Protein evolution: intrinsic preferences in peptide bond formation: a computational and experimental analysis[†]

SUBRAMANIA RANGANATHAN*[,a,b], DINABANDHU KUNDU[c] and S D VUDAYAGIRI[a],

[a]*Discovery Laboratory, Indian Institute of Chemical Technology, Hyderabad 500 007, India*
[b]*Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560 064, India*
[c]*Department of Chemistry, Scottish Church College, Kolkota 700 006, India*

**Corresponding author (Fax, 91-40-27160-757; Email, rangan@iict.ap.nic.in)*

Two possibilities exist for the evolution of individual enzymes/proteins from a milieu of amino acids, one based on preference and selectivity and the other on the basis of random events. Logic is overwhelmingly in favour of the former. By protein data base analysis and experiments, we have provided data to show the manifestation of two types of preferences, namely, the choice of the neighbour and its acceptance from the amino end (left) or the carboxyl end (right). The study tends to show that if the 20 proteinous amino acids were made to combine in water, the resulting profile would be nonrandom. Such selectivity could be a factor in protein evolution.

[Ranganathan S, Kundu D and Vudayagiri S D 2003 Protein evolution: intrinsic preferences in peptide bond formation: a computational and experimental analysis; *J. Biosci*. **28** 683–690]

1. Introduction

Decades ago, Bernal conjectured that were the 20 proteinous amino acids be allowed to condense in water, under ambient conditions and pH, the resulting peptide/protein sequence will be nonrandom, thus setting the stage for the presence of sequence – selective peptides/ proteins, prior to the evolution of the genetic code and the ensuing instructed protein synthesis (Bernal 1967). Over the years, this notion has received indirect support. The sequences arising from thermal polymerization of coded amino acid mixtures was found nonrandom (Harada and Fox 1965). The extensive protein sequences data currently available strongly suggests such preferences. The delineation of such preference profiles in proteins has been used to analyse binding sites in proteins (Villar and Kauvar 1994), short sequence regularities (Rani and Mitra 1994, 1996; Vonderviszt *et al* 1986), possible composition of early proteins (Kolaskar and Ramabrahmam 1982), mapping of evolutionary trees (Doolittle 1989) and classification of organisms (Erhan 1978). X-ray studies of mixed crystals of amino acids have shown intrinsic alignment for selective peptide bond formation (Vijayan 1988). Amino acid neighbour preference profile is the focus of experiments to demonstrate, chiral evolution (Kricheldori *et al* 1985), ambient nonrandom polymerization of peptides (Tyagi and Ponnamperuma 1990; Dose *et al* 1982) and theories relating to the evolution of self-organizing systems (Dose *et al* 1982; Orgel 1992; Ranganathan and Ranganathan 1981; Walder *et al* 1979) and enzyme-directed peptide synthesis (Ranganathan *et al* 1999).

Developments in biochemistry, X-ray crystallography, organic chemistry and analytical methodology, in the ensuing period have enabled us, using computation and experiments, to address this question (Bernal 1967). We have looked for imprints of preference profile in peptide bond formation, amongst present day proteins, in terms of their primary and secondary structures and demonstrate experimentally whether these are intrinsic to the amino acid and not related to it being part of a protein.

Keywords. Analysis; experiments; peptide neighbour preferences

[†]Dedicated to the memory of Darshan Ranganathan.

2. Materials and methods

2.1 *Computer analysis*

The data base for studies reported here was constructed by tandem computer analysis of sixty proteins (present in nature, having ten thousand residues) in the domains of enzymes, inhibitors, globins, cytochromes, ferredoxins, metalloproteins, immunoglobulins, hormones, toxins, ribosomal proteins, muscle proteins, structural proteins and miscellaneous proteins. In all these cases, the 3D structures were established by high resolution X-ray crystallography. [PDB-codes: 1acx, 1aec, 2pab, 2aza, 5cpv, 3cpa, 1crn, 1cse, 1csr, 2cna, 1a5d, 1hrc, 2frc, 1cpt, 351c, 1cyo, 1eb7, 8dfr, 2era, 1a6l, 1ahn, 3grs, 2hmq, 2mhr, 1eca, 1a3n, 3hhb, 1ctf, 1fb4, 1dc1, 1f1g, 1bww, 9ins, 1neh, 2pka, 6ldh, 7lyz, 1681, 1lpr, 2mlt1a6m, 1ovo, 1ppt, 1bxoh, 1plc, 3rpz, 1ruv, 7rsah, 7rxn, 1snb, 1d6t, 1ton, 5tnc, 1tro, 2ptc, 5pti, 1ubi, 1wgt.]

Gross neighbour preference involved in peptidation for each of the 20 proteinous amino acids in the basic set was generated; and at the same time, a parallel set was constructed based on the assumption that there is a non-preference in the choice of neighbours. Analysis of the basic set was also carried out to determine a second element of preference: namely: the placement of the neighbour either to the left or to the right of the central residue. The protocols used in the analysis have been reported by us (Ranganathan *et al* 1999).

2.2 *Experimental*

2.2a *General*: All amino acids used were of L-configuration. ^{1}H NMR spectra were obtained on WP 80 Bruker instrument at 80 MHz in $CDCl_3$. The chemical shifts were recorded in ppm with TMS at 0·00 as the internal standard or as the external reference. IR spectra were recorded on PE 1600 FT instrument either as neat liquids or as KBr pellets. FAB mass was recorded using a Jeol SX-120-/DA-600 instrument using argon (6 kV, 10 mA) as the FAB gas. The accelerating voltage was 10 kV and the spectra were recorded at room temperature with m-nitrobenzyl alcohol as the matrix. Elemental analysis was carried out in automatic C, H, N analyser. Silica gel G (Merck) was used for thin-layer chromatography (TLC) and column chromatography was carried out on silica gel (acme 100–200 mesh) column, which were generally made from slurry in benzene or hexane or ethylacetate. Reactions were monitored whenever possible by TLC. The organic extracts were invariably dried over anhydrous. $MgSO_4$ and solvents were evaporated *in vacuo*. HPLC was performed on a reverse phase column [Shim-pack CLC-ODS(M)], using UV-VIS spectrophotometric detector (Shimadzu SPD-6AV) at 210 nm.

The N-Tosyl-Pro, Gly, Leu, Phe, Trp and their methyl ester hydrochlorides were prepared by reported procedures.

2.2b *General procedure-I: a-amino acid coupling*: *Synthesis of peptides (DCC/HOBt method)*: N-hydroxybenzotriazole (HOBt) (1 mmol) and dicyclohexylcarbodiimide (DCC) (1 mmol) were added sequentially at 0°C either to a stirred solution of N-protected amino acid (1 mmol) in dry CH_2Cl_2 (20 ml) or to a mixture of dry DMF and CH_2Cl_2. After a period of ~ 0·25 h, the reaction mixture was admixed either with the amino acid methyl ester [freshly prepared at 0°C from the corresponding ester hydrochloride (1·2 mmol)] and triethylamine (1·2 mmol) in dry CH_2Cl_2 or in a mixture of dry DMF and CH_2Cl_2. The combined mixture was left stirred at room temperature for 48 h, the precipitated DCU was filtered and the residue was washed with CH_2Cl_2 (2 × 20 ml) and the combined filtrates were washed sequentially with cold 2 N H_2SO_4 (20 ml), water (20 ml) and saturated bicarbonate solution (20 ml). The organic extract was dried over anhydrous $MgSO_4$ and evaporated *in vacuo*. The residue was, in most cases, directly crystallized from ethylacetate-hexane or purified on a short column of silica gel using ethylacetate-benzene or ethylacetate-hexane as eluents.

(i) *Ts-Pro-Pro-OMe (57%)*: Mp. 105°C; IR (KBr) 2957, 2924, 1748, 1664, 1597, 1150 cm^{-1}; ^{1}H NMR ($CDCl_3$) ***d*** 1·59–2·28 (m, 8H, Pro C^bH_2 × 2 + Pro C^gH_2 × 2), 2·37 (s, 3H, tosyl CH_3), 3·37 (m, 2H, Pro C^dH_2), 3·69 (s + m, 5H, $COOCH_3$ + Pro C^dH_2), 4·31–4·78 (m, 2H, Pro C^aH × 2), 7·28, 7·81 (d, d, 4H, aromatic); FAB MS: m/z 381 $(M + H)^+$; anal. calcd. for $C_{18}H_{24}SN_2O_5$: C 56·84, H 6·31, N 7·36; found C 57·32, H 6·42, N 7·62.

(ii) *Ts-Pro-Gly-OMe (56%)*: IR (neat) 3390, 2953, 1749, 1672, 1596, 1527, 1160 cm^{-1}; ^{1}H NMR ($CDCl_3$) ***d*** 1·44–2·06 (m, 4H, Pro C^bH_2 + Pro C^gH_2), 2·53 (s, 3H, tosyl CH_3), 3·0–3·47 (m, 2H, Pro C^dH_2), 3·81 (s, 3H, $COOCH_3$), 3·91–4·44 (m, 3H, Pro C^aH + Gly CH_2), 7·37, 7·78 (d + m, 5H, aromatic + Gly NH)); FAB MS: m/z 341 $(M + H)^+$.

(iii) *Ts-Gly-Pro-OMe (50%)*: Mp. 60–61°C; IR (KBr) 3233, 2954, 2926, 1742, 1646, 1160 cm^{-1}; ^{1}H NMR ($CDCl_3$) ***d*** 1·65–2·34 (m, 4H, Pro C^bH_2 + Pro C^gH_2), 2·47 (s, 3H, tosyl CH_3), 3·34–3·62 (m, 2H, Pro C^dH_2), 3·69 (s + d, 5H, $COOCH_3$ + Gly CH_2), 4·41 (m, 1H, Pro C^aH), 5·53 (m, 1H, Gly NH), 7·31, 7·78 (d, d, 4H aromatic); FAB MS: m/z 341 $(M + H)^+$; anal. calcd. for $C_{15}H_{20}SN_2O_5$: C 52·94, H 5·88, N 8·23; found C 52·69, H 5·92, N 7·83.

(iv) *Ts-Gly-Gly-OMe (55%)*: Mp. 91–92°C; IR (KBr) 3412, 3143, 1746, 1646, 1536, 1165 cm^{-1}; ^{1}H NMR ($CDCl_3$) ***d*** 2·41 (s, 3H, tosyl CH_3), 3·65 (s + m, 5H, $COOCH_3$ + Gly CH_2), 4·0 (d, 2H, Gly CH_2), 6·03 (m, 1H, Gly NH), 7·31, 7·78 (d + m, d, 5H, aromatic + Gly NH); FAB MS: m/z 301 $(M + H)^+$; anal. calcd. for $C_{12}H_{16}SN_2O_5$: C 48·0, H 5·33, N 9·33; found C 48·27, H 5·33, N 9·24.

(v) *Ts-Pro-Leu-OMe (54%)*: Mp.100–102°C; IR (KBr) 3263, 2956, 1746, 1653, 1559, 1157 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 0·94 (d, 6H, Leu CH_3 × 2), 1·34–2·0 (m, 7H, Leu C^bH_2 + Leu C^gH + Pro C^bH_2 + Pro C^gH_2), 2·43 (s, 3H, tosyl CH_3), 3·75 (s + m, 5H, Pro C^dH_2 + $COOCH_3$), 4·06 (m, 1H, Pro C^aH) 4·53 (m, 1H, Leu C^aH), 7·37, 7·78 (d + m, d, 5H, aromatic + Leu NH); FAB MS: m/z 397 $(M + H)^+$; anal. calcd. for $C_{19}H_{28}SN_2O_5$: C 57·57, H 7·07, N 7·07; found C 57·50, H 6·79, N 6·85.

(vi) *Ts-Leu-Pro-OMe (48%)*: Mp.124–126°C; IR (KBr) 3137, 2953, 1758, 1639, 1598, 1167 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 0·94 (d, 6H, Leu CH_3 × 2), 1·18–2·15 (brm, 7H, Leu C^bH_2 + Leu C^gH + Pro C^bH_2 + Pro C^gH_2), 2·41 (s, 3H, tosyl CH_3), 3·37 (m, 2H Pro C^dH_2) 3·65 (s, 3H, $COOCH_3$), 3·72–4·12 (m, 2H, Pro C^aH + Leu C^aH), 5·47 (d, 1H, Leu NH), 7·28, 7·72 (d, d, 4H, aromatic); FAB MS: m/z 397 $(M + H)^+$.

(vii) *Ts-Leu-Leu-OMe (60%)*: Mp.124°C; IR (KBr) 3261, 2957, 1726, 1657, 1597, 1527, 1166 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 0·81 (d, 12H, Leu CH_3 × 4), 1·19–1·87 (m, 6H, Leu C^bH_2 × 2 + Leu C^gH × 2), 2·25 (s, 3H, tosyl CH_3), 3·72 (s + m, 4H, $COOCH_3$ + Leu C^aH), 4·37 (m, 1H, Leu C^aH), 6·03 (d, 1H, Leu NH), 6·75 (d, 1H, Leu NH),7·25, 7·78 (d, d, 4H, aromatic); FAB MS: m/z 413 $(M + H)^+$.

(viii) *Ts-Pro-Phe-OMe (57%)*: IR (neat) 3396, 2953, 1744, 1674, 1597, 1517, 1161 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 1·31–1·66 (m, 4H, Pro C^bH_2 + Pro C^gH_2), 2·41 (s, 3H, tosyl CH_3), 2·88–3·47 (m, 4H, Phe C^bH_2 + Pro C^dH_2) 3·75 (s, 3H, $COOCH_3$), 4·03 (m, 1H, Pro C^aH) 4·81 (m, 1H, Phe C^aH), 7·0–7·81 (m + d, 10H, aromatic + Phe NH); FAB MS: m/z 431 $(M + H)^+$; anal. calcd. for $C_{22}H_{26}SN_2O_5$: C 61·39, H 6·04, N 6·51; found C 61·07, H 6·03, N 6·18.

(ix) *Ts-Phe-Pro-OMe (51%)*: IR (neat) 3190, 2924, 1744, 1638, 1598, 1161 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 1·41–1·87 (m, 4H, Pro C^bH_2 + Pro C^gH_2), 2·28 (s, 3H, tosyl CH_3), 2·66–3·19 (m, 4H, Phe C^bH_2 + Pro C^dH_2), 3·59 (s, 3H, $COOCH_3$), 3·62–4·25 (m, 2H, Pro C^aH + Phe C^aH), 5·53 (d, 1H, Phe NH), 7·07–7·78 (m + d, 9H, aromatic); FAB MS: m/z 431 $(M + H)^+$.

(x) *Ts-Phe-Phe-OMe (58%)*: Mp.131–132°C; IR (KBr) 3349, 3317, 3258, 1741, 1662, 1597, 1541, 1159 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 2.31 (s, 3H, tosyl CH_3), 2·88 (m, 4H, Phe C^bH_2 × 2), 3·59 (s, 3H, $COOCH_3$), 3·78 (m, 1H, Phe C^aH), 4·63 (m, 1H, Phe C^aH), 4·94 (d, 1H, Phe NH), 6·5 (d, 1H, Phe NH), 6·72–7·59 (m + d, 14H, aromatic); FAB MS: m/z 481 $(M + H)^+$.

(xi) *Ts-Pro-Trp-OMe (53%)*: Mp.67–69°C; IR (KBr) 3394, 2924, 1742, 1666, 1596, 1520, 1159 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 1·03–1·63 (m, 4H, Pro C^bH_2 + Pro C^gH_2), 2·31 (s, 3H, tosyl CH_3), 3·0, 3·25 (m, m, 4H, Trp C^bH_2 + Pro C^dH_2) 3·63 (s, 3H, $COOCH_3$), 4·0 (m, 1H, Pro C^aH) 4·75 (m, 1H, Trp C^aH), 6·81–7·69 (m, 10H, aromatic + Trp NH), 8·22 (brs, 1H, Trp ring NH); FAB MS: m/z 470 $(M + H)^+$.

(xii) *Ts-Trp-Pro-OMe (62%)*: Mp.71–73°C; IR (KBr) 3396, 2923, 1742, 1636, 1558, 1160 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 1·47–1·94 (m, 4H, Pro C^bH_2 + Pro C^gH_2), 2·34 (s, 3H, tosyl CH_3), 2·60–3·37 (d + m, 4H, Trp C^bH_2 + Pro C^dH_2) 3·66 (s, 3H, $COOCH_3$), 3·94–4·47 (m, 2H, Pro C^aH + Trp C^aH), 5·84 (d, 1H, Trp NH), 7·0–7·81 (brm, 9H, aromatic), 8·31(brs, 1H, Trp ring NH); FAB MS: m/z 470 $(M + H)^+$.

(xiii) *Ts-Trp-Trp-OMe (47%)*: Mp.120–121°C; IR (KBr) 3402, 2922, 1750, 1671, 1523, 1162 cm^{-1}; ^1H NMR $(CDCl_3)$ *d* 2·31 (s, 3H, tosyl CH_3), 3·12 (m, 4H, Trp C^bH_2 × 2), 3·62 (s, 3H, $COOCH_3$), 3·78–4·18 (m, 1H, Trp C^aH), 4·59–5·09 (d + m, 2H, Trp C^aH + Trp NH), 6·59–8·12 (m, 17H, aromatic + Trp NH + Trp ring NH × 2); FAB MS: m/z 559 $(M + H)^+$.

2.2c *General procedure-II*: *Reaction of proline and target amino acid (AA) with 3 equivalent amounts of water-soluble carbodiimide in water*: Aqueous solutions of proline (1 mmol, 5 ml), target amino acid (1 mmol, 5 ml) and the water-soluble carbodiimide, 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulphonate (3 mmol, 5 ml) were mixed and stirred for 48 h at room temperature. The initial pH of ~ 7·1 showed essentially no change during this period. The reaction mixture was treated with NaOH (0·6 g, 2 mmol) and was cooled (~ 0°C), tosyl chloride (0·39 g, 2 mmol) was added. The reaction mixture was stirred for 4 h at room temperature, filtered (to remove unreacted tosyl chloride). The aqueous portion was cooled in ice, acidified with 5 N HCl to pH 2, saturated with NaCl, extracted with ethylacetate (3 × 20 ml), dried over $MgSO_4$, and evaporated. The residue was dissolved in minimum amount of MeOH, cooled, treated with ethereal CH_2N_2, and evaporated. HPLC was performed on the residues and their composition determined by comparison with authentic possible dipeptides (HPLC). Duplicate runs were performed in each case.

3. Results

A comparison of the percentage occurrence of each of the 20 coded amino acids in the general data set, with those present in the *a*-helix and *b*-sheet regions show conformity with the gross structural make-up of these regions, in terms of the augmented presence of the recognized promoter residues (Chou 1989).

A comparison of gross neighbour preferences involved in dipeptide formation for each of the twenty proteinous amino acids in the basic set and in the parallel set constructed, assuming a nonpreference in the choice of neighbours is presented in table 1. The computed values are presented as percentage deviations from the nonpreference values (NPV) – and the significant deviations (20% and above) from the basic set are highlighted. This analysis (table 1) convincingly demonstrates pervasive neighbour preferences across the total domain.

A second element of preference for neighbour can be expected, that could be illustrated with the formation of a dipeptide of alanine (Ala) with glycine (Gly), which could result in either Gly-Ala or Ala-Gly, depending upon the relative preference for placement of Gly either from the amino side (left) or the carboxyl side (right) of Ala. As could be anticipated, these preferences are unequal. The outcome of this analysis, where significant aberrations are highlighted (table 2), carried out with the basic set, clearly brings out this novel and hitherto not reported profile.

Tables 1 and 2, clearly bring out the preference-profile in neighbour selection. However the factors that may control this are not clear. Intrinsic preferences for peptide formation, where the side chains play an important role, can be shown experimentally. Such options may not be available for a growing peptide chain. What can be done therefore is to determine preferences in dipeptide formation and if preferences exist here, it would have a role in protein evolution. The extent of correlation will reflect possible roles for the side chains of amino acids in the process.

For the experimental verification of the preference profiles (tables 1, 2) for dipeptide formation, proline was chosen as the central residue since this amino acid showed a pervasive and strong preference profile across the 20 proteinous amino acids. To simplify interpretation, glycine (Gly), leucine (Leu), phenyl alanine (Phe), and tryptophan (Trp) – having no functional groups in the side

Table 1. Percentage deviation in neighbour preference observed in a total of 9956 pairs from those expected on the basis of non-preference for neighbours.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	+ 21	+ 5	+ 21	– 12	– 23	+ 5	+ 5	– 20	+ 7	**– 18**	– 20	– 11	– 12	+ 17	– 23	– 5	– 1	+ 12	+ 4	– 15	A
A	141	+ 8	– 7	**– 27**	**– 26**	**+ 56**	**+ 31**	**– 35**	– 16	**– 29**	**– 33**	+ 9	+ 4	+ 16	+ 12	+ 13	– 19	– 9	**+ 29**	+ 11	C
C	43	13	– 11	+ 19	**+ 39**	– 3	**– 25**	0	+ 3	– 20	– 5	– 2	+ 4	– 13	**– 44**	– 8	– 1	+ 7	**+ 31**	+ 8	D
D	94	29	63	**+ 42**	+ 14	– 18	– 24	**– 26**	+ 7	+ 22	+ 17	– 16	+ 9	+ 3	+ 3	– 13	– 16	– 1	+ 7	+ 19	E
E	86	26	57	**52**	– 4	– 10	**– 33**	+ 3	+ 9	– 9	– 15	0	– 6	– 12	**+ 35**	+ 9	**+ 43**	– 18	**– 70**	+ 4	F
F	61	**19**	**19**	37	27	+ 8	+ 5	+ 3	– 15	– 18	– 13	**– 40**	– 9	+ 2	+ 16	**+ 31**	– 7	– 7	0	+ 6	G
G	144	**45**	97	88	63	148	**– 50**	0	+ 24	+ *39*	– 22	**– 29**	**+ 68**	+ 11	– 13	– 3	– 10	– 3	**– 29**	– 17	H
H	41	**13**	**28**	25	**18**	42	**12**	**+ 41**	– 7	+ 6	– 13	– 13	+ 13	– 6	+ 17	– 10	+ 18	+ 5	**+ 46**	– 15	I
I	76	**23**	**51**	**46**	33	78	22	**41**	– 4	+ 24	**+ 52**	– 12	– 8	– 37	**– 34**	– 1	+ 4	+ 5	**– 41**	+ 12	K
K	100	31	76	61	43	103	29	54	71	+ 1	+ 19	+ 8	+ 2	+ 2	**+ 46**	– 2	– 1	+ 6	– 19	**– 32**	L
L	123	**38**	82	74	53	127	**36**	66	87	107	**+ 129**	+ 7	– 13	+ 15	+ 9	– 15	**– 38**	+ 4	0	**– 31**	M
M	30	9	20	18	13	30	9	16	**21**	26	7	**+ 47**	+ 16	**– 26**	0	+ 4	+ 19	+ 10	**+ 25**	+ 26	N
N	73	22	49	44	32	**75**	**21**	39	52	63	15	**38**	**– 28**	– 3	– 14	+ 3	+ 8	– 7	– 15	+ 9	P
P	74	23	49	45	32	76	**22**	40	52	64	16	38	39	**+ 62**	**+ 43**	0	– 23	**– 29**	+ 20	**+ 27**	Q
Q	60	19	40	37	26	62	18	33	**43**	52	13	**31**	32	**26**	+ 20	– 24	**– 26**	+ 19	0	– 17	R
R	53	17	**36**	32	**23**	55	16	29	**38**	**46**	11	28	28	**23**	20	– 5	– 1	– 11	+ 9	+ 11	S
S	129	40	86	78	56	**132**	38	69	91	112	27	67	67	55	49	118	– 18	+ 12	– 6	+ 20	T
T	101	31	68	62	**44**	104	30	55	72	88	**21**	52	53	43	38	93	73	+ 1	**+ 36**	**– 31**	V
V	113	35	76	67	49	116	33	61	81	99	24	59	59	**49**	43	104	82	91	**– 50**	– 10	W
W	24	7	**16**	15	**10**	25	7	**13**	**17**	21	5	**12**	13	10	9	22	17	**19**	**4**	+ 8	Y
Y	60	19	40	37	26	62	18	33	43	**53**	**13**	**31**	32	**26**	23	55	44	**49**	10	26	
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

Non preference value

The bottom half gives the actual numbers based on non-preference values (NPV) (Ranganathan *et al* 1999). This can be read out from the junction of desired neighbours (e.g. FR = 23). The top half provides the deviations actually seen from NPV. This also can be read out from the junction of the desired neighbours (e.g. FR = + 35) denoting the actual number seen as 23 + 23 × 35/100 = 31. The diagonal element places the two values one above the other for identical pairs. The notable deviations are highlighted.

Table 2. Analysis of neighbour (left-right) preferences (62 proteins, 9416 residues).

A	85/85	21/24	56/58	32/44	24/23	78/73	27/16	34/27	52/55	55/46	15/9	30/35	32/33	36/34	21/20	61/62	51/49	64/63	12/13	23/28
C	24/21	7/7	10/17	9/10	4/10	41/29	6/11	7/8	14/12	12/15	4/2	17/7	14/10	13/9	9/10	23/22	9/16	13/19	6/3	7/14
D	58/56	17/10	28/28	32/36	32/25	50/44	6/15	27/24	31/38	37/29	4/15	22/26	23/28	15/20	10/10	47/32	35/32	31/50	8/13	30/13
E	44/32	10/9	36/32	37/37	23/20	37/35	9/10	12/22	29/36	51/39	14/7	21/16	17/32	22/16	15/18	33/33	22/30	32/34	12/4	18/26
F	23/24	10/4	25/32	20/23	13/13	21/36	3/9	16/18	27/20	25/23	6/5	16/16	23/7	11/12	18/13	31/30	32/31	16/24	2/1	15/12
G	73/78	29/41	44/50	35/37	36/21	80/80	24/20	48/32	48/40	53/51	15/11	22/23	25/44	29/34	35/29	80/93	54/43	55/53	9/16	34/32
H	16/27	11/6	15/6	10/9	9/3	20/24	3/3	9/13	22/14	28/22	2/5	5/10	22/15	7/13	6/8	16/21	14/13	16/16	2/3	5/8
I	27/34	8/7	24/27	22/12	18/16	32/48	13/9	29/29	26/24	25/45	7/7	18/16	32/13	15/16	13/21	29/33	35/30	38/26	7/12	19/9
K	55/52	12/14	38/31	36/29	20/27	40/48	14/22	24/26	34/34	54/54	15/17	27/19	19/29	11/16	14/11	53/37	32/43	44/41	3/7	29/19
L	46/55	15/12	29/37	39/51	23/25	51/53	22/28	45/25	54/54	54/54	13/18	33/35	42/23	30/23	37/30	59/51	46/41	47/58	6/11	14/22
M	9/15	2/4	15/4	7/14	5/6	11/15	5/2	7/7	17/15	18/13	8/8	8/8	7/7	7/8	7/5	10/13	7/6	13/12	3/2	5/4
N	35/30	7/17	26/22	16/21	16/16	23/22	10/5	16/18	19/27	35/33	8/8	28/28	20/24	12/11	11/17	32/38	32/30	38/27	8/7	19/20
P	33/32	10/14	28/23	32/17	7/23	44/25	15/22	13/32	29/19	23/42	7/7	24/20	14/14	11/20	8/16	44/25	25/32	33/22	9/2	16/19
Q	34/36	9/13	20/15	16/22	12/11	34/29	13/7	16/15	16/11	23/30	8/7	11/12	20/11	21/21	16/17	22/33	16/17	17/18	7/5	14/19
R	20/21	10/9	10/10	18/15	13/18	29/35	8/6	21/13	11/14	30/37	5/7	17/11	16/8	17/16	12/12	18/19	13/15	28/23	4/5	7/12
S	62/61	22/23	32/47	35/33	30/31	93/80	21/16	33/29	37/53	51/59	13/10	38/32	25/44	33/22	19/18	56/56	47/45	44/49	15/9	35/26
T	49/51	16/9	32/35	30/22	31/32	43/54	13/14	30/35	43/32	41/46	6/7	30/32	32/25	17/16	15/13	45/47	30/30	45/47	10/6	26/27
V	63/64	19/13	50/31	34/32	24/16	53/55	16/16	26/38	41/44	58/47	12/13	27/38	22/33	18/17	23/28	49/44	47/45	46/46	10/15	16/18
W	13/12	3/6	13/8	4/12	1/2	16/9	3/2	12/7	7/3	11/6	2/3	7/8	2/9	5/7	5/4	9/15	6/10	15/10	1/1	4/5
Y	28/23	14/7	13/30	26/18	12/15	32/34	8/5	9/19	19/29	22/14	4/5	20/19	19/16	19/14	12/7	26/35	27/26	18/16	5/4	14/14
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

Central residue

From the bottom line any of the twenty amino acids can be chosen as the central residue and the left-right preferences with respect to any of the twenty amino acids directly read out. For example for the central residue F the profile with respect to P, is PF : FP :: 7 : 23 and conversely for the central residue P with respect to F, is FP : PF :: 23 : 7. The significant results are highlighted.

chain and which should reflect the consequences of increasing steric requirements – were selected as partners for peptidation with proline.

The experimental procedure envisaged the peptidation of proline, individually, with Gly, Leu, Phe, and Trp, in water at pH ~ 7 using a water-soluble carbodiimide condensing agent and subsequent analysis of the resulting dipeptides. Each such peptidation, involving amino acid AA and Pro can give four dipeptides, Pro-Pro, AA-Pro, Pro-AA, and AA-AA. Authentic samples of 13 dipeptides that can arise from the four sets, were prepared, ends protected and conditions standardized for their clean separation and identification by HPLC. The profiles of the standards thus secured, are presented in table 3.

Products from each peptidation, involving Pro and Gly, Leu, Phe and Trp in water at pH 7·1, with the water-soluble condensing agent, were processed precisely as with the standards, and their yields and percentage distribution were determined, by HPLC. The results are presented in table 4 together with left-right preferences from the data base.

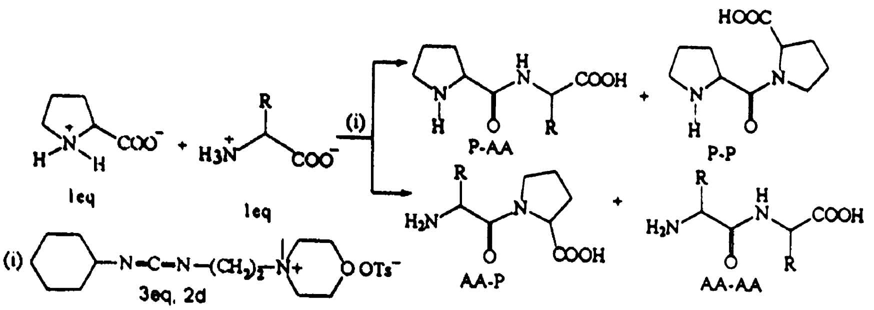
Analysis of the results would be facilitated by examination of likely pathways involved in the peptidation (figure 1). The first step in the peptidation leading to the formation of the activated ester involving zwitterionic AA and Pro, proceeds most likely by proton transfer-mediated carboxylate addition (top of figure 1). In both cases the activated ester, expected to be present in low concentrations at any time, is stabilized against polymerization by equilibrium favoured nitrogen protonation. The crucial peptidation step (figure 1), generally considered as proceeding through a six-membered transition state (Bodanzky and Bodanzky 1994), as envisaged here, would be encouraged by favourable electrostatic interaction with the termini of the emerging dipeptides. On the other hand, increasing steric requirement of the amino acid side chain would not only make such electrostatic interactions difficult, but also the actual peptidation involving nucleophilic addition followed by elimination of the condensing agent.

The complete absence of Pro-Pro in the entire set (table 4) must be due to steric overcrowding at the peptidation step contributed by the rigid proline frame (figure 1). The highly negative (– 28%) preference for Pro-Pro peptidation, seen in the data base (table 1) is in good agreement with the experimental finding. A superficial ordering of

Table 3. HPLC profile of authentic N, C-protected dipeptides.

Sub set	Dipeptide	Mobile phase	Flow rate	Retention time (min)
A	Ts-Pro-Pro-OMe	MeOH : H_2O		8·85
	Ts-Pro-Gly-OMe	: :	0·8 ml/min	6·86
	Ts-Gly-Pro-OMe	60 : 40		6·39
	Ts-Gly-Gly-OMe			5·25
B	Ts-Pro-Pro-OMe	MeCN : H_2O		6·28
	Ts-Pro-Leu-OMe	: :	0·8 ml/min	10·55
	Ts-Leu-Pro-OMe	60 : 40		9·29
	Ts-Leu-Leo-OMe			12·58
C	Ts-Pro-Pro-OMe	MeCN : H_2O		6·26
	Ts-Pro-Phe-OMe	: :	0·8 ml/min	11.16
	Ts-Phe-Pro-OMe	60 : 40		9·25
	Ts-Phe-Phe-OMe			13·67
D	Ts-Pro-Pro-OMe	MeCN : H_2O		6·24
	Ts-Pro-Trp-OMe	: :	0·8 ml/min	9·3
	Ts-Trp-Pro-OMe	60 : 40		8
	Ts-Trp-Trp-OMe			10·66

Table 4. Experimentally determined proline (P) dipeptide distribution in the condensation in water of proline with amino acids (AA) having side chains without functional groups mediated by water-soluble carbodiimide.



	Experimental				Data base	
Amino acid (AA)	P-P	AA-P	P-AA	AA-AA	AA-P	P-AA
Glycine	0	0	100 (11·3%)	0	36	64
Leucine	0	59 (16%)	41 (11%)	0	65	35
Phenylalanine	0	74 (12·7%)	26 (4·4%)	0	77	23
Tryptophan	0	100 (17·.6%)	0	0	18	82

The isolated yields of the dipeptides are given in brackets. The relevant database information is also presented.

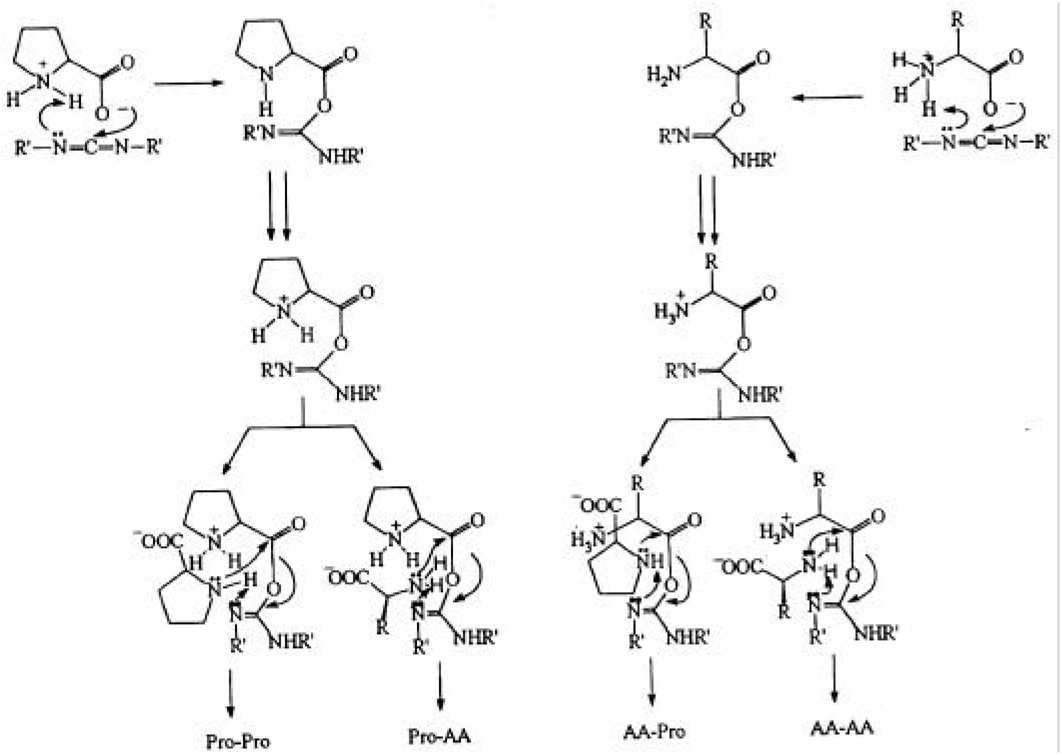


Figure 1. Suggested mechanistic explanation of the preference profile seen in the process, Pro + AA → Pro-Pro + Pro-AA + AA-Pro + AA-AA, using water soluble carbodiimide in water at pH 7·1.

the four transition states envisaged in the peptidation step (figure 1), in terms of crowding, would tend to show that whilst that leading to Pro-Pro is crowded, those to Pro-AA and AA-Pro are more relaxed compared to AA-AA; and that in the competitive runs, in no case, AA-AA was formed (table 4).

A consistent experimental finding is that with increasing steric needs of the AA, AA-Pro overtakes Pro-AA (table 4) and find excellent examples from the basic set (table 2), a good one being that arising from a single methyl substitution at the *g* position namely, Pro-Val: Val: Pro :: 60 : 40; Pro-Ile: Ile-Pro :: 29 : 71. Amongst Gly, Leu, Phe and Trp used in experiments, tryptophan being the most sterically demanding, the ratio Trp-Pro: Pro-Trp, should be highest, as actually seen (table 4). However, the basic set shows a reverse trend. An explanation could be the generally accepted view that Trp is a late addition to the proteinous amino acid complement, arising from pressure on evolution (Creighton 1984).

The experimental findings presented here are in no way comprehensive and the conclusions can be interpreted as useful pointers. The creation of a larger mural will call for an incredible amount of experimentation, involving the synthesis and characterization of a very large number of compounds. Within these limitations, we have shown that neighbour preferences exist, and could be amenable for mechanistic predictions.

4. Conclusion

Returning to the poser by Bernal (1967), we have presented data that strongly suggests that: Were the 20 proteinous amino acids made to combine in water, the resulting peptide sequences would be nonrandom.

Data base analysis has clearly brought out preferences in the choice of neighbours. Intrinsic preferences, if any, were tested experimentally in a limited manner and here again the dipeptide distribution was nonrandom. Whilst the existence of preferences can be seen, the factors that control this can be varied and subtle.

The intrinsic preferences seen can be put to practical advantage in the creation of peptide libraries by sequential condensation of the unprotected monomers. The preference profile is likely to restrict the formation to a gradient pattern.

Acknowledgements

We are grateful to Prof. C Ramakrishnan, Molecular Biophysics Unit, Indian Institute of Science, Bangalore, for help and discussions. Financial assistance from the Indian National Science Academy and Department of Science and Technology, New Delhi is gratefully acknowledged.

References

Bernal J D 1967 *The origin of life* (London: Weidenfeld and Nicolson)

Bodanzky M and Bodanzky A 1994 *The practice of peptide synthesis* (New York: Springer Verlag)

Chou P Y 1989 Prediction of protein structural classes from amino acid composition; in *Prediction of protein structure and the principles of protein conformation* (ed.) G D Fasman (New York: Plenum Press) pp 549–558

Creighton T E 1984 *Proteins: Structure and molecular principle* (New York: Freeman)

Doolittle R 1989 Similar amino acid sequences revisited; *TIBS* **14** 244–245

Dose K, Hartmann J and Brand M C 1982 Formation of specific amino acid sequences during carbodiimide mediated condensation of amino acids in aqueous solutions; *Biosystems* **15** 195–200

Erhan S 1978 Systematics: Use of protein nearest neighbor frequency distribution as an objective key; *Int. J. Bio-Med. Comput.* **9** 115–125

Harada K and Fox S W 1965 Thermal polycondensation of free amino acids with polyphosphoric acid; in *The origins of prebiological systems and of their molecular matrices* (ed.) S W Fox (New York: Academic Press) pp 289–298

Kolaskar A S and Ramabrahmam V 1982 Obligatory amino acids in primitive proteins: *Biosystems* **15** 105–109

Kricheldori H R, Au M and Mang T 1985 Models of molecular evolution; *Int. J. Peptide Protein Res.* **26** 149–157

Orgel L E 1992 Molecular replication; *Nature* (*London*) **358** 203–209

Ranganathan S, Kundu D and Tamilarasu N 1999 Protein evolution: The deciphering of latent facets – correlation of synthesis profiles of ribosomally directed proteins and enzyme directed peptides; *J. Biosci*. **24** 103–113

Ranganathan S and Ranganathan D 1981 Self organizing systems: Evolution of the genetic apparatus; *Trans. Bose Res. Inst.* **44** 109–116

Rani M and Mitra C K 1994 Periodicities in protein sequences; *J. Biosci*. **19** 255–266

Rani M and Mitra C K 1996 Pair preferences: A quantitative measure of regularities in protein sequences; *J. Biomol. Struct. Dynam.* **13** 935–944

Tyagi S and Ponnamperuma C 1990 Non randomness in prebiotic peptide synthesis; *J. Mol. Evol.* **30** 391–399

Vijayan M 1988 Molecular interactions and aggregation involving amino acids and peptides and their role in chemical evolution; *Prog. Biophys. Mol. Biol.* **52** 71–99

Villar H O and Kauvar M 1994 Amino acid preferences at protein binding sites; *FEBS Lett.* **349** 125–130

Vonderviszt F, Matrai G Y and Simon I 1986 Characteristic sequential residue environment of amino acids in proteins; *Int. J. Peptide Protein Res.* **27** 483–492

Walder J A, Walder R Y, Heller M J, Freier S M, Letzinger R and Klotz I M 1979 Complementary carrier peptide synthesis: General strategy and implications for prebiotic origin of peptide synthesis; *Proc. Natl. Acad. Sci. USA* **76** 51–55

MS received 9 July 2003; accepted 16 August 2003

Corresponding editor: DIPANKAR CHATTERJI