

Multiple regression analysis of the beta-sheet propensity of amino acids

Tomoko Sotomatsu Niwa*, Akio Ogino

Research Laboratory, Nippon Shinyaku Co., Ltd, Nishiohji, Hachijo, Minami-ku, Kyoto, 601 Japan

Received 1 November 1996; accepted 12 February 1997

Abstract

To study the nature of the beta-sheet propensities of naturally occurring amino acids, multiple regression analyses were performed with the use of various physicochemical parameters and experimentally determined physicochemical properties of amino acids. The beta-sheet propensities used for the analyses were free energies changes in the zinc-finger peptide and IgG-binding domain of protein G taken from the literature. Correlation equations of high quality were obtained for all data sets. Steric parameters of the amino acid side chains and the nuclear magnetic resonance (NMR) chemical shift of the alpha-carbon of amino acids were statistically highly significant in all cases. The addition of an indicator variable expressing the electronic effects of amino acids improved the correlations in some cases. From the correlation equations, the beta-sheet propensities were shown to be determined mainly by the electronic and steric effects of the amino acid side chains. The results obtained also serve as a basis for evaluating more precise beta-sheet propensity scales. © 1997 Elsevier Science B.V.

Keywords: Beta-sheet; Propensity; Amino acids; Protein structure; Statistical analysis

1. Introduction

The beta-sheet structure is one of the major secondary structures of proteins, and the beta-sheet propensity scale is indispensable to conformational studies of proteins [1]. The beta-sheet propensity is also important in drug design studies, because proteolytic enzymes such as aspartic acid and serine proteases bind both substrates and inhibitors as beta-pleated sheets. In spite of these findings, the physical basis of beta-sheet propensities is not well understood.

The objective of this study was to elucidate the factors determining the beta-sheet-forming propensities of amino acids. Their propensities may be governed by steric, electronic, and hydrophobic factors, and it would thus be difficult to analyze them by the

linear correlation method. Hence, we selected the multiple correlation method using various kinds of independent variables.

The statistically determined beta-sheet propensity scale by Chou and Fasman (P_{beta}) [2] is well known. However, their P_{beta} values were the averaged values derived from the known structures of different kinds of proteins. Here we chose experimentally determined free energy changes of proteins, and analyzed the data from different sources separately, in order to make clear the environmental and positional effects on the beta-sheet propensities.

2. Methods

The thermodynamic data used in the present analyses were taken from the literature [3–6]. Statistical

* Corresponding author.

analyses were performed with the StatView program [7] run on a Macintosh PowerPC 8500. In regression analyses, we used various hydrophobic, steric, and electronic parameters of amino acid side chains and experimentally determined values such as nuclear magnetic resonance (NMR) chemical shifts and pK_a values of amino acids. When necessary, indicator variables were applied. These parameters (except for the hydrophobic parameters) were taken from the literature [8,9]. The hydrophobic parameters were those determined by us with the use of experimentally determined 1-octanol/water partition coefficients of oligopeptides [10].

3. Results

3.1. Zinc-figure peptide

Propensities for residues in the beta-sheet have been studied with the use of various host–guest systems [3–6]. Kim and Berg measured the thermodynamic beta-sheet propensities for each of the 20 commonly occurring amino acids by using a zinc-figure peptide host system in which amino acids were substituted into a guest site (residue 3), a solvent-exposed position in an antiparallel beta-sheet [3]. Because these peptides are unfolded in the absence of bound metals but are folded in their presence, they assumed that the thermodynamics of metal binding fully reflect peptide-folding energy. In order to determine these energies with high precision, they used a competitive cobalt(II)-binding assay. For the relative free energies, $\Delta\Delta G$, listed in Table 1, Eq. (1) was given. Pro was excluded from the analyses throughout the present study. The $\Delta\Delta G$ values and parameters used in analyses are listed in Table 1.

$$\Delta\Delta G = -0.492B1 - 0.037B5 + 0.483 \quad (1)$$

(0.111) (0.013) (0.172)

$$n = 19, s = 0.040, r = 0.949, F = 72.560, p < 0.0001$$

$B1$ and $B5$ are the STERIMOL parameters by Verloop [8,9], and express the minimum and maximum width of amino acid side chains. In Eq. (1) and the following equations, n is the number of compounds, s is the standard deviation, r is the correlation coefficient, F

is the ratio of variances between the calculated and observed values, and p is the p -value, probability of the null hypothesis. The figures in parentheses are the 95% confidence intervals. Eq. (1) shows that sterically bulky amino acid side chains increase the beta-sheet propensities.

Next, we applied the NMR chemical shift of the alpha-carbon of amino acids [9], ΔHc , and formulated Eq. (2).

$$\Delta\Delta G = -0.265\Delta Hc/10 - 0.126 \quad (2)$$

(0.074) (0.089)

$$n = 19, s = 0.058, r = 0.879, F = 57.705, p < 0.0001$$

The ΔHc values were divided by 10 to make the coefficient with ΔHc comparable to those with other parameters. Eq. (2) shows that the higher the $\Delta Hc/10$, the higher the beta-sheet propensities.

3.2. IgG-binding domain of protein G

Minor and Kim measured the beta-sheet propensity for the naturally occurring amino acids in a variant of the small, monomeric, beta-sheet-rich, IgG-binding domain from protein G [5]. This protein exhibits a reversible two-state thermal denaturation transition. Amino-acid substitutions were made at a guest site (residue 44) on the solvent-exposed surface of the beta-sheet. The stability of each protein was measured by thermal unfolding as monitored by circular dichroism. The measured $\Delta\Delta G$ values listed in Table 1 were analyzed as shown below.

$$\Delta\Delta G(44) = 0.965B1 + 0.735I(ST) - 1.572 \quad (3)$$

(0.596) (0.343) (0.921)

$$n = 19, s = 0.214, r = 0.840, F = 19.140, p < 0.0001$$

$I(ST)$ is an indicator variable which is one when the guest amino acid is Ser or Thr, and is zero for the other amino acids. Next, we applied the NMR chemical shift of the alpha-carbon of amino acids [9], ΔHc , and formulated Eq. (4).

$$\Delta\Delta G(44) = 0.508\Delta Hc/10 + 0.627I(ST) - 0.654 \quad (4)$$

(0.261) (0.326) (0.306)

$$n = 19, s = 0.196, r = 0.867, F = 24.302, p < 0.0001$$

Table 1
Free energy differences and physicochemical parameters used in the multiple regression analyses

Amino acid	ddG	$ddG(44)$	$ddG(53)$	$ddG(53)^*$	P_{beta}
Ala	-0.35	0.00	0.00	0.00	0.79
Arg	-0.44	-0.43	0.45	-0.40	0.94
Asn	-0.38	-0.24	-0.08	-0.52	0.66
Asp	-0.41	-0.10	-0.94	0.85	0.66
Cys	-0.47	0.08	0.52	-0.78	1.07
Gln	-0.40	0.04	0.23	-0.38	1.00
Glu	-0.41	0.31	0.01	-0.23	0.51
Gly	0.00	-0.85	-1.20	1.21	0.87
His	-0.46	-0.01	-0.02	-0.37	0.83
Ile	-0.56	0.02	1.00	-1.25	1.57
Leu	-0.48	-0.24	0.51	-0.45	1.17
Lys	-0.41	-0.40	0.27	-0.35	0.73
Met	-0.46	-0.02	0.72	-0.90	1.01
Phe	-0.55	0.16	0.86	-1.08	1.23
Pro	-0.23	0.62			
Ser	-0.39	0.63	0.70	-0.87	0.94
Thr	-0.48	0.83	1.10	-1.36	1.33
Trp	-0.48	-0.17	0.54	-1.04	1.24
Tyr	-0.50	0.11	0.96	-1.63	1.31
Val	-0.53	0.17	0.82	-0.94	1.64

Amino acid	$B1$	$B5$	$dHc/10$	pai_b
Ala	1.52	2.04	0.73	0.16
Arg	1.52	6.24	1.11	-0.50
Asn	1.52	4.37	0.80	-0.59
Asp	1.52	3.78	0.92	-0.03
Cys	1.52	3.41	1.44	0.56
Gln	1.52	3.53	1.06	-0.58
Glu	1.52	3.31	1.14	-0.05
Gly	1.00	1.00	0.00	0.00
His	1.52	5.66	1.02	0.10
Ile	1.90	3.49	1.61	1.37
Leu	1.52	4.45	1.01	1.47
Lys	1.52	4.87	1.09	0.53
Met	1.52	4.80	1.04	0.94
Phe	1.52	6.02	1.39	1.72
Pro	1.52	4.31	1.78	0.76
Ser	1.52	2.70	1.31	-0.27
Thr	1.73	3.17	1.67	0.01
Trp	1.52	5.90	1.32	1.97
Tyr	1.52	6.72	1.39	1.06
Val	1.90	3.17	1.72	0.95

ddG : $\Delta\Delta G$ (kcal mol⁻¹); $dHc/10$: $\Delta Hc/10$.

The results for the IgG-binding domain of protein G were similar to those obtained for the zinc-finger peptide, except for the $I(ST)$ term. Note that, owing to the differences in the methods of studying the beta-sheet propensities, the $\Delta\Delta G$ values of the

IgG-binding domain of protein G are opposite in sign to those of the zinc-finger peptide.

Minor and Kim also measured the beta-sheet propensity of the IgG-binding domain of protein G, in which amino acid substitutions were made at a guest

site (residue 53) at the central strand bordered on both sides by other beta-sheet strands [4]. The neighboring environment of the guest site was modified to minimize the potential for artifactual interactions. Statistical analyses gave Eq. (5), where $I(\text{anion})$ is an indicator variable which is one when the amino acid at the position 53 is Glu or Aps, and is zero for the other amino acids; $I(\text{anion})$ expresses the anionic charge effect of amino acid side chains.

$$\text{delta}.\text{delta}G(53) = 1.276\text{delta}Hc/10 - 0.734I(\text{anion}) - 1.046 \quad (5)$$

(0.300) (0.376) (0.370)

$$n=19, s=0.236, r=0.933, F=53.488, p < 0.0001$$

As in the case of Eq. (4), $\text{delta}Hc/10$ was statistically highly significant. It should be noted that anionic side chains decreased the beta-sheet propensities in this case.

Smith and co-workers also studied the beta-sheet propensity for the naturally occurring amino acids in the B1 domain of staphylococcal IgG-binding protein G [6]. Amino acid substitutions were made at the same guest site (residue 53). The protein's thermal stability was determined to measure the beta-sheet forming propensities. The obtained $\text{delta}.\text{delta}G(53)^*$ values for beta-sheet formation were analyzed as shown below.

$$\text{delta}.\text{delta}G(53)^* = -1.429\text{delta}Hc/10 + 0.779I(\text{anion}) + 1.003 \quad (6)$$

(0.397) (0.496) (0.487)

$$n=19, s=0.311, r=0.908, F=37.699, p < 0.0001$$

As in Eq. (5), $\text{delta}Hc/10$ and $I(\text{anion})$ were statistically significant. Because of differences in the methods, the $\text{delta}.\text{delta}G(53)$ values obtained by Minor and Kim are opposite in sign to the $\text{delta}.\text{delta}G(53)^*$ values by Smith and co-workers.

3.3. Correlations between B1 and $\text{delta}Hc/10$

In the above analyses, both B1 and $\text{delta}Hc/10$ were statistically highly significant. We then examined the correlations between B1 and $\text{delta}Hc/10$, and obtained eqns (7) and (8).

$$\text{delta}Hc/10 = 1.857B1 - 1.721 \quad (7)$$

(0.577) (0.896)

$$n=19, s=0.211, r=0.855, F=46.128, p < 0.0001$$

$$\text{delta}Hc/10 = 1.818B1 + 0.360I(\text{STC}) + 0.381I(\text{aro}) - 1.778 \quad (8)$$

(0.313) (0.154) (0.153) (0.486)

$$n=19, s=0.112, r=0.966, F=68.893, p < 0.0001$$

As shown in eqns (7) and (8), there are high correlations between B1 and $\text{delta}Hc/10$. Since B1 mainly expresses the branching effects on the beta-carbon of amino acid side chains (see Table 1), the beta-branching structure is demonstrated to increase the $\text{delta}Hc/10$ values. $I(\text{STC})$ is an indicator variable which is one when the amino acid is Ser, Thr, or Cys, and is zero for the other amino acids, and expresses the contribution of inductive electronic effects of hetero atoms of the amino acid side chains upon the alpha-carbon. $I(\text{aro})$ is an indicator variable which is one when the amino acid is Phe, Tyr, or Trp, and is zero for the other amino acids. Aromatic residues also increase the $\text{delta}Hc/10$ values. Table 2 lists the correlations among the parameters used above in eqns (1)–(6); high colinearity was found.

4. Discussion

4.1. Statistical quality

As shown in Table 1, the ranges of the free energy differences were small, except for Pro, which has no hydrogen atom on the backbone amide group. However, we were able to formulate correlation equations of high quality. The correlation coefficients, r , were larger than 0.80 in all cases. The multiple regression method proved to be more suitable than the usual simple regression method to analyze the beta-sheet propensity.

4.2. Contributing factors

4.2.1. Hydrophobic effect

We previously evaluated the hydrophobic parameters of amino acids by using experimentally determined 1-octanol/water partition coefficients of oligopeptides [10]. These parameters are free from conformational factors such as beta-turn, and have been shown to work well to rationalize the thermal stability of proteins [10]. Our hydrophobic parameters, pai_b , were statistically insignificant throughout the present analyses. The hydrophobic factor is hence unimportant for the beta-sheet preference.

Table 2
Correlation matrix of free energy differences and physicochemical parameters

	<i>ddG</i>	<i>ddG</i> (44)	<i>ddG</i> (53)	<i>ddG</i> (53)*	<i>B1</i>	<i>B5</i>	<i>dHc</i> /10
<i>ddG</i>	1.000						
<i>ddG</i> (44)	−0.529	1.000					
<i>ddG</i> (53)	−0.786	0.595	1.000				
<i>ddG</i> (53)*	0.782	−0.596	−0.959	1.000			
<i>B1</i>	−0.828	0.570	0.692	−0.654	1.000		
<i>B5</i>	−0.593	−0.058	0.406	−0.491	0.164	1.000	
<i>dHc</i> /10	−0.879	0.704	0.855	−0.839	0.855	0.342	1.000

ddG: $\Delta\Delta G$; *dHc*/10: $\Delta H_c/10$.

4.2.2. Steric effect

Compared with the alpha-helix structure, the beta-sheet structure is less crowded. We did not find critical steric interactions among the side chains of amino acids around the guest site of the IgG-binding domain of protein G (PBD code: 2GB1) from molecular graphics studies. Additionally, the guest site of zinc-finger peptide is located near the terminal position (residue 3), and the neighboring environment of the guest site was modified to minimize the potential for artifactual interactions in the IgG-binding domain of protein G. The steric interaction between the amino acid side chains is thus unimportant. In fact, as is shown in eqns (1), (2) and (3), bulky and beta-branched side chains increased the beta-sheet propensities. Hence it appears that the steric effects by bulky and beta-branched side chains restrict the access of water and protein groups and thereby enhance the hydrogen-bonding interactions of backbone amide groups or electrostatic interactions between main chains.

4.2.3. Electronic effect

The $\Delta H_c/10$ values are roughly regarded as the inductive effect of amino acid side chains, since the degree of shielding of the alpha-carbon nucleus depends on the density of the circulating electrons. Eqs. (2), (4)–(6) thus show that the inductive effect largely determines the beta-sheet propensities. The significance of the $I(ST)$ term also suggests the importance of the inductive effect.

Recently, Avbelj and Moulton [11] reported that the main-chain electrostatics are important to determine the conformational preferences of amino acids. Maccallum and co-workers proposed that Coulombic interactions between charged main-chain atoms not

hydrogen-bonded to each other influence the conformations of antiparallel beta-sheets [12,13]. The significance of the $\Delta H_c/10$ term in the present eqns (2), (4)–(6) supports both models. Of course, our results do not deny the hydrogen-bonding interactions between main-chain amide groups.

In eqns (5) and (6), an indicator variable, $I(\text{anion})$, was highly significant, but not in eqns (3) and (4). This can be explained in two ways. First, residue 44 locates on the solvent-exposed surface of the edge beta-sheet, the anionic side chains are well solvated, and the charged effects are weakened. In contrast, residue 53 is at the central strand bordered on both sides by other beta-sheet strands, is not fully solvated, and still interacts with protein groups. Second, the electrostatic interactions with nearby amino acids decrease the beta-sheet propensities of residue 53. Because electrostatic interactions are long-range interactions, we could not at present identify the amino acids that interact with residue 53.

Because *B1* is highly correlated with $\Delta H_c/10$ (eqns (7) and (8)), separation of the steric effect from the electronic effect is unfortunately impossible. Steric parameters *B1* and *B5* gave the best equation for zinc-finger peptides, and $\Delta H_c/10$ gave the best equations for the IgG-binding domain of protein G. It is thus more reasonable to conclude that both the steric and electronic effects operate cooperatively to support the beta-sheet structures.

4.2.4. Environment effects

B1 or $\Delta H_c/10$ was significant in all cases, and *B1* was highly correlated with $\Delta H_c/10$. Hence *B1* and $\Delta H_c/10$ can be understood to express the intrinsic beta-sheet preferences of amino acids. We

compared the coefficients with $B1$ and $\Delta H_c/10$ in eqns (1)–(6). The absolute values of the coefficients for zinc-finger peptides were much smaller than those for the IgG-binding domain of protein G. Additionally, these coefficients changed with the change of the guest position for the IgG-binding domain of protein G. The beta-sheet propensities were thus demonstrated to be largely dependent on the kind of peptide or protein and the position of the guest amino acid.

4.3. Evaluation of propensity scales

The Chou–Fasman probability values are averaged over all possible beta-sheet environments, such as middle and edge positions as well as partially and fully hydrogen-bonded positions. Actually, we failed to formulate statistically significant correlation equations for their probability values. This is easily understood when eqns (3)–(6) are compared. To gain beta-sheet propensity scales by a statistical survey of experimental structural data, selection of the data is critically important; only central and fully hydrogen-bonded beta-sheets should be selected. Another way to evaluate beta-sheet propensity scales is to prepare good model proteins. The IgG-binding domain of protein G appears to be an appropriate, when further amino acid mutations are done to elucidate the electronic contributions expressed by the $I(\text{anion})$ term in eqns (5) and (6).

5. Summary

In spite of the difficulties described above, we were able to formulate correlation equations of high quality through use of the multiple correlation method and make clear the basis of the beta-sheet propensities. Such analyses are expected to be useful to study other structural properties of proteins.

References

- [1] C.L. Nesloney, J.W. Kelly, *Bioorg. Med. Chem.* 4 (1996) 739.
- [2] P.Y. Chou, G.D. Fasman, *Biochemistry* 13 (1974) 211.
- [3] C.A. Kim, J.M. Berg, *Nature* 362 (1993) 267.
- [4] D.L. Minor Jr., P.S. Kim, *Nature* 367 (1994) 660.
- [5] D.L. Minor Jr., P.S. Kim, *Nature* 371 (1994) 264.
- [6] C.K. Smith, J.M. Withka, L. Regan, *Biochemistry* 33 (1994) 5510.
- [7] StatView, Abacus Concepts, Inc., Berkeley, CA, 1992.
- [8] A. Verloop, J. Tipker, *Pharmacochem. Libr.* 10 (QSAR Drug Des. Toxicol.) (1987) 97.
- [9] J.-L. Fauchere, M. Charton, L.B. Kier, A. Verloop, V. Pliska, *Int. J. Pept. Protein Res.* 32 (1988) 269.
- [10] T.S. Niwa, A. Ogino, paper 10 presented at ECCS-2; *J. Mol. Struct. (Theochem)* 392 (1997) 43–53.
- [11] F. Avbelj, J. Moul, *Biochemistry* 34 (1995) 755.
- [12] P.H. Maccallum, R. Poet, E.J. Milner-White, *J. Mol. Biol.* 248 (1995) 361.
- [13] P.H. Maccallum, R. Poet, E.J. Milner-White, *J. Mol. Biol.* 248 (1995) 374.