Protein Engineering vol.4 no.2 pp.155–161, 1990

Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence

Kunchur Guruprasad, B.V.Bhasker Reddy and Madhusudan W.Pandit[1]

Centre for Cellular and Molecular Biology, Hyderabad - 500 007, India

[1]To whom correspondence should be addressed

Statistical analysis of 12 unstable and 32 stable proteins revealed that there are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. Based on the impact of these dipeptides on the unstable proteins over the stable ones, a weight value of instability is assigned to each of the dipeptides. For a given protein the summation of these weight values normalized to the length of its sequence helps to distinguish between unstable and stable proteins. Results suggest that the *in vivo* instability of proteins is possibly determined by the order of certain amino acids in its sequence. An attempt is made to correlate metabolic stability of proteins with features of their primary sequence where weight values of instability for a protein of known sequence could thus be used as an index for predicting its stability characteristics.

Key words: instability index/primary structure/protein instability/stability prediction/turnover rate

Introduction

Proteins are known to degrade rapidly when their conformations are altered by mutations, incorporation of amino acid analogs, denaturation or premature chain terminations, etc. (Goldberg and St John, 1976). All such modifications are likely to prevent proper folding or to disrupt tertiary structure which can make the resulting aberrant polypeptide prone to degradation. However, normal cellular proteins also display a wide range of half-lives; turnover rates of individual proteins can differ as much as 1000-fold (Rechsteiner *et al.*, 1987). These observations have led to many speculations about the features of proteins that elicit proteolysis. Several hypotheses have been proposed to explain the intracellular stability (or instability) of proteins; these hypotheses have been reviewed recently (Rechsteiner *et al.*, 1987). Current status suggests that among several factors, sequence specific properties (Bachmair *et al.*, 1986; Rogers *et al.*, 1986), global features and the location of a protein in the cell are important in deciding the intracellular stability of a protein. In this communication, we report on an interesting observation that has emerged from the analysis of primary sequences of a set of proteins that are known to degrade rapidly (Rogers *et al.*, 1986), and the comparison of these data with those obtained from a set of stable proteins (i.e. those with comparatively high *in vivo* half-lives). The proteins which are known to degrade very rapidly are found to contain certain dipeptides in relatively large proportions; these might be directly or indirectly involved in the rapid degradation of proteins. Our analysis also points out that there is yet another set of dipeptides, the existence of which could be equally important for the stable proteins. The statistical analytical procedure leading to the dipeptide instability weight value reported here indicates the possibility of its use in predicting whether a particular protein would be unstable (or stable) *in vivo*.

Materials and methods

The strategy and the development of the instability index

As classified by Rogers *et al.* (1986), a set of 32 proteins with an *in vivo* half-life of > 16 h was taken as a class of stable proteins and a set of 12 proteins with an *in vivo* half-life of < 5 h was taken as an unstable class of proteins (Table I). The sequences of these two sets of proteins were analysed separately.

The frequency of occurrence of each of the 20 amino acids was calculated in the unstable as well as stable class of proteins and was compared with the frequency of occurrence of various amino acids in the total Protein Sequence Database of the PIR (Release 12.0). These data (Figure 1) explicitly bring out certain notable differences between the frequency of occurrence of various amino acids in unstable and stable proteins. It can be seen from Figure 1, that in the case of unstable proteins, amino acids Met(M), Gln(Q), Pro(P), Glu(E) and Ser(S) are found to occur with a relatively high frequency. The PEST hypothesis proposed earlier by Rogers *et al.* (1986) reports the presence of regions consisting of amino acids (Pro)P, (Glu)E, (Ser)S and (Thr)T in the unstable proteins. Our observation indicates that, unlike the PEST hypothesis, Thr(T) does not occur more frequently in unstable proteins compared with stable proteins. In fact, the presence of amino acids such as Met(M) is more frequent in unstable proteins. Similarly it was noted that amino acids Asn(N), Lys(K) and Gly(G) occur with relatively higher frequencies in stable proteins when compared with the unstable ones. Therefore, it appears that the PEST hypothesis could partly be a reflection of the consequence of differential occurrence of certain amino acids. However, the PEST hypothesis talks about regions of negatively charged amino acids that tend to be clustered and generally flanked by positively charged amino acids. Therefore, it appears that instability or stability characteristics are probably governed by an arrangement of certain amino acids in a specific order. Hence the smallest unit that defines order in a sequence being a dipeptide, the occurrence of an amino acid in juxtaposition to the other might be a significant factor in determining the stability of the protein. Therefore we chose to search for the occurrence of all 400 possible dipeptides in the two classes of proteins.

The expected (probable) occurrences of dipeptides were calculated, assuming the constituents of dipeptides as independent events, by equation:

$$N^e(xy) = [N^o(x)/T] * [N^o(y)/T] \sum_{x=1}^{20} \sum_{y=1}^{20} N^o(xy) \qquad (1)$$

where $N^e(xy)$ and $N^o(xy)$ are the expected and the observed occurrence respectively, of dipeptide xy; and $N^o(x)$ and $N^o(y)$ are the observed occurrences of amino acids x and y respectively. T is the total number of amino acids in a particular class.

Table I. Proteins studied, their half-life and II

Serial no.	Protein[a]	Half-life[b] (h)	PIR/EMBL* Code	II
(A) Stable proteins				
1	ADK	133	KIHUA	29.4
2	ADH	139	DEHUAB	17.4
3	AAT	66	XNPGDC	36.3
4	CAI	72	CRHU1	25.1
5	CPA	137	CPRTA	21.5
6	CAT	80	CSBO	27.5
7	CHY	50	KYBOA	15.3
8	CIS	128	YKPG	20.8
9	CCC	26	CCHU	11.4
10	AAD	118	OXPGDA	34.4
11	DPA	80	ADECD	14.2
12	DHF	97	RDBOD	27.4
13	ELA	46	ELPG	23.0
14	FER	61	FRHUL	23.8
15	GPD	84	DEHUGL	16.1
16	HEM	209	HAHU	7.0
17	HEM	209	HBHU	6.1
18	ABP	65	JGECA	23.1
19	LDH	171	DEPGLH	25.6
20	LYS	16	LZCH	19.9
21	MYO	127	MYHO	9.1
22	PAR	41	PVRYC	22.8
23	PGK	207	KIBYG	22.2
24	PLA	78	PSKF2U	15.1
25	PYK	181	KICHPM	21.2
26	RNA	61	NRGPA	52.2
27	SOD	41-200	DSBOCZ	7.0
28	STI	40	TISY	27.4
29	SUB	84	SUBSC	12.0
30	THI	155	TXEC	5.6
31	TPI	122	ISBYT	19.7
32	TRY	44	TRBOTR	20.6
(B) Unstable proteins				
33	E1A	0.5	AD5001*	100.3
34	α-Casein	2-5	KABOSB	57.7
35	β-Casein	2-5	KBBOA2	96.5
36	c-fos	0.5	HSCFOS*	78.8
37	c-myc	0.5	HSMYC1*	92.2
38	v-myb	0.5	GGCMYB1*	74.5
39	HSP 70	1-2	HHFF72	44.1
40	HMG-CoA	1.5-3	RDHYE	54.5
41	ODC	0.5	DCMSO	52.3
42	P730	1.0	ASAP3R*	50.4
43	TAT	0.5	RNTATR*	53.9
44	p53	0.5	MMP53*	70.0

[a] Abbreviations: ADK, adenylate kinase 1; ADH, alcohol dehydrogenase; AAT, aspartate amino transferase; CAI, carbonic anhydrase 1; CPA, carboxypeptidase A; CAT, catalase; CHY, chymotrypsinogen; CIS, citrate synthase; CCC, cytochrome C; AAD, D-amino acid oxidase; DPA, deoxyribose phosphate aldolase; DHF, dihydrofolate reductase; ELA, elastase; FER, ferritin; GPD, glyceraldehyde 3-phosphate dehydrogenase; HEM, haemoglobin; ABP, L-arabinose binding protein; LDH, lactate dehydrogenase; LYS, lysozyme *c*; MYO, myoglobin; PAR, parvalbumin; PGK, phosphoglycerate kinase; PLA, phospholipase A2; PYK, pyruvate kinase; RNA, ribonuclease A; SOD, superoxide dismutase; STI, soybean trypsin inhibitor; SUB, subtilisin; THI, thioredoxin; TPI, triosephosphate isomerase; TRY, trypsinogen; E1A, adenovirus early protein; HSP70, heatshock protein 70; HMG-CoA, hydroxymethyl glutaryl-CoA; ODC, ornithine decarboxylase; P730, phytochrome; TAT, tyrosine amino transferase.

[b] References for half-life are taken from Rogers *et al.* (1986).

The chi-square values, $\chi^2(xy)$ between observed and expected occurrences of dipeptides (xy) for each class of proteins were calculated by equation:

$$\chi^2(xy) = [N^o(xy) - N^e(xy)]^2/N^e(xy) \qquad (2)$$

An average value of χ^2 for each class of protein was calculated using equation:

$$\chi^2 = (1/400) \sum_{xy=1}^{400} \chi^2(xy) \qquad (3)$$

χ^2 was then used as the confidence limit to select significant dipeptides for each class of protein. The condition used for selecting significant dipeptides for unstable and stable classes of proteins respectively are:

$$\chi^2_{us}(xy) \geq \chi^2_{us}$$

and

$$\chi^2_{s}(xy) \geq \chi^2_{s}$$

$\chi^2_{us-s}(xy)$ was also calculated from values of observed occurrences of dipeptides in unstable and stable classes of proteins respectively, using the following equation:

$$\chi^2_{us-s}(xy) = [N^o_{us}(xy) - N^o_s(xy)]^2/N^o_s(xy) \qquad (4)$$

where $N^o_{us}(xy)$ and $N^o_s(xy)$, are the observed occurrences of dipeptides in unstable and stable classes of proteins respectively. As we were interested in picking up factors which contribute to the instability, in calculating $\chi^2_{us-s}(xy)$, each squared difference in the occurrence of dipeptides in unstable and stable classes was weighted inversely by the occurrence of dipeptides in the stable class. A third set of dipeptides was selected from these $\chi^2_{us-s}(xy)$ values satisfying the condition:

$$\chi^2_{us-s}(xy) \geq \chi^2_{us-s}$$

The potential occurrence, $P(xy)$ for each of the dipeptides, in all the three sets was calculated by equation:

$$P(xy) = N^o(xy) \;/\; N^e(xy) \qquad (5)$$

Each of the above three sets of dipeptides was further classified into two subsets depending upon $P(xy)$ being significantly different from unity (≥ 1.5 or ≤ 0.64). The corresponding potential values for dipeptides from all the three sets were used to formulate the conditions given in Table II. We classified the dipeptides that satisfy each of the conditions given in Table II based on the chi-square value and its $P(xy)$ of every dipeptide for each class separately. Weight value of instability, corresponding to each of the conditions, was obtained by summing up $N^o(xy)$ for all dipeptides (xy) satisfying that condition. The impact factor for ith condition (IF_i) was estimated by equation:

$$\mathrm{IF}_i = (\mathrm{Vus}_i - \mathrm{Vs}_i) \;/\; \mathrm{Vs}_i \qquad (6)$$

where Vus_i and Vs_i are the normalized values of occurrence of dipeptides satisfying ith condition in the unstable and stable class of proteins respectively. The impact factor for the ith condition

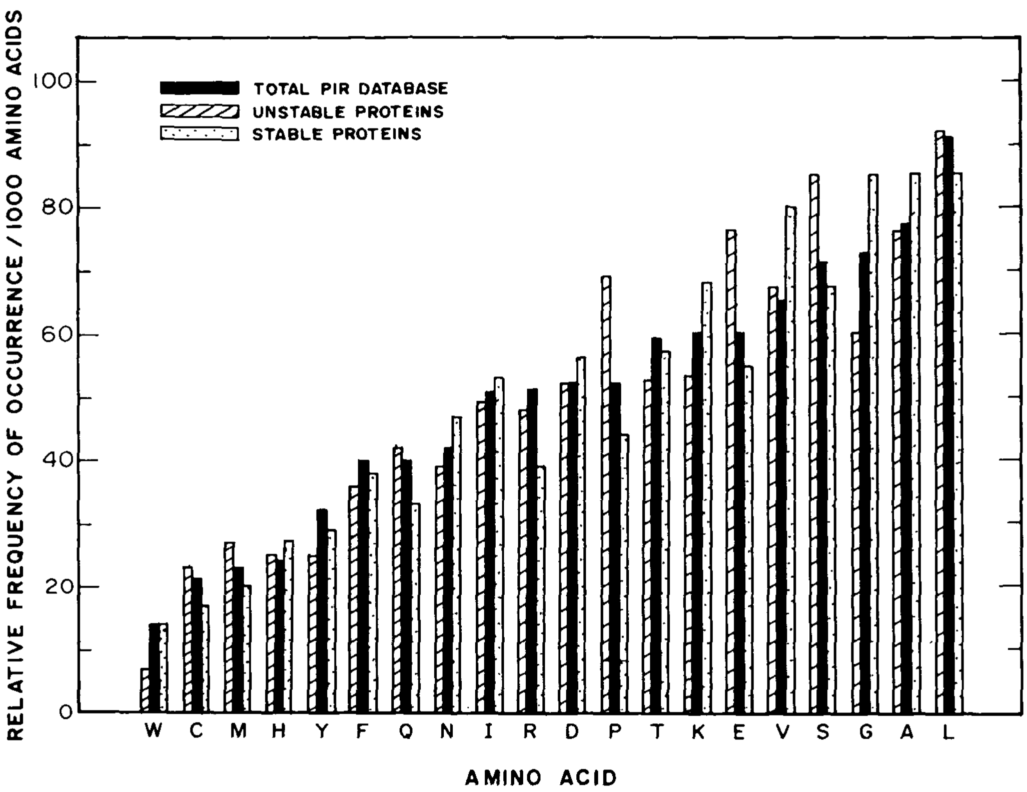

Fig. 1. Relative frequency of occurrence of individual amino acids per 1000 amino acids in various sets of proteins; PIR—Protein Sequence Database is taken from Release 12.

was operated on, to bring it into a positive range, and was termed as the instability weight value (IWV_i) given by equation:

$$IWV_i = 2 + (IF_i \;/\; |LIF|) \qquad (7)$$

where LIF was the lowest impact factor observed. The contribution of each of the dipeptides towards instability was obtained by summing the instability weight values corresponding to the condition(s) satisfied by the dipeptide and termed as the dipeptide instability weight value (DIWV). The DIWVs for all 400 combinations are presented as a matrix (GRP matrix) in Table III. The instability index (II) for a protein was then computed using the DIWV by equation:

$$II = (10/L) \sum_{i=1}^{L-1} DIWV(x_iy_{i+1}) \qquad (8)$$

where x_iy_{i+1} is a dipeptide, L is the length of the sequence and 10 is a scaling factor. Instability indices for various proteins in the stable as well as in the unstable class are given in column 5 of Table I.

Results and discussion

Dipeptides involved in instability of proteins

The observed and the expected frequency of occurrence of dipeptides in the sets of both stable and unstable proteins helped to identify the dipeptides which are predominant in either of these sets. GRP-matrix (Table III) indicates that out of the 400 possible dipeptides 81 (i.e. 20%) with DIWV > 1 may be involved in the instability of the protein, whereas ~72 (i.e. 18%) with DIWV < 1 may contribute to the stability. We do not know at present how the presence (or the absence) of dipeptides such as these render a protein stable or susceptible to degradation. However, the net effect of their combined presence in a protein appears to reflect in its instability index that serves as a useful indicator in deciding the stability (or instability) characteristics of the protein.

Table II. Rule-based weightages

Conditions		Instability weight value
If the dipeptide satisfies:		
1.	$P_{us} \geq 1.50$ and $P_s < 1.50$	+13.34
2.	$P_{us} < 1.50$ and $P_s \geq 1.50$	− 1.88
3.	$P_{us} \leq 0.64$ and $P_s > 0.64$	− 6.54
4.	$P_{us} > 0.64$ and $P_s \leq 0.64$	+24.68
5.	$P_{us\text{-}s} \geq 1.50$	+20.26
6.	$P_{us\text{-}s} \leq 0.64$	− 7.49
7.	None of the above conditions	+ 1.0

In vivo half-life of a protein versus its instability index

We have presented various proteins that are used in our analysis along with the values of their half-lives in Table I. It can be seen from the results that all the unstable proteins have an II > 40, whereas all the stable proteins, with the only exception of RNase A, have an II < 40. The segregation of the two classes of proteins based on the II is illustrated in Figure 2. This figure clearly brings out differences between the two classes of proteins. The relatively high value of the II (= 52.2) for RNase A puts this protein into the class of unstable proteins; however, it is known that RNase A in its native form is held together by four disulphide bridges between Cys groups. Such disulphide bridges are known to impart significant stability to the protein making it resistant to degradation

Table III. GRP matrix of condition-based instability values for 400 possible dipeptides

First amino acid of dipeptide	Second amino acid of dipeptide																			
	W	C	M	H	Y	F	Q	N	I	R	D	P	T	K	E	V	S	G	A	L
W	1.0	1.0	24.68	24.68	1.0	1.0	1.0	13.34	1.0	1.0	1.0	1.0	−14.03	1.0	1.0	−7.49	1.0	−9.37	−14.03	13.34
C	24.68	1.0	33.6	33.6	1.0	1.0	−6.54	1.0	1.0	1.0	20.26	20.26	33.6	1.0	1.0	−6.54	1.0	1.0	1.0	20.26
M	1.0	1.0	−1.88	58.28	24.68	1.0	−6.54	1.0	1.0	−6.54	1.0	44.94	−1.88	1.0	1.0	1.0	44.94	1.0	13.34	1.0
H	−1.88	1.0	1.0	1.0	44.94	−9.37	1.0	24.68	44.94	1.0	1.0	−1.88	−6.54	24.68	1.0	1.0	1.0	−9.37	1.0	1.0
Y	−9.37	1.0	44.94	13.34	13.34	1.0	1.0	1.0	1.0	−15.91	24.68	13.34	−7.49	1.0	−6.54	1.0	1.0	−7.49	24.68	1.0
F	1.0	1.0	1.0	1.0	33.6	1.0	1.0	1.0	1.0	1.0	13.34	20.26	1.0	−14.03	1.0	1.0	1.0	1.0	1.0	1.0
Q	1.0	−6.54	1.0	1.0	−6.54	−6.54	20.26	1.0	1.0	1.0	20.26	20.26	1.0	1.0	20.26	−6.54	44.94	1.0	1.0	1.0
N	−9.37	−1.88	1.0	1.0	1.0	−14.03	−6.54	1.0	44.94	1.0	1.0	−1.88	−7.49	24.68	1.0	1.0	1.0	−14.03	1.0	1.0
I	1.0	1.0	1.0	13.34	1.0	1.0	1.0	1.0	1.0	1.0	1.0	−1.88	1.0	−7.49	44.94	−7.49	1.0	1.0	1.0	20.26
R	58.28	1.0	1.0	20.26	−6.54	1.0	20.26	13.34	1.0	58.28	1.0	20.26	1.0	1.0	1.0	1.0	44.94	−7.49	1.0	1.0
D	1.0	1.0	1.0	1.0	1.0	−6.54	1.0	1.0	1.0	−6.54	1.0	1.0	−14.03	−7.49	1.0	1.0	20.26	1.0	1.0	1.0
P	−1.88	−6.54	−6.54	1.0	1.0	20.26	20.26	1.0	1.0	−6.54	−6.54	20.26	1.0	1.0	18.38	20.26	20.26	1.0	20.26	1.0
T	−14.03	1.0	1.0	1.0	1.0	13.34	−6.54	−14.03	1.0	1.0	1.0	1.0	1.0	1.0	20.26	1.0	1.0	−7.49	1.0	1.0
K	1.0	1.0	33.6	1.0	1.0	1.0	24.68	1.0	−7.49	33.6	1.0	−6.54	1.0	1.0	1.0	−7.49	1.0	−7.49	1.0	−7.49
E	−14.03	44.94	1.0	−6.54	1.0	1.0	20.26	1.0	20.26	1.0	20.26	20.26	1.0	1.0	33.6	1.0	20.26	1.0	1.0	1.0
V	1.0	1.0	1.0	1.0	−6.54	1.0	1.0	1.0	1.0	1.0	−14.03	20.26	−7.49	−1.88	1.0	1.0	1.0	−7.49	1.0	1.0
S	1.0	33.6	1.0	1.0	1.0	1.0	20.26	1.0	1.0	20.26	1.0	44.94	1.0	1.0	20.26	1.0	20.26	1.0	1.0	1.0
G	13.34	1.0	1.0	1.0	−7.49	1.0	1.0	−7.49	−7.49	1.0	1.0	1.0	−7.49	−7.49	−6.54	1.0	1.0	13.34	−7.49	1.0
A	1.0	44.94	1.0	−7.49	1.0	1.0	1.0	1.0	1.0	1.0	−7.49	20.26	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
L	24.68	1.0	1.0	1.0	1.0	1.0	33.6	1.0	1.0	20.26	1.0	20.26	1.0	−7.49	1.0	1.0	1.0	1.0	1.0	1.0

(Creighton, 1988). Therefore it appears that cross-linking between Cys groups may override the impact of dipeptides. In our approach we have not taken into account such higher-order elements, which could modify the intrinsic stability characteristics of the protein. It is also interesting to note that none of the unstable proteins which we have studied has such a high degree of cross-linking (disulphide bridges per 100 amino acids) between Cys groups. Thus there is a direct correlation between the half-lives of various proteins and the II arrived at by the method described. The II of a protein, therefore, could be used directly to predict whether a given protein is stable or unstable based on the II being either <40 or >40 respectively, provided the sequence of the protein is known.

Validity of the II of a protein in determining its stability

In order to test the merits of our method in predicting the stability characteristics of proteins, we used another set of proteins which was not a part of our database used in developing the II but for which the data on *in vivo* half-life became available. These proteins are listed in Table IV along with their IIs. It can be seen from the last two columns of Table IV that, in every case, the prediction based on the II of the protein is in complete agreement with the conclusion based on the *in vivo* half-life of the protein. This shows clearly that it is possible to predict the stability characteristics of a protein with reasonable success by using its II.

Comparison with earlier hypotheses

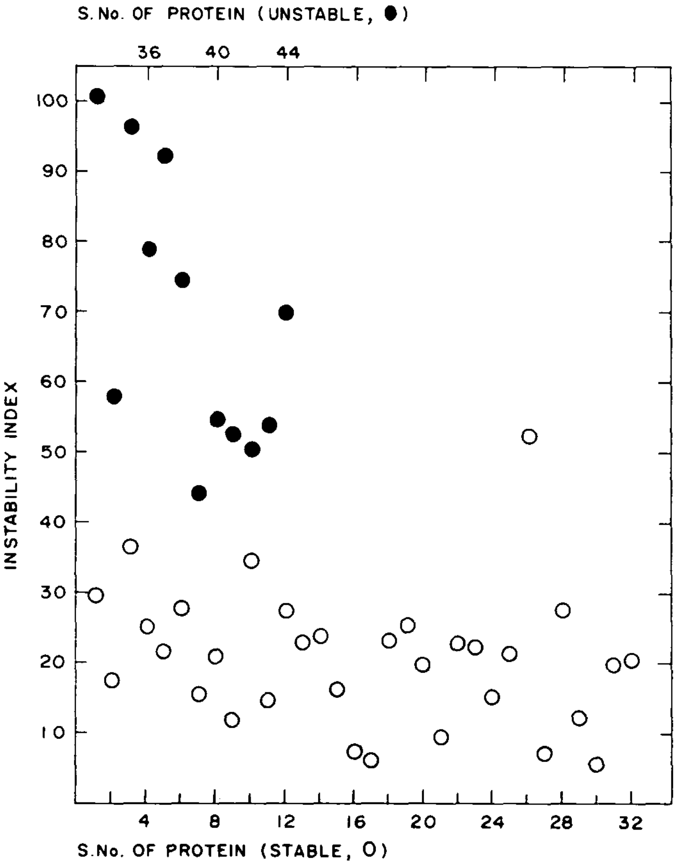

Fig. 2. IIs of various proteins used in the analysis. The proteins could be identified by their serial (S) nos. given in Table I. ○, Protein with S. nos. 1–32; ●, proteins with S. nos. 33–44.

The results of our analysis indicate that, although there is a significant number of proteins which shows overlap between the predictions made by the II described here and those governed by PEST hypothesis, there exists a considerable number of proteins in which there is no correlation between the PEST hypothesis and the experimental observations. For example, protein P730 (phytochrome) which contains PEST regions with a negative PEST-FIND (PF) score between −2.21 and −30.47 is expected to be stable according to the PEST hypothesis. However, this protein has been shown to be unstable (Pratt *et al.*, 1974). The II calculated by our method for this protein is 50.4, and categorizes it as an unstable protein in agreement with experimental observations. There are a few other proteins, e.g. AAT (aspartate amino transferase), DHF (dihydrofolate reductase) and PYK (pyruvate kinase), which are reported to contain at least one PEST region with a positive PF score and

Table IV. Prediction of the stability of a few proteins whose half-life is known

Protein	PIR Code	Half-life (h)	II	Whether stable (S) or unstable (US) from	
				II	Half-life
Metallothionein (gold)	SMRT2	<2[a]	66.82	US	US
Protein kinase C	KIRTC2	2-8[a]	40.29	US	US
Phosphoenol pyruvate carboxylase	QYEC	5[b]	46.8[c]	US	US
Glucose-6-phosphate dehydrogenase	DESHGC	15[b]	27.18[c]	S	S
Ornithine amino transferase	XNRTO	19[b]	34.21	S	S
Frustose-1,6-diphosphatase	PAPGF	36[b]	29.16[c]	S	S
Cytochrome P450	O4RTP2	48[b]	35.89	S	S
Cytochrome *b*5	CBRT5	55[b]	32.86	S	S
Fatty acid synthetase	FZRTI	71[b]	12.37	S	S
Malate dehydrogenase	DEPGMM	96[b]	27.8[c]	S	S
Arginase	WZBYR	96[b]	28.78[c]	S	S
Cytochrome *b*	CBRT	130[b]	32.86	S	S
Ubiquitin	UQBY	9-40[a]	30.33[c]	S	S
Catalase	CSRT	9-40[a]	33.82	S	S
Na^+/K^+-ATPase	PWSHNA	9-40[a]	31.67[c]	S	S
Transferrin receptor	JXHU	9-40[a]	22.06[c]	S	S
Cytochrome *c* oxidase	OBMS2	41-200[a]	36.25	S	S
β-Galactosidase	GBECE	41-200[a]	37.62[c]	S	S
Actin	ATHU	41-200[a]	36.19	S	S
c-AMP protein kinase	OKBO1R	2-8[a]	49.3[c]	US	US
c-GMP protein kinase	OKBOG	2-8[a]	41.4[c]	US	US
Calmodulin	MCHU	9-40[a]	27.53	S	S

[a] Rechsteiner *et al.* (1987).
[b] Dice and Goldberg (1975).
[c] The II was calculated from the available sequence as the source of the original protein was not mentioned in the cited reference.

therefore are expected to be unstable according to the proponents of the PEST hypothesis. Experimental results, however, are not in agreement with these expectations as these proteins have been reported to be stable *in vivo* (Rogers *et al.*, 1986). It is interesting to note that IIs for these proteins predict them to be stable (Table I) in keeping with the experimental results. These examples indicate that instability of a protein is governed by factors other than those suggested by the PEST hypothesis.

When we examined PF scores for the PEST regions for the set of proteins given in Table IV, it was observed that predictions based on the PEST hypothesis and II did not tally in two cases. Sodium potassium ATPase (PWSHNA) has three PEST regions with positive PF scores of 6.93, 5.78 and 5.05 and therefore should be unstable. However, the prediction on the II suggests that the protein is stable and is in accordance with the experimental observation (Rechsteiner *et al.*, 1987). Metallothionein (SMRT2) completely lacks PEST regions with a positive PF score. But this protein has an II, which predicts that the protein is unstable—the conclusion is again in agreement with experimental observation (Rechsteiner *et al.*, 1987). The II appears therefore to pick some features in the sequence which are overlooked by the PEST hypothesis. It may be pointed out that out of the 81 dipeptides significant for the unstable class (DIWV > 1 in Table III) only 39 are originating from dipeptides containing either P, E, S or T. More than half of the dipeptides are exclusive of any of these amino acids. These facts establish the significance of dipeptides devoid of P, E, S or T. In addition, it is interesting to note that there are 25 dipeptides containing either P, E, S or T which have DIWV < 0 indicating that in fact, they contribute significantly to the stability of the protein, in contrast to the proponents of the PEST hypothesis. It is possible to carry out rigorous experimental checks by selective clipping of residues, which contribute significantly to the high PF score, and following the fate of the residual protein in terms of stability. Rogers *et al.* (1986) have reported that β-casein at its N-terminal end (residues 1−25) has only one PEST region with a positive PF score. The clipping of this region should lead to a residual protein which should be stable; however, the II suggests that the residual protein will remain unstable. Similarly these workers have reported another protein, v-myb, where residues 1−16 at the N-terminal end contain a PEST region with a positive PF score (1.85). As proposed by them, removal of this region should make the protein stable. The II for the residual v-myb, however, suggests that it would be unstable. Experimental checks such as these will help in arriving at discrete features which play a crucial role in conferring stability on the protein.

Another hypothesis proposed by Bachmair *et al.* (1986), based on the studies on β-galactosidase and termed as the N-end rule, suggests that a specific amino acid at the N-terminus can serve as one of the determinants for *in vivo* half-life of a protein. As the N-end rule is derived from studies on a single protein, one cannot be certain about its applicability to all the proteins in general. In addition, it has already been suggested that the N-end rule may be primarily cotranslational (Rechsteiner *et al.*, 1987), and therefore has a limited use in determining the stability characteristics of a protein in general. We have, therefore, not exhaustively compared our data with that given by the N-end rule. However, preliminary examination of such data indicates that predictions based on the N-end rule are not in anyway better than those based on the PEST hypothesis.

Analysis was carried out for all the proteins whose sequences are known from the SWISS-PROT Protein Sequence Database (Release 13) in order to pick up proteins that have very low II values (<10). Similarly, proteins that have very high II values (>90) were also picked up. For the reasons discussed earlier, proteins that have a cysteine content higher than RNase A were eliminated. Out of these, unique proteins showing extremely high as well as low II values are listed in Table V. These proteins would prove to be good candidates to provide the experimental validity to the predictions in regard to their instability based on II.

The results allow us to make certain plausible conjectures about point mutations altering the stability of a protein. The GRP-matrix consists of certain dipeptides with high positive as well as negative values. Point mutations in an unstable protein may lead to a change from one dipeptide with a high positive value to another with a negative value, consequently making a protein stable.

A single residue change in a protein can affect the contribution of instability weight values of two dipeptides, which arise due to residues on either side of it. Therefore all possible tripeptides were first generated where the dipeptides comprising them had DIWV values either >1 or <0 (refer to Table III). The difference in the II value (ΔII_{tri}) for a pair of tripeptides '*axb*' and '*ayb*' obtained by changing the central residue and keeping the neighbouring residues the same was estimated by the relation:

$$\Delta II_{tri} = [DIWV(ax) + DIWV(xb)] - [DIWV(ay) + DIWV(yb)] \quad (9)$$

We have analysed the data on substitutions in the central

Table V. Unique proteins with extreme values of IIs

Protein name	SWISS-PROT Code
(A) Proteins with II ≥ 90 and length ≥ 50	
Amelogenin	AMEG$BOVIN
Calspermin	CALS$RAT
Beta casein precursor	CASB$BOVIN
Core antigen	CORA$HPBV4
Contiguous repeat polypeptide (CRP) precursor	CRPP$RAT
Early E1A 11 kd protein	E111$ADEM1
Early E1A 32 kd and 26 kd proteins	E1A$ADEN2
Early E1A 6 kd protein	E1A6$ADEN5
EBNA-2 Nuclear protein	EBN2$EBV
Extensin precursor	EXTN$DAUCA
Alpha/beta-gliadin precursor (Prolamin)	GDA0$WHEAT
Gamma/gliadin precursor	GDB2$WHEAT
low mol. wt glutenin precursor	GLTA$WHEAT
B1-Hordein	HOR1$HORVU
Sperm histone (Protamine)	HSP$COTJA
Involucrin	INVO$LEMCA
Myc proto-concogene protein	MYC$FELCA
Sperm-specific protein PHI-0	PHI0$HOLTU
Salivary proline-rich protein precursor	PRP1$HUMAN
Rev protein anti-repression transactivator	REV$HIV13
Ribosomal protein S18	RS18$MARPO
40S Ribosomal protein S26	RS264DROME
Spermatid-specific protein	SSS1$SCYCA
Transition protein 2	STP2$MOUSE
Female-specific transformer protein	TRSF4DROME
Tequment phosphoprotein	USO9$HSV11
11 kd protein	V11K$PRV
Probable E3 protein	VE3$BPV2
(B) Proteins with II ≤ 10 and length ≥50	
Acylphosphatase, erythrocyte	ACYE$HUMAN
Antifreeze protein IIA7 precursor	ANPX$SEAM
ATP synthase lipid-binding protein	ATPH$SOYBN
Azurin iso-1	AZU1$METJ
Cytochrome *C*556	C556 $AGRTA
Cytochrome *C*	CYC$ANAPL
DNA-binding protein (7 kd)	DN7A$SULAC
Fatty-acid-binding protein	FABH$RAT
Type-1 fimbrial protein	FM1A$ECOLI
Nitrogen regulatory protein PII	GLNB$ECOLI
Haemoglobin alpha-A chain	HBA$AEGHO
Hisactophilin	HIAP$DICDI
Small histidine-alanine-rich protein precursor	HRP$PLAFF
Histidine-rich-protein precursor	HRP1$PLAFA
Acrosin inhibitor II	IAC2$BOVIN
Insulin	INS$BALBO
Light-harvesting protein, alpha chain (LH-2)	LHA2$RHOCA
Mammary-derived growth inhibitor	MDGI$BOVIN
Retinoic-acid-induced differentiation factor	MK1$MOUSE
Major outer membrane lipoprotein precursor	MULI$ERWAM
Myoglobin	MYG$HORSE
Outer membrane protein precursor (porin)	OMPC$SALTY
Outer mitochondrial membrane protein (porin)	PORI$NUECR
Profilin	PROF$ACACA
Retinol-binding protein II, cellular (CRBP-II)	RET2$RAT
50S Ribosomal protein L1	RL1$BACST
30S Ribosomal protein S14	RS14$MYCCA
S-100 protein, alpha chain	S10A$BOVIN
S-Antigen protein precursor	SANT$PLAFW
Probable early protein GP5	VG5$BPPH2
Gene 5 protein	VG5$SPV4
Early protein GP5B	VG5B$BPPZA
Gene I5 Protein	VI5$VACCV
Lysis protein	VLYS$BPT7

Table VI. Substitutions of central amino acid in a tripeptide (*axb*) that are likely to alter ΔII by ≥75 as a result of the replacement of *x* by *y* (values of ΔII are given in brackets)

a	$x \rightarrow y$	*b*	*a*	$x \rightarrow y$	*b*
Trp	Met→Ala(105)	His			
	His→Val(84)	Tyr	Asn	Ile→Gln(75)	Glu
	His→Gly(87)	Tyr		Ile→Thr(77)	Glu
	His→Thr(77)	Asn		Ile→Gly(111)	Glu
	His→Gly(87)	Ile	Ile	Glu→Pro(98)	Cys
	Asn→Gly(75)	Ile		Glu→Lys(80)	Ile
Cys	His→Gln(92)	Tyr		Glu→Val(87)	Asp
	His→Val(92)	Tyr		Glu→Lys(78)	Pro
Met	His→Met(80)	Tyr	Arg	Trp→Try(75)	His
	His→Gln(116)	Tyr		Trp→Gly(87)	Asn
	His→Arg(116)	Tyr		His→Gly(80)	Tyr
	His→Arg(75)	Asn		His→Gly(80)	Ile
	His→Thr(99)	Asn		Arg→His(99)	Trp
	Pro→Gln(78)	Phe		Arg→Tyr(133)	Trp
	Pro→Gln(78)	Val		Arg→Asn(113)	Trp
	Ser→Gln(92)	Cys		Arg→Pro(99)	Trp
	Ser→Gln(75)	Pro		Arg→Gly(111)	Trp
	Ser→Arg(75)	Pro		Arg→Gly(87)	Asn
His	Tyr→Pro(98)	Met		Arg→Tyr(139)	Arg
	Tyr→Gly(75)	Try		Arg→Pro(103)	Arg
	Tyr→Pro(78)	Asp		Ser→Tyr(87)	Arg
	Tyr→Trp(86)	Ala		Ser→Tyr(83)	Pro
	Tyr→Gly(87)	Ala		Ser→Asn(78)	Pro
	Asn→Gly(87)	Ile		Ser→Tyr(78)	Glu
	Ile→Thr(75)	Glu		Ser→Gly(78)	Glu
	Ile→Gly(106)	Glu	Asp	Arg→Thr(80)	Trp
Tyr	Met→Trp(87)	His		Ser→Lys(78)	Pro
	Met→Tyr(75)	His	Thr	Glu→Gln(78)	Cys
	Met→Arg(99)	His		Glu→Asn(81)	Cys
	Met→Glu(116)	His	Lys	Met→Ile(86)	His
	Met→Ala(86)	His		Met→Ile(87)	Pro
	Met→Arg(92)	Tyr		Arg→Pro(99)	Trp
	Met→Gly(84)	Tyr		Arg→Gly(86)	Trp
	Met→His(78)	Pro		Arg→Leu(75)	Trp
	Met→Arg(86)	Pro		Arg→Pro(105)	Arg
	Met→Glu(75)	Pro		Arg→Leu(78)	Arg
	Met→Glu(75)	Ser	Glu	Cys→His(78)	Trp
	Met→Trp(81)	Ser		Cys→Trp(107)	Thr
	His→Arg(80)	Tyr		Cys→His(92)	Thr
Gln	Ser→Cys(78)	Gln	Ala	Cys→His(78)	Trp
	Ser→Tyr(87)	Arg		Cys→His(93)	Thr
	Ser→Cys(75)	Pro		Cys→Asp(101)	Thr
	Ser→Tyr(83)	Pro			
	Ser→Phe(75)	Pro			
	Ser→Val(75)	Pro			
	Ser→Tyr(78)	Glu			

position of a tripeptide along with their associated ΔII_{tri} values, and listed only those amino acid replacements (Table VI) where the ΔII_{tri} value changes significantly (>75). The formula that can be used for estimating the change in the II of a protein would be:

$$\Delta II_{protein} = (10/L) * \Delta II_{tri} \qquad (10)$$

where $\Delta II_{protein}$ is the difference in II of a protein of length L before and after the replacement of residue x by residue y.

Therefore depending upon the length of the protein and $\Delta II_{protein}$ value, one can choose effective substitutions from ΔII_{tri} values given in Table VI, so that the II of the protein assumes any value ≤ 40 and hence becomes stable.

As $\Delta II_{protein}$ is inversely proportional to the length of the protein, a single substitution may not be sufficient to change II to the expected level; hence, more than one substitutions may be necessary. It should also be kept in mind that some of the substitutions which suggest improvement in the stability may not be desirable in terms of the functional requirements of the protein. Only those substitutions which would enhance the stability of a protein without affecting its function would be useful.

Conclusions

Our analysis and the method developed for predicting the *in vivo* stability of a protein suggest that the primary determinants of the stability of the protein probably reside in its primary structure and is an intrinsic property of a protein. There appears to be a correlation between the sensitivity of a protein to *in vivo* degradation and the presence of certain dipeptides in it. The overall influence of such dipeptides appears to contribute to instability or stability characteristics of proteins. Apart from these sequence-dependent characteristics several factors, e.g. structure-dependent features, the presence of disulphide bridges, ligand binding, protease recognition mechanisms, etc., are known to determine the *in vivo* protein stability. Therefore stability of a protein as manifested *in vivo* could be a net effect of the contributions made by several such factors. Our work indicates that sequence-specific elements is one of the significant factors which play an important role in determining the stability of a protein. The observations presented here might help open new vistas, where the knowledge about the dipeptides having specific characteristics could be used in the modification of existing proteins or in the design of novel protein molecules of a desired stability.

References

Bachmair,A., Finley,D. and Varshavsky,A. (1986) *Science*, **234**, 179–186.

Creighton,T.E. (1988) *BioEssays*, **8**, 57–63.

Dice,J.F. and Goldberg,A.L. (1975) *Arch. Biochem. Biophys.*, **170**, 213–219.

Goldberg,A.L. and St John,A.C. (1976) *Annu. Rev. Biochem.*, **45**, 747–803.

Pratt,L.H., Kidd,G. and Coleman,R. (1974) *Biochim. Biophys. Acta*, **365**, 93–107.

Rechsteiner,M., Rogers,S. and Rote,K. (1987) *Trends Biochem. Sci.*, **12**, 390–394.

Rogers,S., Wells,R. and Rechsteiner,M. (1986) *Science*, **234**, 364–368.

Received on May 10, 1990; accepted on August 30, 1990