

Theoretical Analysis and Computational Predictions of Protein Thermostability

Angel Mozo-Villiarías^{*,1} and Enrique Querol²

¹Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, E25198 Lleida, Spain

²Institut de Biotecnologia i Biomedicina i Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, E08193 Barcelona, Spain

Abstract: The interest in finding the keys to the thermal stabilization of proteins has remained constant and unquestionable throughout the last twenty years. This article reviews the most recent theoretical and computer advances related to the problem of thermally stabilizing proteins. Although comparison between mesophilic and thermophilic sequences has suggested some thermostabilization mechanisms, it has not been able 'per se' to provide unambiguous thermostabilization rules applicable for every case. Two of the mechanisms used by nature are seen as the major factors governing thermostability: the electrostatic forces of charged amino acids within a protein and the packing of its hydrophobic core on the other. Other mechanisms that have also been implicated (i.e. hydrogen bonding, α -helix stabilization, backbone rigidifying, etc), may play a refining role, based on the principle that nature has punctually and opportunistically thermostabilized proteins in each particular case, thereby solving each specific problem. How electrostatic and hydrophobic forces affect each other is still remains a largely open question and some recently developed criteria based on these two effects have been analyzed in the review.

Keywords: Protein thermostability, mesophiles, thermophiles, electrostatics, hydrophobicity, algorithms, theory.

INTRODUCTION

The possibility of maintaining the structure and function of a protein at a temperature above that of its native state, has been the objective of by many researchers ever since mutating a protein became a relatively easy process.

Interest in modifying the thermal properties of a protein is very broad since it ranges from purely theoretical aspects, such as those related to folding, through matters of biotechnological interest, to others directly applicable to food processing and/or the pharmaceutical industry. During the last twenty years a great quantity of experimental data about protein thermostabilization (see [1] for a review) has been produced and the reviewing of these results reveals the many different interests and approximations that lead to them. However, there has been no parallel progress with respect to theoretical predictions about thermostabilization. Most of the "positive" examples of thermostabilizations achieved were obtained on a case-by-case basis, employing semi empirical and/or computational methods seldom generalizable to other cases. Nevertheless, during the last five or six years, new generalizable algorithms and even new paradigms have appeared in this field and have begun to yield some systematic results.

For many authors, the most practical way of presenting data on the thermostabilization of a protein from an experimental point of view is with respect to transition temperature T_m . This corresponds to the temperature at which $\Delta G = 0$. When dealing with stabilizing (or

destabilizing) mutations, data are therefore expressed in terms of changes in T_m , as ΔT_m . Other authors prefer to present their stabilization data directly in terms of ΔG at a given temperature. When dealing with mutations, the same authors describe changes in terms of $\Delta\Delta G$, as changes in differences in free energy between the folded and unfolded states. $\Delta\Delta G$ (thermodynamic stability) is certainly related to T_m (thermal stability) as it is known that T_m depends on both ΔH and ΔS (and therefore so does ΔG at any temperature). When these magnitudes are changed (due to the introduction of a mutation), $\Delta\Delta G$ also changes. The use of ΔG to characterize protein stability is more general since it can be associated with any type (thermal, chemical, pH, etc) of destabilization process. It should also be also noted that most methods may prove seriously compromised at the highest range of temperatures (80°-90°C or higher). At these temperatures the thermostability is not the only problem: chemical damage may also occur, producing the irreversible inactivation of proteins. The best known mechanisms currently include: hydrolysis of peptide bonds in aspartic residues, thiol-catalyzed disulphide interchange, deamidation of asparagine and glutamine residues, β -elimination of disulphides, oxidation of free cysteines and methionines, etc.

This review deals with the state of the most relevant theoretical, informatic and predictive models during the last few years. The review is divided into three blocks: (i) Methods that directly compare mesophilic and thermophilic proteins in order to obtain general thermostabilization rules; (ii) Theoretical and computer algorithms that deal with all of the aspects that specifically confer upon a protein its thermal characteristics; (iii) Theoretical methods specifically dealing with electrostatic and/or hydrophobic effects. This is certainly an arbitrary division as there is a great deal of unavoidable overlap between the different blocks, but it is one that helps to make this subject more comprehensible to

*Address correspondence to this author at the Departament de Ciències Mèdiques Bàsiques, Facultat de Medicina, Universitat de Lleida, Avda Rovira Roure 44, E25198 Lleida, Spain; Tel: +34 973 702412; Fax: +34 973 702426; E-mail: angel.mozo@cmb.udl.es

the reader. As a matter of fact, the three blocks are no more than three aspects of the same question and all three are mentioned in virtually all theoretical and experimental papers published on the subject. The third block certainly contains aspects of the two and it has been singled out because of the growing sense of preponderance that these interactions have with respects to all aspects of protein stability. This evidence has grown with the increasing number of experimental instances of mutations being performed in the hydrophobic core of proteins.

The reader becomes increasingly conscious of the heterogeneity of approaches which is by no means a reflection on the fact that theorization about protein stability (and particularly, thermostability) still remains a relatively young field.

COMPARISON BETWEEN MESOSTABLE AND THERMOSTABLE PROTEINS

Because of its immediacy, one of the earliest strategies, from both a theoretical and an experimental point of view, involved the comparison of the structures and amino acid compositions of mesostable and thermostable protein isoforms especially in the case of homologous pairs. This is certainly a logical approach to the problem if one considers that it may constitute a set of strategies designed by nature to produce sturdier molecules capable of thriving under harsh conditions. In the world of protein engineering this has produced empirical procedures that have yielded relative good results. Most of the work associated with the development of these methods has centered upon statistical procedures. One classic and also pioneering work in this field is that of Vogt *et al.* [2]. These authors performed a statistical test upon 16 different families of proteins that each contained at least one thermophile. In 80% of the cases studied, there was a clear correlation with the thermostable character of the family in question found: the number of hydrogen bonds and the polar surface fraction related to the number of hydrogen bonds to water molecules are increased in thermophiles in comparison to mesophiles. In contrast, the correlation for electrostatic interactions was only that of two thirds in all the protein families. The analysis carried out in this paper has also provided a table listing the number of substitutions found that lead to thermostabilization (positive numbers) or destabilizations (negative numbers), for both α -helices and β -sheets. According to this table the most numerous favourable substitutions in α -helices were Gly to Ala and Val to Ala, whereas the most numerous unfavourable substitutions were Leu to Phe and Arg to Lys. In β -sheets, they are Val to Ile, and Ala to Gly, respectively. Along these lines, Karshikoff and Ladenstein produced another classic work comparing meso-thermo pairs of proteins [3]. This type of work was based on comparing global physical properties of proteins such as core packing and relationships between their size, volume and molecular weights. In this case in a comparative study of 80 proteins from mesophile organisms and a further 24 proteins from thermophiles, researchers found that the packing densities did not significantly differ, thereby effectively discarding the possibility that any observed difference in thermostability between proteins might be attributable to packing. In the opinion of these authors, only electrostatic interactions are responsible for regulating of the thermal stability of proteins.

More recently Parthasarathy and Murthy [4] have extended this comparison to the distribution and dispersion of B values in both type of proteins. In this study 96 mesostable and 21 thermostable proteins were used and it was concluded that in thermostable proteins, Ser and Thr have less flexibility than in mesostable proteins, whereas the proportion of Glu and Lys residues in high B-values is higher in thermostables than in mesostables. This is the opposite of what happen with Ser and Thr residues. These authors suggest that it should be possible to design thermostabilization strategies on the basis of these findings. A similar but more general conclusion was reached by Fukuchi and Nishikawa [5]. These authors, worked with a set of 144 mesophilic intracellular proteins, 47 thermophilic intracellular proteins, 59 mesophilic extracellular proteins and 29 nuclear proteins. After surveying interior and exterior protein composition in amino acids, the authors concluded that there is a bias towards a greater number of charged amino acids in thermostable than in mesostable proteins. A similar work with a somewhat more far reaching scope was conducted by Gianese *et al.* [6]. These authors treated a set of 28 proteins with related structures which include mesophilic, thermophilic and psychrophilic (adapted to low temperatures) organisms. Using a series of computer programs to compare sequences and fine structural data, these authors concluded that, generally speaking, each protein family tends to adopt its own strategy or mechanism. However, a common trend could also be observed: the lower the optimal temperature, the fewer ion pairs were present in the protein, the less side chain contributed to the exposed surface and the smaller the apolar fraction of the buried surface. More recently, after comparing all the relevant parameters in a set of 18 non redundant families of proteins, Kumar *et al.* [7] came to the conclusion that the significant differences observed between mesophilic proteins and thermophilic proteins do not lie in properties related with hydrophobicity or compactness or with other core-related characteristics. In general, it is the electrostatic properties and the number and arrangement of ion pairs that provide proteins with their resistance to temperature.

All these studies, some of which have been carried out for almost ten years, explain why electrostatic interactions play a preponderant role in protein thermostability. They have also been accompanied by a multitude of experimental work (see papers quoted in [1, 8, 9]).

ALGORITHMS AND THEORY

The development of special computer algorithms has been one of the most prolific fields for the development of protein stability issues has found more ground, as clearly evidenced by the articles commented bellow. In general, they are ways of generating informatic procedures aimed at optimizing a set of parameters related to the positions, the interactions, the boundaries and the character of the amino acids in order to obtain better stabilizing energies. This means that these algorithms range from the very simple, starting with relatively crude assumptions, to the very detailed and sophisticated in which numerous parameters and quite stringent conditions are applied. In the majority of cases their applicability is usually limited to the protein that specifically interests each author, although an increasing number of models now claim more general applicability.

Until now the predominant local scope of their applicability has tended to limit their success.

The work of Topham *et al.* [10], is very representative of the early attempts in designing new computer algorithms to detect differences between mesophilic and thermophilic proteins. These authors provided a simple strategy including the creation of a table of known amino acid replacements, which they translated into a table showing propensities, or rather the probability of propensities. The authors used this table to compute the stability difference score between the wild type and the mutant protein, in a similar way to which free energy differences are calculated. They applied this method to study 159 mutants of T4 lysozyme. Comparing their theoretical predictions with actual experimental results they obtain a score of 86%. In their attempt to explain the thermostability of α -helices, Petukhov *et al.* [11] use the AGADIR algorithm [12] which calculates the free energies associated with intrahelix interactions. These interactions include hydrogen bonds, helicity propensities, N- and C-capping interactions and electrostatic interactions, including the macrodipole constituted by the helix. They concluded that the most important factor in helical proteins was due to the ensemble action of all the helices rather than to that of any particular helix.

Malakausakas and Mayo [13] developed a general type of algorithm in order to design a hyperthermostable variant of immunoglobulin Gb1 from *Streptococcus*. This model is general in the sense that potentials for the relevant interactions (core packing, burial of hydrophobic surface, hydrogen bonds, helix-dipoles) take part, and may thus be applicable to other cases. This method is based on a fast discrete search algorithm, using the Dead-End Elimination Theorem (DEET) [14] with which the authors are able to make predictions relating to melting temperatures beyond 100°C. These authors argue that a whole set of factors take part in the higher stability of proteins.

Gilis and Rooman [15] have recently designed a statistical procedure called PoPMuSiC [16] based on the performance of all possible mutations in a protein. These authors estimated the stability of each mutant by means of linear combinations of data-derived potentials. These potentials were derived from observations of the frequencies of sequence and structure patterns of high resolution X-ray structures. These potentials must be combined with coefficients that in turn depend on the water accessibility of the mutated residues. They assessed the performance of this algorithm by comparing calculated and experimental $\Delta\Delta G$. The errors reported were within the order of magnitude of those obtained experimentally, except in the case of buried residues for which errors increased with the differences in size between native and mutated residues.

Carter *et al.* [17] worked in a similar line of research, relating the packing of a protein to the hydrophobic character of the core. These authors searched for correlations between four-body likelihood potentials obtained from what they call "Simplicial Neighbourhood Analysis of Protein Packing", (SNAPP) and $\Delta\Delta G$ of unfolding (both computed and measured). In five chosen proteins (lysozyme, barnase, nuclease, CI2 and calbindin), they found acceptable linear correlations ($r \approx 0.8$ on average) between these magnitudes, but with specifically different slopes for each protein. This

method allowed these authors to make predictions about the effect of mutations on the hydrophobic core that are based on observations relating to variations in SNAPP.

López de la Paz *et al.* [18] used the protein-design algorithm PERLA (Protein Engineering Rotamer Library Algorithm) in order to design stabilizing and destabilizing mutations in the 'de novo' designed protein *betanova*. This algorithm builds selected amino acids in a polypeptide structural template, by using a custom-made side-chain rotamer library, built from the analysis of a protein database. This algorithm computes interaction energies contributed by van der Waals, hydrogen bonding, and electrostatics. The authors found good level of agreement between their predictions and the experimental data for most of the peptides under study, and specially for those with a high β -sheet content.

The work of Mooers *et al.* [19] follows the line of the automatic design (implemented with program ORBIT: Optimization of Rotamers By Iterative Techniques). This program is based on the Dead-End Elimination Theorem [14] for the optimization of energies. This program which uses van der Waals interactions, hydrogen bonds, electrostatic and solvation energies, generates mutations with more favourable energies of stabilization through designing alternative cores in lysozyme T4. These mutants were also experimentally created but only a few of them resulted in species that were more stable than the wild type. These authors concluded that core packing is protein specific and that most changes affecting protein cores have a destabilizing effect.

Lee and Wand [20] analyzed the importance of internal movements within a protein pertinent to basic aspects controlling folding, thermal stability and function. Analyzing the amplitudes of these movements they discovered a heterogeneous spectrum that could be divided into three general classes. They also found that the distribution of the residual entropy was heterogeneous, a fact that revealed the microscopic origin of the heat capacity of proteins. These observations help explain the low-temperature glass transition of proteins. This transition is associated with the motional modes which is what gives support to biological activity.

Another basic study of interest with respect to the energies implicated in transitions is that carried out by D. Poland [21]. This theoretical work describes the relationship between enthalpy in proteins and the variations in free energy associated with temperature and heat capacity. The work develops a free energy function which describes the thermodynamic behaviour throughout the denaturation process.

There are several reports of successful approaches for large numbers of proteins and mutants. Based on the well known Connolly method [23], which computes accessible surface area (ASA) and cavity volumes, Funahashi *et al.* [22], calculated the $\Delta\Delta G$ of 110 variants of T4 and human lysozymes. The correlation between estimated and experimental $\Delta\Delta G$ s provided fairly good correlations. One of the most general purpose algorithms created to predict the stabilization of a protein is Guerois *et al.*'s FOLDEF (FOLD-X energy function) [24]. This algorithm, which was checked

over 42 proteins and more than 1000 mutations, calculates $\Delta\Delta G$ ($\Delta G_{WT} - \Delta G_{mut}$). Its contributions to ΔG include packing (van der Waals), solvation, electrostatic and entropic terms. These energy contributors are scaled to what the authors call atomic occupancies, according to a scale estimated by Holm and Sander [25]. When comparing calculated $\Delta\Delta G$ s with those experimentally determined, Guerois *et al.* obtained linear regressions factors in the order of 0.8 or better. This is perhaps one of the most complete and general algorithms tested with a large set of proteins. These authors began an interesting discussion by comparing their results with those of other authors, such as Funahashi *et al.* [22], who applied their algorithms to large numbers of proteins and mutants. These authors mainly based their calculations on ASA (accessible surface area) and the reason for their results not being as good as those obtained by FOLDEF is that the FOLDEF method performed a more complete calculation with respect to all aspects of the occupancy of the atoms. When comparing the FOLDEF results with those obtained by Gilis and Rooman (PoPMuSiC [15]), it was noted that the latter placed special emphasis on the solvent accessibility when computing ΔG . This probably explains why PoPMuSiC performs better than FOLDEF when dealing with residues with accessible surface areas higher than 50%, whereas the opposite is true for accessible surface areas below 50%.

In this strictly theoretical work, England *et al.* [26] developed a criterion for what they called “folding designability”; in other words, the number of amino acid sequences that can adopt a certain fold in a stable manner. They used a new theoretical concept: contact trace, in order to compare mesostable protein folds with their thermostable counterparts. This contact trace is related to the contact free energy found in specific sequences of amino acids. As a consequence, larger trace contacts imply more negative values for free energies for those sequences. In thermostable proteins they found folds with larger contact traces. These authors argue that the adaptation to high temperatures is achieved through more designable folds.

Ramos de Armas *et al.* [27] developed a sophisticated computational method, Markovian Backbone Negentropies (MBN) in order to predict protein stability. This method is based on the description of entropies for charge distributions over the whole protein molecule and throughout time. The method was designed to classify proteins by mutating the alanines of the Arc repressor protein according to their respective thermal stabilities. It had a predictive success of 81% with the method correctly predicting the stability of 71% of proteins whose stabilities were close to those of the wild type, and predicting the stability of 92% of those proteins of reduced stability. This algorithm fared well with respect to other algorithms also mentioned in their paper.

The work of Burioni *et al.* [28] should be mentioned within the field of theoretical research considering the energy involved in stability. They computed the harmonic spectrum of energy by applying the Gaussian Network Model (GNM). This work pinpointed a particularly significant parameter: the spectral dimension that describes the low frequency behaviour of the energy. This low frequency energy turned out to be proportional to the number of amino acids in the protein. The authors concluded that a

protein of a given length tends to adopt the most swollen state compatible with its thermodynamic fluctuations. In a study on folding thermodynamics, Irbäck and Mohanty [29] reported a relationship between folding and thermostability through a simplified interaction at the atomic level. Its potential was tested in a set of small peptides (about 20 residues) which contained helices, sheets and Trp cages. Contrasting folding simulations of the model with experimental data, these researchers were able to obtain relative good predictions for the thermal dependence of the Trp cages and helices studied (depending on whether they contrasted CD or IR data). However, the best results obtained were those that referred to beta sheets.

Bloom *et al.* [30] proposed an interesting method for predicting about how the number of mutations carried out on a particular protein affect its subsequent stability and even its own viability. These authors describe an algorithm that uses what they call the ‘*m*-neutrality’. This relate to the probability of the $\Delta\Delta G$ produced in a mutation being more destabilizing than that associated with the wild-type, over any number of mutations performed on the protein. The most interesting aspect of their observations is the notion that a protein can gain thermodynamic strength from the first few mutations.

In the field of sequence alignment, Zakrzewska *et al.* [31] described a homology consensus approach for the thermostabilization of the human acidic fibroblast growth factor (FGF-1). The study of this protein promises certain applications in biomedicine but its applicability implies serious difficulties on account of its instability. These authors managed to stabilize FGF-1 on the assumption that the rapidly growing number of available protein sequences and their consensus multi-alignment could help to predict protein stability better than the non-consensus amino acids.

This section is summarized in Table 1.

Molecular Dynamics (MD)

We decided to insert this subsection on Molecular Dynamics within the section on algorithms and computer simulations of protein thermal behaviour, as it constitutes an extended methodology that has created its own idiosyncrasy. The increase in computational power will further promote this approach which calls for important computer resources.

The group of Horii *et al.* [32] employed the free energy perturbation method and molecular dynamics to study the consequences of mutating the amino acids buried most deeply in the core of a protein. They then compared the results obtained using their own method with crystallographic data relating to goat α -lactalbumin. In a simulation of the denaturation of the native protein and its stabilizing mutations: T29V and T29I these authors compared theoretical RMSF profiles (by MD) with X-ray experimental results in order to explain the role of the O-H group of Thr29 in maintaining the stability of this protein.

Funahashi *et al.* [33], basing their work on MD simulations computed all of the components of $\Delta\Delta G$ changes in order to explain the different stabilities exhibited by human lysozyme and its three mutants, I56A, I56V and I56F. These authors described differences in free energy in terms of local potentials and longer range potentials. In explaining

Table 1. Summary of Algorithms Used by the Authors Reviewed in this Paper

| Algorithm | Authors | Protein(s) | Features | Accession web address |
|----------------------|---|--|--|---|
| Propensities | Topham <i>et al.</i> [10] | T4 lysozyme, 32 mutants | Computation of stability scores vs. ΔG | SDM program available from the the authors upon request |
| AGADIR | Muñoz, Serrano [12] | Helices of Bacterial RecA | Comparison of meso-thermo families | |
| DEET | Malakauskas, Mayo [13] | Streptococcal G81 | Comparison of $\Delta\Delta G$ | |
| PoPMuSiC | Gilis and Rooman <i>et al.</i> [15, 16] | 296 mutants in 7 proteins and synthetic peptide, | Comparison of $\Delta\Delta G$ | http://babylone.ulb.ac.be/popmusic |
| SNAPP | Carter <i>et al.</i> [17] | 5 proteins, 76 mutants | Comparison of $\Delta\Delta G$ | http://mmlsun4.pha.unc.edu/3dworkbench.html |
| PERLA | López de la Paz <i>et al.</i> [18] | Betanova, 13 mutants | Comparison of $\Delta\Delta G$ | http://ProteinDesgn.EMBL-Heidelberg.de |
| ORBIT | Mooers <i>et al.</i> [19] | T4 lysozyme, 15 mutants | Repacking the core | |
| | Lee, Wand [20] | Calmodulin, Human adipocyte Fatty-acid-binding protein, Human Ubiquitin Phospholipase | Temperature dependence of internal motions | |
| | Poland [21] | Barnase, Tendamistat | Enthalpy computations | |
| FOLDEF | Guerois <i>et al.</i> [24] | 42 proteins, 1033 mutants | Comparison of $\Delta\Delta G$ | http://fold-x.embl-heidelberg.de |
| | England <i>et al.</i> [26]□ | ~20 amino acid sequences | Comparison of meso-thermo sequences | |
| MBN | Ramos de Armas <i>et al.</i> [27] | Arc repressor, 27 mutants | Entropy changes | |
| GNM | Burioni <i>et al.</i> [28] | 59 proteins | Low frequency of density modes | |
| | Irbäck, Mohanty [29] | 2 α -helices, 5 β -sheets | Energy contours | |
| <i>m</i> -neutrality | Bloom <i>et al.</i> [30] | TEM1 β -lactamase, lattice proteins | Comparison of <i>m</i> -neutrality scores | |
| | Zakrzewska <i>et al.</i> [31] | Fibroblast growth factor, 16 mutants | Comparison with CD measurements | |

the discrepancies in stability values (in terms of $\Delta\Delta G^{\text{sp}}$ and $\Delta\Delta G^{\text{emp}}$) between the three mutants the researches concluded that their model overestimated the values for the accessible surface area (ASA) in some of the mutants.

Within this subsection we also include the work of Kovalsky *et al.* [34]. These authors investigated the variations in pH behaviour associated with MD. HIV-1 protease was most active under low pH (3.5-6.5). As pH increased, the stability of this protein declined. Working with MD simulations these authors were able to reach certain conclusions about the dynamic differences created by either monoprotonation or diprotonation of the catalytic dyad. In this way they proved that Na^+ stabilizes the protease at neutral pH.

THE MOST RELEVANT INTERACTIONS

The most important interactions identified as crucial for maintaining of the three dimensional structure of a protein are the electrostatic and hydrophobic forces. This is demonstrated by the amount and quality of work that this question has recently received. This suggested the need for a special section in this article in which to review the most relevant theoretical and computational articles that have

appeared during the last five years. The majority of these studies consider the independent action of each one of these forces and it is only very recently that the mutual influence and combined action of the two forces has begun to attract much attention. Not surprisingly, most of the work relating to electrostatic interactions involves a great deal of theory as well as computer simulations, and it is more difficult to find parallel trend when dealing with hydrophobic effects. As a consequence, there is a larger number of papers dealing with the theoretical aspects of electrostatics than dealing with hydrophobicity theory. This is not meant to imply that there is less likelihood of hydrophobic effects contributing to protein thermal stability, it only reflects the fact that there is no easy analytical formalism for treating hydrophobicity in a similar way to electrical forces.

Electrostatics

In the last ten years numerous theoretical and semitheoretical articles have claimed an important role for electrostatic interactions in maintaining the thermal stability of proteins. Two such examples of this are the papers by Elcock [35] and Xiao and Honig [36]. Through the calculation of all the electric contributions to ΔG , using

continuum electrostatic models [37], these authors helped to establish the idea that ion pairs and specially optimally-distributed ion-pair networks, play a key role enhancing thermal stability. Zhou and Dong [38] developed a theoretical model of electrostatic forces to explain the differences in stability between the mesostable cold shock protein and its thermostable counterpart. This model explored the differences between the folded and extended states of the proteins in such a way as to allow its authors the possibility to make predictions about changes in the energy associated with the mutation of proteins. The most important contribution to ΔG has come from the entropy change which is larger in the mesostable protein.

In spite of the large-scale development of models dealing with electrostatics in proteins, the precise role of the dielectric constant is still a considerable source of uncertainty and controversy. In two lucid discussions on this issue, Schutz and Warshell [39] and Pitera *et al.* [40] questioned the very definition of the dielectric constant at the microscopic level. Also Wisz and Hellinga [41] also addressed the question of the distribution of the dielectric constant, making a distinction between the core, boundary and surface of a protein in order to contrast calculated and experimentally obtained pK values. Mostly basing their analysis on continuum electrostatics, Dominy *et al.* [42] carried out a computational study into variations in the dielectric constant in mesophilic proteins and their homologous thermophilic. These authors used the generalized Born theory (GB) in order to calculate the electrostatic interaction energy and its dependence on temperature variations. In doing so, they found that upon increasing temperature there was an increase in the dielectric response as simulated by molecular dynamics, which was also responsible for increasing the dipole moment. They concluded that electrostatic interactions increased the stability of proteins at high temperatures while the dielectric constant was increased by increasing the atomic fluctuations. However, applying another strategy that is not very frequently used, Makhatadze *et al.* [43] produced mutants in which certain exterior charges were reversed and concomitant free energy changes were computed. This was applied to ubiquitin, in which K11 and E34 residues (which are externally opposed) were changed for E11 and K34. In this case, in spite of the dielectric constant remaining basically unchanged, amino acid swapping did not lead to an increase in protein stability. According to these researchers, this clearly showed that protein stability is strongly context dependent. The notion of context dependency is supported by the work of Mozo-Villariás *et al.* [8]. Using a semiempirical approach, which took into account the global character of the protein, these authors defined an eclectic quasi-dipole moment that was independent of the origin of coordinates (considering the fact that most proteins are not neutral, this quasi dipole moment did not coincide with the true dipole moment) as the product of either the positive or the negative charge by the distance between the electric centroids of the protein. Using a set of 37 meso-thermostable protein pairs and a set of 55 point mutations describer for other proteins, these authors showed a correlation between the relative decrease in the quasi-dipole moment and the increase in T_m . By computing the quasi-dipole moment obtained by substituting of each amino acid in a protein

using either a positive, negative or neutral amino acid, These researchers obtained the relative quasi-dipole moment profile for a protein, and it was also possible to identify those amino acids that were most susceptible to mutation. Fig. 1 provides an example of a substitution quasi-dipole moment computed on β -glucanase.

Hydrophobicity

Theoretical studies of hydrophobic forces call for a large numbers of statistical treatments based on protein family structures with data being available from the PDB. Several aspects of hydrophobic effects in proteins are used by researchers to characterize them and if possible make predictions relating to thermostabilization. One of them is the balance between local and non-local interactions within the protein core. Gilis and Rooman [44] carried out an interesting study on protein stability based on comparing computed $\Delta\Delta G$ differences and the experimental measurements relating to 238 mutations carried out in the core of 16 proteins. These stabilizations were described in terms of a balance between potentials describing local and non-local interactions, with local interactions being most prevalent in the less exposed domains. The most prominent of the local forces used by these authors was torsion (which in turn was associated with the propensities of amino acids to associate with torsion angles of the backbone). The non-local interactions were mostly dependent on distances. For mutations involving fully buried amino acids the best potential was obtained with a combination of distance potential weighted by a factor of 1, plus a torsion potential weighted by a factor of 0.4. For partially buried amino acids the optimum proportion was obtained by combining a torsion potential weighted by a factor of 1 plus a distance potential with a weight of 0.7. Fully exposed amino acids are best represented by a torsion potential. Correlations between computed and measured $\Delta\Delta G$ lie around 0.8 or better.

The issue of clustering also ties in with the one that considers local *vs.* long-range interactions. The idea of hydrophobic clustering is certainly not new [45] but it has only recently seen a revival with the development of new informatic tools. In a comment relating with folding, R. Baldwin asked whether a network of hydrophobic clusters might explain long-range interactions. Selvaraj and Gromiha [47], who analyzed the three-dimensional structures of 36 (α/β)₈ barrels, reported factors responsible for providing common folds and also increments in thermal stability due to segments with hydrophobic clusters. They claimed that these clusters are stabilized through long-range interactions networks. Lu and Hodges [48], working with hydrophobic mutations in α -helices searched for the minimum size of hydrophobic clusters in two stranded coiled-coils. Using muscle tropomyosin, they designed a series of mutants in which alanines were substituted by amino acids of a more hydrophobic character (leucine, isoleucine) in a given order. These authors established that the subsequent thermostabilizations depended on the order in which the substitutions were made. As a consequence, the hydrophobicity ended up as an extremely context-dependent phenomenon.

As pointed out by Liao *et al.* [49]. There is an unquestionable link between packing density and hydrophobicity,

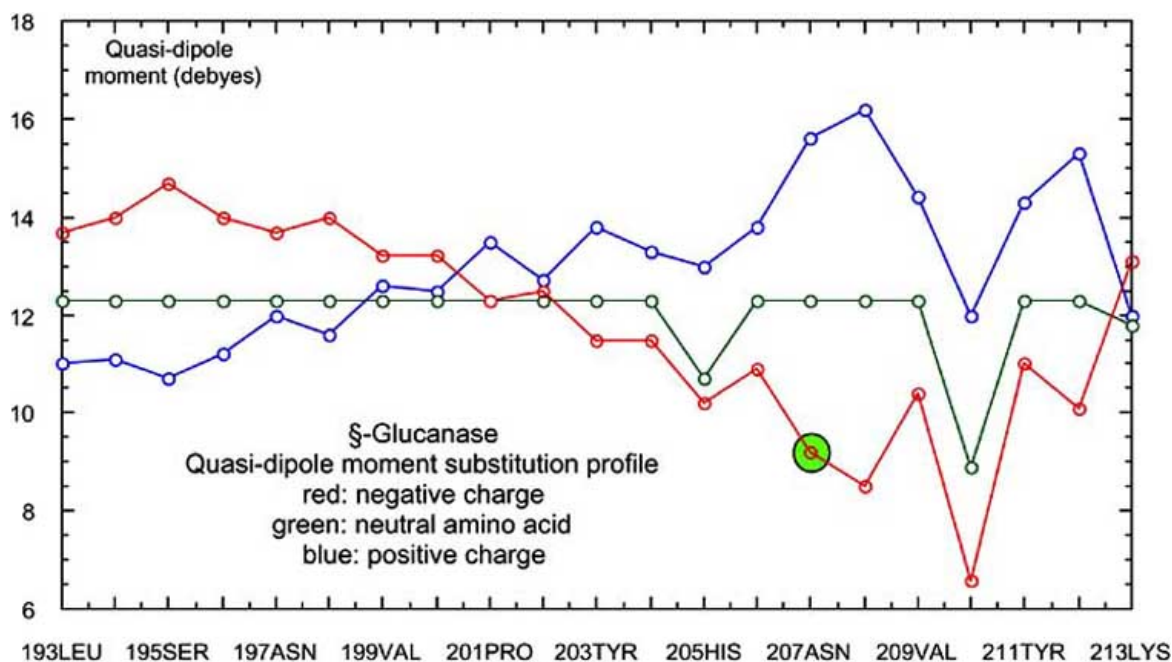


Fig. (1). Quasi-dipole moment of β -glucanase obtained by substituting each amino acid by either a negative charge (red), a neutral amino acid (green) or a positive charge (blue) and recalculating the total quasi-dipole moment for each substitution [8]. Lower values of the quasi-dipole moment imply protein variants less thermostable than the wild type. The green circle correspond to the N207D mutant obtained experimentally.

and also between sequential entropy and stability. These authors obtained a set of sequence entropies as result of aligning all of the residues from a set of 130 proteins. The linear relationship between these computed entropies and the inverse of packing densities and hydrophobicities, led them to the conclusion that the capacity to produce mutations is directly related to the cavities left by each amino acid and its propensity to appear buried in the protein. This relationship between the inverse on the packing density and hydrophobicity is very strong and the authors inferred a relationship between all these magnitudes and the thermal resistance of a protein.

Following an similar procedure to that described for the electrostatic interactions, Mozo-Villarias *et al.* [9], computed the profile of the relative change in hydrophobicity density (defined as total hydrophobicity per unit volume) in a set of 24 of proteins when subjected to a total of 81 point mutations. These authors found a linear regression between the relative (to wild type) increase in the hydrophobic density of a protein and the half-point transition temperature increase.

Mutual Influence Between Electrostatic and Hydrophobic Forces

A very promising line of research both from the theoretical and experimental points of view is that dealing with the balance between electrostatic and hydrophobic effects. Kegel and van der Schoot [50] analyzed this aspect to explain the temperature dependence of the spontaneous assembly of hepatitis B virus coat protein into capsids. Although the specific problem dealt with was not thermal stability, the model used and the conclusions reached may have a great deal of relevance to protein stability. They

developed a free energy association model which depends on two opposing terms, one of which is a function of the number of hydrophobic contacts, while the other is an electrostatic contribution in terms of the Debye-Hückel approximation. Taking all the information together, these factors were able to explain the dependences of ionic strength and the neutralizing union on nucleic acids.

CONCLUSION

Future research employing a holistic approach to the problem is likely to improve our ability to predict the stabilization of a protein. There seems to be increasing general interest in theoretical electrostatic and hydrophobic effects as evident from the increasing success and refinement of these approaches. Since structure and function are so intricately related in proteins, any future success in theoretical predictive methods relating to thermostabilization must also bring also new insights in related issues such as folding and ligand-binding. New theoretical and informatic methods are currently under way and should provide future new insights.

ACKNOWLEDGEMENTS

This research was supported by Grants BFU2004-06377-C02-01 from the MCYT (Ministerio de Ciencia y Tecnología, Spain) and by the Centre de Referència de R+D de Biotechnologia de la Generalitat de Catalunya and by Grant X0157 from La Paeria (Lleida City Hall).

REFERENCES

- [1] Querol E, Perez-Pons JA, Mozo-Villarias A. Analysis of protein conformational characteristics related to thermostability. *Protein Eng* 1996; 9:265-271.

- [2] Vogt G, Woell S, Argos P. Protein Thermal Stability, Hydrogen Bonds and Ion Pairs. *J Mol Biol* **1997**; 269: 631-643.
- [3] Karshikoff A, Ladenstein R. Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Eng* **1998**; 11: 867-872.
- [4] Parthasarathy S, Murthy MRN. Protein thermal stability: insights from atomic displacement parameters (B factors). *Protein Eng* **2000**; 13: 9-13.
- [5] Fukuchi S, Nishikawa K. Protein Surface Amino Acid Compositions Distinctively Differ Between Thermophilic and mesophilic Bacteria. *J Mol Biol* **2001**; 309: 835-843.
- [6] Gianese G, Bossa F, Pascarella S. Comparative Structural Analysis of Psychrophilic, and Meso- and Thermophilic Enzymes. *Proteins* **2002**; 47: 236-249.
- [7] Kumar S, Tsai CH, Nussinov R. Factors enhancing protein thermostability. *Protein Eng* **2000**; 13: 179-191.
- [8] Mozo-Villarias A, Cedano J, Querol E. A simple electrostatic criterion for predicting the thermal stability of proteins. *Protein Eng* **2003**; 16: 279-286.
- [9] Mozo-Villarias A, Cedano J, Querol E. Hydrophobic density profiles to predict thermal stability enhancement in proteins. Submitted.
- [10] Topham CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity. *Protein Eng* **1997**; 10: 7-21.
- [11] Petukhov M, Kil Y, Kuramitsu M, Lanzov M. Insights Into Thermal Resistance of proteins From the Intrinsic Stability of Their α -Helices. *Proteins* **1997**; 29: 309-320.
- [12] Muñoz V, Serrano L. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* **1997**; 41:495-509.
- [13] Malakaukas SM, Mayo SL. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct Biol* **1998**; 5: 470-475.
- [14] Desmet J, De Maeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**; 356: 539-542.
- [15] Gilis D, Rooman M. PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins. *Protein Eng* **2000**; 13: 849-856.
- [16] Kwasigroch JM, Gilis D, Dehouk Y; Rooman M. PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics* **2002**; 18: 1701-1702.
- [17] Carter C, LeFebvre B, Cammer S, Tropsha A, Edgell M. Four-body Potentials Reveal Protein-specific Correlations to Stability Changes caused by Hydrophobic Core Mutations. *J Mol Biol* **2001**; 311: 625-638.
- [18] López de la Paz M, Lacroix E, Ramirez-Alvarado M, Serrano L. Computed-aided Design of β -sheet Peptides. *J Mol Biol* **2001**; 312: 229-246.
- [19] Mooers BHM, Datta D, Baase W, Zollars ES, Mayo SL, Matthews BW. Repacking the Core of T4 Lysozyme by Automated Design. *J Mol Biol* **2003**; 332: 741-756.
- [20] Lee AL, Wand J. Microscopic origins of entropy, heat capacity and the glass transition in proteins. *Nature* **2001**; 411: 501-504.
- [21] Poland D. Free Energy Distribution in Proteins. *Proteins* **2001**; 45: 325-336.
- [22] Funahashi J, Takano K, Yutani K. Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? *Protein Eng* **2001**; 14: 127-134.
- [23] Connolly ML. The molecular surface package. *J Mol Graph* **1993**; 11:139-41.
- [24] Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J Mol Biol* **2002**; 320: 369-387.
- [25] Holm L, Sander C. Evaluation of Protein Models by Atomic Solvation Preference. *J Mol Biol* **1992**; 225: 93-105.
- [26] England JL, Shakhnovich BE, Shakhnovich EI. Natural selection of more designable folds: A mechanism for thermophilic adaptation. *Proc Natl Acad Sci* **2003**; 100: 8727-8731.
- [27] Ramos de Armas R, González-Díaz H, Molina R, Uriarte E. Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. *Proteins* **2004**; 56: 715-723.
- [28] Burioni R, Cassi D, Cecconi F, Vulpiani A. Topological Thermal Instability and Length of Proteins. *Proteins* **2004**; 55: 529-535.
- [29] Ibárcak A, Mohanty S. Folding Thermodynamics of Peptides. *Biophys J* **2005**; 88: 1560-1569.
- [30] Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci* **2005**; 102: 606-611.
- [31] Zakrzewska M, Krowarsch D, Wiedlocha A, Otlewski J. Design of fully active FGF-1 variants with increased stability. *Protein Eng Des Selec* **2005**; 17: 603-611.
- [32] Horii K, Saito M, Yoda T, et al. (2001). Contribution of Thr29 to the Thermodynamic Stability of Goat-Lactalbumin as Determined by Experimental and Theoretical Approaches. *Proteins* **2001**; 45: 16-29.
- [33] Funahashi J, Sujita Y, Kitao A, Yutani K. How can energy component analysis explain the difference in protein stability caused by amino acid substitutions? Effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme. *Protein Eng* **2003**; 16: 665-671.
- [34] Kovalsky D, Dubyna V, Mark AE, Kornelyuk A. A Molecular Dynamic Study of the Structural Stability of HIV-1 Protease Under Physiological Conditions: The Role of Na⁺ Ions in Stabilizing the Active Site. *Proteins* **2005**; 58: 450-458.
- [35] Elcock A. The Stability of Salt Bridges at High Temperatures: Implications for Hyperthermophilic Proteins. *J Mol Biol* **1988**; 284: 489-502.
- [36] Xiao L, Honig B. Electrostatic Contributions to the Stability of Hyperthermophilic Proteins. *J Mol Biol* **1999**; 289: 1435-1444.
- [37] Honig B, Nicholls A. Classical electrostatics in biology and chemistry. *Science* **1995**; 268: 1144-1149.
- [38] Zhou HX, Dong F. Electrostatic Contributions to the Stability of a Thermophilic Cold Shock Protein. *Biophys J* **2003**; 84: 2216-2222.
- [39] Schutz CN, Warshel A. What are the "Dielectric Constants" of Proteins and How To Validate Electrostatic Models. *Proteins* **2001**; 44: 400-417.
- [40] Pitera JW, Falta M, van Gunsteren WF. Dielectric Properties of Proteins from Simulation: The Effects of Solvent, Ligands, pH and Temperature. *Biophys J* **2001**; 80: 2546-2555.
- [41] Wisz M, Hellinga H. An Empirical Model for Electrostatic Interactions in Proteins Incorporating Multiple Geometry-Dependent Dielectric Constants. *Proteins* **2003**; 51: 360-377.
- [42] Dominy B, Minoux H, Brooks C. An Electrostatic Basis for the Stability of Thermophilic Proteins. *Proteins* **2004**; 57: 128-141.
- [43] Makhatadze GI, Loladze VV, Ermolenko EN, Chen XF, Thomas ST. Contribution of Surface Salt Bridges to Protein Stability: Guidelines for Protein Engineering. *J Mol Biol* **2003**; 327: 1135-1148.
- [44] Gilis D, Rooman M. Predicting Protein Stability Changes upon Mutation Using Database-derived Potentials: Solvent Accessibility Determines the Importance of Local versus Non-Local Interactions Along the Sequence. *J Mol Biol* **1997**; 272: 276-290.
- [45] Zehfus MH. Automatic recognition of hydrophobic clusters and their correlation with protein folding units. *Protein Science* **1995**; 4: 1188-1202.
- [46] Baldwin RL. Making a Network of Hydrophobic Clusters. *Science* **2002**; 295: 1657-1658.
- [47] Selvaraj S, Gromiha MM. Role of Hydrophobic Clusters and Long-range Contact Networks in the Folding of (a/b)₈ Barrel Proteins. *Biophys J* **2003**; 84: 1919-1925.
- [48] Lu SM, Hodges RS. Defining a minimum size of a hydrophobic cluster in two-stranded α -helical coiled-coils: Effects on protein stability. *Protein Sci* **2004**; 13: 714-726.
- [49] Liao H, Yeh W, Chiang D, Jernigan R, Lustig B. Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Eng* **2005**; 18: 59-64.
- [50] Kegel WK, van der Schoot P. Competing Hydrophobic and Screened-Coulomb Interactions in Hepatitis B Virus Capsid Assembly. *Biophys J* **2004**; 86: 3905-3913.