Protein Engineering vol.11 no.12 pp.1137–1145, 1998

Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability

Boojala V.B.Reddy[1], Sunando Datta[2] and Shrish Tiwari

Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500007, India

[2]Present Address: Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

[1]To whom correspondence should be addressed at the University of California, #367, San Diego Supercomputer Center, MC 0505 9500 Gilman Drv., La Jolla, CA 92093-0505, USA

Advances in site-directed mutagenesis and other genetic engineering techniques have made it possible to create novel proteins of interest. A challenging aspect of these studies is to understand the effect of substitution mutations on folding and stability of natural proteins. We present an analysis of protein structure data, available from the literature, for which substitution mutations have been made and changes in stability characteristics are reported. Amino acid structural environment parameters have been computed for a set of 304 non-homologous best resolved protein structures. The structural environment parameters were used to calculate each of the 20 amino acid propensities to a given structural environment. The observed increase or decrease in stability upon mutation was found to be correlated with the average residue structural environment propensity of wild-type residue versus mutant residue. The analysis presented here helps identification of less optimally placed residues in a given protein structure, and suggests possible substitution mutations to a residue with higher propensity to the corresponding local structural environment. We propose that such substitution mutations, suggested based on amino acid propensities to local structural environments, should bestow higher stability to the protein structure.

Keywords: point mutations/free-energy change/residue structural propensity/protein stability/protein engineering

Introduction

A variety of theoretical and experimental techniques are being used to further understanding of the physical principles, forces and mechanistic pathways leading to protein folding and stability. Yet a quantitative understanding of the role of individual amino acids in both the direction of protein folding and the stabilization of protein structure is lacking (Baldwin, 1994). At present the problem of protein stability is gaining widespread interest not only among the researchers interested in basic studies of protein structure and folding but also among those interested in using enzymes as practical catalysts under different experimental conditions (Gupta, 1993; Braxton, 1996).

X-Ray studies have shown that the substitution of one amino acid for another generally results in small conformational changes that are restricted to the immediate vicinity of the substituted site (Lesk and Chothia, 1986; Shortle, 1992; Shortle and Sondek, 1995). However, the thermodynamic stability of the single site mutant can change significantly from that of the corresponding native protein (Querol *et al.*, 1996). Therefore genetic engineering techniques have been employed to modify the amino acid sequence of enzymes of industrial interest at specific sites on the polypeptide chain (site-directed mutagenesis) in order to improve their stability towards heat, pH and other common protein denaturants (Bryan, 1995). A limitation to the successful design of these new enzyme properties by site-directed mutagenesis is that the effects of amino acid replacements are not easy to predict—thus it is difficult to decide which amino acid substitutions should be made.

Effects of substituted mutations—theoretical consideration

We use the term 'protein stability' to designate the effect of reversible denaturation as many proteins behave in solution as if there are only two states of the chain: the native state *N* and a reversibly denatured state *D*. The *D* state is operationally defined as the state that the polypeptide chain enters upon a major cooperative breakdown of the *N* state by the loss of a large fraction of its structure. In this context, the best measure of the stability of the *N* state is the free-energy change, $\Delta G_{N\rightarrow D}$ or simply ΔG, on conversion from *N* to *D*. The free-energy term, therefore, is of greatest interest for understanding the functional aspects of the reaction mechanism and stability of a protein. The free-energy difference between the *N* and the *D* states, $\Delta G = G_D - G_N$, can be obtained by measurement of the appropriate equilibrium constant *K* ($[S_D]/[S_N]$) in the equation $\Delta G = \Delta G_0 - RT \ln K$.

Amino acid substitutions provide a very precise way to alter chemical structure of a particular side-chain, allowing comparison of the reaction energetics for the mutant protein with that of the wild-type. A thermodynamic cycle can be defined revealing the four free energy terms as follows:

$$\Delta G_{wt}$$

$$S_{wt,\,N} \longleftrightarrow S_{wt,\,D}$$

$$\Delta G_N \updownarrow \qquad\qquad \updownarrow \Delta G_D$$

$$S_{mut,\,N} \longleftrightarrow S_{mut,\,D}$$

$$\Delta G_{mut}$$

$\Delta G_{wt} = G_{D,wt} - G_{N,wt,\,\Delta} G_{mut} = G_{D,mut} - G_{N,mut}$ and $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt.}$

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt.} = \Delta G_D - \Delta G_N$$

The understanding of the physical and chemical origins of the changes in free energy provides a direct route to an improved understanding of the workings of the wild-type proteins. However there are numerous difficulties such as (i) what fraction of $\Delta\Delta G$ arises from mutant effects on state N (ΔG_N) and state D (ΔG_D)? And how to calculate these components? (ii) Secondly, the free-energy (G) of each of the four states has multiple components, $G = G_{\text{electrostatics}} + G_{\text{hydrogen bonds}} + G_{\text{van der Waals attractions}} + G_{\text{steric repulsions}} + G_{\text{hydrophobicity}} \ldots$ A dissection of $\Delta\Delta G$ would have to address these components in different structures. It is, therefore, difficult to address every detail of the mechanism(s) responsible for the mutant effects. Consequently, results to date have not provided any reliable quantitative insights into the details of the interactions that stabilize the native state. Instead, a number of general statistical trends in the data provide indications of the relative importance of different types of interactions.

Determinants of protein stability—statistical analyses

Analysis of homologous proteins from the group of thermophiles and mesophiles provided some significant clues for related proteins, performing the same functions, having very different stability and helped in identifying candidate amino acids for point mutations with the aim of increasing protein stability (Argos *et al.*, 1979; Imanaka *et al.*, 1986; Menéndez-Arias and Argos, 1989; Mrabet *et al.*, 1992; Querol *et al.*, 1996). Several studies on the analysis of protein structures with substituted point mutations have enabled the quantification of the contribution of the different interactions that take place in a protein, resulting in some general rules about possible ways to increase protein stability (Shortle, 1992; Gupta, 1993; Mathews, 1993; Serrano *et al.*, 1993; Fersht and Serrano, 1993; Filippis, 1994). However, it is often difficult to identify the important determinants involved in a specific case, since the sequences of related proteins diverge significantly, each thermozyme or substituted mutant is stabilized by a unique combination of different mechanisms (Vieille and Zeikus, 1996; Querol *et al.*, 1996).

There have been reports indicating contribution of amino acid exposure to the solvent, its hydrophobicity, secondary structural preferences, hydrogen bonding and electrostatic interactions, disulfide bonds, substrate–metal cofactor binding etc. to protein stability (Shortle, 1992; Filippis *et al.*, 1994; Murphy, 1995; Munoz *et al.*, 1996; Pace *et al.*, 1996). From the analyses of mutant protein structures, a more statistical account of these and many other physical and chemical contributions to protein stability have been reported (Querol *et al.*, 1996). Through a comparative study of sequence data of stable versus less-stable proteins a method was proposed to suggest substitution mutations to increase intracellular stability of proteins (Guruprasad *et al.*, 1990; Reddy, 1993, 1996; Reddy *et al.*, 1998). There are also some recent reports which suggest substitution mutations solely from amino acid sequence (Varadarajan *et al.*, 1996), from database derived potentials (Gilis and Rooman, 1997) and by rational design of α-helix stability using helix/coil transition theory (Villegas *et al.*, 1995). There are useful predictions of free-energy changes using amino acid substitution-based information for different mutant proteins (Bordo and Argos, 1991; Lopez-Hernandez and Serrano, 1995; Topham *et al.*, 1997).

We have recently examined a statistical approach, based on residue propensities, to predict the effects of a substitution mutation, and the preliminary results were presented in a short report (Reddy and Datta, 1997). In this communication we discuss the approach in greater detail; we use a series of residue propensity values separately for each of the physico-chemical structural parameters that contribute to overall *in vitro* protein stability. We calculate an average propensity of each amino acid in a given structural environment profile, defined based on the physico-chemical properties of the amino acids in protein structural environment. We demonstrate the use of such a structural profile of amino acid residues to predict the experimentally observed qualitative changes in the stability characteristics. We further show that these environment-dependent amino acid propensity values could be used to identify non-optimally placed residues and suggest substitution point mutations to engineer stability characteristics for a given protein structure.

Materials and methods

United States National Library of Medicine database (MEDLINE) for the years 1986–1998 was used to collect information on the proteins for which substituted point mutation(s) were carried out and where a change in the stability was reported. We have used protein structures for which 2.5 Å or better X-ray or NMR resolved structures of the parent proteins were available to calculate environment related parameters of amino acids. About 376 substituted single residue mutants have been identified for which parent protein structures and altered *in vitro* stability (thermal, solvent or *pH* induced) information is available (Table I), which from now on will be referred to as the 'mutant data set'. We have sub-divisions of this data set as mutant data set-1 (a 'sample data set' collected initially) and mutant data set-2 (additional 'test data set' collected subsequently). A data set of 304 non-homologous (⩽25% sequence identity) best resolved (⩽2.0 Å resolution X-ray defined) protein structures (Hobohm *et al.*, 1992), referred to as the 'natural data set' hereafter, is used to calculate general propensities of residues for a given structural environment.

The parameters used to define the structural environment of amino acid residues are secondary structure types, solvent accessibility, hydrogen bonding and packing density (Ooi number: Nashikawa and Ooi, 1980). Residue volume and hydrophilicity of the residues are also used to weigh against the residue occurrences for final calculation of the amino acid propensities. The parameters are computed using the methods described below.

Secondary structure

Secondary structural definition of Kabsch and Sander (1983) as summarized by Smith (1989) in his SSTRUC program was used to define the secondary structure type taken by the residue in the wild type protein such as helices (α-helices + 3_{10} helices), β-strands and the remaining as random coils. Alternatively, the φ and ψ angles together define the residue secondary structure.

Packing density (Ooi number)

A contact number of other residues within a 8 and 14 Å radius (Ooi value) was computed using the method of Nashikawa and Ooi (1980). Since the longest distance from C_i^{α} to C_{i+1}^{α} is only about 4 Å, the nearest neighbour residues on either side of the dipeptide were omitted in the counting. Ooi numbers calculated in both, 8 and 14 Å radius are used together as a structural environment parameter (SEP).

Hydrogen-bond formation

Hydrogen-bond formation was defined based on the criterion of a donor–acceptor distance of ⩽ 3.5 Å (Baker and Hubbard, 1984). Angular criteria were not considered because the side-chain atoms are generally not well positioned by crystallography

Table I. Data of proteins with substitution point mutations that alter *in vitro* stability

(a) 'Mutant data set -1' initially collected and used to optimize the prediction method

1BNIA:	T 6 S-G-A-Q-E-N-D+; T 26 N-S-G-V-Q-E-; R 83 K–; I 109 V-A- I 4 V-A-; N 5 A-; D 8 A-; V 10 T-A-; Y 13 A-; L 14 A-; Q 15 I+; T 16 S-; Y 17 A-; H 18 Q-; N 23 A-; I 25 V-A-; E 29 G-; L 33 Q-; V 36 T-A-; N 41 D-; V 45 T- A-; I 51 V-; D 54 N-A-; I 55 A-T-; N 58 A-; K! 62 R-; I 76 V-A-; N 77 A-; Y 78 F-; N 84 A-; I 88 V-A-; L 89 V-T-; S 91 A-; S 92 A-; I 96 A-V-; T 99 V-; T 105 V
1LYSA	S 91 T+V-A-D-Y-; T 40 S-I-; I 55 V-M-F-A-T-; H 15 L+; A 31 V+; D 101 S+; R 114 H+
2LZM	P 86 A-S-R-T-D-C-H-I-L-G-; M 102 K-; L 133 D-A-; T 157 R-I-; S 38 D+N-; D 92 N-; T 109 D+N+; T 115 E+; N 116 D+; R 119 M+; N 144 D+H+E+; Q 123 E+; N 101 D-; E 128 A+; D 127 A+; V 131 A+; N 132 A+
1SBT	Q 19 E+; V 26 R-; T 164 R-; N 218 S+; L 235 R-; Q 271 E-
2RN2	D 134 A+C+E+F+G+I+K+L+M+N+P+Q+S+T+V+W+Y+

1BTL T 71 S-	1MPP A 102 T-; G 176 D-
1BGH I 47 V-; V 35 I-	1WSYA E 49 F+; D 60 F-
1ALD D 128 G-	2TRXA D 26 A+; P 34 S=; P 76 A-
1ALKA D 101 A-	1YCC K 73 M-Y-F-W-; F 82 S-
7APIA E 264 V-	1DHFA W 24 F-
3DRCA G 121 V-L-	2HIPA Y 12 F-H-N-
1POW P 178 S+; S 188 D+; A 458 V+	1HGU S 71 A-V-Q-T+
1IFL P 30 A-G-C-S-	1HUW E 74 D-Q-S-T-L-A-
1LDB R 171 Y+W+	1FXAA C 41 S-;C 46 S-; C 49 S-; C 79 S-
1NDK R 109 A-; N 119 A-Q-	1NDPA R 109 A-; N 119 A-Q-
1NDC R 109 A-; N 119 A-Q-;	1CLF P 19 K-; P 48 K-
9PAP N 175 A+Q+	

(b) 'Mutant data set -2' additional data collected to test the prediction accuracy

1CB1 L 6 V-; F 10 A-; L 23 A-G-; L 28 A-; V 61 A-G-; F 66 A-W-; V 70 L-; I 73 V-
1SUP K 43 N+; M 50 F+; A 73 L+; Q 206 V+; Y 217 K+; N 218 S+; Q 271 E+
1A2WA A 19 P-; Q 28 L-; K 31 C-; S 32 C-; Y 97 A-F-G-

1RCH	D 10 N+; E 48 Q+; A 52 C+D-E-F-G-H-I+K-L+M+N-P-Q-S-T-V+Y-; D 70 N+; D 134 N+
1AZF	H 15 L+; A 31 V+; T 40 I-S-; I 55 A-F-L-M-T-V-; S 91 A-D-T+V-Y-; D101S+; R 114 H+
1LYD	M 6 A-I-; L 7 A-; I 17 A-; I 27 A-; I 29 A-; L 33 A-; L 46 A-; I 50 A-; I 58 A-; L 66 A-; F 67 A-; V 71 A-; I 78 A-; L 84 A-; V 87 A-; L 91 A-; V 94 A-; L 99 A-F-I-M-V-; I 100 A-; M 102 T-; V 103 A-; F 104 A-; M 106 A-; V 111 A-; L 118 A-; L 121 A-; L 133 A-; V 149 A-; F 153 A-L+M-V-
1REX	I 23 V-; I 56 V-; I 89 V-; I 106 V-
1STN	I 15 G-; L 36 G-; L 37 G-; L 38 G-; V 39 G-; L 108 G-
1GAI	G 121 T-; R 122 Y-; P 123 G-; Q 124 H+; R 125 K+
1XOA	S 1 H-; D 26 A+; W 28 A-; E 30 H-; Q 62 H-
1DDRA	G 67 A+C-D-L-S+T-V-; G 121 H-L-V-Y-; A 145 F-G-H-R-S-T-V-

1SCD F 189 P+	1CLL F 92 A-
1HDGO R 20 A-N-	1CYDA T 38 D+
1AZI L 104 N-	1BTL C 77 S-
1TPFA H 47 N-	1FHB N 52 A+I+; Y 67 F+
1CDKA E 230 A-;	1A2PA Q 15 I+; H 18 Q-; N 58 A-
1CHKA E 22 A+D-Q-	1MYLA N 29 A+; S 44 A+; E 48 A+
1EGDA K 304 E-	1MYKA N 29 A-; S 44 A-; E 48 A-
1DPO D 189 S+	2SAK M 26 A+
1BURA L 290 F-	1OMD S 55 D+; G 98 D+
1BGAA E 96 K+; M 416 I+	1IPD A 172 V+
1APS C 21 A-S-	1YGW A 21 F-G-I-L-M-V-
1GVP V 35 F-; V 45 T-; I 78 C-	4GCR F 56 A-D-W-
1ISCA Y 34 F-	1AP6A I 58 T-

The PDB code followed by the residue in the wild type protein, its position and the residue(s) to which it is mutated. The sign following by the residue to which mutation was made indicates whether there was an increase(+), decrease(–) in *in vitro* stability as observed in the original reports. A table with more exhaustive information is available on request.

and not all hydrogen atom positions are fixed by the positions of the heavier atoms. Hydrogen bonding was examined from a side-chain at positions i to the residues other than those at positions $i - 1$, i and $i + 1$. The average number of hydrogen bonds (dipole interactions) that could be formed by the residue in a given protein structure were computed.

Solvent accessibility

Solvent accessible contact area of amino acids was calculated using the method of Lee and Richards (1971), as coded by Sali and Blundell (1990) in their PSA program, with a probe radius of 1.4 Å. The percentage accessibility contact area of the residue for side-chain, main-chain, polar side-chain, non-polar side-chain and total atoms are used.

Occurrence of amino acids in structural class intervals

The structural parameters are calculated for each of the 20 amino acids in the natural data set. Similarly, for every wild type residue which has been subjected to a mutation with an observed change in $\Delta\Delta G$, structural environment parameters

Table II. Representative examples of the data of proteins with substitution point mutations that alter the *in vitro* protein stability. RSEP of the wild type residue of mutant protein with increase/decrease in stability information, used as input data file to get propensities for the 20 amino acids in the natural data set

RSEP	PDB code	AA	AA S. No.	Solvent accessibility					SS	ϕ	ψ	Ooi No.		MHB	THB	No. of Mt	Mt. AA
				TA	PS	NS	TS	TM				8 Å	14 Å				
SA1	1BNIA	V	10	0.0	0.0	0.0	0.0	0.0	H	–76.9	–39.7	12	49	3	3	2	T-A-
SA2	1FXAA	C	46	19.1	25.0	0.0	25.0	0.0	S	–144.3	–170.7	9	36	0	0	1	S-
SA3	2LZM	T	109	82.7	107.2	93.8	104.2	13.2	H	–61.1	–52.7	7	31	3	3	2	D+N+
HB1	1BNIA	R	83	28.5	24.4	21.8	22.9	63.7	–	–70.1	150.2	8	43	0	0	1	K-
HB2	1BNIA	V	36	46.4	58.4	0.0	58.4	0.0	g	–120.9	113.2	9	26	3	5	2	T-A-
HB3	1ALKA	D	101	16.4	34.1	7.6	22.0	0.0	h	–104.1	–173.6	11	63	1	7	1	A-
Oi1	1POW	S	188	67.5	93.8	52.9	80.2	38	G	–93.6	–4.5	5	20	0	0	1	D+
Oi2	2LZM	S	38	43.2	61.5	58.5	60.5	2.8	h	–89.4	130.1	6	23	3	3	2	D+N-
Oi3	1SBT	Q	271	47.4	43.4	65.4	56.4	9.7	H	–80.3	–38.4	10	45	1	2	1	E-
SS1	1BNIA	D	8	61.9	70.4	95.9	82	2.8	H	–67	–43.1	6	35	2	4	1	A-
SS2	1BNIA	I	25	11.3	13.8	0	13.8	0	E	–141.2	144.2	12	42	2	2	2	V-A-
SS3	1BNIA	N	23	10.2	15.8	8.5	12	4.8	e	–81.3	–6.1	11	39	1	1	1	A-
SS4	1BNIA	N	58	16.9	8.3	16.5	12.6	30.1	t	51.6	42.2	8	35	2	4	1	A-
SS5	1BNIA	T	6	53.4	77.6	42.7	70.0	0.2	h	–101.9	154.2	8	41	3	3	7	S-G-A-Q-E-N-D+

RSEP, residue structural environment profile; PDB code; AA S. No., residue position in the sequence; AA, amino acid in the wild type protein; percentage solvent accessibilities for total atoms (TA), polar side chain atoms (PS), non-polar side chain atoms (NS), total side chain atoms (TS) and total main chain atoms ™; secondary structure type (SS); hydrogen bonding with main chain–main chain atoms (MHB) and total possible hydrogen bonding interactions (THB) respectively; Ooi number in 8 and 14 Å sphere of radii. The substituted residue and the corresponding effect on increase/decrease *in vitro* stabilities are indicated by the + or – signs.

(SEPs) in their native structures have been calculated and the example, see Table II for structural parameters calculated for a few wild type residues in their respective PDB structures. The calculated values of the structural parameters are divided into different class intervals and a total occurrence of residues in each class interval is computed. These occurrences are shown in histograms for the natural and for the mutant data set (Figure 1). Based on these occurrence values window sizes have been set for each structural parameter as described below.

We have assigned small window sizes spanning around a given value of each structural parameter; for example, for given ϕ and ψ values we consider residues within ±20 angular intervals as having a similar secondary structural environment. Similarly, for other structural parameters we have set appropriate window sizes as described in Table III. These window sizes are set based on the structural parameter value and the total occurrence of residues in the 'natural data set' with similar value evaluated. For example, in the case of the hydrogen bonding parameter we have set a zero window size for the main-chain—main-chain hydrogen bonding (MM) value and three window sizes for all the side-chain hydrogen bonding values. These are described in Table III(b). It can be seen from Figure 1A(ii) that occurrence of residues with 0–3 hydrogen bonds is significantly high, correspondingly the window size for all these values is set to a value of ±1. The occurrence of residues with 4–5 hydrogen bonds and with >5 hydrogen bonds is decreasing, correspondingly the window sizes for these values is set as ±2 and ±3, respectively. That is, as total occurrence is decreasing we have increased the window sizes. These window adjustments gives statistically meaningful residue occurrences for calculating their propensities by slightly compromising with similarity in the structural environment. A similar logic is followed to set window sizes for other structural environment parameters, see Table III.

Residue structural environment profile (RSEP)

The calculated structural parameters of each wild type residue with corresponding assigned window size (see Table III) is defined as the residue structural environment profile (RSEP). For example, the structural profile of 1BNIA-R83 (row 4 of Table II) is {28.5 ± 10, 24.4 ± 10, 21.8 ± 30, 22.9 ± 20, 63.7 ± 20, –, –70.1 ± 20, 150.2 ± 20, 8 ± 1, 43 ± 5, 0 ± 0, 0 ± 1}. The parameters are in the same order as given in Table II. In a natural data set any residue having a structural parameter value within the window size is said to be present in the similar structural environment. We take each (or a combination) of these structural values separately and count the occurrence of all the 20 residues in that RSEP in the natural data set of 304 non-homologous proteins and compute their propensity defined as:

$$Pe(x) = (ne(x)/Ne)/(n(x)/N)$$

where *Ne* is the total occurrence of all amino acids in the given RSEP of the wild type amino acid (R in the example given above), *ne*(*x*) is the total occurrence of a particular amino acid *x* in environment *e*, *n(x)* is the occurrence of amino acid *x* and *N* is the total number of amino acids in the natural data set. As an example the propensity values calculated for all the 20 amino acids in the *RSEP* of residues in Table II are given in Table IV.

Results and discussion

In protein structures, it is consistently observed that hydrophobic residues prefer buried regions and hydrophilic residues prefer surface regions of the structure. To study such preferences more precisely, many varied local structural environments could be defined in protein structures. Each of the 20 amino acids, having different side-chain properties, prefer an optimal local structural environment in the protein selected through a natural process. Thus, in a natural data set for every amino acid one can find a structural environment that shows a highest propensity and the remaining amino acids may also prefer that structural environment but with a lesser propensity. This observation (or assumption) is the basis for this analysis.

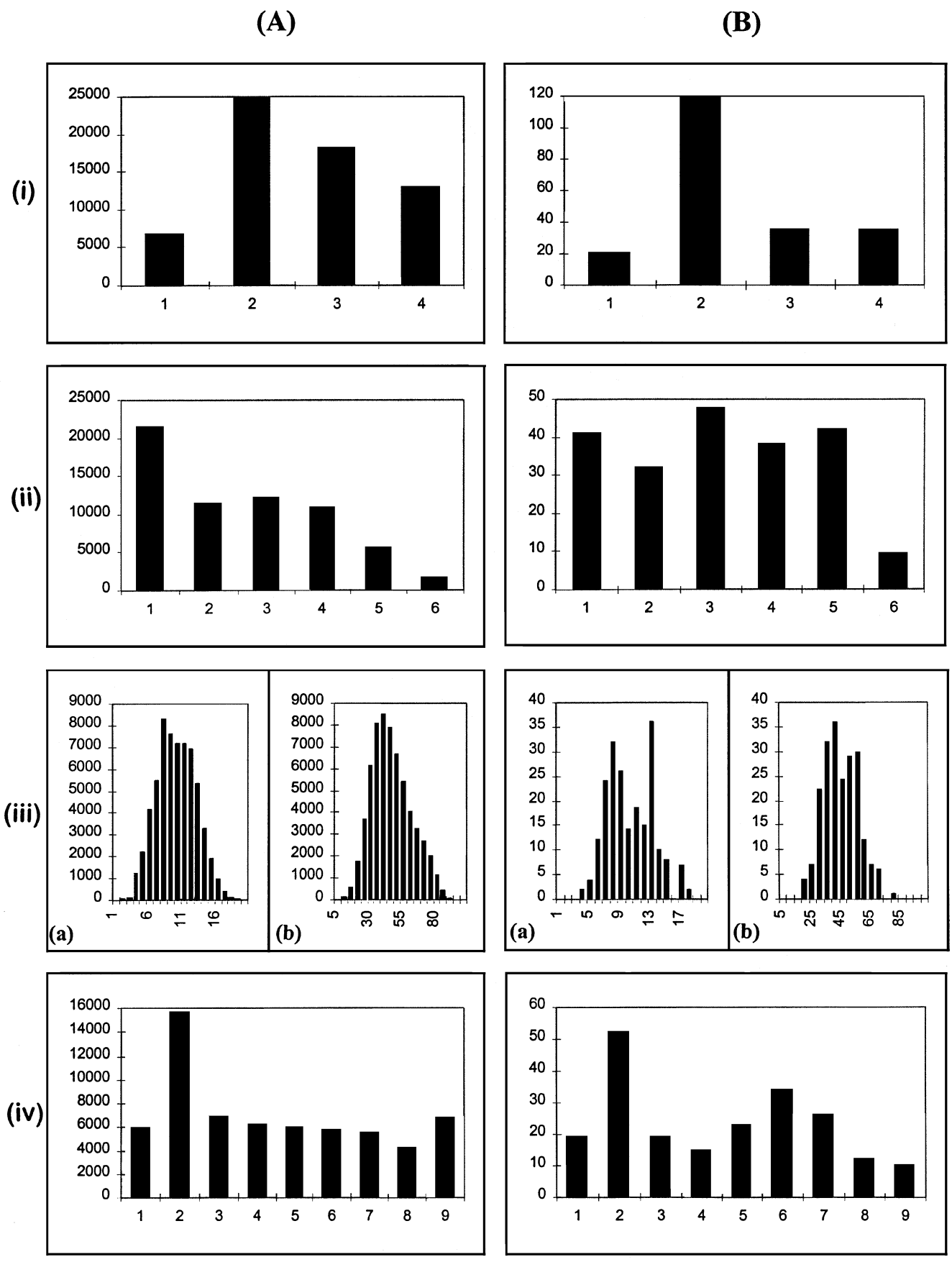

Fig. 1. Amino acid occurrence in the four structural classes: (**A**) 'natural data set', (**B**) wild type amino acids in 'mutant data set-1'. (i) Secondary structural classes: 1, coils; 2, helices; 3, β-strands; 4, turns. (ii) Occurrence with total number of possible hydrogen bonds: 1, no hydrogen bonds (HB); 2, 3 and 4, 1, 2 and 3 HBs, respectively; 5, 4–5 HBs; and 6, 6 or more HBs. (iii) Ooi number in 14 and 8 Å radii spheres; (iv) all atom solvent accessibility (SA) classes: 1, 0% SA; 2, ⩾0–10% SA; 3, ⩾10–20% SA;..., 8, ⩾60–70% and 9, ⩾70% accessibility. Window sizes have been set based on these occurrences—high occurrences takes smaller window sizes and lower occurrences takes larger window sizes.

Table III. Window sizes used to calculate occurrence of residues in a given structural environment of wild type amino acid

(1) Secondary Structure (SS): a window size of ±20° to the observed ϕ and ψ angles is used.
(2) Hydrogen Bonding (HB): (i) main chain—main chain HB number should be identical to the observed value of wild type residue and (ii) if the total number of remaining hydrogen bonds is: (a) ⩽3, the window size is chosen to be ±1; (b) 4–5, the window size is chosen to be ±2; (c) >5, the window size is chosen to be ±3.
(3) Ooi Number: the window sizes for Ooi numbers of amino acid for 8 and 14 Å are used as tabulated below:

(i) 8 Å Radius		(i) 14 Å Radius	
No. amino acids	Window size	No. amino acids	Window size
⩽5 or >14	±3	⩽20 or >65	±15
6,7,13,14	±2	21–30, 51–65	±10
8–12	±1	31–50	±5

(4) Solvent Accessibility: the window size used for given interval of percentage of solvent accessibility of wild type residue is tabulated below:

All Atoms		Non-polar SC		Polar SC		Total SC		Total MC	
(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
<10%	±5%	<10%	±5%	0%	±0%	<10%	±5%	<10%	±5%
10–50%	±10%	10–40%	±10%	0–20%	±20%	10–20%	±10%	10–20%	±10%
>50%	±15%	>40%	±15%	>20%	±30%	>20%	±20%	>20%	±20%

Table IV. All 20 amino acid propensities computed for structural environments of the wild-type residues (given in bold) in their respective structures (as in Table III). Note the significant variation in propensity values of amino acids for different structural environments

AA	RV	Hpo	SA1	SA2	SA3	HB1	HB2	HB3	Oi1	Oi2	Oi3	SS1	SS2	SS3	SS4	SS5
A	92	1.8	1.51	0.76	0.07	1.02	0.87	0.60	0.90	0.91	0.84	1.51	0.90	1.06	0.43	0.71
C	118	2.5	1.95	**1.49**	0.00	1.02	0.96	1.24	0.17	0.22	0.86	0.73	1.19	1.01	1.15	1.57
D	125	–3.5	0.18	0.01	1.03	0.95	1.16	**1.33**	1.72	1.67	0.90	**0.93**	0.27	1.51	1.89	0.57
E	155	–3.5	0.11	0.03	5.37	0.98	1.12	0.99	1.62	1.71	0.95	1.45	0.81	1.20	0.64	0.87
F	203	2.7	1.90	2.64	0.05	1.03	0.89	1.15	0.37	0.39	1.27	0.94	1.45	0.79	0.70	1.43
G	66	–0.4	1.02	0.60	0.03	1.03	0.95	0.79	1.65	1.54	0.72	0.42	0.25	0.53	2.34	0.42
H	167	–3.2	0.44	0.10	0.96	0.97	1.07	1.01	0.91	0.88	0.95	0.80	1.09	1.11	2.10	0.90
I	169	4.5	2.39	2.73	0.08	1.03	0.84	0.79	0.30	0.30	1.07	1.02	**1.43**	0.43	0.00	1.05
K	171	–3.9	0.09	0.06	2.16	1.02	1.00	0.65	1.36	1.54	1.06	1.19	1.00	1.10	1.15	1.20
L	168	3.7	2.07	2.34	0.02	1.02	0.82	0.68	0.31	0.31	1.26	1.25	0.77	1.01	0.24	1.39
M	171	1.9	2.05	2.28	0.00	1.03	0.79	0.86	0.45	0.44	1.02	1.25	1.25	0.92	0.50	1.04
N	135	–3.5	0.25	0.03	0.73	0.95	1.16	1.81	1.49	1.38	0.92	0.66	0.45	**1.35**	**5.47**	0.75
P	129	–1.6	0.62	2.41	0.00	1.04	0.94	0.91	1.60	1.41	0.92	0.74	0.00	1.33	0.00	0.43
Q	161	–3.5	0.16	0.00	3.48	0.97	1.08	1.23	1.13	1.25	**1.12**	1.33	0.94	1.00	1.25	1.14
R	202	–4.5	0.09	0.02	0.96	**0.91**	1.08	2.27	0.88	0.92	1.28	1.19	1.15	1.02	0.91	1.01
S	99	–0.9	0.57	0.18	2.07	0.98	1.19	0.91	**1.46**	**1.44**	0.77	0.80	1.34	1.66	0.64	1.06
T	122	–0.7	0.58	0.40	**1.83**	1.00	1.19	1.11	1.03	1.04	1.02	0.73	1.58	1.11	0.10	**1.48**
V	142	4.2	**2.16**	2.81	0.09	0.97	**1.10**	0.83	0.34	0.36	1.04	0.92	2.03	0.35	0.10	1.10
W	238	–0.9	1.03	0.75	0.27	0.99	1.08	0.96	0.35	0.32	1.31	1.00	1.22	0.97	0.26	1.53
Y	2.4	–1.3	0.58	0.16	0.22	0.65	2.01	0.87	0.33	0.39	1.16	0.88	1.76	0.67	0.63	1.53

AA, amino acid; RV, residue volume in $Å^3$; Hpo, hydrophobicity value; SA, all atom solvent accessibilities (0.0, 19.1 and 82.7 %, respectively); HB, hydrogen bonds (0, 5 and 9, respectively); Oi, Ooi numbers for 8 and 14 Å [(5, 20), (6, 23) and (10, 45) respectively]; SS, ϕ, ψ angles denoting different secondary structures: (–101.9, 154.2) and (–67.0, –43,1) of helices, (–141.2, 144.2) and (–81.3, –6.1) of β-sheets and (51.6, 42.2) of coil structure, respectively.

Amino acid occurrence in different structural environments

Figure 1 gives the general distribution of amino acids in different structural environments of the natural data set and the mutant data set-1. In the natural data set, a high percentage of residues are from helix secondary structural regions followed by β-strand, turn and coil regions. The occurrence of residues with no hydrogen bonding interactions is very high followed by residues with 1, 2, 3, 4–5 or ⩾6 hydrogen bonds. The occurrence of residues with Ooi number 7–13 in 8 Å radius and 35–65 in 14 Å radius is very high. We also find that the residue occurrence having 0–10% total atom solvent accessibility is very high in the data base. A similar pattern of wild type residue occurrence is observed in the mutant data set. To get a statistically significant value of residue occurrence and also to define a more similar structural environment for calculation of propensity values, window size for the different ranges of structural parameters is used.

The amino acid propensity values, calculated using protein structures from the 'natural data set', gives the optimally preferred structural environment for each of the 20 amino acids for any given RSEP. Correspondingly, these values are expected to be proportional to stability characteristics contributed by the physico-chemical nature of the residue. In this analysis a statistical combination of such calculated propensity values for different structural parameters are used to optimize and predict stability characters of protein due to specific mutations. The aim of the analysis was twofold: (i) to show how the natural propensity of each amino acid varies with the given local structural environment of the residue and (ii) to identify the best combination of structural environment

Table V. Predictions showing the correlation with average propensity using different combinations of structural parameters on mutation data sets

Structural type	Correlation score Number of +ves	–ves	% of +ves	Structural type	Correlation score Number of +ves	–ves	% of +ves
Mutation data set-1							
SS	113	78	59.2	HB+SA	126	65	66.0
OI	127	64	66.5	All-SS	133	58	69.6
HB	112	77	58.6	All-OI	124	67	64.9
SA	125	64	66.1	All-HB	129	62	67.5
SS+OI	129	62	67.5	All-SA	128	63	67.0
SS+HB	112	79	58.6	All	135	56	70.7
SS+SA	125	66	65.4	All+RV	140	51	73.2
OI+HB	125	66	65.4	All+HP	140	51	73.2
OI+SA	131	60	68.5	All+RV+HP	141	50	73.8
All+RV+HP	144	41	77.8	All+RV+HP			
mutation data set-2				mutation data set 1 & 2	283	93	75.3

The positive correlation is for those mutations where increase in stability is corresponding to higher propensity of the mutant residue, or vice versa, in the corresponding structural environment. SS, secondary structure; SA, solvent accessibility; HB, hydrogen bonding; OI, Ooi number; All, all the four parameters; RV, residue volume; HP, hydrophobicity.

Table VI. Mutations suggested to engineer stability for lipase (1LBT) and proteinase K (2PRK). Residue propensities of suggested mutation for all the four structural parameters are also given along with the average (Ave) propensity value

Substitution	SS	Oi	HB	SA	Ave	Diff
1LBT-L001M	9.99	0.51	0.89	9.65	4.17	3.04
1LBT-S105G	8.79	1.2	0.74	1.16	2.42	1.6
1LBT-T159P	4.9	0.83	1.17	2.14	1.97	1.06
1LBT-S243P	5.59	1.16	0.89	1.08	1.76	0.84
1LBT-T186P	2.32	0.95	0.72	2.92	1.51	0.7
1LBT-D075P	5.8	0.7	0.72	0.54	1.5	0.67
1LBT-S161P	5.48	0.7	0.72	1.16	1.63	0.62
1LBT-Q175P	5.7	0.77	0.65	0.94	1.38	0.61
1LBT-D200P	4.1	0.68	0.82	0.54	1.19	0.58
1LBT-Q193N	2.92	0.82	1.81	0.25	1.29	0.57
1LBT-T040P	4.4	0.61	0.63	0.62	1.37	0.55
1LBT-N259H	0.89	1.07	1.03	4.54	1.57	0.5
2PRK-S262T	2.12	1.09	0.98	9.82	3.09	1.61
2PRK-Q089H	0.8	1.11	0.99	4.61	1.77	0.78
2PRK-T088P	4.95	0.67	1.2	0.66	1.63	0.74
2PRK-S190P	5.17	0.87	0.86	0.63	1.52	0.69
2PRK-D039G	8.35	1.44	1.42	0.88	1.49	0.65
2PRK-S176G	0.5	0.96	1.29	5.27	1.63	0.65
2PRK-D187N	4.09	0.86	0.98	0.51	1.54	0.6
2PRK-A001M	9.65	0.49	0.41	2.19	2.1	0.56
2PRK-E043H	1.47	1.23	1.44	2.31	1.48	0.51
2PRK-E050N	4.4	1.19	1.36	1.71	1.98	0.5

The suggested mutations are arranged in the decreasing order of the difference in propensity (Diff) between the suggested mutation and the wild-type residue at that position of the protein structure.

dependent parameters useful for suggesting substitution mutations to optimize the stabilizing interactions of a given protein structure.

Variation in amino acid propensities to different RSEPs

As described in the method, we discuss a few randomly selected examples of environment-dependent propensity values calculated for each of the 20 amino acids having different RSEP with respect to only one parameter. Table II gives such randomly selected residue structural environment parameters and Table IV gives their calculated propensity values. The *SA1* of 1BNIA-V10 (in Table IV) is a completely buried environment and the backbone conformation for V is favoured to be in helices with three main-chain—main-chain hydrogen bonding possibilities. As can be seen from the propensity values of other residues to this structural environment (see Table IV) only isoleucine (I) has a better propensity to this structural environment. Residues V, L, M, C and F seem to have equally good preference to this structural environment. *SA2* and *SA3* are considerably different structural environments for solvent accessibility parameter with some variations in other structural parameters (see Table II). The *SA2* environment appears to be preferred by V, I, F, P, L, M and C and very much avoided by residues D, E, H, K, N, Q, R and Y, whereas in the case of the *SA3* environment, which is more hydrophilic, we find that the residue preferences are reversed. There is a considerable variation in the propensity of amino acids for Oi1, Oi2 and Oi3 environments where the difference is primarily in the Ooi number. Similarly one can see variation in residue propensities where primary differences are in the secondary structural type (*SS1* to *SS5*) or in hydrogen bonding (HB1, HB2 and HB3). There could be compensatory variations among different structural parameters leading to similar residue propensity values, such as SA1, SS1 have similar propensity of 1.5 and HB1, SS3 have similar propensity of 1.0 for amino acid A.

It is clear from these examples that each amino acid has a definite propensity to any given local structural environment in proteins. It is also clear that amino acids show more preference to a structural environment optimally suitable to its physicochemical nature. In other words a residue may prefer many other environments with certain energetic constraints on the protein which is reflected on natural propensity of amino acid (*Pe*) calculated from the natural data set. This analysis, therefore, suggests a possible method to identify at least a few amino acids, in every wild type protein, not optimally suitable to be present in their existing locations. Thus a more preferred residue for that environment could be suggested to replace such non-optimally placed residues due to natural selection. In other words, the amino acid propensity values from the natural data set help us to suggest substitution point mutations to engineer protein stability.

Protein stability as a function of amino acid propensities in their RSEP

In order to test the proposed rationale, that amino acid propensity to a given RSEP is proportional to the stability

characteristics of the amino acid in that environment, the natural propensity (*Pe*) of all 20 amino acids was computed for each of the wild type amino acid RSEP of 'mutant data set-1'. The propensity value of the wild type residue was compared with the propensity value of each of the mutant residues at that RSEP. In most of the cases where a decrease in stability upon mutation was observed, there was also a decrease in *Pe* of the mutant residue compared with that of the wild type residue in that RSEP. Similarly, increase in stability was correlated with the increase in *Pe* of the mutant residue in the corresponding RSEP. However, such a positive correlation was observed for about 65% of cases when a single structural environment parameter (SEP) was used to calculate *Pe*. When different combinations of SEPs are used to calculate an average *Pe* the percentage of positive correlation is improved (Table V). The average *Pe* calculated by using all the SEPs gives about 71% positive correlation, and the highest positive correlation, 74%, was observed using weighted average propensity values against residue volume and hydrophobicity.

In order to test the correlation observed on mutant data set-1 we collected additional mutational data from the literature as mutation data set-2 [Table I(b)]. We found 185 additional substitution mutations in the literature that were not used in the sample mutant data set-1. When we tested correlation on mutation data set-2 we observed about 78% positive correlation of residue propensity values versus the increase/decrease in protein stability. The equally interesting positive correlation in mutation data set-2 gives clear indication that natural propensities of residues to a given structural environment determines the stability characteristics of a protein.

The mutant data set of proteins which gives information on change in stability due to substitution mutation has been collected from the literature. This information was generated through investigations of a different nature and through unrelated sets of experimental conditions on different kinds of proteins (Table I). But still the propensity values could help us to predict the effect of substitution mutation on protein stability with 75.3% positive correlation with the literature data.

The analysis, therefore, clearly indicates that we can precisely quantify the various stabilizing interactions in the form of structural environment profile of residue, as described and demonstrated, by using the few structural parameters discussed here. Though we have demonstrated a qualitative correlation, the study can be extended to quantitative estimations by including a few other structural environment parameters that contribute to stability interactions. Using such parameters in RSEP one could derive weighted average propensity values by optimizing on an existing mutation data set. We are now planning to use the available $\Delta\Delta G$ values to identify a linear correlation with propensity-dependent function, which should help us to develop a method to suggest possible substitution mutations to engineer stability in a given protein structure/sequence.

Suggesting substitution mutations to engineer protein stability

In order to suggest substitution mutations for a given protein structure, we first identify residues with low propensities in their local structural environment and see if there are any residues with higher propensity to that RSEP in the remaining 19 amino acids. We then suggest any residue having much higher propensity to that RSEP as a possible substitute for mutation to optimize the structure, which should increase the stability of the mutant protein. Such mutant proteins should be more stable than their parent proteins due to the placement of more optimally accepted residues in that physico-chemical structural environment. We have taken two industrially important proteins, lipase (1LBP) and proteinase K (2PRK), from the PDB data bank to identify less optimal residues and to suggest higher propensity residues as mutations for the respective RSEP (Table VI).

In the case of lipase, the residue propensities suggest that methionine is more preferred than leucine at position 1, glycine is preferred more than serine at position 105 and proline shows more propensity than the corresponding wild-type residues at positions 159, 243, 186, 75, 161, 175, 200 and 40. In the case of proteinase K, T has more propensity than S at 262 and 216 positions and also few other residues as listed in Table VI. We have suggested about 11 mutations in each of these proteins in decreasing order of the propensity difference (up to 0.5) between the suggested mutation and the wild-type residues (Table VI).

Summary

In summary, the analysis presented here reports the following: (i) each residue has a different propensity to a given local structural environment in the protein structures; this has also been reported earlier by others on many occasions. (ii) The representative data set of non-homologous, best resolved protein structures could be used as a natural data set to help calculate natural propensity of amino acids to a given structural environment. (iii) Similar structural environments in proteins are defined, not on the basis of standed cut-off class intervals, but based on the given parameter value with appropriate value-dependent window sizes, to define the most similar structural environment and also to get a statistically more significant number for analysis. (iv) A statistical occurrence of residues in different structural parameters is analysed, which helped us to set window sizes to classify similar structural environments. (v) The 'mutant data set-2' was used to verify correlation between the observed changes in stability characters and the local structural environment dependent amino acid propensities. (vi) The mutant data set-1 was helpful to optimize the average residue structural propensity calculations. The mutant data set-2 was useful to test the whole observations. (vii) The use of amino acid volumes and hydrophobicity values were useful for the standardization of propensity value calculations. (viii) We finally present predictions made to identify residues with less propensity in two proteins of industrial importance, and suggested possible substitution mutations to engineer higher stability characteristics. (ix) We also discuss use of a few more stability determining structural environment parameters of amino acids to suggest a better set of substitution mutations to engineer stability to a given protein structure.

Acknowledgements

We would like to thank Dr T.R.K.Murthy for critical review and for useful suggestions. The authors are grateful to the DST for grant (SP/SO/D-01/95/ dated: 21.6.1996). DIC-Bioinformatics (at CCMB) for the databases and Internet facilities. As a visiting scientist BVBR acknowledges the European Bioinformatic Institute (EBI), Cambridge, UK for use of their computational facilities. ST is supported by a fellowship from DST grant.

References

Argos,P., Rossmann,M., Grau,U., Zuber,H. and Frank,G. (1979) *Biochemistry*, **25,** 5698–5703.

Baker,E.N. and Hubbard,R.E. (1984) *Prog. Biophy. Mol. Biol.*, **44,** 97–179.

Baldwin,R.L. (1994) *Nature*, **369,** 183–184.

Braxton,S. (1996) In Cleland,J.L. and Craile,C.S. (eds), *Protein Engineering: Principles and Practice*. Wiley-Liss, New York, pp. 299–316.

Bordo,D. and Argos,P. (1991) *J. Mol. Biol.*, **217**, 721–729.

Bryan,P.N. (1995) In Shirley,B.A. (ed.) *Methods in Molecular Biology: Protein Stability and Folding—Theory and Practice*, Vol. 40. pp. 271–290.

Fersht,A.R. and Serrano,L. (1993) *Curr. Opin. Struct. Biol.*, **3,** 75–83.

Filippis,V.D., Sander,C. and Vriend,G. (1994) *Protein Engng*, **7,** 1203–1208.

Gilis,D. and Rooman,M. (1997) *J. Mol. Biol.*, **272,** 276–290.

Gupta,N.M. (1993) *Thermostability of Enzymes*. Springer, Berlin.

Guruprasad,K., Reddy,B.V.B. and Pandit,M.W. (1990) *Protein Engng*, **4,** 155–161.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1,** 409–417.

Imanaka,T., Shibazaki,M. and Takagi,M. (1986) *Nature*, **324**, 695–697.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Lee,B. and Richard,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.

Lesk,A.M. and Chothia,C. (1986) *Phil. Trans. Roy. Soc. London.*, **317,** 345–356.

Lopez-Hernandez,E. and Serrano,L. (1995) *Proteins*, **22,** 340–349.

Mathews,B.W. (1993) *Annu. Rev. Biochem.*, **62,** 139–160.

Menéndez-Arias,L. and Argos,P. (1989) *J. Mol. Biol.*, **206,** 397–406.

Mrabet,N.T., Van den Broeck,A., Van den brande,I., Stanssens,P., Laroche,Y., Lambeir,A.M., Matthijssens,G., Jenkins,J., Chiadmi,M. and van Tilbeurgh,H. (1992) *Biochemistry*, **31**, 2239–2253.

Munoz,V., Cronet,P., Lopez-Hernandez,E. and Serrano,L. (1996) *Folding Des.*, **1,** 167–178.

Murphy,K.P. (1995) In Shirley,B.A. (ed.), *Protein Stability and Folding—Theory and Practice*. Humana Press, New Jersey, pp. 1–34.

Nashikawa,K. and Ooi,T. (1980) *Int. J. Pept. Protein Res.*, **16**, 19–32.

Pace,C.N., Shirley,B.A., Mc Nutt,M. and Gajiwala,K. (1996) *FASEB J.*, **10**, 75–83.

Querol,E., Perez-Pons,J.A. and Mozo-Villarias,A. (1996) *Protein Engng*, **9**, 265–271.

Reddy,B.V.B. (1993) *Protein Engng*, **6** (Supplement), 23.

Reddy,B.V.B. (1996) *J. Biomol. Str. Dyn.*, **14**, 201–210.

Reddy,B.V.B. and Datta,S. (1997) *J. Biomol. Str. Dyn.*, **14**, 772.

Reddy,B.V.B., Ramesh,P. and Tiwari,S. (1998) *Bioinformatics*, **14**, 225–226.

Šali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.

Serrano,L., Day,A.G. and Fersht,A.R. (1993) *J. Mol. Biol.*, **233**, 305–312.

Shortle, D. (1992) *Qurt. Rev. Bioph.*, **25**, 205–250.

Shortle,D., Sondek,J. (1995) *Curr. Opin. Biotech.*, **6**, 387–393.

Smith,D. (1989) *SSTRUC: A program to calculate Secondary Structural Summary.* Department of Crystallography, Birkbeck College, University of London.

Topham,C.M., Srinivasan,N. and Blundell,T.L. (1997) *Protein Engng*, **10**, 7–21.

Varadarajan,R., Nagarajaram,H.A. and Ramakrishnan,C. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 13908–13913.

Vieille,C. and Zeikus,G. (1996) *Trends Biotech*, **14**, 183–190.

Villegas,V., Viguera,A.R., Aviles,F.X., Serrano,L. (1995) *Folding Des.*, **1,** 29–34.

Received April 9, 1998; revised August 24, 1998; accepted September 18, 1998