Elucidation of determinants of protein stability through genome sequence analysis

Suvobrata Chakravarty[a], Raghavan Varadarajan[a,b,*]

[a] *Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India*
[b] *Chemical Biology Unit, Jawaharlal Nehru Center for Advanced Scientific Research, Jakkur, Bangalore 560004, India*

Received 1 December 1999; received in revised form 9 February 2000

Edited by Matti Saraste

Abstract Sequences of putative soluble proteins from complete genomes of eight thermophiles and 12 mesophiles were analyzed to gain insight into determinants of protein thermostability. The *predator* algorithm was used to assign secondary structures to each protein sequence. Based on simple statistical tests, a set of stabilizing factors was identified. These include reduced protein size, increases in number of residues involved in hydrogen bonding, β-strand content and helix stabilization through ion pairs. There are also significant increases in the relative amounts of charged and hydrophobic β-branched amino acids and decreases in uncharged polar amino acids in proteins from thermophiles relative to mesophilic organisms. Factors such as the relative proportion of residues in loops, proline and glycine content and helix capping do not appear to be important.

Key words: Genome; Thermostability; Secondary structure prediction; *t*-Test

1. Introduction

Understanding the molecular basis of protein thermal stability is an important fundamental problem with obvious practical applications. One approach to this problem involves the comparison of structures and sequences of homologous proteins from thermophilic and mesophilic organisms. [1,2]. Previous analyses have suggested that factors that may contribute to enhanced thermostability include improved hydrogen bonding, better hydrophobic packing, enhanced secondary structure propensity, helix dipole stabilization, removal of residues sensitive to oxidation or deamination, and improved electrostatic interactions [3]. Although proteins can be engineered to achieve greater stability by utilizing one or more of these strategies, it is clear that no single and preferred mode of stabilization occurs [4]. The limited number of crystal structures of proteins from thermophiles and hyperthermophiles has hampered detailed structural comparisons with mesophilic proteins.

All the information needed to create thermotolerance is encoded in the protein sequence as proteins of both thermophiles and mesophiles are composed of the same 20 amino acids [5,6]. Recent comparisons of sequences of thermophilic proteins with their homologues from meshophilic species have shown a decrease in content of polar uncharged amino acids and shortening of loops in the thermophilic proteins relative to mesophilic homologues [7,8]. While focusing on homologous proteins has certain advantages, it greatly reduces the number of sequences available for analysis. In the present work we have adopted an alternative approach. This involves analysis of sequences of all putative soluble proteins from eight thermophilic and 12 mesophilic organisms to identify factors that contribute to the enhancement of protein thermostability (Table 1). The present work confirms the importance of some of the factors identified from earlier analyses and, in addition, identifies several new factors responsible for enhanced protein thermostability.

2. Materials and methods

2.1. Membrane protein prediction

Putative membrane proteins were identified through the presence of a membrane spanning helix as proposed earlier [9]. The average hydrophobicity of the most hydrophobic protein segment, maxH, (for a 19 residue window) was used to detect membrane spanning helices [10,11]. The Kyte and Doolittle hydrophobicity scale was used [9]. The histogram of maxH values, calculated for all proteins of every genome, shows a bimodal distribution [11]. The data indicate that proteins are separated into two distinct groups of soluble and membrane proteins with mean maxH values of 1.28 and 2.61, respectively (unpublished results). The minimum between the two peaks occurs at a value of 2.0. In order to predict the percentage of membrane proteins in each genome, any protein having a value of maxH $\geq$ 2.0 was considered to be a membrane protein. The accuracy of the procedure was checked on a data set of 413 non-homologous globular proteins [12] and 105 membrane proteins [13,14]. It resulted in 95 correctly predicted membrane proteins out of 105 (90.5%) ($\sim$10% under prediction, i.e. membrane proteins are predicted to be soluble proteins) and nine incorrectly predicted membrane proteins out of 413 soluble proteins ($\sim$2% over prediction). In order to minimize the inclusion of membrane proteins in the data set of putative soluble proteins, we used a more stringent cutoff of maxH$\geq$1.58 [10] to remove putative membrane proteins. Under prediction of membrane proteins reduces to $<$5% with this criterion.

2.2. Amino acid composition

The ratio between the total number of occurrences of a particular amino acid and the sum of sizes of all soluble (non-membrane) proteins in a genome is the proportion of that residue (fA_g) in the soluble proteins of a genome. The set of values of fA_g of a particular amino acid from the eight thermophilic organisms comprise the thermophilic

*Corresponding author. Fax: (91)-80-3341683 or (91)-80-3348535.
E-mail: varadar@mbu.iisc.ernet.in

Abbreviations: fA_g, proportion of a particular residue in genome; fr-h, fraction of residues in helix; fr-b, fraction of residues in sheet; fr-l, fraction of residues in loops; Sb3, number of salt bridges of type i ± 3 per helix; Sb4, number of salt bridges of i ± 4 per helix; dN, net charge at N-terminus of helix; dC, net charge at C-terminus of helix; Nc, fraction of helices with N-capping boxes; β-b, number of β-branched residues per helix; ORF, open reading frame; X_T, average value of a trait in thermophilic sample; X_M, average value of a trait in mesophilic sample

PII: S0014-5793(00)01267-9

sample for that amino acid. Statistical tests (discussed below) on both thermophilic and mesophilic samples of fA_g are carried out for each amino acid to determine which amino acids are present in different relative amounts in thermophiles and mesophiles. Amino acids were grouped into three classes based on relative polarity. These consisted of hydrophobic residues (Ala, Val, Ile, Leu, Met, Pro, Trp, Phy, Tyr), charged residues (Arg, Lys, His, Glu, Asp), and polar (uncharged) residues (Ser, Thr, Glu, Asp, Cys).

2.3. Secondary structure prediction

Secondary structure prediction was carried out using the *predator* program [15,16]. The algorithm has an accuracy of 70% [15,16]. We carried out statistical tests on the accuracy of the algorithm to ensure that prediction accuracy is similar for proteins from thermophiles and mesophiles (unpublished results).

2.4. Secondary structure comparison

From the predicted secondary structures a number of traits listed below were investigated for enhancement of thermal stability. The average values of a specific trait X from each thermophilic organism constitute the thermophilic sample for X while corresponding average values from mesophilic organisms constitute the mesophilic sample for X. The *t*-test and sign test were carried out for thermophilic and mesophilic samples of each of the following traits: (a) Secondary structure content: The fraction of residues of soluble proteins in each genome in helix, sheet and loop (fr-h, fr-b and fr-l). (b) Average loop, helix and strand length for each genome. (c) Salt bridges/helix: Sb3 and Sb4 represent the number of salt bridges of the type (i, i ± 3) and (i, i ± 4) per helix from the predicted secondary structural assignments. Arg, Lys, His, Glu and Asp are considered for Sb3 and Sb4 calculations. Sb3* and Sb4* represent values of Sb3 and Sb4 normalized to the fraction content of charged residues of soluble proteins in the genome. (d) Helix charge-dipole interaction (net charge at N- and C-terminus of helix, dN and dC): Negatively (Asp and Glu) and positively (Arg, Lys and His) charged residues are preferentially found at the N and C termini of helices respectively to stabilize the helix dipole [17]. In the present work, the net negative charge in the first four N-terminal residues and the net positive charge in the last four C-terminal residues of every predicted helix (longer than eight residues) were determined. dN and dC represent the average charge at the N and C termini of helices respectively in a particular genome. (e) β-Branched residues per helix (β-b): β-branched residues are known to destabilize helices [18,19]. Hence the number of β-branched residues (Val, Ile and Thr) per helix was determined for each genome. β-b* represents the normalized value of this trait. (f) Fraction of helices with N-capping boxes (Nc): The Ncap residue is defined to be the residue adjacent and N terminal to N1, the first residue of an α-helix. In an Ncap box, the side chain at N3 is hydrogen bonded to the amide group of the Ncap and vice versa. Ncap boxes are thought to be helix stabilizing. A potential Ncap box is said to occur if Ser, Thr, Asp, Asn, His, Glu or Gln are present at both the Ncap and N3 positions [20]. The fraction of helices with Ncap boxes (Nc) for each genome was calculated by dividing the number of predicted helices with Ncap boxes by the total number of predicted helical segments in the genome. Nc* represents the normalized value of this trait.

2.5. Statistical tests of significance

Let X_{jT} be the average value of a trait X in genome j of a thermophile. The set of X_{jT} for the eight thermophilic genomes constitutes the thermophilic sample of X with a sample mean of X_T. Similarly the set of X_{jM} for 12 mesophilic genomes constitutes the mesophilic sample of X with a sample mean of X_M. The assumption here is that the thermophilic sample X_{jT} ($j=1$–8) is a sample of the total population that consists of values X_{jT} ($j=1$–n where $n \gg 8$) from all n thermophilic genomes that exist. Similar arguments apply to the mesophilic sample. The question we would like to address is whether the population means of X are different for the thermophilic and mesophilic populations based on data from the available samples. Student's *t*-test was carried out for comparison of population means of a particular trait, with the assumption that average values of the traits are independent and follow a normal distribution. For a particular trait X, observed to have a higher value in thermophilic proteins than in mesophilic proteins the appropriate null hypothesis would be that the average value of the traits are equal in both groups (H_o: $X_T - X_M = 0$) against an alternative hypothesis (H_1: $X_T > X_M$). X_M and X_T represent mesophilic and thermophilic sample means, respectively. For purposes of illustration we consider the case where the trait X is the average fraction of residues in α-helices, fr-h. The t statistic is written in the following manner:

$$t = [\mathrm{fr-hT} - \mathrm{fr-hM}]/\sqrt{[S_T^2/(N_T-1) + S_M^2/(N_M-1)]},$$

$$df = N_T + N_M - 2 \quad (1)$$

S_T^2 and S_M^2 represent sample variances of the thermophilic and mesophilic groups for the particular trait. N_T (8 genomes) and N_M (12 genomes) represent the thermophilic and mesophilic sample sizes respectively and df, the number of degrees of freedom is 18. For a one tailed t test (with df = 18) at a 1% level of significance, H_o is rejected for $t > 2.55$ or $t < -2.55$. If t is > 2.55 then the probability that fr-hT is greater than fr-hM is > 0.99. If t is < -2.55 then the probability that fr-hT is less than fr-hM is > 0.99. We have also carried out t-tests taking into account the errors in secondary structure prediction and membrane protein prediction (unpublished results). Inclusion of these prediction errors does not change any of the reported statistics or results. The t-test assumes a normally distributed sample. In order to avoid this assumption a simple non-parametric test (sign test) was carried out on all the traits. The results of the sign test were very similar to those of the t-tests (unpublished results).

3. Results

3.1. Membrane protein prediction and amino acid composition

The average percentage of predicted membrane proteins present were 22.6 ± 3.3 for thermophiles and 23.9 ± 2.7 for mesophiles, respectively. These are in good agreement with other analyses of membrane protein content for bacterial genomes [11,21].

3.2. Amino acid composition

The results of the amino acid composition analysis of putative soluble proteins are shown in Table 2. A positive value in the t-test result indicates that the trait has a higher numerical value in the thermophiles than in mesophiles while a negative value indicates the opposite result. The bold underlined scores are significantly different at 1% level of significance and those in bold italics are significantly different at a lower (5%) level of significance. In the remainder of the discussion we focus primarily on differences that are significant at the 1% level in both the t-test and the sign-test but also mention those are significant at the 1% level in only one of the two tests. Of the individual amino acids, Val and Glu are enriched while His, Ser, Thr and Gln contents are depleted in thermophiles relative to mesophiles. Thermophiles also show an increase in the charged amino acid content and a decrease in polar uncharged amino acids. The hydrophobic amino acid content is marginally higher in thermophiles but not at the 1% level (Table 2). The observed increase may be due to the increased rigidity and high hydrophobicity of these amino acids. The increase in charged amino acid content is probably due to the enhanced occurrences of salt bridges and ion pairs in thermophilic proteins [22]. The increase in Val content may be due to the increased rigidity of β-branched amino acids [23], which results in a smaller conformational entropy increase upon unfolding than for unbranched amino acids. Surprisingly, there are only small increases in the content of Ile, another β-branched amino acid and Pro a rigid amino acid which has been used to increase protein stability in several mutational studies [24,25]. The decreased content of uncharged polar residues is likely to minimize deamidation and backbone cleavages involving Asn and Gln, which are catalyzed by Serine and Threonine [26]. The reduced proportion

Table 1
List of twenty organisms whose genomic sequences are analyzed

Organism	Website URL	ORFs	ID
Aquifex aeolicus	www.ncgr.org/microb	1522	AA
Archaeoglobus fulgidus	www.tigr.org/mdb/afdb	2409	AF
Aeropyrum pernix	www.mild.nite.go.jp	2694	AP
Methanoccus jannaschii	www.tigr.org/mdb/mjdb	1771	MJ
Methanobacterium thermoautotrophicum	www.b.osci.ohio-state.edu/–genomes/mthermo	1871	MT
Pyrococcus abyssi	www.genoscope.cns.fr/Pab	1765	PA
Pyrococcus horikoshii	www.tigr.org/mbd/tmdb	2061	PH
Thermotoga maritima	www.tigr.org/mbd/tmdb	1864	TM
Borrelia burgdorferi	www.tigr.org/mdb/bbdb	1638	BB
Bacillus subtilis	www.ncgr.org/microb	4100	BS
Chlamydia pneumoniae	www.stdgen.lanl.gov/bacteria/cpneu	1052	CP
Escherichia coli	www.genetics.wisc.edu	4290	EC
Helicobacter pylori	www.tigr.org/mdb/hpdb	1577	HP
Haemophilus influenzae	www.tigr.org/mdb/hidb	1707	HI
Mycoplasma genitalium	www.tigr.org/mdb/mgdb	479	MG
Mycoplasma pneumonia	www.zmbh.uni-heidelberg.de	672	MP
Mycobacterium tuberculosis	www.tigr.org/mdb/mtdb	3924	MT
Synechocystis sp. strain PCC6803	www.kazusa.or.jp	3168	SP
Treponema pallidum	www.tigr.org/mdb/tpdb	1030	TP
Rickettsia prowazekii	www.evolution.bmc.uu.se/–thomas/Rickettsia	837	RP

of Gln and Asn in thermophiles is consistent with the observation that temperature induced deamidation of these residues has acted against the selection of these residues in the thermophilic genomes [7]. A recent study [7] showed similar results with a few differences. This study consisted of a comparison of 115 complete or partial protein sequences form mesophiles (*Methanococcus sp.*) and their high temperature homologues (*Methanococcus jannaschii*). It was observed that thermophilic proteins show decreased contents of Ser, Asn, Thr and Met and increased contents of Ile, Arg, Glu, Lys and Pro. Although this work and the earlier study [7] employed quite different data and methodology, the results in regard to compositional differences are quite similar. This is strong evidence that amino acid compositions are significantly different in thermophiles and mesophiles.

3.3. Size dependence

The average soluble protein sizes in the thermophilic and mesophilic groups are 268 ± 38 and 310 ± 16 residues, respectively. The mean values are significantly different at the 1% level of significance (t-statistic $= -3.44$). The smaller average size of thermophilic proteins has been noted in an earlier analysis, though the relevance of this observation to increased stability was not clarified [8]. It is important to note that thermophilic organisms have significantly higher proportions of smaller proteins than mesophilic organisms. Smaller proteins have al lower value of the thermodynamic parameter ΔC_p [27,28]. A plot of the free energy of unfolding, $\Delta G°(T)$ as a function of T is known as the stability curve [29]. The curve is completely specified by the three parameters $\Delta G°(T_o)$, $\Delta H°(T_o)$ and ΔC_p, where $\Delta G°(T_o)$ and $\Delta H°(T_o)$ are the free energy and enthalpy changes upon unfolding at some reference temperature T_o. The curvature of the stability curve is determined by the magnitude of $\Delta C_p/T$ [28]. A decrease in ΔC_p results in a lower curvature and a higher value of the heat denaturation temperature, T_m.

3.4. Secondary structure prediction

The accuracy of secondary structure prediction using the *predator* program is comparable for both thermophiles and mesophiles (unpublished results). The results of statistical comparison of various traits observed from predicted secondary structures are shown in Table 3. The bold underlined scores are significantly different at a 1% level of significance and those in bold italics are significantly different at a 5% level of significance. The fraction of residues per protein in β-strands is significantly higher in thermophiles. This might result in increased hydrogen bonding which in turn should lead to enhanced thermal stability. Surprisingly, there is no significant difference in helical content between the two groups. According to the t-test, the fraction of residues in the unstructured loops is not different in the two groups. This is in contrast to a recent study, which suggested that

Table 2
Amino acid compositions in thermophiles and mesophiles[a]

Amino Acid	Thermophile	Mesophile	t-statistic
Gly	7.15 ± 0.71	6.30 ± 1.59	1.42
Ala	6.54 ± 1.33	7.57 ± 2.33	−1.56
Val	7.83 ± 0.59	6.31 ± 0.93	**4.12**
Ile	7.44 ± 1.48	6.84 ± 1.85	0.76
Leu	9.48 ± 0.59	10.00 ± 0.62	***−1.87***
Met	2.37 ± 0.34	2.05 ± 0.37	***1.91***
Pro	4.46 ± 1.06	3.84 ± 0.97	1.29
Trp	0.95 ± 0.20	0.97 ± 0.31	0.14
Phe	3.92 ± 0.70	4.43 ± 1.11	−1.14
Tyr	3.61 ± 0.37	3.25 ± 0.48	***1.76***
Arg	6.29 ± 1.32	4.89 ± 1.62	2.03
Lys	7.79 ± 2.37	7.05 ± 2.69	0.63
His	1.79 ± 0.25	2.22 ± 0.43	**−2.60**
Glu	9.42 ± 1.02	6.79 ± 0.79	**6.51**
Asp	5.18 ± 0.73	5.38 ± 0.40	−0.80
Ser	5.22 ± 0.94	6.28 ± 0.76	**−2.76**
Thr	4.16 ± 0.29	5.19 ± 0.52	**−5.00**
Cys	1.01 ± 0.29	1.14 ± 0.33	−0.85
Gln	1.84 ± 0.23	4.23 ± 1.01	**−6.58**
Asn	3.48 ± 0.85	5.11 ± 1.90	***−2.26***
Hydrophobic	51.01 ± 2.13	49.10 ± 2.22	***1.90***
Charged	30.46 ± 1.97	26.33 ± 1.49	**5.34**
Polar	15.72 ± 0.93	21.99 ± 2.18	**−7.47**

[a]In the last column values that are statistically significant at the 5% level are in bold italics and those at the 1% level are in bold and underlined.

Table 3
Comparison of various structural traits from predicted secondary structures[a,b]

Trait	Thermophile	Mesophile	*t*-statistic
Fr-h	0.3871 ± 0.032	0.3923 ± 0.021	−0.47
Fr-b	0.1645 ± 0.011	0.1467 ± 0.012	**3.65**
Fr-l	0.4604 ± 0.046	0.4714 ± 0.015	−0.77
Loop length	7.1738 ± 0.769	7.7889 ± 0.466	***−2.23***
Helix length	12.091 ± 0.357	12.134 ± 0.438	0.23
Sheet length	4.850 ± 0.076	4.758 ± 0.066	**2.85**
Sb3	0.8162 ± 0.176	0.4941 ± 0.070	**5.74**
Sb3*	2.6560 ± 0.454	1.8700 ± 0.181	**5.43**
Sb4	0.8962 ± 0.187	0.5558 ± 0.083	**5.43**
Sb4*	2.9170 ± 0.467	2.1000 ± 0.218	**5.21**
dN	−0.0440 ± 0.068	0.0366 ± 0.055	**−2.92**
dC	0.2537 ± 0.059	0.2700 ± 0.066	−0.55
Nc	0.1200 ± 0.009	0.1350 ± 0.017	**−2.23**
Nc*	0.3870 ± 0.019	0.3830 ± 0.040	0.26
β-br	1.5512 ± 0.212	1.6300 ± 0.189	−0.86
β-br*	7.9575 ± 0.844	8.8758 ± 0.745	**−2.57**

[a]In the last column values that are statistically significant at the 5% level are in bold italics and those at the 1% level are in bold and underlined.
[b]Descriptions of all the traits are indicated in the abbreviations and in Section 2.

shortened loops in proteins are one of the main contributors to enhanced thermal stability [8]. The average loop length is different at the 5% level of significance but not at the 1% level of significance. The smaller loop length in thermophiles may just be a reflection of the fact that the average size of proteins from thermophilic organisms is smaller than that of mesophilic proteins. Shortened loops are thought to contribute to thermostability by reducing the conformational entropy associated with folding of a polypeptide. If indeed such contributions are important, it is the fraction of residues in loops rather than the average loop length, which should be lower in thermophiles. This is not what is observed.

The number of salt bridges per helix (Sb3 and Sb4) and the number of salt bridge per helix normalized to the percentage of charged residues (Sb3* and Sb4*) in the respective genomes are both significantly higher in thermophiles. Thus the increased proportion of charged residue in helices from thermophiles is not simply due to an increase in the overall content of charged residues in thermophiles. The charge-dipole interactions at the N termini of helices are substantially larger in thermophilic proteins though there is no difference in charge at the C termini of helices. The composition of the N-capping box (Nc) is significantly more favorable for mesophilic proteins. However, this is simply due to an increase in the content of polar, uncharged residues in mesophiles. The normalized value Nc* does not differ between the two groups. Finally, helices from thermophilic proteins contain a smaller fraction of helix destabilizing β-branched residues than helices in mesophilic proteins.

4. Discussion

Analysis of protein sequences from complete genomes of thermophiles and mesophiles reveals some of the factors responsible for enhancement of thermostability in proteins. There are several clear differences in amino acid composition, size and secondary structure between proteins from thermophiles and mesophiles and simple statistical tests can be used to identify these differences. The availability of several genomes in each category (thermophiles and mesophiles) makes it possible to carry out the appropriate statistical tests. The thermophile dataset includes six archaeal and two eubacterial (AA and TM) genomes. Similar trends are observed in both archaeal and eubacterial thermophilic genomes. This suggests that the observed differences between thermophiles and mesophiles are correlated with thermostability and are not simply differences between archaeal and eubacterial genomes. The present work confirms the importance of compositional differences identified from an earlier analysis [30]. In addition, several new factors that are responsible for enhanced protein thermostability are identified. These include reduced overall size (rather than reduced loop length) and reduced content of β-branched residues in helices, and increases in β-strand content and length and in the number of intrahelical salt bridges.

Acknowledgements: We thank the Bioinformatics Center and the SERC at the Indian Institute of Science for access to databases and computational facilities. All the genome sequences taken from the Bioinformatics Center at the Indian Institute of Science. We thank Dr. N.V. Joshi for several helpful suggestions and discussions with regard to the statistical analysis. This work was supported by grants from DST and DBT to R.V.

References

[1] Perutz, M. and Raidt, H. (1996) Nature 255, 256–259.
[2] Argos, P., Rossmann, M., Grau, U., Zuber, H., Frank, G. and Tratschin, J. (1979) Biochemistry 25, 5698–5703.
[3] Querol, E., Perez-Pons, J.A. and Mozo-Villaria, A. (1996) Protein Eng. 9, 265–271.
[4] Vogt, G., Woell, S. and Argos, P. (1997) J. Mol. Biol. 269, 631–643.
[5] Vieille, C., Burdette, D.S. and Zeikus, J.G. (1996) Biotechnol. Annu. Rev. 2, 1–83.
[6] Bohm, G. and Jaenicke, R. (1994) Int. J. Pept. Protein Res. 43, 97–106.
[7] Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I., Woose, C.R. and Olsen, G.J. (1999) Proc. Natl. Acad. Sci. USA 96, 3578–3583.
[8] Thompson, M.J. and Eisenberg, D. (1999) J. Mol. Biol. 290, 595–604.
[9] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[10] Klein, P., Kanehisa, M. and DeLisi, C. (1985) Biochim. Biophys. Acta 815, 468–476.
[11] Boyd, D., Schierle, C. and Beckwith, J. (1998) Protein Sci. 7, 201–205.
[12] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[13] Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) Protein Sci. 4, 521–533.
[14] Cserzo, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997) Protein Eng. 10, 673–676.
[15] Frishman, D. and Argos, P. (1997) Proteins Struct. Funct. Genet. 27, 329–335.
[16] Frishman, D. and Argos, P. (1996) Protein Eng. 9, 133–142.
[17] Richardson, J.S. and Richardson, D.C. (1988) Science 240, 1648–1652.
[18] Piela, L., Nemethy, G. and Scheraga, H.A. (1987) Biopolymers 26, 1273–1286.
[19] Creamer, T.P. and Rose, G.D. (1994) Proteins 19, 85–97.
[20] Harper, E.T. and Rose, G.D. (1993) Biochemistry 32, 7605–7659.
[21] Wallin, E. and von Heinje, G. (1998) Protein Sci. 7, 1029–1038.
[22] Xiao, L. and Honig, B. (1999) J. Mol. Biol. 289, 1435–1444.
[23] Lee, K.H., Xie, D., Freire, E. and Amzel, L.M. (1994) Proteins 20, 68–84.

[24] Veltman, O.R., Vriend, G., Middelhoven, P.J., van den Burg, B., Venema, G. and Eijsink, V.G. (1996) Protein Eng. 9, 1181–1189.
[25] Van den Burg, B., Vriend, G., Veltman, O.R., Venema, G. and Eijsink, V.G. (1998) Proc. Natl. Acad. Sci. USA 95, 2056–2060.
[26] Tomazic, S.J. and Klibanov, A.M. (1988) J. Biol. Chem. 263, 3092–3096.
[27] Myers, J.K., Pace, C.N. and Scholtz, J.M. (1995) Protein Sci. 4, 2138–2148.
[28] Ganesh, C., Eswar, N., Srivastava, S., Ramakrishnan, C. and Varadarajan, R. (1999) FEBS Lett. 454, 31–36.
[29] Becktel, W.J. and Schellman, J.A. (1987) Biopolymers 26, 1859–1877.
[30] Gerstein, M. (1997) J. Mol. Biol. 274, 562–576.