Bioinformatics, 35(15), 2019, 2569–2577
doi: 10.1093/bioinformatics/bty1011
Advance Access Publication Date: 10 December 2018
Original Paper

OXFORD

Structural bioinformatics

Insights on protein thermal stability: a graph representation of molecular interactions

Mattia Miotto[1,2,3], Pier Paolo Olimpieri[1], Lorenzo Di Rienzo[1], Francesco Ambrosetti[1,4], Pietro Corsi[5], Rosalba Lepore[6,7], Gian Gaetano Tartaglia[8,9,10,*] and Edoardo Milanetti[1,2]

[1]Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00184, Rome, Italy, [2]Center for Life Nano Science@Sapienza, Instituto Italiano di Tecnologia, Viale Regina Elena, 291 00161 Roma (RM), Italy, [3]Soft and Living Matter Laboratory, Institute of Nanotechnology, Consiglio Nazionale delle Ricerche, 00185 Rome, Italy, [4]Bijvoet Center for Biomolecular Research, Faculty of Science – Chemistry, Utrecht University, Padualaan 8, Utrecht 3584CH, the Netherlands, [5]Department of Science, Università degli Studi "Roma Tre", via della Vasca Navale 84, 00146 Rome, Italy, [6]Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland, [7]SIB Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland, [8]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader St. 88, 08003 Barcelona, Spain, [9]Institucio' Catalana de Recerca i Estudis Avancats (ICREA), 23 Passeig Lluìs Companys, Barcelona 08010, Spain and [10]Department of Biology and Biotechnology, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 1, 2018; revised on October 29, 2018; editorial decision on November 26, 2018; accepted on December 7, 2018

Abstract

Motivation: Understanding the molecular mechanisms of thermal stability is a challenge in protein biology. Indeed, knowing the temperature at which proteins are stable has important theoretical implications, which are intimately linked with properties of the native fold, and a wide range of potential applications from drug design to the optimization of enzyme activity.

Results: Here, we present a novel graph-theoretical framework to assess thermal stability based on the structure without any *a priori* information. In this approach we describe proteins as energy-weighted graphs and compare them using ensembles of interaction networks. Investigating the position of specific interactions within the 3D native structure, we developed a parameter-free network descriptor that permits to distinguish thermostable and mesostable proteins with an accuracy of 76% and area under the receiver operating characteristic curve of 78%.

Availability and implementation: Code is available upon request to edoardo.milanetti@uniroma1.it

Contact: gian@tartaglialab.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Temperature is one of most crucial factors organisms have to deal with in adapting to extreme environments (Rothschild and Mancinelli, 2001) and plays a key role in many complex physiological mechanisms (Chen and Shakhnovich, 2010). Indeed a fundamental requirement to ensure life at high temperatures is that the organisms maintain functional and correctly folded proteins (Chen and Shakhnovich, 2010; Mozhaev *et al.*, 1996; Talley and Alexov, 2010). Accordingly, evolution shapes energetic and structural placement of each residue–residue interaction for the whole protein to withstand thermal stress. Studying thermostability is fundamental for several reasons ranging from theoretical to applicative aspects (Huang *et al.*, 2016), such as gaining insight on the physical and chemical principles governing protein folding (Amadei *et al.*, 2017;

Brinda and Vishveshwara, 2005; Robinson-Rechavi and Godzik, 2005), and improving the thermal stability of enzymes to speed up chemical reactions in biopharmaceutical and biotechnological processes (Chen *et al.*, 2017; Daniel, 1996).

Despite the strong interest in thermostability (Argos *et al.*, 1979; Bischof and He, 2005; Razvi and Scholtz, 2006), its prediction remains an open problem. As pointed out by Pucci *et al.*(2014) and Alfano *et al.*(2017), a complete characterization of the thermal properties of a protein is given by the knowledge of two contributions: (i) the thermodynamic stability, defined as the difference in free energy between the folded and unfolded states (ΔG) and (ii) thermal resistance, described by the melting temperature (T_m).

Here, we focus on the thermal resistance, distinguishing high and low thermal stable proteins on the basis of their T_m, experimentally defined as the temperature at which the concentration of the protein in its folded state equals the concentration in the unfolded state. To date, computational approaches, both sequence- and structure-based, have exploited statistical analysis (Amadei *et al.*, 2017; Pucci *et al.*, 2016, 2017), molecular dynamics (Manjunath and Sekar, 2013; Tavernelli *et al.*, 2003) and machine learning (Ku *et al.*, 2009; Wu *et al.*, 2009) to predict the melting temperature. Most of the studies are based on comparative analyses between pairs of homologs belonging to organisms of different thermophilicity (Mozo-Villarías *et al.*, 2003; Vogt *et al.*, 1997).

Predicting the stability of a protein *ab initio* using a structure-based approach has never been achieved so far. Lack of success in this area is mostly due to limitations in our knowledge about the relationship between thermal resistance and role of the interactions that stabilize a protein structure (Folch *et al.*, 2010). Some differences in terms of amino acid composition or spatial arrangement of residues have been reported (Amadei *et al.*, 2017; Vijayabaskar and Vishveshwara, 2010; Vishveshwara *et al.*, 2002). One of most notable differences involves the salt bridges: hyperthermostable proteins have stronger electrostatic interactions than their mesostable counterparts (Lee *et al.*, 2014). Recently Folch *et al.*(2010, 2008) reported that distinct salt bridges may be differently affected by the temperature and this might influence the geometry of these interactions as well as the compactness of the protein. Core packing seems related to thermal resistance at least to some extent (Vogt and Argos, 1997). Yet, a lower number of cavities and a higher average relative contact order (i.e. a measure of non-adjacent amino acid proximity within a folded protein) have been also observed while comparing thermostable proteins with their mesostable paralogs and orthologs (Robinson-Rechavi and Godzik, 2005). Noteworthy, the hydrophobic effect and residue hydrophobicity seem to play a rather marginal role on protein stabilization (Priyakumar, 2012; Van den Burg *et al.*, 1994), while they are considered the main forces driving protein folding.

Here, we present a new analysis based on the graph theory that allows us to reveal important characteristics of the energetic reorganization of intramolecular contacts between mesostable and thermostable proteins. In light of our results and to promote their application, we have designed a new computational method able to classify each protein as thermostable or as mesostable without using other information except for the 3D structure.

2 Materials and methods

2.1 Datasets

The T_m dataset, composed of proteins with known melting temperature (T_m), was obtained from the ProTherm database (Kumar *et al.*, 2006). Each protein of the dataset was accurately manually checked, in order to guarantee both the completeness of the structure and the reliability of the associated experimental melting temperature. The second dataset, consisting of proteins from hyperthermophilic organisms manually collected, is referred to as the T_{hyper} dataset. The union of the two dataset, referred as the T_{whole} dataset, accounts of 84 proteins (see Supplementary Material for details) and constitutes the largest structural dataset, to the best of our knowledge, of protein with well-defined thermal resistance, experimentally measured at physiological conditions.

2.2 Structural analysis

Proteins from both the T_m and T_{hyper} datasets were analyzed for their secondary structure content and architecture according to the CATH Protein Structure Classification database (Sillitoe *et al.*, 2015). Per residue secondary structure assignment was done using the DSSP software (Kabsch and Sander, 1983). See section in Supplementary Material for details.

2.3 Network representation and analysis

In this work, protein structures are represented as Residue Interaction Networks (RINs), where each node represents a single amino acid aa_i. The nearest atomic distance between a given pair of residues aa_i and aa_j is defined as d_{ij}. Two RIN nodes are linked together if $d_{ij} \leq 12$ Å (Phillips *et al.*, 2005; Vanommeslaeghe and MacKerell, 2012). Furthermore links are weighted by the sum of two energetic terms: Coulomb (C) and Lennard-Jones (LJ) potentials (see Supplementary Material for more details). Network analysis has been performed using the i-graph package (Csardi and Nepusz, 2006) implemented in R (Ihaka and Gentleman, 1996). For each RIN, the Strength local parameter (Barrat *et al.*, 2004) is defined as:

$$s_i = \sum_{j=1}^{N_{aa}^i} E_{ij} \tag{1}$$

where the Strength s_i of the i-esime residue is calculated as the sum of the energetic interactions (E_{ij}) between the residue i and all the other j residues contacting it (N_{aa}^i).

2.4 Network randomization

In order to distinguish mesostable from thermostable proteins, we compare the Strength calculated in the real RIN against the same parameter obtained from a random RINs. We defined a T_s score as:

$$T_s = \bar{s}_{protein} - (\bar{s} - \sigma) \tag{2}$$

to estimate how much the original RIN mean Strength value deviates from the expected mean value of rRIN distribution. $\bar{s}_{protein}$ is the average of the Strength parameter for the RIN; $\bar{s}$ and σ are the mean and standard deviation of the average values of the rRIN distribution. See Supplementary Material for details.

2.5 Performance evaluation

We evaluated the performance of the T_s score in discriminating between thermostable and mesostable proteins by a seven cross-validation. The mesostable proteins of the T_m dataset were divided in seven groups, guaranteeing that number of residues and T_m values were as broad distributed as possible. Details are reported in Supplementary Material.

2.6 Clustering and principal component analysis

We clustered the T_s descriptors using the Euclidean distance and the Ward method as linkage function (Ward, 1963) via the 'hclust' function of the 'Stats' package of R (Ihaka and Gentleman, 1996). Principal component analysis (PCA) was performed over eight graph-based descriptors using 'princomp' function of R software and the correlation matrix was used for the analysis (Venables and Ripley, 1997). Each descriptor has been computed using a specific function available in the R i-graph package. We refer to Supplementary Material for details.

3 Results

3.1 Uncovering the differences in energetic organization

Aiming at the comprehension of the basic mechanisms that allow proteins to remain functional at high temperature, we focused on the non-bonded interactions that play a stabilizing role in structural organization (Chakrabarty and Parekh, 2016). In particular, we considered only residue–residue interactions neglecting protein–solvent ones since a quantitative appraisal of their role would require a dynamical approach (Chong *et al.*, 2016).

To investigate how different thermal properties are influenced by the energy distribution at different layers of structural organization, we analyzed the interactions occurring in proteins of the T_{whole} dataset (see Section 2). To describe the role of single residues in the complex connectivity of whole protein, we adopted a graph-theory approach describing each protein by the RIN: each residue is represented as a node and links between residues are weighed with non-bonded energies (as described in Section 2).

At first, we investigated the relationship between thermostability and energy distribution of intramolecular interactions. To this end, the T_m dataset was divided into eight groups according to protein T_m and for each group the energy distribution was evaluated, as shown in Figure 1a. The general shape of the density functions is almost identical between the eight cases, independently from the thermal properties of the macromolecules, and this is clearly due to the general folding energetic requirements.

A strong dependence between thermal stability and the percentage of strong interactions is evident looking at the disposition of the density curves (Fig. 1a): the higher the thermal stability the higher the probability of finding strong interactions. Yet, less thermostable proteins possess a larger number of weak interactions. In particular, as shown in Figure 1a–c, it is possible to identify three ranges of energies that correspond to three peaks of probability density, i.e. a very strong favorable energy region ($E < -70$ kcal/mol), a strong favorable energy region between -70 and -13 kcal/mol, and a strong unfavorable interaction region ($E > 11$ kcal/mol). More formally, for a protein the probability of having an interaction with energy E, $P(E)$, in the three ranges linearly depends on the protein melting temperature with correlation coefficients of 0.90, 0.85, 0.87, respectively (Fig. 1b).

In order to have strong-signal sets, we reduced the division in just two groups, classifying proteins as mesostable or thermostable if their melting temperatures are, respectively, lower or higher than

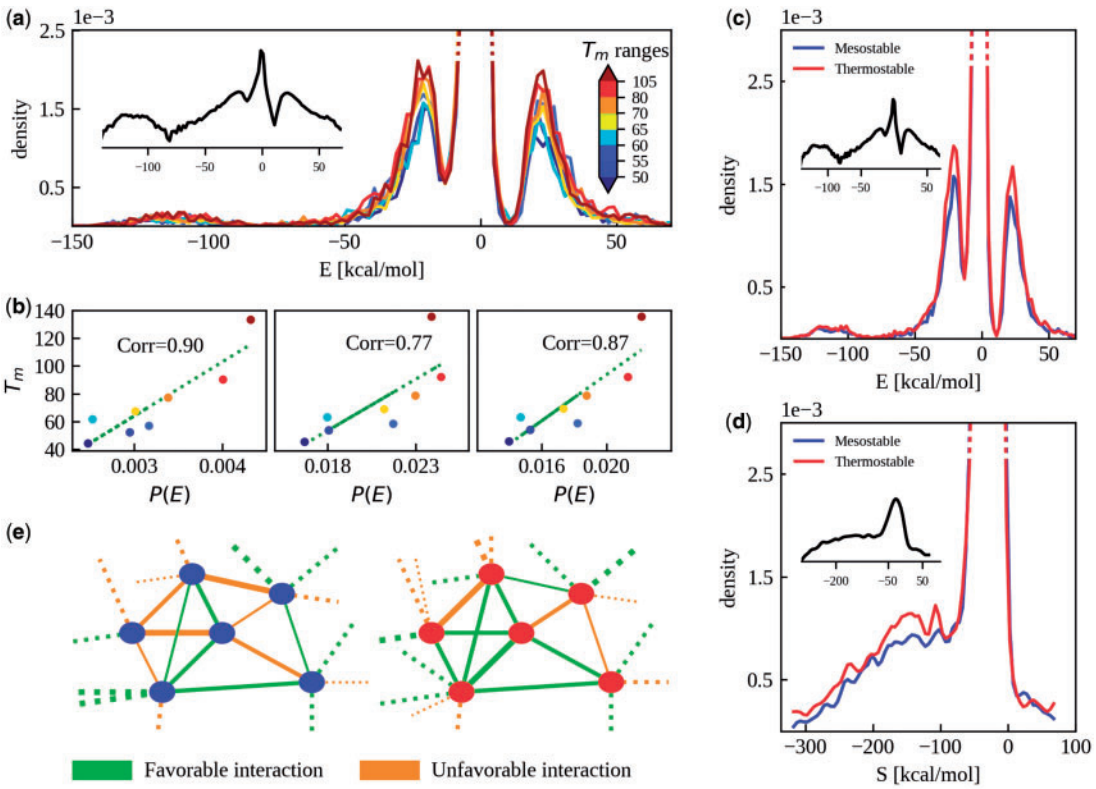

Fig. 1. (**a**) Probability density distributions of total interaction energies for the eight subsets defined in the T_m dataset from lower (dark blue) to higher (dark red) T_m. Each distribution is built using a group of proteins whose melting temperatures lie in the same range. The density functions exhibit a dependence with the melting temperatures ranges and peak heights increase with the temperatures. (**b**) Correlation between the area of each density peak and the average T_m for the eight groups. (**c**) Probability density distributions in log-scale of total interaction energies for mesostable (blue) and thermostable (red) proteins belonging to the T_{whole} dataset. (**d**) Probability density distributions in log-scale of Strength network parameter for mesostable (blue) and thermostable (red) proteins belonging to the T_{whole} dataset. Insets show the distributions in log-scale obtained using all proteins. (**e**) Schematic representation of the strong favorable and unfavorable interactions both for a mesostable (left) and a thermostable network (right) (Color version of this figure is available at *Bioinformatics* online.)

70^oC, which is the optimal reaction temperature of thermophilic enzymes (Brock, 1985; Serre and Duguet, 2003). In this way the energy distributions in Figure 1a are calculated only for the mesostable and thermostable distributions in Figure 1c (T_{whole} dataset). The two-group division allows us to include the hyperthermophilic proteins in our analysis, since their T_m is higher than the threshold. The two resulting distributions, found to be significantly different with a P-value of 4.2×10^{-46} [non-parametric test of Kolmogorov–Smirnov (Marsaglia *et al.*, 2003)], have an expected value at −0.5 kcal/mol and negative interactions have a probability of more than 60% to be found. Regions below −13 kcal/mol and above 11 kcal/mol represent the 6.6 and 5.2% of the total energy for thermostable and mesostable proteins, respectively. Typically, such energies require the presence of at least one polar or charged amino acid and in particular Arg, Asp, Glu and Lys are involved in more than 90% of the interactions. Noteworthy, the small fraction of energies centered near −120 kcal/mol (see Fig. 1a) is due to polar or charged amino acid interactions taking place at short distance. Next, we investigated the residue Strength, defined for each node as the sum of the weights of its links (see Section 2). The two Strength distributions for mesostable and thermostable proteins are shown in Figure 1d. Even in this case, they are different according to Kolmogorov–Smirnov test with a P-value of 1.9×10^{-9}.

For the first time, our analysis provides both a general intuition on the protein folding and a specific insight on thermal stability. Even if strong positive and strong negative peaks have a comparable height (Fig. 1c), the rearrangement of protein side chains masks the positive interactions, substantially preventing the condensation of unfavorable interactions in a single residue, as testified by the small probability of finding a residue with a positive Strength. Indeed, for the whole dataset there is more than 97% of probability of finding a residue with negative Strength. The most frequent value is found at −27 kcal/mol, with a change in the slope of the density functions around −70 and 5 kcal/mol, corresponding to the regions with negative and positive Strengths. At the Strength level of organization, a difference between thermostable and mesostable proteins is found. Indeed, residues belonging to the group of thermostable proteins show a higher probability of having high negative Strength values with respect to the mesostable ones, testifying an overall higher compactness of thermostable protein fold.

Figure 1e shows a schematic representation of the organization of strong energies both for mesostable proteins and thermostable proteins. In fact, the most important finding is that thermostable proteins have more favorable energies concentrated in a few specific residues. In contrast, mesostable proteins tend to have a less organized negative residue–residue interactions network. Given this different way to rearrange amino acidic side chains between proteins with different thermal properties, we mapped the energetic interactions between the protein secondary structures (helix–helix, helix–strand, helix–loop, strand–strand, strand–loop and loop–loop) in order to study how energetic allocation is reflected on a higher level of organization.

Looking at the difference in energy of a specific class of interaction with respect to the average, we found that thermostable proteins preferentially gather their energy through helix–loop interactions. These results suggest a stabilizing role for this class (see section in Supplementary Material for details).

3.2 Assessing protein thermal stability

In the light of our findings on the energetic difference between mesostable and thermostable proteins, we looked for a way to assess the thermal resistance of a protein given its structure. The simplest way to quantify the impact of energy distribution on the thermal resistance is the comparison with a protein of same structure but different energy organization, i.e. a homolog (Yang *et al.*, 2015). Ideally, differences between two homologous proteins with different thermal stability are attributable only to their different thermal resistance. The pronounced reorganization of the interactions in thermostable proteins confirms that they undergo an evolutionary optimization process which introduces fold-independent correlations in the spatial distribution of the interactions. By contrast, mesostable proteins do not have these correlations, thus with respect to thermal stability, their energy organization can be considered more random.

We designed a procedure that compares a given protein with modified versions of itself where protein structure is preserved, while chemical interactions have energies typical of mesostable proteins and randomly assigned in a physical way, i.e. maintaining residue–residue distance information (see Section 2). This randomization strategy provides a way to compare each real protein network with an ensemble of re-weighted cases, having the same number of nodes and links but with new weights (i.e. energies). These energies are extracted from the mesostable energy distribution using the interaction distance as constraint for the sampling. This procedure has the purpose of disrupting the effects of evolutionary optimization and is expected to have a larger effect on the highly organized network of thermostable proteins. By virtue of the different energy distribution between mesostable and thermostable proteins, sampling mesostable energies allows to properly assess the difference between the real thermostable protein network and its randomized counterpart. All steps of our method are schematically illustrated in Figure 2. In particular, given a link characterized by an energy weight E_{ij} and by a distance of interaction d_{ij}, we replaced the energy with a new one (E'_{ij}) extracted from an energy distribution defined for the specific distance interval d_{ij} belongs to. For each distance interval k, we generated a probability density function $\rho_k(E)$, using only the energies values observed in such interval in the mesostable proteins. At the end of the process, for each real RIN, we generated an ensemble of random networks (rRINs). The randomization allows us to develop a classifier based on the distance between the real network Strength and the random Strength distribution. The T_s score, defined in Equation (2) (see Section 2), is a measure of how much the original RIN average Strength value deviates from the expected average value of the rRIN distribution. Note that our descriptor is general and parameter-free and can be computed for every kind of weighted graph. The T_s score can be used as a thermal stability classifier setting the threshold value at 0; substantially considering true all predictions for which the T_s score is higher (resp. lower) than 0 and the protein T_m is higher (lower) than 70^oC. A so defined method is completely parameter-free. It only requires a probability density of mesostable protein interactions. In order to evaluate a possible dependence of the method from the chosen dataset, we performed a cross-validation (7-folds see Section 2) using the T_s score computed with total energy Strength. The method achieves an average accuracy of 72 ±3% with a mean receiver operating characteristic (ROC) curve characterized by an area under the curve (AUC) value of 80 ±2%. The small error on both the performances (due to the dimensions of the dataset) indicates the independence of the method from the input information.

Classifying on the threshold of the T_s score, i.e. considering the T_s as a binary variable, does not satisfactorily match with the information contained in the descriptor. In order to have a more sensible classification, we evaluated three different scores, using the total energy and specific interaction terms, i.e. the C and LJ interactions (see Supplementary Material), and performed a clustering analysis.

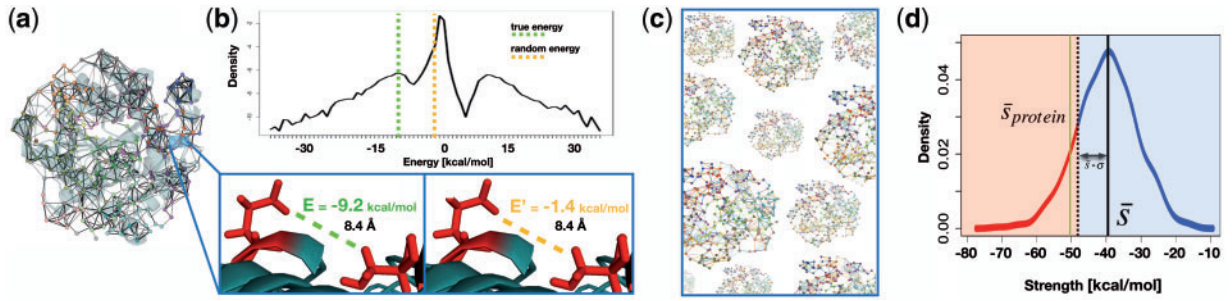

Fig. 2. Given a protein structure, our method represents it as a RIN (**a**). (**b**) The minimal atom–atom distance (8.4 Å in the example), for each residue pair, is calculated. The energy value (green line on the sketch) related to each contact is replaced with another one (yellow), randomly extracted from the energy distribution of mesostable protein contacts lying in the same distance interval (8–8.5 Å in the example). Performing this procedure for each pair, a new network of intramolecular interactions is established characterized by a new energy organization. Reiterating the process, we obtain an ensemble of random networks (**c**). (**d**) Finally, for each random network the average Strength parameter is calculated, obtaining a Strength distribution. Green line represents the mean Strength value of the real network, while red and blue region in the random Strength distribution show the classification criterion: if real Strength lies in red (resp. blue) region the protein is classified as thermostable (resp. mesostable) (Color version of this figure is available at *Bioinformatics* online.)

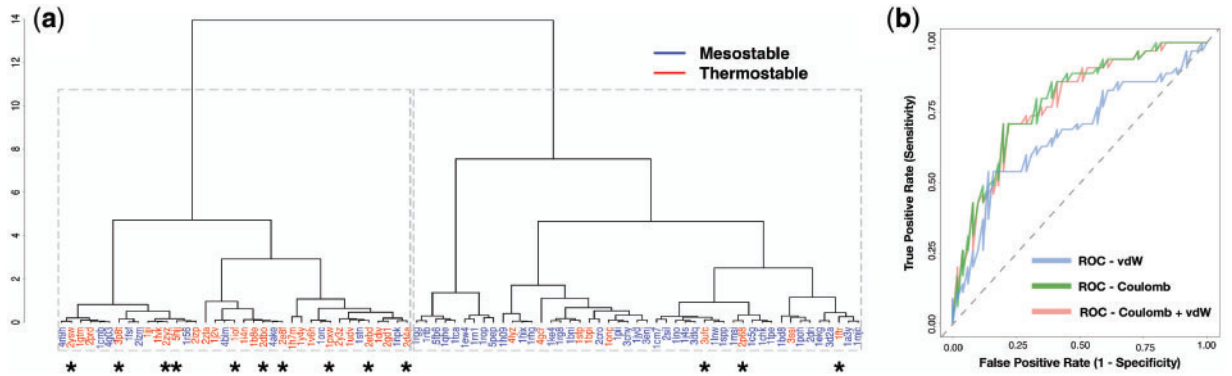

Fig. 3. (**a**) Cluster of the T_{whole} dataset proteins with three Strength based descriptors, i.e. C, LJ and total energy. Stars indicate proteins on the T_{hyper} dataset. The two groups are discriminated with a *P*-value of 2.6×10^{-6} (Fisher's exact test). (**b**) ROC curves of the three descriptors with the whole network T_s scores

Figure 3a shows the hierarchical clustering obtained clustering all the proteins of our T_{whole} dataset using the Ward method as linkage function while the Manhattan distance among the three descriptors was used as distance metric. We also tested different metrics and clustering methods obtaining very similar results (data not shown). The optimal clustering cut was estimated using both the Connectivity, Dunn and Silhouette parameters, which indicates the two group division as the optimal one. We called these groups 'Mesostable' (right group in Fig. 3a) and 'Thermostable' (left group). Indeed, the right cluster, containing 47 proteins, includes almost exclusively mesostable proteins (38), while the left cluster contains 26 thermostable proteins over the total 37 proteins. The overall accuracy of the method is 76%. We correctly assign the right thermal stability to 64 out of 84 proteins. The AUC of the ROC curve for the three T_s descriptors are 78, 79 and 68% (see Fig. 3b).

3.3 Key residues identification

Here, we investigated the thermal resistance properties of proteins at the residue level. As protein stability is the result of the cooperative effects and the synergic actions of several residues, assessing the specific contribution of each amino acid is difficult (Sadeghi *et al.*, 2006). We define the T_s^i score (see Supplementary Material), creating two groups of residues for each protein: with T_s^i lower or higher than zero. We consider residues belonging to the first group to have a more stabilizing role than the ones in the second group. Consequently, along the lines of the global-protein classification procedure, we defined 'thermostable' (respectively 'mesostable') residues belonging to the first (second) group. Using a total energy-based score, thermostable residues are the (11±4)% of total residues. In the C network (see Fig. 4a), the most frequent thermostable amino acids are the four charged amino acids: Arg, Asp, Glu and Lys, which cover the 96.6 and 96.1% of thermostable residues in thermostable and mesostable proteins, respectively. Apolar and aromatic residues (Leu, Met, Phe and Tyr) are typically thermostable residues of the van der Waals (vdW) network, including 53 and 54% of the total residues in mesostable and thermostable proteins, respectively (see Fig. 4b).

In order to investigate the role of each residue in the complexity of the whole system, we analyzed the properties of all residues using a graph-theory approach, calculating eight network parameters (see Section 2). A PCA was performed in both kinds of network. In Figure 4c and d, all residues were projected along the first two principal components. Thermostable residues are neatly separated from others if we consider the largest eigenvalue of the PCA in the C network and more weakly if we take into account the second and third ones (see Supplementary Material for details). Generally, charged residues form highly energetic electrostatic cages which prevent water inclusion (Levy and Onuchic, 2004; Sabarinathan *et al.*, 2011) while apolar and aromatic amino acids form short-ranged vdW interactions that confer stability to the overall structure (Lanzarotti *et al.*, 2011; Paiardini *et al.*, 2008). Here we identify key residues whose peculiar spatial disposition confers them a particular

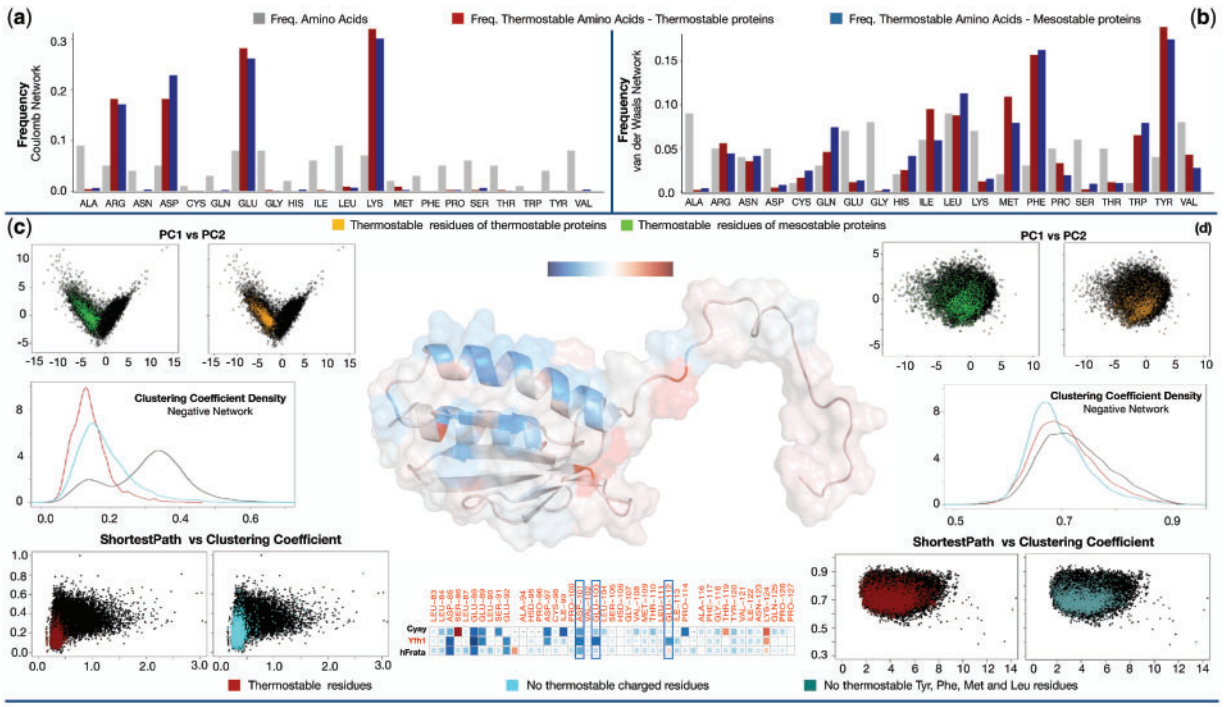

Fig. 4. (**a**, **b**) Frequencies of thermostable amino acids for the thermostable (red) and mesostable (blue). Frequencies of all the amino acids are shown in gray. (**c**, **d**) Projection along the first two principal components of all residues. Thermostable residues for mesostable (resp. thermostable) proteins are indicated in green (orange) dots. All residues are mapped in LC space. In red Arg, Asp, Glu and Lys amino acids are shown as the most frequent thermostable residues of the C network. In yellow dots, Tyr, The, Leu and Met are shown as the most frequent thermostable amino acids of vdW one. In the middle, cartoon representation of Yfh1 and multiple alignment with thermostable and mesostable residues colored in shades of red and blue (Color version of this figure is available at *Bioinformatics* online.)

role in the stabilization of the protein. Notably, our approach, based on a heterogeneous dataset, permits us to confirm and generalize the stabilizing role of both the charged and apolar/aromatic residues formerly suggested by homologous-based studies.

The mean shortest path (L) and the clustering coefficient (C) are able to catch the effect of the thermostable residues on maintaining these important structural motifs. The former provides information about the position of the residue in the network with the most central residues, having higher shortest path values. The latter quantifies the residue surrounding packing, being a ratio between the actual links and maximal number of possible links (Atilgan *et al.*, 2004; Vendruscolo *et al.*, 2001).

In Figure 4c (left panel), we projected all residues in the LC plane coloring in dark red the charged thermostable residues and in cyan the charged non-thermostable residues. Charged residues are concentrated in the region characterized by both small L and C values, with their thermostable subset tending to possess the smaller possible value of C. This means that thermostable residues have both to be exposed and surrounded by residue that makes low energetic interaction between each other's. In analogy with Coulombian networks, we projected in the LC plane the four kinds of key residues identified in the vdW networks. Even if the signal is weaker, key residues in the thermostable vdW network (Leu, Met, Phe, Tyr) tend to possess a higher clustering coefficient, testifying the packing stabilizing effect of vdW interactions. Densities of C parameter are found to be different with a P-value $< 10^{-16}$ (non-parametric test of Kolmogorov–Smirnov).

These findings allow us to divide residues in eight groups: four groups are identified by the C interaction, i.e. thermostable charged/uncharged residues and non-thermostable charged/uncharged residues; while vdW interaction networks divide residue according to thermostable/non-thermostable being or not being in the Leu-Met-Phe-Tyr group. For each protein of the T_{whole} dataset it is possible to compute the sum of the T_s^i scores in each of the eight possible groups, obtaining a vector of eight descriptors for each protein. Performing a linear regression with the four C-based vector component, the four vdW-based ones and with the whole eight-component vector we end up with a preliminary AUC of the ROC curves of 81, 77 and 83%, respectively (see Supplementary Material), and we are currently developing a residue-specific approach for T_m prediction.

3.4 Frataxin: a particular case of study

As a further application of our method, we investigated the stability of Yfh1, the yeast ortholog of frataxin. This highly conserved family of proteins is being deeply studied since in human it is responsible for the Friedreich's ataxia neurodegenerative disease. Furthermore, Yfh1 displays a very peculiar behavior in its thermal stability properties (Adrover *et al.*, 2010, 2012; Pastore *et al.*, 2007). In fact, several experimental studies show that both Yfh1' cold and heat denaturation occurs at experimentally accessible temperature in physiological conditions, at 5 and 35 °C, respectively. This is very rare since usually cold denaturation occurs at very low temperatures, below freezing water one, making this phenomenon very rarely observed in wild-type proteins.

To investigate the origins of the marginal stability exhibited by Yfh1, we compared the global and local thermal resistance analysis of Yfh1 with its bacterial (CyaY) and human (hFrata) orthologs, which are thermally stable until 54 and 58 °C under physiological conditions. Our global descriptor correctly classifies all three proteins as mesostable with a positive global T_s score. We then

proceeded to assess the local stability by computing the T_s^i for each residue (see Supplementary Table S5). In particular, we focused on one cluster of charged amino acids experimentally identified by Sanfelice *et al.* (2015). This cluster is composed by residues D101, E103 and E112, that interact between two different strands of the beta sheet (β_1 and β_2) and therefore are regarded as responsible for the structural stability. Interestingly, according to our T_s^i scores (see Fig. 4), all these residues are less stabilizing in Yfh1 with respect to CyaY and hFrata, despite their evolutionarily conservation. Another feature, the local analysis unveiled, is the presence of near-neighbor highly 'mesostable' and 'thermostable' residues, such as the ones at the beginning of the destabilizing flexible region of Yfh1 N-terminal loop (Adrover *et al.*, 2010) which are absent in CyaY and hFrata (see 3D structure in Fig. 4). The good agreement between our predictions and experiments shows the capability of the method to determine thermal stability properties both from a global and local point of view.

4 Discussion

Our work aims to represent a step toward the understanding of the thermal properties of a protein given its 3D structure. While the axiom thermophilic organisms have thermostable proteins is certainly correct, some mesophilic proteins may as well be thermostable (Pucci and Rooman, 2017). Knowledge on the organism optimal growth temperature, T_{env}, used to classify mesophiles and thermophiles, may be misleading with high value of correlation due to the fact that T_{env} is always a lower-bound for T_{m}.

The basic idea behind our method relies on the assumption that thermostable proteins undergo an optimization process during evolution that leads to specific structural arrangement of their energy interactions. Our analysis is based on a RIN in which the 3D structure of a protein is schematized as a graph with the residues acting as nodes and the molecular interactions as links. In our definition of network, links are weighted according to the sum of two non-bonded energetics terms: electrostatic and LJ potential. The analysis of the distribution of energies (links) highlighted the correlation between the thermal stability of protein sets (grouped according to their T_{m}) and the probability of finding high intramolecular interactions, with a highest correlation of 0.90 considering eight groups of proteins (Fig. 1). Unfortunately, neither it is possible to further divide the dataset in more groups due to the dataset dimension, nor we could not consider the energy distribution for the single protein because the small number of links makes the statistics noisy, especially in strong energy regions. Moreover, moving to higher orders of organization, e.g. considering the individual residual energies (Strength parameter), further reduces the data. For this reason, the next-up analysis was performed with a two-group division of the dataset.

Interestingly, we found that not only strong negative energies determine the thermal stability of a protein, but also strong positive interactions play a role. Such finding confirms the complex nature of the protein interaction network and in fact the stabilizing role of repulsive energies can be explained in cases where repulsion between a couple of residues results in a better spatial rearrangement of protein regions. To better grasp the role of favorable and unfavorable energies disposition, we determined the stabilizing contribution of each amino acid, defining the residue Strength [(Equation (1)]. Indeed, this parameter gives an estimate of the residue significance in the overall protein architecture and can be used both as a local property of each individual amino acid and as a global average network feature of the entire protein. Moving to the higher level of organization we investigated the biological role of the secondary structure interactions in thermal stability. The interactions between residues belonging to alpha helixes and loops concentrate more energy in thermostable proteins than mesostable ones. Those results suggest that the thermal stability of a given protein is deeply linked both to the intensity of interactions and to their spatial disposition, and that both are fine-tuned during the evolutionary process. In order to assess the thermal stability, we investigated the network energy organization and compared it against an ensemble of randomized networks. The ensemble comparison has two main purposes: The first consists in overcoming the limitation of the need of pairs of homologous proteins for direct comparison. The second purpose, raised from the observation that thermostable proteins are enriched of high connected nodes (hubs) and have more organized networks of interactions respect mesostable proteins (Jonsdottir *et al.*, 2014; Kumar *et al.*, 2000; Pucci and Rooman, 2016), relies in the need introducing a quantitative measure of the evolutionary optimization process thermostable proteins underwent, i.e. the distance between real protein interaction network and a randomized one, in which we disrupt the optimization of energy achieved by thermostable proteins during evolution. As described in the method section, the energies of a network are always obtained from a distribution of mesostable protein interactions. In this way, the more the original network diverts from the ensemble, the higher the probability that the protein belongs to the thermostable class. Moreover, the comparison allows us to assess in a quantitative way the effect of the energetic topology of the protein. Using this protocol to build up the T_s parameter-free descriptor and performing a cluster analysis, we are able to discriminate between mesostable and thermostable proteins, with a maximum accuracy of 76% and a maximum AUC of 78%.

At last, we investigated whether evolution acts on particular residues to optimize protein thermal stability or if stability is given by a cooperative effect with evolution acting on the whole protein. Our analysis identifies two sets of key (thermostable) residues according to the kind of energetic interactions the network is built with (C or vdW). Surprisingly, thermostable residue frequency in thermostable and mesostable proteins is comparable and they represent only a small subset of all residues.

This single residue approach allows us explore the local contributions to global stability and sheds light on peculiar cases of marginal thermal stability. In particular, we investigate the case of Yfh1 protein, the yeast ortholog of frataxin. Our global descriptor correctly classifies the protein as mesostable while our residue-based T_s^i descriptors allow us to identify stabilizing/destabilizing regions in agreement with previous works (Sanfelice *et al.*, 2015). In general, a complete description of the cold denaturation processes needs to explicitly include the water–residue interactions since it has been postulated (Privalov, 1990) and partially confirmed through molecular dynamics simulations at the specific unfolding temperature (Adrover *et al.*, 2012) such interactions play a paramount role in driving denaturation.

In order to better understand the theoretical aspects of thermostability and improve the classification to be used in more applicative fields, we created a new parameter dependent T_s score given by a linear combination of the T_s score of the eighth possible set of residues (see Section 3). The improved performance of 83% of ROC's AUC highlighted the promising features of the single residue approach.

Acknowledgements

The authors dedicate this article to the memory of Prof. Anna Tramontano, whose striking ideas lied the basis of the present work.

Funding

The research leading to these results was supported by Epigenomics flagship project EPIGEN. G.G.T. is funded by European Research Council [*RIBOMYLOME_309545*]; Spanish Ministry of Economy and Competitiveness [*BFU2014 – 55054 – P*, *BFU2017 – 86970 – P*]; and 'Fundació La Marató de TV3' [*PI043296*].

Conflict of Interest: none declared.

References

Adrover,M. *et al.* (2010) Understanding cold denaturation: the case study of yfh1. *J. Am. Chem. Soc.*, **132**, 16240–16246.

Adrover,M. *et al.* (2012) The role of hydration in protein stability: comparison of the cold and heat unfolded states of Yfh1. *J. Mol. Biol.*, **417**, 413–424.

Alfano,C. *et al.* (2017) An optimized strategy to measure protein stability highlights differences between cold and hot unfolded states. *Nat. Commun.*, **8**, 15428.

Amadei,A. *et al.* (2017) Density discriminates between thermophilic and mesophilic proteins. *J. Biomol. Struct. Dyn.*, **36**, 3265–3273.

Argos,P. *et al.* (1979) Thermal stability and protein structure. *Biochemistry*, **18**, 5698–5703.

Atilgan,A.R. *et al.* (2004) Small-world communication of residues and significance for protein dynamics. *Biophys. J.*, **86**, 85–91.

Barrat,A. *et al.* (2004) The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA*, **101**, 3747–3752.

Bischof,J.C. and He,X. (2005) Thermal stability of proteins. *Ann. N. Y. Acad. Sci.*, **1066**, 12–33.

Brinda,K.V. and Vishveshwara,S. (2005) A network representation of protein structures: implications for protein stability. *Biophys. J.*, **89**, 4159–4170.

Brock,T.D. (1985) Life at high temperatures. *Science*, **230**, 132–138.

Chakrabarty,B. and Parekh,N. (2016) NAPS: network analysis of protein structures. *Nucleic Acids Res.*, **44**, W375–W382.

Chen,P. and Shakhnovich,E.I. (2010) Thermal adaptation of viruses and bacteria. *Biophys. J.*, **98**, 1109–1118.

Chen,Y.-C. *et al.* (2017) Thermal stability, storage and release of proteins with tailored fit in silica. *Sci. Rep.*, **7**, 46568.

Chong,Y. *et al.* (2016) Protein dynamics and thermodynamics crossover at 10°C: different roles of hydration at hydrophilic and hydrophobic groups. *Chem. Phys. Lett.*, **664**, 108–113.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ.*, **1695**, 1–9.

Daniel,R. (1996) The upper limits of enzyme thermal stability. *Enzyme Microb. Technol.*, **19**, 74–79.

Folch,B. *et al.* (2008) Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. *J. Chem. Inf. Model.*, **48**, 119–127.

Folch,B. *et al.* (2010) Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. *Biophys. J.*, **98**, 667–677.

Huang,P.S. *et al.* (2016) The coming of age of de novo protein design. *Nature*, **537**, 320–327.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Jonsdottir,L.B. *et al.* (2014) The role of salt bridges on the temperature adaptation of aqualysin I, a thermostable subtilisin-like proteinase. *Biochim. Biophys. Acta*, **1844**, 2174–2181.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Ku,T. *et al.* (2009) Predicting melting temperature directly from protein sequences. *Comput. Biol. Chem.*, **33**, 445–450.

Kumar,M.D. *et al.* (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.

Kumar,S. *et al.* (2000) Factors enhancing protein thermostability. *Protein Eng.*, **13**, 179–191.

Lanzarotti,E. *et al.* (2011) Aromatic–aromatic interactions in proteins: beyond the dimer. *J. Chem. Inf. Model.*, **51**, 1623–1633.

Lee,C.W. *et al.* (2014) Protein thermal stability enhancement by designing salt bridges: a combined computational and experimental study. *PLoS One*, **9**, e112751.

Levy,Y. and Onuchic,J.N. (2004) Water and proteins: a love-hate relationship. *Proc. Natl. Acad. Sci. USA*, **101**, 3325–3326.

Manjunath,K. and Sekar,K. (2013) Molecular dynamics perspective on the protein thermal stability: a case study using SAICAR synthetase. *J. Chem. Inf. Model.*, **53**, 2448–2461.

Marsaglia,G. *et al.* (2003) Evaluating Kolmogorov's distribution. *J. Stat. Softw.*, **8**, 1–4.

Mozhaev,V.V. *et al.* (1996) High pressure effects on protein structure and function. *Proteins*, **24**, 81–91.

Mozo-Villarías,A. *et al.* (2003) A simple electrostatic criterion for predicting the thermal stability of proteins. *Protein Eng. Des. Sel.*, **16**, 279–286.

Paiardini,A. *et al.* (2008) "Hot cores" in proteins: comparative analysis of the apolar contact area in structures from hyper/thermophilic and mesophilic organisms. *BMC Struct. Biol.*, **8**, 14.

Pastore,A. *et al.* (2007) Unbiased cold denaturation: low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Am. Chem. Soc.*, **129**, 5374–5375.

Phillips,J.C. *et al.* (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.

Privalov,P.L. (1990) Cold denaturation of proteins. *Crit. Rev. Biochem. Mol. Biol.*, **25**, 281–305.

Priyakumar,U.D. (2012) Role of hydrophobic core on the thermal stability of proteins - molecular dynamics simulations on a single point mutant of Sso7d abstract. *J. Biomol. Struct. Dyn.*, **29**, 961–971.

Pucci,F. and Rooman,M. (2016) Improved insights into protein thermal stability: from the molecular to the structurome scale. *Philos. Trans. A Math. Phys. Eng. Sci.*, **374**, 20160141.

Pucci,F. and Rooman,M. (2017) Physical and molecular bases of protein thermal stability and cold adaptation. *Curr. Opin. Struct. Biol.*, **42**, 117–128.

Pucci,F. *et al.* (2014) Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. *PLoS One*, **9**, e91659.

Pucci,F. *et al.* (2016) Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoTMuSiC. *Sci. Rep.*, **6**, 23257.

Pucci,F. *et al.* (2017) SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, **33**, 3415–3422.

Razvi,A. and Scholtz,J.M. (2006) Lessons in stability from thermophilic proteins. *Protein Sci.*, **15**, 1569–1578.

Robinson-Rechavi,M. and Godzik,A. (2005) Structural genomics of thermotoga maritima proteins shows that contact order is a major determinant of protein thermostability. *Structure*, **13**, 857–860.

Rothschild,L.J. and Mancinelli,R.L. (2001) Life in extreme environments. *Nature*, **409**, 1092–1101.

Sabarinathan,R. *et al.* (2011) Water-mediated ionic interactions in protein structures. *J. Biosci.*, **36**, 253–263.

Sadeghi,M. *et al.* (2006) Effective factors in thermostability of thermophilic proteins. *Biophys. Chem.*, **119**, 256–270.

Sanfelice,D. *et al.* (2015) Cold denaturation unveiled: molecular mechanism of the asymmetric unfolding of yeast frataxin. *Chemphyschem*, **16**, 3599–3602.

Serre,M.-C. and Duguet,M. (2003) Enzymes that cleave and religate DNA at high temperature: the same story with different actors. *Prog. Nucleic Acid Res. Mol. Biol.*, **74**, 37–81.

Sillitoe,I. *et al*. (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res*., **43**, D376–D381.

Talley,K. and Alexov,E. (2010) On the pH-optimum of activity and stability of proteins. *Proteins*, **78**, 2699–2706.

Tavernelli,I. *et al*. (2003) Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation. *Biophys. J*., **85**, 2641–2649.

Van den Burg,B. *et al*. (1994) Protein stabilization by hydrophobic interactions at the surface. *Eur. J. Biochem*., **220**, 981–985.

Vanommeslaeghe,K. and MacKerell,A.D. (2012) Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model*., **52**, 3144–3154.

Venables,W. and Ripley,B. (1997) *Modern Applied Statistics with S-Plus*. 2nd edn. Springer-Verlag New York, New York.

Vendruscolo,M. *et al*. (2001) Three key residues form a critical contact network in a protein folding transition state. *Nature*, **409**, 641–645.

Vijayabaskar,M. and Vishveshwara,S. (2010) Interaction energy based protein structure networks. *Biophys. J*., **99**, 3704–3715.

Vishveshwara,S. *et al*. (2002) Protein structure: insights from graph theory. *J. Theor. Comput. Chem*., **01**, 187–211.

Vogt,G. and Argos,P. (1997) Protein thermal stability: hydrogen bonds or internal packing? *Fold. Des*., **2**, S40–S46.

Vogt,G. *et al*. (1997) Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol*., **269**, 631–643.

Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*., **58**, 236–244.

Wu,L.-C. *et al*. (2009) An expert system to predict protein thermostability using decision tree. *Expert Syst. Appl*., **36**, 9007–9014.

Yang,H. *et al*. (2015) Rational design to improve protein thermostability: recent advances and prospects. *ChemBioEng Rev*., **2**, 87–94.