Structural Genomics of *Thermotoga maritima* Proteins Shows that Contact Order Is a Major Determinant of Protein Thermostability

Marc Robinson-Rechavi* and Adam Godzik
Joint Center for Structural Genomics
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92033

Summary

Despite numerous studies, understanding the structural basis of protein stability in thermophilic organisms has remained elusive. One of the main reasons is the limited number of thermostable protein structures available for analysis, but also the difficulty in identifying relevant features to compare. Notably, an intuitive feeling of “compactness” of thermostable proteins has eluded quantification. With the unprecedented opportunity to assemble a data set for comparative analyses due to the recent advances in structural genomics, we can now revisit this issue and focus on experimentally determined structures of proteins from the hyperthermophilic bacterium *Thermotoga maritima*. We find that 73% of *T. maritima* proteins have higher contact order than their mesophilic homologs. Thus, contact order, a structural feature that was originally introduced to explain differences in folding rates of different protein families, is a significant parameter that can now be correlated with thermostability.

Introduction

Whereas most organisms do not survive at temperatures above 50°C, a small number of hyperthermophilic bacteria and archaean thrive over 80°C and in some cases greater than 110°C (Stern and Liebl, 2001). This adaptation implies that molecular components of the cell, notably the proteins, should be stable at such high temperatures. Experimental studies of proteins from thermophilic organisms have demonstrated that they are usually stable and functional at high temperatures in vitro (Daniel and Danson, 2001). Understanding the molecular basis of thermostability is a very interesting and important research problem, both for our fundamental knowledge of protein structures and for the potential biotechnological applications of thermostable proteins, such as in bioremediation and high-temperature industrial processes.

Many structural features have been linked to protein thermostability, but the complexity of structural information and the heterogeneous sources of data have led to a confused picture (Petsko, 2001). One feature which seems intuitively to be related to stability is compactness of the protein, yet its characterization has proved elusive. Several statistical studies of protein structures from thermophiles have addressed this question using different proxies of “compactness.” The simplest measure is protein length (longer proteins have larger and better packed buried volumes), but it differs little or not at all among homologous proteins from thermophiles and mesophiles (Das and Gerstein, 2000; our unpublished data). Various groups focused on compactness of thermostable proteins, with Kumar et al. (2000) using accessible surface area and Szilagyi and Zavodszky (2000) using cavity size as measured by total surface area of cavities, but none have found significant differences. Several of these studies report more residues in α helices and less residues in loops or disordered regions in proteins from thermophiles (Kumar et al., 2000; Szilagyi and Zavodszky, 2000; Chakravarty and Varadarajan, 2002), but this difference was not the main feature in any of these studies, and we were unable to confirm it independently (our unpublished data). Chakravarty and Varadarajan (2002) found some variation in solvent accessibility by distinguishing polar and non-polar residues, and many groups identified an increase in the number of ion pairs inside thermostable proteins, which may point out to stronger packing in such proteins.

It seems then that either protein compactness is not such an important feature of proteins from thermophiles, or that an appropriate measure has not yet been found. In this study, we have investigated contact order (Plaxco et al., 1998), a structural parameter describing packing topology in proteins and found to correlate well with folding rates. We found a very significant trend: in 73% of the protein pairs analyzed, the thermostable homolog has a higher contact order. Apart from the choice of relevant parameters to compare, a drawback in such studies until recently was the paucity of data. To obtain experimental data sets for 25 pairs of structures (e.g., Szilagyi and Zavodszky, 2000), proteins from thermophilic archaean and bacteria were grouped together and compared to a mixed set of bacterial, archeel, and sometimes euakaryotic homologs. To increase sample size, some authors made models of structures of proteins from thermophiles (e.g., Chakravarty and Varadarajan, 2002), which has the added risk that the modeling step may itself introduce bias in the observations. We have taken advantage of the recent and rapid increase in structural data from *Thermotoga maritima*, a hyperthermophilic bacteria with an optimum growth temperature of 80°C. Most of these new structures come from efforts at our PSI structural genomics center, the Joint Center for Structural Genomics (Lesley et al., 2002). For each *T. maritima* structure, we have identified a homolog with an experimentally determined structure and estimated the type of the homology relationship by phylogenetic analysis. Only orthologs or paralogs were retained, while laterally transferred genes were eliminated from further analysis. Moreover, we have limited our sampling of mesophiles to bacteria. Seventy-three pairs (47 orthologs and 26 paralogs) fulfilling our criteria were identified, with an average rmsd of 2.2 Å. This is the largest data set of this type studied so far and also the only one to our knowledge with well defined homology relationships between proteins.

*Correspondence: marc@sdsc.edu*
Table 1. Output of Statistical Tests on Contact Order

<table>
<thead>
<tr>
<th>Contact Threshold</th>
<th>Homology Relation</th>
<th>N</th>
<th>T. maritima Mean</th>
<th>Mesophilic Mean</th>
<th>Mean Difference</th>
<th>T Test</th>
<th>Wilcoxon Signed Ranks Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Å</td>
<td>orthologs</td>
<td>47</td>
<td>0.0877</td>
<td>0.0842</td>
<td>0.0034</td>
<td>p = 0.0036</td>
<td>p = 0.001</td>
</tr>
<tr>
<td></td>
<td>paralogs</td>
<td>26</td>
<td>0.0919</td>
<td>0.0868</td>
<td>0.0050</td>
<td>p = 0.0028</td>
<td>p = 0.007</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>73</td>
<td>0.0892</td>
<td>0.0852</td>
<td>0.0040</td>
<td>p = 0.0003</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>4.5 Å</td>
<td>orthologs</td>
<td>47</td>
<td>0.0630</td>
<td>0.0591</td>
<td>0.0038</td>
<td>p = 0.0003</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>paralogs</td>
<td>26</td>
<td>0.0655</td>
<td>0.0609</td>
<td>0.0046</td>
<td>p = 0.0043</td>
<td>p = 0.007</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>73</td>
<td>0.0639</td>
<td>0.0598</td>
<td>0.0041</td>
<td>p &lt; 10^-4</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>

Results and Discussion

More than 70% of protein pairs we analyzed have higher relative contact order in T. maritima (53 out of 73). The average relative contact order of the T. maritima proteins is 0.0892, whereas the average of their close mesophilic homologs is 0.0852, a small but highly significant difference (paired t test, p = 0.0003; Table 1; and see Table S1 in the Supplemental Data available with this article online). The difference in contact order correlates to the solvent accessibility only in a subset of proteins (our unpublished data), and it is not related to any significant difference in secondary structure. The trend is not dependent on the type of homology relation: higher contact order is observed for orthologs as well as paralogs (Table 1). Such a significant difference is not recovered for distant homology pairs, detected by fold recognition, probably because major changes in function are dominant over any trends associated with thermostability at this scale (data not shown). We verified that these differences were not specific to T. maritima by comparing 31 pairs of close homologs between T. maritima and other thermophiles (18 bacterial and 13 archal proteins); contact order is not significantly different in other thermophiles (0.0800 versus 0.0798; p = 0.95). Thus the difference we observed between T. maritima and mesophiles is related to the thermostability of T. maritima proteins. Of note, the increase in compactness is not due to any significant shortening of the loops in our data set (p = 0.47), in contradiction with previous reports (Thompson and Eisenberg, 1999).

Higher contact order has never been reported previously in association with thermostability, at least to our knowledge. Importantly, contact order is correlated to the folding rate of proteins that fold by a two-state kinetics (Plaxco et al., 1998; Makarov et al., 2002). This correlation has been confirmed by comparing homologs (Bemporad et al., 2004) or experimental mutants (Mason et al., 2002). Thus, the addition of one cysteine bridge can be enough to change significantly both contact order and folding rate (Mason et al., 2002), proving the possibility of significant differences between otherwise similar proteins, such as those we compared. Indeed, homologous proteins in our data set differ significantly in contact order despite very similar structures (average rmsd = 2.2 Å). There are specific cases with direct experimental evidence that a few key point "mutations" can change the flexibility, the thermostability, and the contact order in a manner consistent with our observations. For instance, a modified version of the artificial miniprotein BBAT1, "peptide 1" (PDB: 1sn9), was reported (Ali et al., 2004), which has higher thermostability (40°C to 64°C). The thermostable “peptide 1” has a higher contact order than BBAS, a variant of this protein that is not thermostable (PDB: 118).

The highest difference in contact order in our data set is observed for GAPDH (glyceraldehyde 3-phosphate dehydrogenase), with an rmsd of only 1.18 Å between T. maritima and E. coli for a difference in contact order of 0.039 (Figure 1). Interestingly, this enzyme has been a highly studied example of thermostability (Korndorfer et al., 1995; Tanner et al., 1996; Song et al., 1998), yet a difference in contact order was not described previously, although support for higher compactness and rigidity was reported (Korndorfer et al., 1995). Over all bacterial GAPDHs, there appears to be a positive relation between contact order and growth temperature.

Figure 1. Comparison of GAPDH Structures from Bacteria

Relation between the optimal growth temperature and the relative contact order of GAPDH from different bacterial species. The x axis represents the mean optimal growth temperature for each species (Huang et al., 2004), and the y axis represents the relative contact order computed on one chain of the GAPDH homotetramer. Each point corresponds to one experimental structure from the PDB; points are color coded according to growth temperature. Chain O of the structures used from E. coli and T. maritima are represented. E. coli GAPDH-A is the mesophilic bacterial GAPDH with the lowest rmsd to T. maritima GAPDH; they are paralogs. No structure is available for mesophilic bacterial orthologs of T. maritima GAPDH (i.e., GAPDH-C), and T. maritima does not have GAPDH-A. The PDB entries used are: 1obf (A. xylosidans; 28°C–37°C), 1gad (E. coli; 37°C), 1gd1 (B. stearothermophilus; 37°C–65°C), 1cer (T. aquaticus; 70°C–72°C) and 1hdg (T. maritima; 80°C); in parenthesis, the range of optimal growth temperatures.
Contact Order Is Correlated to Thermostability

GAPDHs from two thermophilic archaea do not follow this trend, with contact orders of 0.074 and 0.077 for optimal growth temperatures of 87°C and 83°C; it has already been reported that archaeal and bacterial GAPDHs have different thermostability strategies (Charron et al., 2002).

Galitskaya et al. (2003) have shown that the correlation between contact order and folding rate holds only for proteins with two-state folding kinetics, but not for proteins with three-state folding kinetics; instead, folding rate in the latter is inversely correlated to protein length. Interestingly, the 20 proteins that have a decreased contact order in T. maritima also have greater average length (232 versus 224 amino acids). This is in contrast to generally shorter chains in T. maritima proteins than in mesophiles, including those with increased contact order. The difference in length variation between pairs with increased or decreased contact order is significant (Wilcoxon test: p = 0.0038). This suggests that lower folding rates in T. maritima than in mesophiles may be ubiquitous, through higher contact order for some proteins and through increased chain length in others, although this remains quite hypothetical. Of note, some experimental studies suggest similar folding rates, but lower unfolding rates, of proteins from thermophiles (Hollien and Marqusee, 2002). And as a note of caution, it should be kept in mind that the correlation between contact order and folding rate was established at room temperature (20°C–37°C; Table 1 in Plaxco et al. [1998]), not at 80°C.

The original definition of contact order, used in these calculations, sets a limit of 6 Å between atoms to define them as being “in contact” (Plaxco et al., 1998). This definition was proposed as a proxy for topological complexity, and the resulting contact order parameter has the obvious merit of being correlated to experimental measures of protein folding rates. But we wanted to test a more stringent definition of contact order, limited to those residues that are effectively in physical contact. Thus, we set a limit of 4.5 Å, which was previously established as the optimal value for the development of empirical interaction parameters in proteins (Godzik et al., 1992, 1995). This modified contact order definition varies even more significantly than the original (Table 1).

Interestingly, T. maritima proteins from this study also show a systematic and statistically significant difference in energy-like score calculated using the contact-based empirical energy parameters (Godzik et al., 1992, 1995). On the set of pairs analyzed here, in 52 out of 73 pairs the Thermotoga structure had a better score, with an average difference of −0.044 kT (p ≤ 0.0001). Empirical scores like this are typically used to validate models in comparative modeling, and their correlation to experimentally measured protein stability was discussed in the literature, but never convincingly proven.

In conclusion, contact order is a major structural determinant of protein thermostability. It allows us to demonstrate that a specific structural feature distinguishes proteins from a thermophile from their mesophilic homologs. This feature is related to compactness of the structure, but also to its topological complexity. What is even more important, it indirectly relates thermostability to the protein folding/unfolding rate. Further functional and structural implications of this finding should be explored both experimentally and by bioinformatic analyses.

Experimental Procedures

All sequences associated with experimentally determined protein structures were downloaded from the PDB database (Bourne et al., 2004) version from November 1, 2004. The subset of entries from Thermotoga maritima was compared by BlastP (Altschul et al., 1997) to all other sequences. For each T. maritima entry which had at least one hit with an E value under e−4, aligned homologous proteins from completely sequenced genomes were recovered from Hogenom (Perrière et al., 2003). These alignments were edited to add sequences of homologous PDB entries that are absent from Hogenom (typically from organisms whose genome is not sequenced) and to merge protein families that were classified separately in Hogenom but were homologous according to the results of Blast on the PDB sequence set.

For each alignment a phylogenetic tree was built by PhyML (Guindon and Gascuel, 2003), with the JTT model and a γ distribution between sites (parameter α estimated by PhyML with eight categories). These trees were used to assess homology relations between the proteins from the PDB entries: orthology, paralogy, or xenology (lateral gene transfer). Five cases of suspected lateral transfer were excluded from the final data set. When several homologous proteins were available from T. maritima, only one was used.

Structure coordinates were obtained from the PDB (Bourne et al., 2004), and only single chains were used in further analysis. For homo-oligomers, only the first chain in the file was used in calculations. For hetero-oligomers, chains were treated separately according to their homology relations as established by phylogenetic analysis. When there were several mesophilic bacterial proteins orthologous to a same T. maritima proteins, the one with the lowest rmsd was used. The same was done for paralogous proteins. When structures of both orthologous and paralogous proteins were available, orthologs were preferred.

Relative contact order was calculated by the program contactOrder.pl (http://depts.washington.edu/bakerpg/contact_order.pl), which implements the definition of Plaxco et al. (1998): any nonwater atoms separated by less than 6 Å are considered “in contact.” In addition, we calculated contact order defining contact with a more stringent distance of 4.5 Å, which optimizes the detection of specific interactions (Godzik et al., 1992, 1995).

Supplemental Data

Supplemental Data can be found online at http://www.structure.org/cgi/content/full/13/6/857/DC1/.

Acknowledgments

We thank Iddo Friedberg, Olga Kirillova and Ian A. Wilson for helpful suggestions, as well as all members of the Joint Center for Structural Genomics.

Received: December 20, 2004
Revised: March 17, 2005
Accepted: March 17, 2005
Published: June 7, 2005

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST:


