Amino acid doublets and triplets in protein sequences – A database analysis

K. Veluraja and S. A. Mugilan

Department of Physics, Manonmaniam Sundaranar University, Tirunelveli 627 002, India

Sequence analysis on the aspect of amino acid doublet and triplet repeats in 36,000 proteins comprising a total of 1,24,96,420 amino acid entries was carried out to study the structural aspects of proteins. Any deviation (normalized) of the doublet and triplet occurrence from their anticipated occurrence based on the frequency of the individual amino acid entries was attributed either to preferential selection (+ve deviation) or to non preferential selection (–ve deviation). All the homodoublets and homotriplets, except Ile–Ile–Ile, show positive deviation, implying the preferential selection for homomers. Most of the heterodoublets show marginal deviation indicating the expected distribution. A few doublets do show high positive and high negative deviations, containing the less propensity amino acids in the repeat. Most of the triplets which contain two identical amino acids in the repeat show positive deviations. In this category too, the high positively deviated triplets mainly contain less propensity amino acids, substantiating the economical usage of less propensity amino acids. A few triplets which contain two proline residues have large difference in observed and expected frequencies, in addition to high positive deviation indicating important structural role of proline residues. Most of the heterotriplets which have three different amino acid entries in the repeat show marginal deviations. The triplets with large deviation in this category also are dominated by less propensity amino acids. High frequency doublet pairs containing high propensity amino acid in the repeat (Glu–Lys, Lys–Glu), contribute structurally to α-helical conformation.

THREE-dimensional structural information (3-D information) about protein is buried in the one-dimensional amino acid sequences (1-D information). The number of available protein sequences is much larger than that of known three-dimensional structures. These protein sequences are available in SWISSPROT database. The three-dimensional structural information about the proteins is stored in the Brookhaven Protein Data Bank (PDB). These data banks contain valuable information on the structure of proteins. Exploiting these databases for information, regarding aspects of protein structure, is an integral part of research in structural biology[1–6]. Such analysis can be in aspects such as sequence analysis, protein fold investigation, secondary structure prediction studies, structure comparisons and knowledge-based model building of proteins[7–11]. This paper deals with the analysis of amino acid doublets and triplets in protein sequences and their implications to protein structure.

For sequence analysis the SWISSPROT Protein Sequence, Release 28 March 94 database was used. 36,000 proteins containing a total of 1,24,96,420 amino acid entries were considered for calculation.

The probability of occurrence of individual amino acid was calculated as

$$P(A) = \sum N_i(A)/\sum T_i,$$

where $N_i(A)$ is the number of entries for A in protein i, and T_i is the total number of amino acid entries in the corresponding protein.

The probability of occurrence of amino acid doublet repeat in the whole database was calculated as

$$P_{\text{comp}}(AB) = \sum N_i(AB)/\sum (T_i - 1),$$

where $N_i(AB)$ is the number of amino acid doublet entries AB in the protein i.

The probability occurrence of amino acid triplets in datum was calculated as

$$P_{\text{comp}}(ABC) = \sum N_i(ABC)/\sum (T_i - 2).$$

Theoretical estimates for the probabilities of amino acid doublets and triplets distribution were calculated using the following,

$$P_{\text{Theo}}(AB) = P(A)*P(B) = P_{\text{Theo}}(BA),$$

$$P_{\text{Theo}}(ABC) = P(A)*P(B)*P(C) = P_{\text{Theo}}(XYZ),$$

where X, Y, Z can be A or B or C. Expected frequency of occurrence (theoretical count) of doublet repeat AB is then $= P(A)*P(B)\Sigma_i\ (T_i-1)$. Expected frequency of occurrence (theoretical count) of triplet repeat ABC is then $= P(A)*P(B)*P(C)\ \Sigma_i\ (T_i-2)$.

The difference in count between the observed frequency (computed count) and the theoretical frequency (expected count) is normalized as follows:

For amino acid doublets

$$PD(AB) = ([P_{\text{comb}}(AB) - P_{\text{Theo}}(AB)]*100)/P_{\text{Theo}}(AB).$$

For amino acid triplets

$$PD(ABC) = ([P_{\text{comp}}(ABC) - P_{\text{Theo}}(ABC)]*100)/P_{\text{Theo}}(ABC).$$

The magnitude of the percentage of deviation could be treated as an index for the significance of observation.

The positive deviation is an indication of the preferential selection and the negative deviation is an indication of the non preferential selection of this repeat. In general, high percentage of deviation on the positive scale is likely to be an indication of an important role in the three-dimensional structure of proteins. On the other

Table 1. Probability of occurence of individual amino acids

Amino acid	Probability * 100
Leu	9.2
Ala	7.7
Ser	7.2
Gly	7.0
Val	6.5
Glu	6.3
Lys	5.8
Thr	5.8
Ile	5.6
Asp	5.3
Arg	5.3
Pro	5.0
Asn	4.5
Gln	4.0
Phe	4.0
Tyr	3.2
Met	2.4
His	2.2
Cys	1.8
Trp	1.3

Table 2. Deviation of homodoublets

Homo-doublets	Computed counts	Expected counts	Difference in counts	Deviation in %
His–His	10579	6330	4249	67
Cys–Cys	6517	3938	2579	66
Gln–Gln	31039	20211	13017	65
Glu–Glu	68819	48945	19875	41
Arg–Arg	47013	34167	12847	38
Trp–Trp	2879	2094	785	38
Ala–Ala	98462	72457	26005	36
Pro–Pro	42004	31375	10629	34
Lys–Lys	56521	42366	14155	33
Ser–Ser	82625	63424	19202	30
Asn–Asn	31687	24958	6728	27
Tyr–Tyr	16136	12934	3202	25
Met–Met	8560	6978	1583	23
Gly–Gly	73380	60832	12548	21
Thr–Thr	48197	42266	5931	14
Phe–Phe	22765	20061	2704	14
Asp–Asp	39288	34777	4511	13
Val–Val	58888	53057	5831	11
Leu–Leu	115209	105664	9548	9
Ile–Ile	41655	38951	2704	7

hand, repeats showing high negative deviation (doublet or triplet) may be detramental to protein structure.

The probability distribution of individual amino acids in the selected proteins is shown in Table 1. The results are in agreement with our earlier result[9], where the analysis was confined to a smaller database. As far as the individual amino acid occurrence is concerned, leucine occupies the highest position and tryptophan the last. Most of the amino acid entries have the probability distribution of more than 0.05 (mean probability distribution) with some exceptions asparagine, glutamine, phenylalanine, tyrosine, methionine, histidine, cysteine and tryptophan have probability of occurrence less than 0.05 while all other amino acids occur more frequently. Amino acids which have aromatic group in their side chain fall under the category of less propensity amino acids.

The computed and the theoretical probability distribution of the 400 amino acid doublets is shown in Figure 1. Figure 1 indicates that the computed probability of homomers always overshoots the theoretical, implying preferential selection of homomers. The range of deviation of homodimers varies from 7% to 67% (Table 2). This table and subsequent ones list the computed counts (actual counts), expected counts (theoretical counts) and the deviation in counts along with the percentage of deviation (normalized). The first three doublets showing more than anticipated frequency correspond to the less propensity amino acids histidine, cysteine and glutamine. The percentage of deviation was marginal (in between –5% and +5%) for most of the heterodoublets. But few heterodoublets have deviation ranging from +15% to +25% on the positive side, and –15% to –24% on the negative side. These heterodoublets are listed in Tables 3 *a* and 3 *b*. A roughly similar calculation was carried out by Rani *et al.*[11]. They have suggested that pair preference might be important for the three-dimensional structure of proteins. In the heterodoublets which show high positive deviation, one or both of the amino acids fall in the category of less propensity amino acids. The pairs that deviate from this rule are Glu–Lys and Lys–Glu (P(Glu) = 0.063 and P(Lys) = 0.058). A preliminary investigation on the conformational properties of these doublets was worked out by analysing the Ramachandran (Φ, Ψ) angles, (for 72 Glu–Lys repeats in 46 protein structures and 69 Lys–Glu repeats in 50 proteins). In Lys–Glu repeat, 72% and 66% of Ramachandran angles for Lys and Glu respectively are in the helical region. Similarly, in the Glu–Lys pair 71% for Glu and 63% for Lys are in helical region. Hence, these heterodoublets play a significant role in the three-dimensional structure of proteins especially in the α-helical segments. A (Φ, Ψ) distribution map related to these pairs is shown in Figures 2 and 3. In the doublets with high negative deviation also (Table 3 *b*), one or both of the amino acids of the repeats fall in the category of less propensity amino acid with exceptions being Glu–Pro, Glu–Ser and Pro–Ile. These doublet repeats have highest count difference in the negatively deviated heterodoublets. The reason for the deviated heterodoublets dominated by the less propensity amino acids may be due to their least occurrence, nature has used them economically by aligning them in proteins in a preferential (positive deviation) or in a non-preferential (negative deviation) way more than the average.

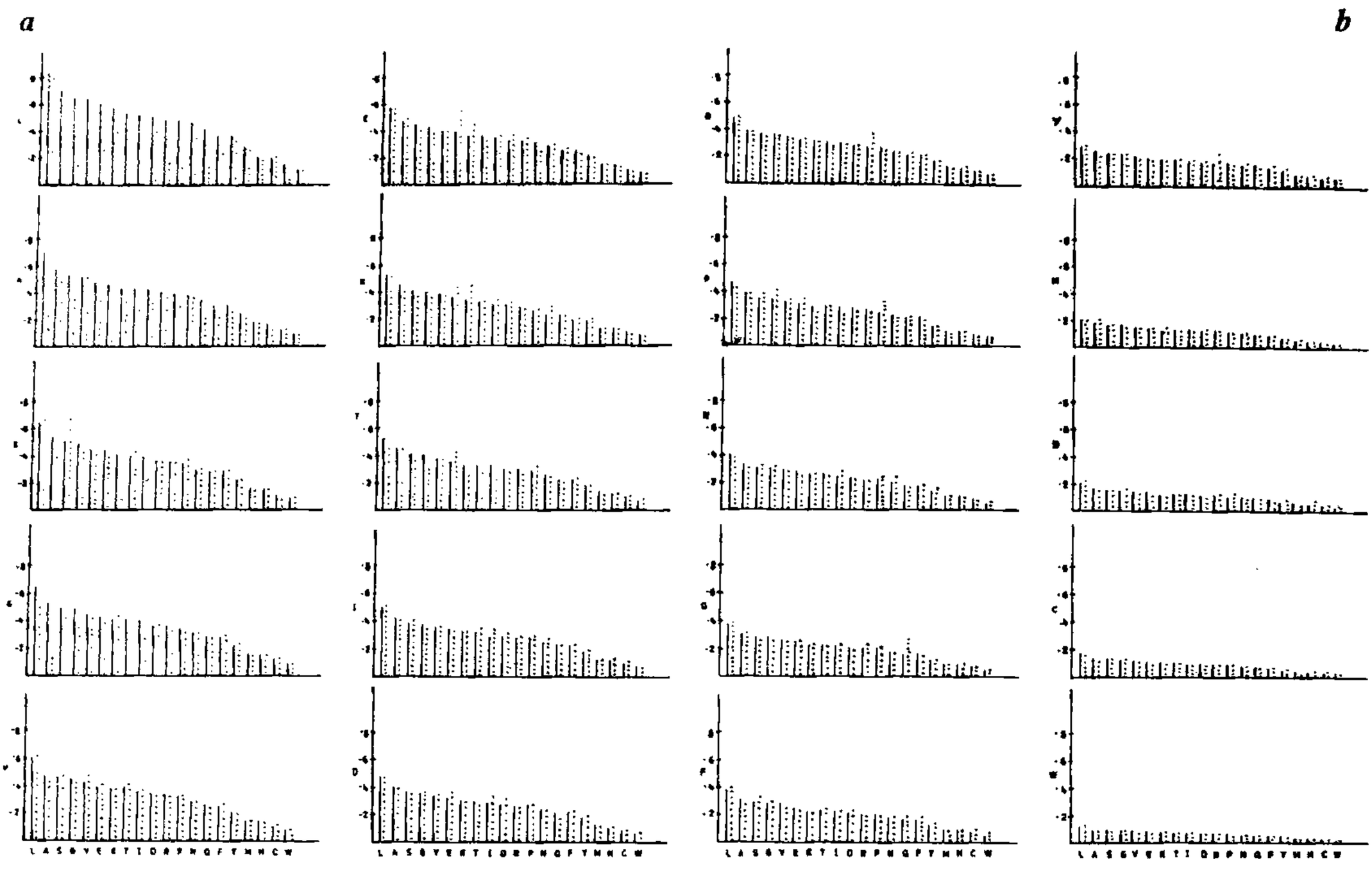

Figure 1. Theoretical (——) and computational (...) probability distributions. *a*, For amino acids Leu, Ala, Ser, Gly, Val, Glu, Lys, Thr, Ile, Asp. *b*, For amino acids Arg, Pro, Asn, Gln, Phe, Tyr, Met, His, Cys, Trp.

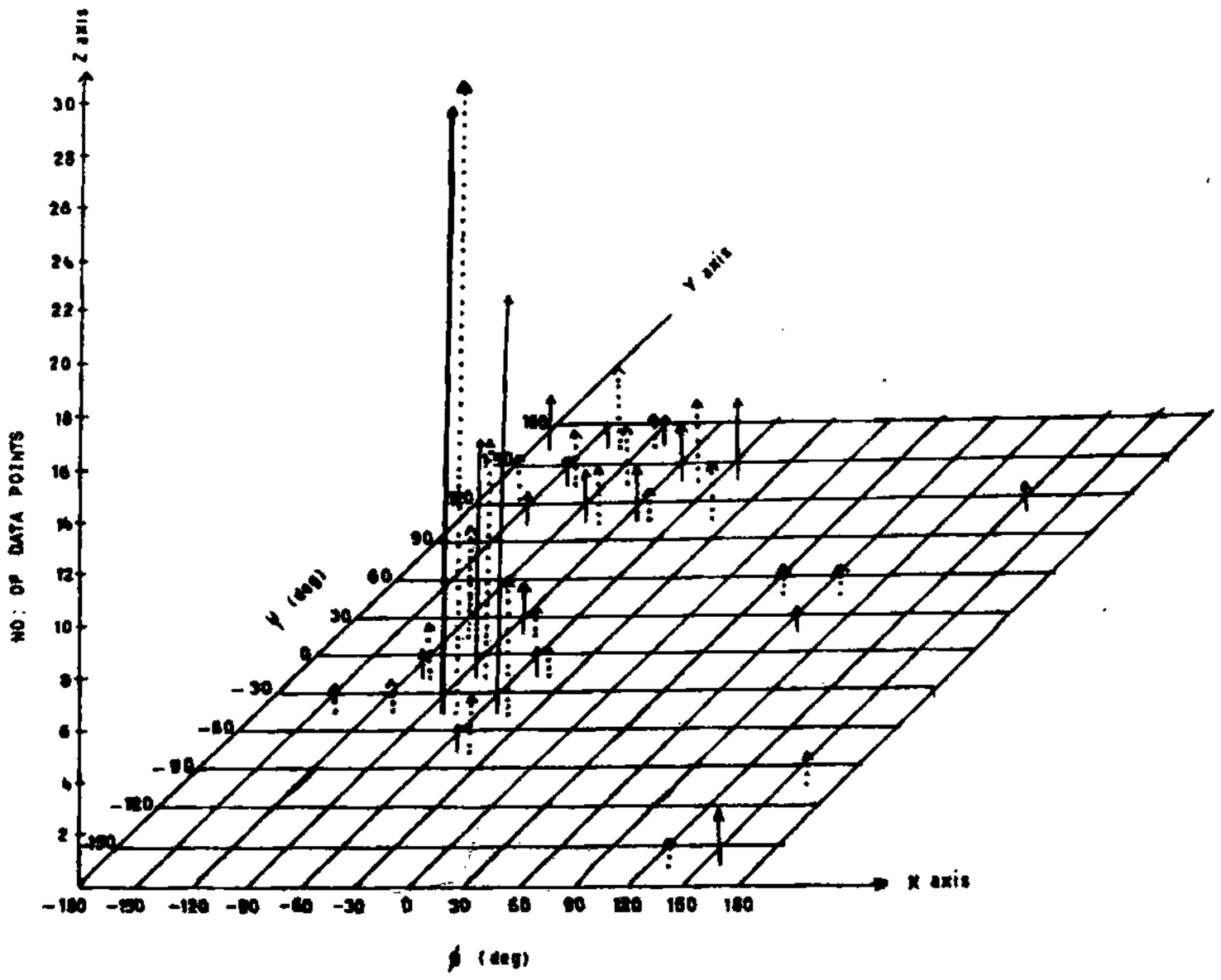

Figure 2. Ramachandran's (Φ, Ψ) angle distribution for Glu–Lys. ——, for Glu; ...., for Lys.

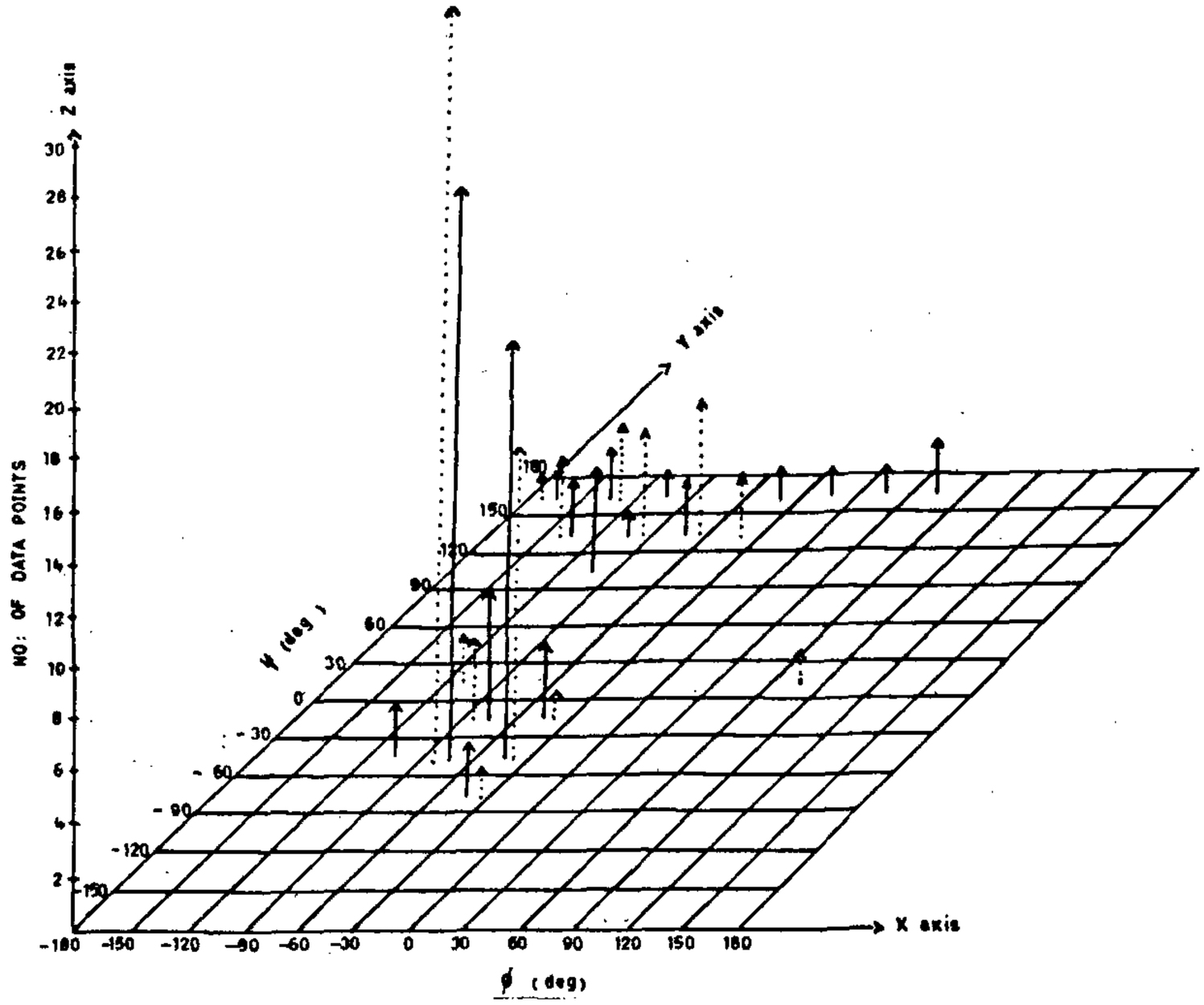

Figure 3. Ramachandran's (Φ, Ψ) angle distribution for Lys-Glu. ——, for Lys; ...., for Glu.

Table 3 *a*. Heterodoublets having positive deviation

Hetero-doublets	Computed counts	Expected counts	Difference in counts	Deviation in %
Cys–His	6243	4997	1246	25
His–Cys	6193	4997	1196	24
Glu–Lys	55275	45543	9732	21
Cys–Tyr	7564	7140	424	20
His–Pro	16946	14093	2853	20
Tyr–Cys	8523	7140	1383	19
His–Phe	13345	11264	2080	19
Met–Ala	26191	22491	3701	17
Lys–Glu	53181	45543	7638	17
Tyr–Trp	6056	5209	847	16
Tyr–Phe	18541	16111	2430	15
Asp–Tyr	24472	21208	3265	15
His–Trp	4187	3639	548	15

Table 3 *b*. Heterodoublets having negative deviation

Hetero-doublets	Computed counts	Expected counts	Difference in counts	Deviation in %
Trp–Pro	6156	8112	–1956	–24
Glu–Pro	30042	39188	–9146	–23
Pro–Met	11364	14803	–3439	–23
Cys–Met	4087	5246	–1159	–22
Glu–Ser	44733	55711	–10978	–19
Tyr–Ala	25070	30615	–5545	–18
His–Asp	12149	14840	–2692	–18
His–Glu	14641	17607	–2965	–17
His–Lys	13582	16386	–2804	–17
Asp–Gln	22192	26516	–4323	–16
Pro–Ile	29643	34952	–5308	–15

Table 4. Deviation of homotriplets

Homo-triplets	Computed counts	Expected counts	Difference in counts	Deviation in %
His–His–His	1089	143	946	662
Gln–Gln–Gln	6089	811	5278	651
Pro–Pro–Pro	5876	1569	4307	274
Cys–Cys–Cys	241	70	171	246
Arg–Arg–Arg	5503	1784	3719	208
Asn–Asn–Asn	3297	1113	2184	196
Glu–Glu–Glu	8727	3059	5668	185
Ser–Ser–Ser	11588	4513	7075	157
Ala–Ala–Ala	13616	5509	8107	147
Gly–Gly–Gly	9932	4238	5694	134
Trp–Trp–Trp	58	27	31	114
Met–Met–Met	353	165	188	114
Lys–Lys–Lys	5229	2464	2765	112
Asp–Asp–Asp	3779	1831	1948	106
Thr–Thr–Thr	4715	2455	2260	92
Tyr–Tyr–Tyr	764	416	348	84
Phe–Phe–Phe	1179	803	376	47
Val–Val–Val	4748	3452	1296	39
Leu–Leu-Leu	13507	9702	3805	38
Ile–Ile–Ile	2166	2172	–6	–0.28

When these studies were extended to triplet distribution (possible combination of 8000 triplets), all the homotriplets except, Ile–Ile–Ile, occurred with increased frequency, indicating a preferential selection of homomers (Table 4). For the homotriplets the percentage of deviation varies from –0.28 to 662. The highly deviated homotriplet is His–His–His and the least one is

Table 5. Deviation of triplet repeat which contains two identical amino acids in it

Triplet repeat with two identical amino acids	Computed counts	Expected counts	Difference in counts	Deviation in %
Cys–Trp–Cys	152	51	101	198
His–His–Pro	849	280	569	168
Cys–His–Cys	222	88	134	152
Cys–Ser–Cys	693	281	412	147
Cys–Cys–His	216	88	128	145
Tyr–Tyr–Cys	557	230	327	142
Cys–Cys–Phe	321	158	163	140
Cys–Tyr–Cys	302	127	175	138
Cys–Cys–Pro	452	198	254	127
Trp–Trp–Asn	211	93	118	127
Cys–Asn–Cys	387	176	211	119
Cys–Cys–Ser	611	281	330	118
His–His–Gln	548	255	293	115
His–Ile–His	755	353	402	114
Cys–Lys–Cys	489	230	259	113
Cys–Ala–Cys	634	299	335	112
Cys–Cys–Tyr	267	127	140	111
Pro–Pro–Gly	4584	2181	2403	110
His–Pro–His	650	317	333	109
Cys–Arg–Cys	430	206	224	108
Gln–Gln–His	945	455	490	108
Cys–Val–Cys	534	257	277	108
His–His–Arg	683	331	352	107
Cys–Gln–Cys	324	158	166	106
Cys–Asp–Cys	424	208	216	104
Cys–Thr–Cys	450	229	221	97
Gln–Gln–Pro	1964	1011	953	94
Cys–Cys–Arg	388	206	182	88
Cys–Cys–Gly	514	275	239	87
Pro–Gly–Pro	4042	2186	1856	85
Gln–Gln–Leu	3394	1856	1538	83
His–His–Met	272	150	122	83
Cys–Cys–Asn	321	176	145	82
Cys–Cys–Gln	285	158	127	80
His–Trp–His	148	82	66	80
Tyr–Tyr–Ser	1658	921	737	80
Trp–Trp–His	83	47	36	76
Cys–Ile–Cys	386	220	166	76
His–His–Phe	445	254	191	76
Arg–Trp–Arg	772	442	330	74
Pro–Thr–Pro	3170	1821	1349	74
Lys–Tyr–Lys	4595	2648	1947	74
His–Gln–His	441	254	187	73
Asn–Pro–Asn	2138	1249	889	71
Gln–Gln–Met	820	477	343	71
Trp–Pro–Trp	178	104	74	70
Glu–Glu–Met	1963	1156	807	70
Pro–Pro–His	1197	706	491	70
Gln–Pro–Gln	1695	1011	684	68
His–Phe–His	424	254	170	67
Asn–Asn–Met	978	589	389	66
Trp–Asn–Trp	154	93	61	65
Pro–Ala–Pro	3934	2386	1548	65
Pro–Asn–Pro	2310	1400	910	65
Pro–Ser–Pro	3638	2157	1481	63
Trp–Phe–Trp	136	83	53	63
His–His–Leu	943	582	361	62
His–His–Trp	133	82	51	62
Gln–Gln–Arg	1702	1056	647	61
Glu–Glu–Ile	4378	2728	1650	61
Asp–Asp–Glu	3481	2173	1308	60

Ile–Ile–Ile. Here the top two triplets which have a deviation of more than 600% are occupied by the less propensity amino acids, histidine and glutamine. It is also observed that most of the heterotriplets, which contain two identical amino acids in the repeat, always show increased frequency. Triplets which show deviation of more than 60% are listed in Table 5. In this group also (Table 5), the highly deviated triplets mostly contain one, two or three less propensity amino acids. A few exceptional triplets in this group are Pro–Pro–Gly (110%), Pro–Gly–Pro (85%), Pro–Thr–Pro (74%), Pro–Ala–Pro (65%), Pro–Ser–Pro (63%), Glu–Glu–Ile (61%) and Asp–Asp–Glu (60%). The above triplets, except Glu–Glu–Ile and Asp–Asp–Glu, contain proline in the repeat suggesting the importance of proline in the protein structures. The triplets which have different amino acids in the repeat invariably have marginal deviations (within + or –10%). A few triplets have deviations ranging from 60% to 196% on the positive side and –40% to –55% on the negative side (Table 6 and Table 7). All except four triplets showing positive deviation, contain two less propensity amino acids. This

Table 6. Heterotriplets having positive deviation

Hetero-triplets	Computed counts	Expected counts	Difference in counts	Deviation in %
Trp–His–Met	254	86	168	196
His–Trp–Tyr	284	117	167	144
His–Pro–Trp	442	183	259	142
Arg-Cys-Trp	354	150	204	136
His–Tyr–Cys	376	160	216	135
Cys-Tyr-His	349	160	189	118
Tyr–Thr–Cys	824	415	409	99
His–Cys–Asp	523	263	260	99
Ile–Gly–Glu	5893	3042	2851	94
Tyr–Cys–Asn	612	319	293	92
Cys–His–Thr	555	291	236	91
Tyr–Pro–Trp	500	261	239	91
Tyr–Cys–Arg	706	373	333	89
Trp–Cys–Tyr	173	92	81	88
Cys–Gly–Lys	1635	901	734	82
Cys–Trp–Ala	396	219	177	81
Arg–Tyr–Trp	490	272	218	80
Trp–Tyr–Gln	376	209	167	80
Ile–Trp–His	365	203	162	80
Trp–Phe–Gln	466	261	205	79
Tyr–Trp–Asp	468	275	193	71
Cys–Trp–Met	115	68	47	69
Glu–Tyr–Trp	551	326	225	69
His–Pro–Glu	1476	881	595	68
Thr–Cys–Trp	278	167	111	67
His–Val–Trp	394	237	157	66
His–Pro–Asp	1236	743	493	66
Tyr–His–Cys	266	160	106	66
His–Phe–Asp	978	594	384	65
Met–Arg–Trp	329	200	129	65
Gln–Met–Trp	254	154	100	65
Tyr–Glu–Cys	734	446	288	64
Val–Trp–His	384	237	147	62

Table 7. Heterotriplets having negative deviation

Hetero-triplets	Computed counts	Expected counts	Difference in counts	Deviation in %
Trp–Pro–Met	85	191	–106	–55
Cys–Asp–Trp	75	152	–77	–51
Lys–Trp–Pro	232	471	–239	–50
Pro–Met–Arg	398	773	–375	–49
Glu–Pro–Asn	903	1748	–845	–48
Ala–Met–Trp	153	291	–138	–47
Trp–Cys–Met	36	68	–32	–47
Met–Trp–Pro	101	191	–90	–47
Phe–Lys–Trp	204	377	–173	–46
Glu–Trp–Pro	273	507	–234	–46
Cys–Glu–Met	179	328	–149	–46
His–Glu–Pro	482	881	–399	–45
Trp–Ser–Met	152	272	–120	–44
Gly–Met–Trp	153	266	–113	–43
Arg–Met–Trp	114	200	–86	–43
Pro–Met–Lys	493	861	–368	–43
Glu–Pro–Asp	1190	2063	–873	–42
Trp–Pro–Cys	83	144	–61	–42
Met–Cys–Pro	153	262	–109	–42
Pro–Met–Trp	113	191	–78	–41
Glu–Pro–Met	548	924	–376	–41
Ile–Trp–Pro	267	452	–185	–40
Phe–Lys–Met	410	768	–358	–40
Trp–Pro–Ile	269	452	–183	–40

also substantiates our earlier interpretation that nature has used the least occurring amino acids in an economical way by aligning them in proteins in a preferential or in a non-preferential way. The triplets Cys–Gly–Lys, His–Pro–Glu, His–Pro–Asp, contain two high propensity amino acids. In the triplets His–Pro–Glu, His–Pro–Asp, there may be a possibility of hydrogen bonding interaction between the side chain of histidine and the aspartic acid or glutamic acid. The triplet which has a deviation of 94% containing three high propensity amino acid is Ile–Gly–Glu, and is the only listed high positive deviated triplet having high count difference. High negative deviated heterotrimeric triplets having deviation ranging from –40% to –55% are listed in Table 7. Most of the high deviated triplets in the repeats have one, two or three less propensity amino acid except Glu–Pro–Asp.

In general, all the homodoublets and homotriplet repeats (except Ile–Ile–Ile) show positive deviation implying preferential selection. Most of the heterodoublets show marginal percentage of deviation (±5%). High deviated (+ve) heterodoublets except Glu–Lys and Lys–Glu mainly contain less propensity amino acids, indicating that nature has used the less propensity amino acids, in a preferential (+ve deviation) manner or in a non preferential (–ve deviation) manner in proteins. In the heterotriplets which contain two identical amino acids, most of the repeats show positive deviation and most of the high positive deviated triplets in this group contain one, two or three less propensity amino acids in the repeat, substantiating the economical usage of low propensity amino acids. A few triplets in this category contain two proline residues which have high difference in counts along with high positive deviation, indicating that proline plays an important role in the structure of proteins. Heterotriplets which contain different amino acids show marginal deviation (±10%). Most of the high deviated triplets in this category contain at least two less propensity amino acids. In the triplets His–Pro–Glu and His–Pro–Asp which contain two high propensity amino acids with high positive deviation (Table 6), we expect a structural role in the form of hydrogen bonding interaction between the side chain histidine and the glutamic acid or aspartic acid. In the high deviated heterodoublet Glu–Lys and Lys–Glu pair, more than 60% of the Ramachandran (Φ, Ψ) angles fall in the α-helical region, indicating the importance of these doublets in the α-helical structures in proteins. Discussions on the conformational aspect here have been restricted to Glu–Lys and Lys–Glu doublets. In depth analysis on the conformational significance of the highly deviated doublets and triplets and their role in the secondary structural elements of proteins will be considered in a later paper. As the database considered for the sequence analysis is very large (36,000 proteins containing a total amino acid 1,24,96,420), it is believed that multiple entries of families of proteins will not bias the result presented in this communication.

1. Anfinsen, C. B., *Science*, 1973, **181**, 223–230.
2. Cbou, P. Y. and Fasman, G. D., *Biochemistry*, 1974, **13**, 222–244.
3. Suzuki, H., Kolaskar, A. S., Samuel, L. S., Otsuka, J. and Tsugita, A., *Protein Seq. Data Anal.*, 1991, **4**, 97–104.
4. Bajorath, J., Stenkamp, R. and Aruffo, A., *Protein Sci.*, 1993, **2**, 1798–1810.
5. Barton G. J., *Methods Enzymol.*, 1990, **183**, 403–427.
6. Naor, D., Fischer, D., Jernigan, R. L., Wolfson, H. J. and Hussinor, R., *J. Mol. Biol.*, 1996, **256**, 924–938
7. Kolaskar, A. S. and Kulkarni-Kale, U., *J. Mol. Biol.*, 1992, **223**, 1053–1061.
8. Barton G. J. and Sternberg, M. J. E., *J. Mol. Biol.*, 1987, **198**, 327–337.
9. Veluraja, K. and Priyadarshini, S., *Curr. Sci.*, 1993, **65**, 633–636.
10. Kannan, K., Babu, J., Veluraja, K. and Rajamanicam, C., *Curr. Sci.*, 1995, **68**, 819–825.
11. Rani, M., Mitra, C. K., Cserzo, M. and Simon, I., *J. Biosci.*, 1995, **20**, 579–590.

ACKNOWLEDGEMENTS. The national facility provided by the Bioinformatics Centre, School of Biotechnology, Madurai Kamaraj University is gratefully acknowledged. We thank Dr S. Rajasekar for a fruitful discussion.

Received 16 August 1996; revised accepted 5 March 1997