results reveal that mass and hydrophobicity packing within extremophilic proteins is indeed more compact than that in their mesophilic counterparts. Analysis of structural constraints within them vindicate it. Total mass (and hydrophobicity) content is found to be maximum in α/β thermophilic proteins and minimum for the all-α mesophilic proteins.

Keywords Inhomogeneously packed protein interior · Mass packing · Hydrophobicity packing · Hydrophobic center · Radial partitioning · Structural classification of proteins

Electronic supplementary material The online version of this article (doi:10.1007/s00249-009-0409-1) contains supplementary material, which is available to authorized users.

A. Banerji
Bioinformatics Centre, University of Pune,
Pune 411007, Maharashtra, India

I. Ghosh (✉)
School of Information Technology,
Jawaharlal Nehru University, New Delhi 110067, India
e-mail: indirag@mail.jnu.ac.in

Introduction

The need to search for an inclusive computational framework that provides objective information about numerous aspects of mass profiling in protein interior is ever present. Such automated tool is necessary not only for the study of packing characteristics in protein interior but also to provide clues for understanding protein stability and protein folding. Previous studies of protein packing geometries have indicated that globular proteins are compact and densely packed (Richards 1974) to the extent that their interior is commonly perceived as one mimicking that of solids (Hermans and Scheraga 1961; Richards 1997). Although presence of voids and cavities in protein interiors are reported (Liang and Dill 2001), significance of compact packing is universally acknowledged because it is considered to be highly important for protein stability (Privalov 1996) and for nucleation of protein folding (Ptitsyn 1998; Ting and Jernigan 2002). It is in this context that we can infer the importance of examining the details of inhomogeneous mass distribution in protein interior by developing a new and insightful tool, which can on one hand explore numerous facets of mass packing characteristics, while on the other hand, investigate the causality behind such packing profiles, taking into account all the influencing biophysical factors.

Previous studies in the area include the work by Richards (1974), where he followed Bernal and Finney (1967) in drawing Voronoi polyhedrons of minimal sizes around constituent atoms of proteins to calculate packing density of atoms within the protein. Taking into consideration the overlap between atomic volumes inside proteins, Liang and Dill (2001) made use of Delaunay triangulation to conclude that proteins can only be modeled as organic crystals if studied with respect to average density, otherwise they appear as fluids with respect to their free volume distributions. Building upon the studies due to Tsai et al. (1999) and Hubbard et al. (1994), Frommel (Rother et al. 2003) observed a deep disagreement about packing information in protein interior, while inferring essentially inhomogeneous packing schemes throughout the interior of the protein, confirming Beardsley and Kauzmann (1996). Kuntz (1972) too had mentioned the inhomogeneous density distribution within carboxypeptidase; but an integrative and general framework to analyze protein's mass distribution, density, hydrophobicity distribution and protein stability—did not generally emerge from aforementioned works. Even the 'radius of gyration' (ROG), which connects protein's shape with its mass distribution, is reported at times to be a "poor" marker to characterize protein packing (Zhang et al. 2003).

Mass packing and protein stability studies find interesting test cases in proteins derived from extremophilic organisms, owing to their extraordinary stability. It is known that, proteins extracted from thermophilic bacteria tend to be more stable (with respect to temperature) than their mesophilic counterparts (Jaenicke and Bohm 1998; Szilágyi and Závodszky 2000), generally; though inherent fold characteristics of protein families remain invariant. However, here also we find apparent disagreements between views expressed by Xiao and Honig (1999), Das and Gerstein (2000), Kannan and Vishveshwara (2000) with that of Karshikoff and Ladenstein (1998); where the later had used computational geometry oriented methodology to conclude that packing density is not a dominant factor contributing to the thermal stability; asserting alongside that proteins drawn from mesophilic and thermophilic bacteria essentially do not differ in their degrees of packing. This finding contrasts with implications of Honig's study, where electrostatic contributions to folding free energy of hyperthermophilic proteins and their mesophilic homologues were calculated only to conclude that electrostatic interactions in the thermophilic proteins are more favorable than their mesophilic counterparts. Gerstein's study (Das and Gerstein 2000) also revealed an overall greater content of charged residues in thermophiles than in mesophiles and noted that, intra-helical salt bridges are more prevalent in thermophiles than mesophiles. From a slightly different standpoint, Kannan and Vishveshwara (2000) had reported the presence of additional aromatic clusters and aromatic networks in the thermophilic proteins, in contrast to their mesophilic counterparts. How such favorable frame of electrostatic interactions and aforesaid structural features can produce essentially no difference in packing in thermophilic proteins (Karshikoff and Ladenstein 1998) seemed paradoxical to us. We therefore wanted to study the entire premise of structural features of protein interior by constructing concentric shells around the center of mass (CM) of the biological units of proteins and by systematically analyzing the mass (and density) distribution in each of these shells. The strength of the method of concentric shells is in the very fact that no pre-conceived limit is imposed on it. The larger the radial extent of the protein under consideration, more is the total number of shells to retrieve the mass packing information therein. On exactly the similar logic, smaller number of shells will be required to represent a protein with smaller radial extent. The number of shells to describe the radial extent of a protein is therefore self-adjustable in nature and can be implemented with easy computational techniques based on decision making.

Special attention was provided to represent and study hydrophobicity with as much precision as possible. Such need was not felt merely due to the strong evidence of its role in determining the overall folded structure (Kauzmann 1959; Dill 1990) but also because of its importance in ensuring protein stability and its possible influence on protein's mass distribution. However, observing the anomaly in scales of residue hydrophobicity as was mentioned by Haney et al. (1999), we chose to work with the 'atomic hydrophobicity' magnitudes, as proposed by Kuhn et al. (1995). Apart from the possible biases (Haney et al. 1999), coarse-graining operation at the residue level, we feared, may shield the detailed characteristic of the finer aspects of hydrophobicity profile that is obligatory for the present study. Atomic hydrophobicity quantifies the energy cost of transferring solvent accessible surface area of the atom from an aqueous environment to octanol; positive values indicate that the atom is hydrophobic, i.e., its solvation by octanol is energetically more favorable than by water, and vice versa. The atomic hydrophobicity magnitude, say ah_i (derived from experimental octanol–water partition coefficients of large set of various chemical compounds) of atom i, for the hydrophobic atoms assume $ah_i > 0$, while for the hydrophilic atoms assume $ah_i < 0$. Since in proteins (irrespective of amino acids concerned) certain atoms show predominant hydrophobic features (e.g., C) while certain other atoms (e.g., N and O) show propensities for being hydrophilic or polar, any protein can be considered as an ensemble of atoms with (+ve) or (−ve) hydrophobic nature; where the magnitude of these (+ve) or (−ve) hydrophobicities represent the extent of hydrophobic or hydrophilic (respectively) nature of the atoms under consideration. Using these residue-specific atomic hydrophobicity magnitudes, we could calculate the hydrophobic center (HC) in the same way as we had calculated CM. The HC provides us with an idea as to where, in the interior of the protein, the entire effect of hydrophobicity due to all of its atoms may be assumed to be concentrated. The cumulative effect of the presence of these magnitudes give rise to the "hydrophobic center". Thus, the very perception of "hydrophobic center" accommodates the presence of hydrophobic (and hydrophilic) atoms with varying degrees of hydrophobicity (and hydrophilicity), whether the atoms are residing deep in the interior or completely exposed on the surface.

The concept of HC is novel and it differs notably from a previously conceived idea of 'center-of-the-protein' (Silverman 2005). This particular way of identifying protein's centrality is not only necessary to study protein stability on objective ground but also is capable of throwing light on protein folding. The necessity to include differential effects due to locally different hydrophobic environment within a residue had been felt earlier also (Rackovsky and Scheraga 1977); however capturing the effect of the same with an atomic level hydrophobic feature extraction scheme, to our knowledge, was never tried before. A systematic profiling of collective magnitudes of atomic hydrophobicity across the concentric shells around HC (and their differential distributions) could therefore acquire the necessary capability to analyze the remnants of so-called hydrophobic collapse (Dill et al. 1995; Ptitsyn 1996) from a new light. Furthermore, it could investigate the causality behind inhomogeneous mass packing in proteins.

In order to ensure a comprehensive view, along with the cumulative profiling of mass and hydrophobicity in all the shells, we chose to calculate two kinds of densities to study mass and hydrophobicity distributions. These densities were calculated for every shell of every protein under consideration. The classical measure of density was obtained by measuring mass and (separately) hydrophobicity per unit shell volume, whereas the other measure of density captured the total content of mass and (separately) hydrophobicity within any shell, normalized by number of atoms present in that shell. We denote the classical measure of density viz. mass/volume as 'density1' and the non-classical (normalized) density viz. mass/no. of atoms/volume as 'density2' from here on.

One of the determinants of the folds of the proteins is the (compact or sparse) packing of the units of secondary structures against each other. Increased mass packing can be expected to imply a loss of flexibility and an increment in various structural constraints (Banerjee et al. 2003) too. Hence, in our effort to construct an integrative framework, the distribution of structural constraints for extremophilic and mesophilic proteins, are taken into account too.

We all know that proteins in general can be divided in four broad structural sets with respect to disposition of secondary structural elements in them; the aforementioned structural sets being all-α proteins (comprised of α-helical domains), the all-β proteins (comprised of β-sheet domains), the α/β proteins (which consist of β–α–β structural units or "motifs" that form mainly parallel β-sheets) and $\alpha + \beta$ proteins (in which domains are formed by independent α-helices and mainly antiparallel β-sheets). Since we wanted to study the couplings between stability and structural parameters in their systematic details, segregating the complete set of proteins in four principal structural classes (viz. all-α, all-β, α/β and $\alpha + \beta$), the entire study was repeated. This piece of work produced a unique way to compare and contrast the packing characteristics (as well as the causality behind them) for the protein structural classes. Furthermore, since there is an ever-present need for an objective, and at the same time, automated toolbox for studying various biophysical properties in protein interiors with respect to four major protein structural classes, the present effort assumes importance. Popular methods such as Structural Classification of Proteins (SCOP) (Murzin et al. 1995) and Class, Architecture, Topology, Homologous super-family (CATH) (Orengo et al. 1997) involve visual inspection of structures as one of the key steps, while automated methods [such as knot theory-based techniques (Røgen and Fain 2003; Ramnarayan et al. 2008)] tend to overlook the involved biological aspects while characterizing parts (and whole) of the protein structure. The present framework required neither visual inspection nor abstract mathematics; with a firm footing on biophysical reality it studied mass (and hydrophobicity) packing within protein biological units, taking into consideration their densities (in classical as well as normalized manner) alongside the relevant set of structural constraints.

Materials and methods

The data

We have worked with the 'biological unit' of proteins taken from Protein Data Bank (PDB) (Berman et al. 2003), because it represents a functional protein molecule (the asymmetric unit of a protein, in contrast, is the fraction of the crystallographic unit cell that has no crystallographic symmetry; and thus may not be relevant in operative biological paradigm). Primarily our calculation was carried out on a large database of 373 structurally well-aligned (87.17%) protein pairs (DB1) drawn from thermophilic and mesophilic organisms (Glyakina et al. 2007). Later, we have extended our calculation to a dataset of 185 mesophilic proteins and 135 extremophilic proteins that were collected randomly from PDB (DB2). The predominant randomness in the DB2 dataset ensured the absence of any possible bias in it. DB2 includes datasets of Karshikoff and Ladenstein (1998) and Kannan and Vishveshwara (2000) in their entirety too, furthermore DB2 contained ~6% of DB1 proteins. The DB1 dataset is solely comprised of the thermophilic (and mesophilic) proteins, whereas DB2 set contains various other extremophilic (and mesophilic) proteins. By subjecting structurally aligned as well as structurally non-aligned protein sets to the same set of operations, a framework to capture widest possible set of information on protein interior packing was constructed. (Appendix 1 and 2 in Electronic supplementary material are devoted for information regarding DB1 and DB2 proteins, respectively).

Methodology

Mass and hydrophobicity distributions (with respective density studies)

The sole input for our algorithm was coordinate information of the biological units of the proteins. The CM for each of these proteins was calculated and they are all transformed from Cartesian to be represented in spherical polar coordinate system. The HC was calculated in identical way as that of CM by merely substituting atomic mass with residue-specific atomic hydrophobicity values. Considering CM and HC to be two separate origins, two separate sets of concentric shells were constructed with invariant shell width (we chose to use a fixed magnitude of 5 Å of radius vector as the shell width, to ensure that we do not miss out on any minute piece of information and yet the number of atoms within the shells always satisfies the statistical parametric limit). The atoms within each shell were identified. The cumulative mass content (using periodic table), cumulative atomic hydrophobicity content (Kuhn et al. 1995), density1 (mass/volume of the shell) and density2 (mass/number of atoms present in that shell) within each shell were calculated.

Structural constraints in protein interior

WHATCHECK (Hooft et al. 1996) server was later used to study distribution of structural constraints for all the extremophilic and mesophilic proteins of DB1 and DB2.

Studies of mass and hydrophobicity profiling on SCOP classes

Only proteins from DB1 database of 373 structurally well-aligned (87.17%) protein pairs (Glyakina et al. 2007) were segregated into four major SCOP classes (all-α, all-β, α/β and $\alpha + \beta$). Upon clustering these proteins into respective structural classes, all the studies regarding mass and hydrophobicity profiling (along with density profiles) are repeated. The DB2 proteins are not considered for this part (SCOP classification) of the study, the random nature of DB2 protein ensemble could have diluted the well-defined nature of protein sets (thermophilic and mesophilic) and therefore could have made it difficult to make structured inference out of structural classification based protein interior study.

Results

We clearly observe that for three sets of comparisons in Fig. 1 (viz. between DB1 thermophilic with DB1 mesophilic proteins, between DB2 thermophilic and corresponding mesophilic proteins and Karshikoff set of thermophilic and mesophilic proteins), the total mass content (TMC) for extremophilic proteins are significantly greater than that of the mesophilic ones. Figure 1 reveals first, a signature difference (>2,500 Da) in the magnitude of maximum TMC of extremophilic and mesophilic protein sets and second, the presence of a secondary peak at ~40 Å distance from the CM for DB1 as well as DB2 extremophilic proteins. While the prominent difference in the magnitude of maximum TMC of DB2 proteins can be attributed to the absence of structurally aligned protein pairs in them; existence of such distinct difference in TMC for DB1 proteins (373 structurally aligned pairs) appears to demonstrate the sensitivity of the proposed radial partitioning scheme and ascribe causality to some previous assertions regarding stability of thermophilic proteins (Jaenicke and Bohm 1998; Szilágyi and Závodszky 2000). The difference in shell-wise TMC for DB1 pairs can be seen to be statistically significant (at 99% confidence interval) for shells starting at 70 Å distance from the CM (please refer to 'Appendix 1 in

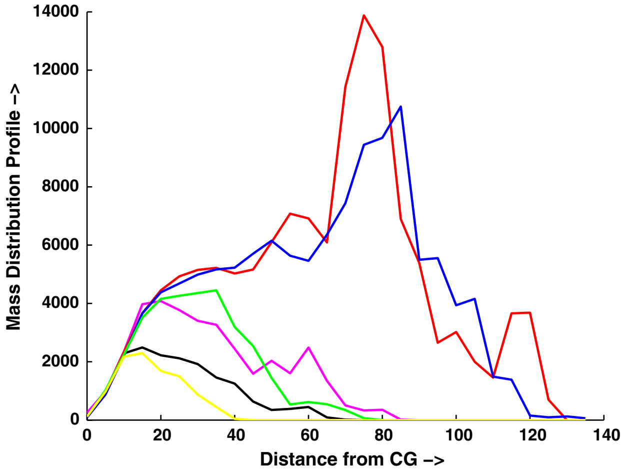

Fig. 1 Mass distribution profile. *Red line* thermophilic profile structurally-aligned database (DB1), *blue line* mesophilic profile structurally-aligned database (DB1), *pink line* extremophilic profile randomly collected database (DB2), *black line* mesophilic profile randomly collected database (DB2), *green line* thermophilic profile Karshikoff database, *yellow line* mesophilic profile Karshikoff database

Electronic supplementary material'). Since DB1 was solely comprised of structurally aligned pairs of mesophilic and thermophilic proteins, the mass and the radial extent for them were matching to each other. However, even under such comparable boundary conditions, the DB1 thermophilic proteins can be observed to have higher magnitude of TMC than their mesophilic counterparts. The capability of extracting such distinguishing information from an inherently congruent system, viz. DB1, proves the reliability of the present methodology.

Moving on to the other observation, while the presence of a secondary peak in extremophilic proteins from DB1 and DB2 appears prominent, even their mesophilic counterparts can be seen to possess the secondary peaks in their profiles of TMC distribution per shell. However the secondary peaks for the mesophilic proteins appear to be less distinct and can be observed to have a far less magnitude than the same for corresponding extremophilic protein sets. It is also evident from the result that even the thermophilic proteins of the Karshikoff dataset show the presence of this secondary peak in their mass distribution profile; however, for the mesophilic proteins of Karshikoff dataset such a secondary peak could not be observed, which might well be attributed to an outstanding difference in the average size of the proteins considered in Karshikoff dataset (Karshikoff and Ladenstein 1998).

Results from the total atomic hydrophobicity profile comparisons (Fig. 2) reveal similar structural details as depicted in Fig. 1. In conformance with TMC profile, the distinct presence of secondary maximas can be seen in the total atomic hydrophobicity content profile for the proteins extracted out of extremophilics (~120 Å for DB1, ~60 Å for DB2 and ~52 Å for Karshikoff extremophilic proteins). Perhaps this is not unexpected, as the correlation coefficient between the total mass content and the total atomic hydrophobicity content across the shells for the extremophilic proteins have been found to be of extremely high order (Table 1).

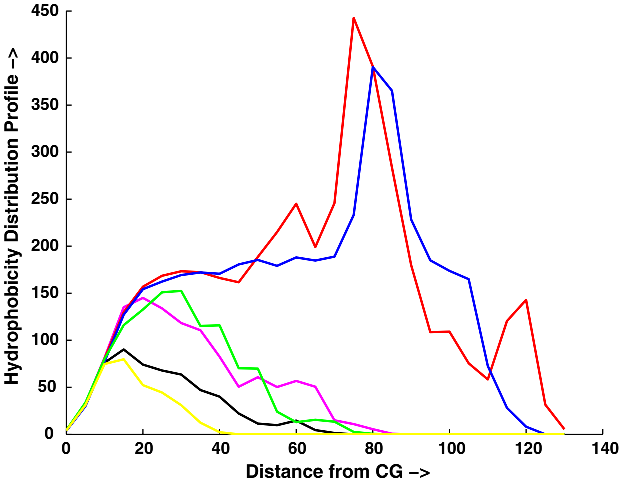

Fig. 2 Hydrophobicity distribution profile. *Red line* thermophilic profile structurally-aligned database (DB1), *blue line* mesophilic profile structurally-aligned database (DB1), *pink line* extremophilic profile randomly collected database (DB2), *black line* mesophilic profile randomly collected database (DB2), *green line* thermophilic profile Karshikoff database, *yellow line* mesophilic profile Karshikoff database

Studies on all the individual proteins of DB1 and DB2 revealed that the HC and CM didn't overlap on each other, although a consistent trend could be observed in their residing very close to each other (<3.0 Å). This tends to suggest that for all the proteins, the point in their interior where the effect of their entire mass content can be supposed to be concentrated (CM), happens to be in close proximity with the point where the effect of their entire hydrophobicity content can be supposed to be concentrated (HC). This observation points unmistakably to the fact that hydrophobicity provides the most important causality behind protein's stability.

The almost identical natures of density1 (Fig. 3 in the main text and Fig. 9 in Appendix 3 in Electronic supplementary material) and separately, density2 (Fig. 4 in the main text and Fig. 10 in Appendix 3 in Electronic supplementary material) distributions, from not only the structurally aligned protein pairs (DB1) but also from random ensemble of proteins (DB2) is remarkable; especially when viewed alongside the observed differences in total mass content (Fig. 1). However this can easily be explained by observing that the volumes of proteins (for which their radial extent is an unambiguous marker) did not show significant difference amongst not only the structurally aligned pairs but in DB2 ensemble too. While the similar nature of decaying density1 profile (Fig. 3 in the main text and Fig. 9 in Appendix 3 in Electronic supplementary material) imply

Table 1 Correlation between mass and hydrophobicity profiles of various distribution profiles

Between total mass profile and total hydrophobicity profile of DB1 extremophilic proteins	0.96
Between total mass profile and total hydrophobicity profile of DB1 mesophilic proteins	0.94
Between total mass profile and total hydrophobicity profile of DB2 extremophilic proteins	0.99
Between total mass profile and total hydrophobicity profile of DB2 mesophilic proteins	0.99
Between density1 mass profile and density1 hydrophobicity profile of DB1 extremophilic proteins	0.99
Between density1 mass profile and density1 hydrophobicity profile of DB1 mesophilic proteins	0.99
Between density1 mass profile and density1 hydrophobicity profile of DB2 extremophilic proteins	0.99
Between density1 mass profile and density1 hydrophobicity profile of DB2 mesophilic proteins	0.99

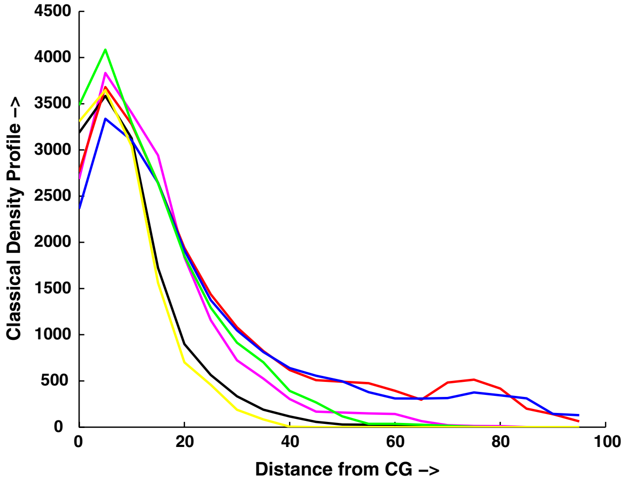

Fig. 3 Density1 distribution profile. *Red line* thermophilic profile structurally-aligned database (DB1), *blue line* mesophilic profile structurally-aligned database (DB1), *pink line* extremophilic profile randomly collected database (DB2), *black line* mesophilic profile randomly collected database (DB2), *green line* thermophilic profile Karshikoff database, *yellow line* mesophilic profile Karshikoff database

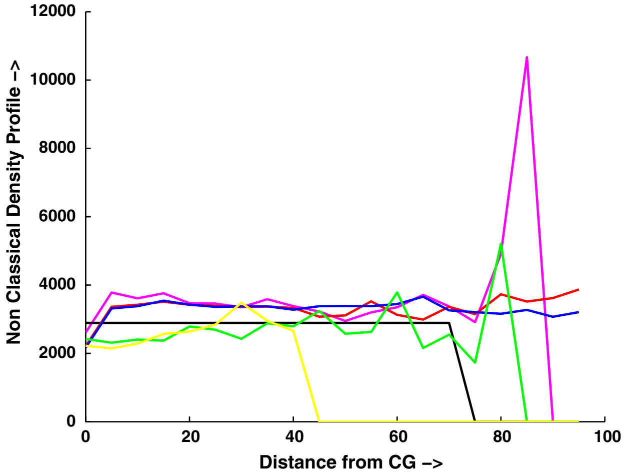

Fig. 4 Density2 distribution profile. *Red line* thermophilic profile structurally-aligned database (DB1), *blue line* mesophilic profile structurally-aligned database (DB1), *pink line* extremophilic profile randomly collected database (DB2), *black line* mesophilic profile randomly collected database (DB2), *green line* thermophilic profile Karshikoff database, *yellow line* mesophilic profile Karshikoff database

a comparable scheme of packing at work for extremophilic and mesophilic proteins alike, the capacity of mass being packed (Fig. 1) seem to differ between the protein sets. This finding points to the possibility of using TMC, instead of any kind of density profile, as potential marker for studying the packing difference between extremophilic and mesophilic proteins. It is relevant here to note that for either of extremophilic or mesophilic sets, density2 profiles (Fig. 4 in the main text and Fig. 10 in Appendix 3 in Electronic supplementary material) for all the compared sets do not show marked variation in abscissa.

The packing characteristics across SCOP classes (all-α, all-β, α/β and $\alpha + \beta$) revealed some extremely interesting results. While the maximum magnitude of total mass content (Fig. 5) [and total hydrophobicity content (Fig. 6)] for all-α thermophilic proteins are found to be significantly higher than the all-α mesophilic proteins (~1,200 Da), the difference between maximum magnitude of total mass content (Fig. 5) [and total hydrophobicity content (Fig. 6)] for all-β thermophilics and all-β mesophilic proteins was not that pronounced (~700 Da). Surprisingly, α/β and $\alpha + \beta$ class of proteins were found to show very little difference in maximum magnitude of TMC between the thermophilic and mesophilic proteins (Fig. 5), although the magnitude of the TMC for the former was found to be consistently higher for the higher order shells (>30 Å). In terms of the magnitude of the TMC, α/β proteins (thermophilic and mesophilic alike) were found to be most massive (~1,000 Da more than all-α and all-β thermophilic proteins), whereas $\alpha + \beta$ proteins (thermophilic and mesophilic alike) were found to be the least massive. This particular observation regarding α/β proteins vindicates the result of a recently conducted protein folding study (Galzitskaya et al. 2008) from a completely different angle.

Interestingly, the density1 (mass per unit shell volume, the classical measure of density) studies appeared to portray a different picture than the TMC distribution profiles. The density1 profiles (Fig. 7) for all-α, all-β and $\alpha + \beta$ proteins show more magnitude of the maximum for density1 of mesophilic proteins than the thermophilic ones, while for

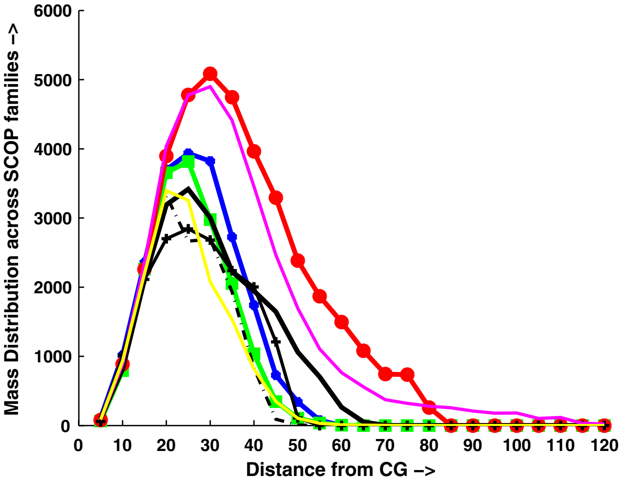

Fig. 5 SCOP-specific mass distribution comparison. *Blue* all_alpha thermophilic proteins, *green* all_beta thermophilic proteins, *red* alpha/beta thermophilic proteins, *black line* alpha + beta thermophilic proteins, *black plus* all_alpha mesophilic proteins, *black dashed dotted line* all_beta mesophilic proteins, *pink line* alpha/beta mesophilic proteins, *yellow line* alpha + beta mesophilic proteins

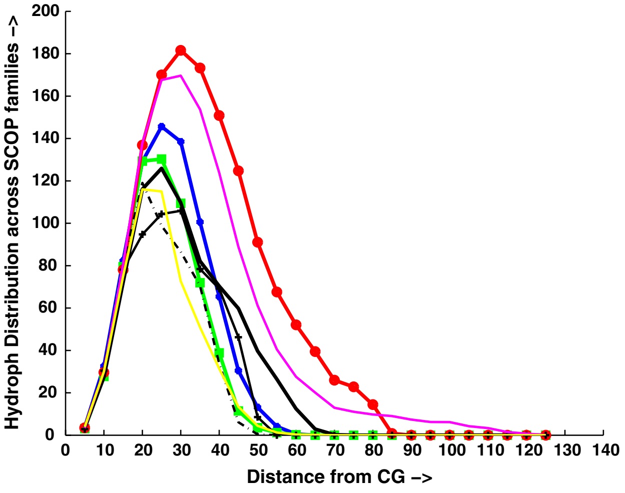

Fig. 6 SCOP-specific hydrophobicity distribution comparison. *Blue* all_alpha thermophilic proteins, *green* all_beta thermophilic proteins, *red* alpha/beta thermophilic proteins, *black line* alpha + beta thermophilic proteins, *black plus* all_alpha mesophilic proteins, *black dashed dotted line* all_beta mesophilic proteins, *pink line* alpha/beta mesophilic proteins, *yellow line* alpha + beta mesophilic proteins

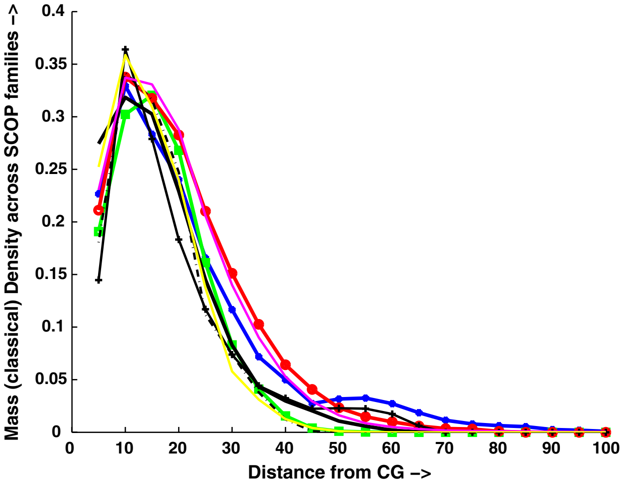

Fig. 7 SCOP-specific mass density1 distribution comparison. *Blue* all_alpha thermophilic proteins, *green* all_beta thermophilic proteins, *red* alpha/beta thermophilic proteins, *black line* alpha + beta thermophilic proteins, *black plus* all_alpha mesophilic proteins, *black dashed dotted line* all_beta mesophilic proteins, *pink line* alpha/beta mesophilic proteins, *yellow line* alpha + beta mesophilic proteins

α/β class of proteins, the maximum for density1 profile from mesophilic and thermophilic was found to be the same; with all of the classes attaining their maximum magnitudes at a distance 10 Å from the CM. Since this is not the same distance from the CM where the maximum of TMC profile is reached, we can infer that the shell in protein interior with maximum magnitude of density1 might not correspond to the shell that can accommodate highest mass and since the density profiles are not sensitive to capture structural differences, the TMCs should be used as the marker for mass packing studies. However this apparent anomaly can be resolved easily by observing the density2 (mass content normalized by number of atoms present in any shell) profiles (Fig. 8). For all-α, all-β and $\alpha + \beta$ proteins, the density2 profiles show a consistent trend of higher (almost enveloping) magnitudes for thermophilic proteins over the mesophilic ones, while for the α/β class of proteins, the mesophilic distribution almost resembles the thermophilic profile.

Discussion

The TMC profile for DB1 proteins can clearly be observed to be more than that of DB2 proteins. The observed shift in maxima of TMC along distances from CM for DB1 protein as compared to that of DB2 set, can be attributed primarily to the difference in molecular weight [(average molecular weight of DB1 extremophilic–average molecular weight of DB2 extremophilic proteins) >8,000 Da]; whereas [(average molecular weight of DB1 mesophilic–average molecular weight of DB2 mesophilic proteins) >15,000 Da]. However the pattern of higher magnitude of TMC for extremophilic proteins (over mesophilic analogues) remains invariant across 3 sets of compared data. This invariance tends to suggest the possibility of using this difference as a physical marker for studying mass packing amongst various protein families. While more magnitude of TMC maxima possibly implies an increased compactness scheme in extremophilic proteins in comparison to the

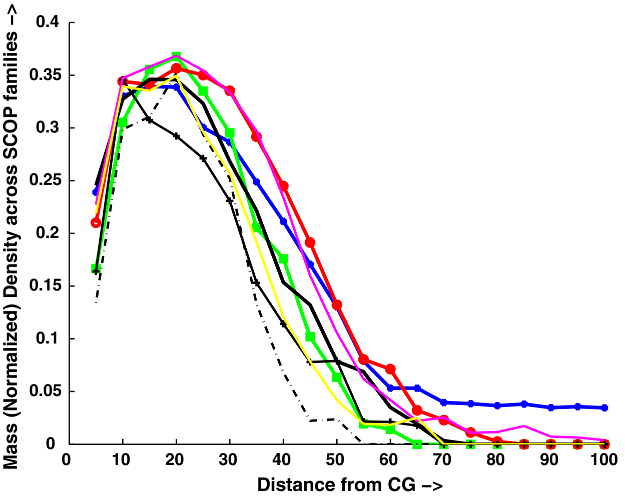

Fig. 8 SCOP-specific mass density2 distribution comparison. *Blue* all_alpha thermophilic proteins, *green* all_beta thermophilic proteins, *red* alpha/beta thermophilic proteins, *black line* alpha + beta thermophilic proteins, *black plus* all_alpha mesophilic proteins, *black dashed dotted line* all_beta mesophilic proteins, *pink line* alpha/beta mesophilic proteins, *yellow line* alpha + beta mesophilic proteins

mesophilic ones, the observation of consistent position of occurrence for secondary maximas in extremophilic TMC profile, clearly show a distinction in the packing schemes between extremophilic and mesophilic proteins. The (almost general) lack of prominence of these secondary peaks for mesophilic proteins (and a complete non-existence in the case of Karshikoff set of mesophilic proteins) further supports this assertion.

The observation of extremophilic proteins having higher magnitude of TMC helps us to view the enhanced stability of thermophilic proteins (Jaenicke and Bohm 1998) from a new perspective. It is interesting to note also, that an earlier study (from different motivation and methodology) had recognized that significantly high degree of compactness in thermophilic proteins to be the most influential physical parameter for their stability (Berezovsky and Shakhnovich 2005). We propose the use of TMC as an independent marker to study compactness (and stability) in proteins.

We note here that our finding of better mass packing in proteins drawn from thermophilic organisms, although contradicts the Karshikoff and Ladenstein (1998) assertion, is convergent with some early (Britton and Yip 1995; Kitty et al. 1998; Yoshihiro et al. 2002) as well as new (Berezovsky and Shakhnovich 2005; Cai et al. 2004) findings. However the essentially atomistic approach (we didn't study proteins at residue level) coupled with a radially symmetric partitioning of protein interior for extremophilic and mesophilic proteins, taking into account their structural classifications, made the present premise significantly different, inclusive and incisive. We suggest the TMC of the proteins to be the best markers for mass packing studies which, as have been observed here, can distinguish even between the structurally aligned protein classes what the density profiles have failed to achieve.

A closer scrutiny of results reveals the presence of extra stress in the realm of structural features of DB1 extremophilic proteins in comparison to that in DB2 extremophilic proteins (Table 2). This disparity in the intensity of structural constraints can well be attributed to first, the difference in molecular weight between DB1 and DB2 extremophilic proteins (noted already); and second, due to the packing of significantly higher magnitude of TMC (>8,000 Da) in DB1 extremophilic proteins. On the other hand, the twin observations that, even with 87% structural alignment, the DB1 mesophilic proteins have a noteworthy difference of TMC (>3,500 Da) between corresponding extremophilic proteins, and, former's enjoying a less stressful structural feature set (Table 2) tend to suggest better evolutionary mechanism in place for the mesophilic proteins as compared to their extremophilic counterparts.

An interesting aspect of our work is the study of atomic hydrophobicity based construction of hydrophobicity profile across the radially partitioned protein interior. Since the measure 'atomic hydrophobicity' takes into account the hydrophobic profile of the neighborhood of any atom within the amino acid residues, the description of hydrophobicity profile obtained with extensive application of such a measure had provided a rather unique view of the protein interior, notably different from an earlier work (Spassov et al. 1995). Investigations into total atomic hydrophobicity distribution (Fig. 2) within the proteins reveals almost identical characteristics observed in mass profile study and in fact, can be thought to attribute the causality to the existence of significantly different TMCs in the extremophilic proteins as compared to that in mesophilic proteins. The ground for this assertion of causality can be established from Table 1 data for correlation coefficient between TMC and total hydrophobicity content for both the DB1 and DB2 proteins. The existence of prominent local maxima in the total atomic hydrophobicity profile (Fig. 2) for extremophilic proteins and a near absence of that in case of mesophilic proteins, tend to suggest that proteins drawn from mesophilic organisms tend to be evolutionary better adapted than extremophilic proteins with respect to hydrophobicity (and mass) distribution preferences. The rationale behind this claim stems from the possible existence of many non-polar amino acids with understandable hydrophobic features in the vicinity of protein surface for

Table 2 Test for structural constraints

Proteins	Impr BL mean	Impr BL Var	Impr BA mean	Impr BA Var	Impr Dhd mean	Impr Dhd Var	Impr Om mean	Impr Om Var
Extremo (DB2)	0.45	0.21	0.78	0.27	0.44	0.67	0.42	0.31
Meso (DB2)	1.03	1.41	1.03	0.49	1.08	0.88	0.76	1.13
Extremo (DB1)	0.37	0.10	0.64	0.11	0.25	0.16	0.47	0.13
Meso (DB1)	0.72	0.65	0.97	0.37	0.57	0.43	0.85	0.49

Results from structural constraint analysis of the entire listing of extremophilic and mesophilic proteins, as provided by the WHATCHECK server are provided here. Ideally the mean scores mentioned should be equal to 1.00, which is almost precisely depicted by the mesophilic proteins. However the significant departure for the same in extremophilic proteins confirms their strained existence. Indeed unpaired 2-tail *t* tests confirm the difference between two sets in 99% interval. The ANOVA results also confirm that even the variances differ at 99% interval

BL Bond length, *BA* bond angle, *Impr* improper, *Dhd* dihedral, *Om* omega extremophilic proteins; which in turn might well be due to, imperfect adaptation and extreme evolutionary circumstances. The almost-nonexistent secondary peaks for the total hydrophobicity content profiles in proteins extracted from mesophilic organisms, by the same logic, can be considered as an evidence for their better evolutionary adaptiveness; because in them the non-polar, hydrophobic amino acids can be seen to be concentrated in the vicinity of respective hydrophobic centers.

The density2 profile offers us with interesting insights too. The remarkable peak(s) in the density2 profiles of DB2 extremophilics or Karshikoff thermophilics, for the very distant shells (>80 Å from CM) was not inconsistent with its definition. It is due to the presence of statistically insignificant number of atoms in these shells, which results in a low magnitude of the denominator, forcing density2 values to become too sensitive. The most notable aspect of density2 profiles is their near-constancy across the shells. While the possibility of smoothening of the finer aspects of mass and hydrophobicity distribution profile (Fig. 4 in the main text and Fig. 10 in Appendix 3 in Electronic supplementary material, respectively) due to huge ensemble of proteins cannot be ruled out, the observation of remarkably steady profiles of density2 can otherwise be explained too. It may well imply the presence of similar atomic hydrophobicity content (causing the nearly-unchanging mass content profile) across the radial extent of the proteins, albeit statistically. Such assertion appears to suggest that, to ensure stability of each shell, roughly same magnitude of total hydrophobicity content is necessary. However, that tends to differ from the so-called ‘hydrophobic collapse’ hypothesis. But this contradiction can easily be resolved in two ways. First, the density2 profiles (Fig. 4 in the main text and Fig. 10 in Appendix 3 in Electronic supplementary material) are obtained from huge datasets. Therefore a significant probability of presence of noise in data from every shell may well contribute to the overwhelmingly consistent profiles. Two, the density2 profiles only state that the proportionality given by the ratio [collective hydrophobicity content in shell (i)/number of atoms in that shell (i)], holds across all the shells, statistically. In fact, further studies on density2 profiles from different datasets have actually been found to reassert the ‘hydrophobic collapse’ hypothesis, and are presented later. The density1 profiles (Fig. 3 in the main text and Fig. 9 in Appendix 3 in Electronic supplementary material), owing to the uniform increment of the denominator (volumes of shells with progressively higher radius) undergoes an almost monotonic decay. However, a careful observation of (mass) density1 profile (Fig. 3) reveals the presence of a local (secondary) maxima in the higher order shells (~75 Å from the CM) for DB1 thermophilic proteins that shows the extraordinary magnitude of maximum mass content in the DB1 thermophilic proteins, so much so that even after being divided by $4/3\pi((75)^3-(70)^3)$ the resultant value remains significant.

Results from SCOP classification based studies demonstrate some novel facts. While the all-α mesophilic proteins are found to be least compactly packed, the α/β thermophilic (and mesophilic) proteins were found to be having most compact packing. The hydrophobicity distribution profiles revealed similar results too. This finding tends to vindicate a previous result (Yoshihiro et al. 2002) where they found that the number of contacts per residue for α/β protein is the maximum. Staying with the findings of the same study (Galzitskaya et al. 2008) we find the causality for attributing slowest folding rate to α/β proteins (probably because of constraints associated with accommodating huge TMC profiles for both thermophilic and mesophilic proteins) and the fastest folding rate to all-α proteins (the all-α mesophilic proteins have least TMC profiles to accommodate out of 8 cases). Consistent with the previous results, the TMC profile for most of shells is found to reveal more mass (and hydrophobicity) content for thermophilic proteins across all-α, all-β, α/β and α + β classes; proving that mass packing in thermophilic proteins is more compact than that in the mesophilic proteins, across all four major structural classes of proteins.

The ‘hydrophobic core’, as hypothesized by Rose (Behe et al. 1991), can clearly be observed for all four structural classes. The magnitude of hydrophobic packing is found to be the highest in the case of α/β class of proteins. For all-α, all-β and α + β class of proteins, although the existence of hydrophobic core can easily be verified, the magnitude of hydrophobic packing is observed to be less than that of α/β proteins. A careful observation reveals that the peak of the hydrophobic packing is reached at 20–25 Å (4th–5th shell) for all the four structural classifications of proteins. As the radial distance from the hydrophobic center increases beyond this limit (20–25 Å), the shell-specific hydrophobicity values decrease in a monotonic manner, for each of the four structural classes of proteins. This tends to indicate that nature has adopted a similar scheme across the structural classifications, which ensures optimal stability for proteins. Another similarity in the form of (almost) consistent trend of more hydrophobicity content for shells away from HC for thermophilic proteins over the mesophilic ones, can easily be explained in the light of formers need for extra stability.

Comparison between the density2 profiles between that in Fig. 4 and Fig. 8 makes an insightful observation. While the presence of large ensemble of proteins had resulted in a near-invariant profile for the density2 profile in Fig. 4, the uniformity of datasets (strict segregation into either of all-α, all-β, α/β and α + β classes) has ensured that minute features of SCOP class-specific behaviors are captured. For example, the density2 profile all-α proteins (thermophilic and mesophilic alike) are found to die down sharply than that for any other class, which tends to suggest the preference of all-α atoms to cluster near the CM, more than that in any other folds. On the other hand, the decay of density2 profiles for all-β and $\alpha + \beta$ class of proteins, tend to suggest in them a uniform rate of absence of atoms in the higher order shells away from the CM. The unchanging trend of enveloping profiles of thermophilic proteins over the mesophilic proteins in most of shells across all the four SCOP classes, imply in yet another way that the packing characteristics in thermophilic proteins is indeed far more compact than the same in mesophilic proteins.

Hence, in this paper, a comprehensive framework to study mass and density distribution with hydrophobicity and hydrophobic density distribution has been presented. Alongside this, the studies of structural constraints due to inhomogeneous packing scheme within protein interior have provided an idea of balancing forces in the system. The importance of working with biological unit of a protein has been elucidated in a recent work (Jefferson et al. 2006). The fact that our entire study was performed on the biological units of all the concerned proteins, lends it with a rigorous biological basis at the first place. Applying a simple yet biologically pertinent algorithm of radial partitioning, the differential nature of mass and hydrophobicity packing has been extensively studied here. This systematic set of studies from various approaches establishes firmly that mass (and hydrophobicity) packing scheme in the interior of thermophilic proteins is indeed more compact than the same in mesophilic proteins, contradicting a previous finding (Karshikoff and Ladenstein 1998). Mass distribution is studied alongside the hydrophobicity distribution throughout the length of this work. Hence a possible causality for typical characteristics of mass distribution profiles within thermophilic and mesophilic proteins, when considered in large ensembles and when segregated in four major SCOP classes, could be established. The very fact that this simple methodology could pinpoint the difference in packing schemes between 373 pairs of structurally aligned thermophilic and mesophilic proteins, and extract valuable information about the prevalent differences from structurally (nearly) indistinguishable system, proves the reliability and usability of it.

On a more important note, the present work constructs a holistic framework to study mass distribution within protein in correlation to the fold characteristic and hydrophobicity profile of them. This work has yielded a more biophysically inclined protein center, namely the hydrophobic center (HC). But rather than being restricted to the introduction of new measures and new markers, more importantly, this study has provided an integrative framework to analyze protein's mass inhomogeneity, interior mass and hydrophobicity distribution profile; so that mass packing, protein stability, protein folding; coupled with evolutionary pressure exerted by fold constraint, amounting to structurally stressed states - can all be studied through one integrative framework. It is therefore, not only new way of looking at structural reality prevalent at protein level but also one that is inclusive.

Acknowledgments One of the authors, Anirban Banerji, would like to thank Professor O. V. Galzitskaya for providing us with the translated copy of the original Russian paper and DBT (Department of Biotechnology, Govertment of India), for supporting this work. Anirban is also grateful to Nilesh Bairagi and Vivek Kumar Singh for helping him with technical know-how and to Dr. Devapriya Choudhury for his critical assessment of the present work.

References

Banerjee R, Sen M, Bhattacharya D, Saha P (2003) The jigsaw puzzle model: search for conformational specificity in protein interiors. J Mol Biol 333:211–226. doi:10.1016/j.jmb.2003.08.013

Beardsley D, Kauzmann W (1996) Local densities orthogonal to beta-sheet amide planes: patterns of packing in globular proteins. Proc Natl Acad Sci USA 93(9):4448–4453. doi:10.1073/pnas.93.9.4448

Behe M, Lattman E, Rose G (1991) The protein-folding problem: the native fold determines packing, but does packing determine the native fold? Proc Natl Acad Sci USA 88:4195–4199. doi:10.1073/pnas.88.10.4195

Berezovsky I, Shakhnovich E (2005) Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci USA 102:12742–12747. doi:10.1073/pnas.0503890102

Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide protein data bank. Nat Struct Biol 10:980. doi:10.1038/nsb1203-980

Bernal J, Finney J (1967) Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. Discuss Faraday Soc 43:62–69. doi:10.1039/df9674300062

Britton K, Yip K (1995) Insights into thermal stability from a comparison of the glutamate dehydrogenase from *Pyrococcus furiosus* and *Thermococcus litoralis*. Eur J Biochem 229:688–695. doi:10.1111/j.1432-1033.1995.tb20515.x

Cai G, Zhu S, Yang S, Zhao G, Jiang W (2004) Cloning, overexpression, and characterization of a novel thermostable penicillin G acylase from *Achromobacter xylosoxidans*: probing the molecular basis for its high thermostability. Appl Environ Microbiol 70:2764–2770. doi:10.1128/AEM.70.5.2764-2770.2004

Das R, Gerstein M (2000) The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics 1:76–88

Dill K (1990) Dominant forces in protein folding. Biochemistry 29:7133–7155. doi:10.1021/bi00483a001

Dill K, Bromberg S, Yue K, Fiebig S, Yee D, Thomas P, Chan H (1995) Principles of protein folding—a perspective from simple exact models. Protein Sci 4:561–602

Galzitskaya O, Reifsnyder D, Bogatyreva N, Ivankov D, Garbuzynskiy S (2008) More compact protein globules exhibit slower folding rates. Proteins 70:329–332. doi:10.1002/prot.21619

Glyakina A, Lobanov M, Galzitskaya O (2007) Search for structural factors responsible for the stability of proteins from thermophilic organisms. Mol Biol J 41:4

Haney P, Badger J, Buldak G, Reich C, Woese C, Olsen G (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc Natl Acad Sci USA 96:3578–3583. doi:10.1073/pnas.96.7.3578

Hermans J, Scheraga H (1961) Structural studies of ribonuclease. VI. Abnormal ionizable groups. J Am Chem Soc 83:3293–3330. doi:10.1021/ja01476a026

Hooft R, Vriend G, Sander C, Abola E (1996) Errors in protein structures. Nature 381:272. doi:10.1038/381272a0

Hubbard S, Gross K, Argos P (1994) Intramolecular cavities in globular proteins. Protein Eng 7:613–626. doi:10.1093/protein/7.5.613

Jaenicke R, Bohm G (1998) The stability of proteins in extreme environments. Curr Opin Struct Biol 8:738–748. doi:10.1016/S0959-440X(98)80094-8

Jefferson E, Walsh T, Barton G (2006) Biological units and their effect upon the properties and prediction of protein–protein interactions. J Mol Biol 364:1118–1129. doi:10.1016/j.jmb.2006.09.042

Kannan N, Vishveshwara S (2000) Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng 1:753–761. doi:10.1093/protein/13.11.753

Karshikoff A, Ladenstein R (1998) Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. Protein Eng 11:867–872. doi:10.1093/protein/11.10.867

Kauzmann W (1959) Some factors in the interaction of protein denaturation. Adv Protein Chem 14:1–63. doi:10.1016/S0065-3233(08)60608-7

Kitty S, Britton K, Stillman T, Lebbink J, Vos W, Robb F, Vetriani C, Maeder D, Rice D (1998) Insights into the molecular basis of thermal stability from the analysis of ion-pair networks in the glutamate dehydrogenase family. Eur J Biochem 255:336–346. doi:10.1046/j.1432-1327.1998.2550336.x

Kuhn L, Swanson C, Pique M, Tainer A, Getzoff E (1995) Atomic and residue hydrophilicity in the context of folded protein structures. Prot Struct Funct Genet 23:536–547. doi:10.1002/prot.340230408

Kuntz I (1972) Tertiary structure in carboxypeptidase. J Am Chem Soc 94:8568–8572. doi:10.1021/ja00779a046

Liang J, Dill K (2001) Are proteins well-packed. Biophys J 81:751–766. doi:10.1016/S0006-3495(01)75739-6

Murzin A, Brenner S, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J (1997) CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108. doi:10.1016/S0969-2126(97)00260-8

Privalov P (1996) Intermediate states in protein folding. J Mol Biol 258:707–725. doi:10.1006/jmbi.1996.0280

Ptitsyn O (1996) How molten is the molten globule? Nat Struct Biol 3:488–490. doi:10.1038/nsb0696-488

Ptitsyn O (1998) Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? J Mol Biol 278:655–666. doi:10.1006/jmbi.1997.1620

Rackovsky S, Scheraga H (1977) Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. Proc Natl Acad Sci USA 74:5248–5251. doi:10.1073/pnas.74.12.5248

Ramnarayan K, Bohr H, Jalkanen K (2008) Classification of protein fold classes by knot theory and prediction of folds by neural networks: a combined theoretical and experimental approach. Theor Chem Accnts Theory Comput Model 119:265–274

Richards F (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 82:1–14. doi:10.1016/0022-2836(74)90570-1

Richards F (1997) Protein stability: still an unsolved problem. Cell Mol Life Sci 53:790–802. doi:10.1007/s000180050100

Røgen P, Fain B (2003) Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci USA 100:119–124. doi:10.1073/pnas.2636460100

Rother K, Preissner R, Goede A, Frömmel C (2003) Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. Bioinformatics 19:2112–2121. doi:10.1093/bioinformatics/btg292

Silverman B (2005) Asymmetry in the burial of hydrophobic residues along the histone chains of Eukarya, Archaea and a transcription factor. BMC Struct Biol 5:20. doi:10.1186/1472-6807-5-20

Spassov V, Karshikoff A, Ladenstein R (1995) The optimization of protein solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions. Protein Sci 4:1516–1527

Szilágyi A, Závodszky P (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure 8:493–504. doi:10.1016/S0969-2126(00)00133-7

Ting K, Jernigan R (2002) Identifying a folding nucleus for the lysozyme/alpha-lactalbumin family from sequence conservation clusters. J Mol Evol 54:425–436. doi:10.1007/s00239-001-0033-x

Tsai J, Taylor R, Chothia C, Gerstein M (1999) The packing density in proteins: standard radii and volumes. J Mol Biol 290:253–266. doi:10.1006/jmbi.1999.2829

Xiao L, Honig B (1999) Electrostatic contributions to the stability of hyperthermophilic proteins. J Mol Biol 289:1435–1444. doi:10.1006/jmbi.1999.2810

Yoshihiro S, Susumu U, Yuji K, Yasuo I, Jun H (2002) Cytochrome c from a thermophilic bacterium has provided insights into the mechanisms of protein maturation, folding, and stability. Eur J Biochem 269:3355–3361. doi:10.1046/j.1432-1033.2002.03045.x

Zhang J, Chen R, Chao T, Liang J (2003) Origin of scaling behavior of protein packing density: a sequential monte carlo study of compact long chain polymers. J Chem Phys 118:6102–6109. doi:10.1063/1.1554395