the polypeptide 20 kD, synthesized by CMS mitochondria, and its effect on microsporogenesis. Nevertheless, it can be inferred that the additional 20 kD mitochondrial protein may inhibit 20 kD protein in anther, leading to pollen abortion. This observation is consistent with the report on the involvement of mitochondrial gene in anther development in *Nicotiana*[20] and structural abonormalities of mitochondria during pollen abortion in maize CMS-T[21], wheat[22], sugarbeet[23], rice[24] and *Brassica*[25]. It has been shown that the protein 13 kD of CMS-T maize is present in the mitochondria of tapetal cells even after the mitochondria have begun to degenerate[26]. Similarly, the expression of *Pcf* gene is greatest in the anthers of CMS *Petunia*[27].

All the studies, including the present one, describe the mitochondrial translation products from the vegetative tissue. However, it has been documented that mitochondria from different organs synthesize different proteins[28], apparently in response to altered cellular milieu or environmental stimuli. Hence, for a better understanding of cytoplasmic male sterility, mitochondrial translation studies should be carried out in anthers where pollen abortion takes place at different developmental stages.

1. Schuster, J. C., Wissinger, B., Hiesel, R., Unseld, M., Gerold, E., Knoop, V., Marchfielder, A., Binder, S., Schobel, W., Scheike, R., Gronger, P., Ternes, R. and Brennicke, A., *Physiol. Plant*, 1991, **81**, 437–445.
2. Levings, C. S. and Brown, G. G., *Cell*, 1989, **56**, 171–179.
3. Hanson, M. R., *Annu. Rev. Genet.*, 1991, **25**, 461–486.
4. Newton, K. J., *Annu. Rev. Plant Physiol. Mol. Biol.*, 1988, **39**, 503–532.
5. Forde, B. G., Oliver, R. J C. and Leaver, C. J., *Proc. Natl. Acad. Sci. USA*, 1978, **75**, 3841–3845.
6. Forde, B. G. and Leaver, C. J., *Proc. Natl. Acad. Sci. USA*, 1980, **77**, 418–422.
7. Nivision, H. T. and Hanson, M. R., *Plant Cell*, 1989, **1**, 1121–1130.
8. Dixon, L. K. and Leaver, C. J., *Plant Mol. Biol.*, 1982, **1**, 89–102.
9. Leaver, C. J., Hack, E. and Forde, B. G., *Methods Enzymol*, 1983, **97**, 476–484.
10. Laemmli, U. K., *Nature*, 1970, **227**, 680–685.
11. Chamberlain, J. P, *Anal. Biochem.*, 1979, **98**, 132–135
12. Lowry, O. H., Rosebrough, N. J., Farr, A. L. and Randall, R. J, *J. Biol. Chem.*, 1951, **193**, 265–275.
13. Leaver, C. J. and Gray, M. W., *Annu. Rev. Plant Physiol.*, 1982, **33**, 373–402.
14. Levings, C. S., *Science*, 1990, **250**, 942–947.
15. Boutry, M. and Briquet, M., *Eur. J Biochem.*, 1982, **127**, 127–135.
16. Powling, A. and Ellis, T. H N., *Theor. Appl. Genet.*, 1983, **65**, 323–328.
17. Boutry, M., Faber, A. M., Charbonnier, M. and Briquet, M, *Plant Mol. Biol.*, 1984, **3**, 445–452.
18. Hakansson, G., Van der Mark, F., Bonnett, H. T. and Glimelius, K., *Theor. Appl. Genet.*, 1988, **76**, 431–437.
19. Manoharan, M. and Rudramuniyappa, C. K, *Theor Appl. Genet*, 1993 (in press).
20. Kofer, W., Glimelius, K. and Bonnett, H. T., *Plant Cell*, 1991, **3**, 759–769.
21. Warmke, H. E. and Lee, S. L. J., *J. Hered.*, 1977, **68**, 213–222
22. Turbin, N V., Pahilova, A N, Atrashenok, N V. and Lyul'kina, E. I, *Dorl ANSSSSR*, 1974, **214**, 721–722.
23. Nakashima, H, *Mem Fac. Agric. Hokkaido Univ*, 1975, **9**, 247-252
24. Li, J Y and Chu, Q., *Acta Agric. Shanghai*, 1987, **3**, 9–14.
25. Polowick, P. L. and Sawhney, V K., *Sex Plant Reprod*, 1990, **3**, 263–276.
26. Hack, E, Lin, C., Yang, H. and Horner, H. T., *Plant Physiol.*, 1991, **95**, 861–870.
27. Young, E. G. and Hanson, M. R, *Cell*, 1987, **50**, 41–49.
28. Newton, K. J. and Walbot, V., *Proc. Natl. Acad. Sci USA*, 1985, **82**, 6879–6883.

ACKNOWLEDGEMENTS. We are grateful to Dr S. K. Mahajan, Head, Molecular Biology and Agriculture Division, BARC, Bombay, for providing facilities and Dr N. B. Gaddagimath, Department of Genetics and Plant Breeding, University of Agricultural Sciences, Dharwad, for materials. Financial assistance from UGC is also acknowledged.

Received 11 November 1992, revised accepted 28 May 1993

Homorepeats of amino acids in proteins – A database search

K. Veluraja* and S. Priyadarshini

Bioinformatics Centre, Department of Biotechnology, Madurai Kamaraj University, Madurai 625 021, India
*Present address: Physics Department, Manonmanian Sundaranar University, Tirunelveli 627 009, India

Occurrence of individual amino acids and their homorepeats has been computed from the available protein sequence database. Significant numbers of homorepeats, up to a maximum of hexamers, were found for most amino acids. Preliminary results indicate that the patterns of percentage occurrence of individual amino acids are similar to those of their homodimers. Higher orders, however, show deviations in the predominance profile. In the case of valine and isoleucine, the frequency of occurrence of homorepeats (beyond trimers) goes down considerably, in sharp contrast to their predominance as individual entities. Tryptophan was not found beyond homotrimeric repeats and tyrosine and cysteine were not found in tandems beyond pentamers. Interspecies distribution profiles reveal some interesting deviations. The cysteine content in *E. coli* proteins was about 50% lower compared to human proteins. Very few *E. coli* proteins have higher-order repeats and the functional importance of these higher-order repeats has been analysed. Plant proteins have very high glutamine tandems in contrast to intermediate frequency of single glutamine occurrences. This suggests a preferential selection of some amino acids and their tandems in the course of evolution to suit diverse functional and structural requirements.

PROTEIN structure is primarily a reflection of its amino acid sequence[1,2]. Automated protein-sequencing methods have generated a vast repertoire of protein sequences[1,3]. These data are made available in various computerized protein sequence databanks[4]. The number of proteins sequenced exceed manyfold those with defined three-dimensional structures solved by X-ray crystallography and NMR

methods. Exploitation of these databases for information regarding aspects of protein structure is an essential part of biological research[5–11]. In this area one aspect is the study of repetitive amino acid sequences in proteins. Highly repetitive sequences are believed to be of great importance in protein structure and function[12,13].

Here we have investigated the percentage occurrence of individual amino acids as well as their homorepeats of the proteins available in the SWISSPROT database[14]. Three sub-databases containing proteins of human, plant and *E. coli* origins were constructed and analysed for an interspecies comparison of the distribution of amino acids and their tandem repeats.

The database used was the SWISSPROT protein sequence database; Release 22 (May 1992). From this database a total of 21,868 proteins comprising 7,294,047 amino acids were selected as the parent database. The total number of proteins of human origin were 1807 (736,335 amino acids), of *E. coli* were 1727 (538,073 amino acids) and of plant origin were 2589 (710,407 amino acids).

Software was developed to compute the relative percentage of occurrence of amino acids as individuals and as their homorepeats, and evaluated as described below.

Percentage occurrence of individual amino acid

$$= \frac{\text{Total number of individual amino acid entries}}{\text{Total number of amino acid entries}} \times 100,$$

Percentage occurrence of amino acids in homorepeats

$$= \frac{\text{Total number of amino acids in repeats}}{\text{Total number of amino acid entries}} \times 100.$$

One trimer includes two dimers and three single amino acids. Similarly, one tetramer includes three dimers and two trimers and four amino acids and so on. The multiple entries of a few proteins in the database will not bias the results as the number of sequences subject to analysis is quite large. A CPU time of 8 h was taken for each computation of the parent database on microvax II.

Table 1 shows the individual amino acid distribution in overall protein entries tabulated on a scale of 0–100%. The profiles of distribution of each amino acid and their respective homorepeats are shown in Figure 1 *a–f*. This

Table 1. Distribution of amino acids in proteins

Amino acids	Percentage range
Leucine	9.18
Alanine, glycine, serine, valine, isoleucine, glutamate	6–7.5
Lysine, threonine, aspartate, arginine, proline	5–6
Asparagine, glutamine, phenylalanine, tyrosine	3–5
Histidine, methionine	2–3
Cystine, tryptophan	1–2

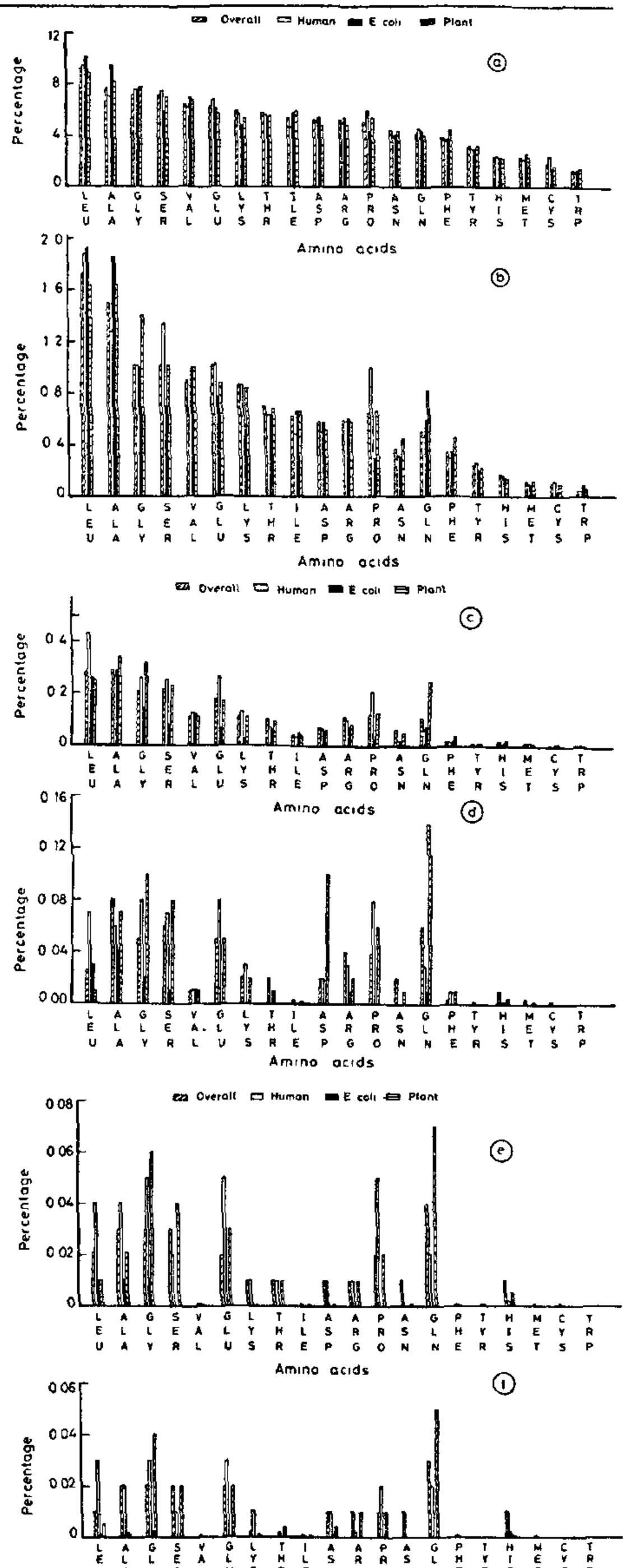

Figure 1. Distribution of amino acids and their homorepeats in overall, human, *E. coli* and plant proteins. *a*, Individual amino acids; *b*, Homodimers, *c*, Homotrimers; *d*, Homotetramers; *e*, Homopentamers; *f*, Homohexamers.

Table 2. Distribution of homorepeats

	Percentage			
	Overall	Human	Plant	*E. coli*
Dimers	13 610	13 730	14.080	12 690
Trimers	2.075	2.280	2.314	1.384
Tetramers	0.514	0.580	0.597	0.140
Pentamers	0.253	0.312	0.283	0 036
Hexamers	0.173	0.190	0 181	0 007

includes a species-wise comparison as well. There are 20 amino acids that contribute to protein structure. Hence, the probability of selection of each amino acid by nature should be 5% (1/20 × 100). However, observations reveal that the hydrophobic amino acids leucine, alanine, glycine, valine and isoleucine predominate, accounting together for 36.02%. This value is significantly higher than the probability of chance selection, which should be 25% for the five amino acids. Table 2 gives the total distribution of higher-order homorepeats in various species. An interesting feature is that *E. coli* proteins have very few higher-order homorepeats (see also Figure 1 *e*, *f*).

A comparison of the interspecies individual amino acid distribution profiles reveals some interesting deviations. *E. coli* proteins show much lower cysteine content compared to human proteins (1.13% in *E. coli* and 2.29% in human). This nearly 50% drop in occurrence can be due to reduced redox potentials in bacterial systems which do not allow extensive disulphide bond formation as in mammalian cells[15]. Hence, the need for cysteine would be less in *E. coli* as disulphide bonds are not the major protein-stabilizing forces here. The relative percentage of serine is also significantly less in *E. coli* (5.8% *E. coli*, 7.03% overall, 7.47% in human and 6.96% in plant). *E. coli* lack the necessary post-translational modification machinery needed for protein glycosylation, hence, the lower frequency of serine, which is one of the potential candidates for glycosylation, could be accounted for.

Figures 1*a* and *b* show that profiles of dimeric repeats have minor perturbations compared to profiles of single entities. For example, glutamine dimers occur at higher frequencies than asparagine (glutamine dimer 0.51%, asparagine dimer 0.47%) though as individual entities the percentage occurrence of asparagine is higher than glutamine (4.09% for glutamine monomers and 4.4% for asparagine monomers). Homotrimeric repeat patterns show higher percentage of alanine triplets compared to leucine in all proteins with the exception of human proteins (Figure 1 *c*). The percentages are: in overall proteins, alanine = 0.29, leucine = 0.28; in *E. coli* proteins, alanine = 0.29, leucine = 0.14; in plant proteins, alanine = 0.34, leucine = 0.25; in human proteins, alanine = 0.23, leucine = 0.43. Predominant leucine triplets in human proteins could have a structural role. The percentage occurrence of valine and isoleucine shows a marked drop as one moves to higher-order repeats, unlike other amino acids which predominated as individual entities. Thus, the higher-order homorepeats of valine and isoleucine may not have essential functional or structural roles in proteins.

Plant protein profiles distinctly show a high percentage of higher-order glutamine homorepeats. The side chain of glutamine acts as a nitrogen donor in a variety of biochemical reactions[16]. We further found that stretches of glutamine tandems predominate in plant seed storage proteins (unpublished observations by authors). We envisage that in addition to being a reserve of nitrogen in plant seed proteins these poly-glutamine domains could also be potential recognition sites for the enzymes of nitrogen metabolism. Another striking feature is the amino acid histidine, which is present as a monomer at a frequency of only 2.3% in overall proteins. However, a significant number of histidine homorepeats up to hexamers are found in all species (overall proteins 0.0002%, human proteins 0.002%, *E. coli* proteins 0.001% and plant proteins 0.0002%). So histidine homorepeats must have an essential structural role in the proteins of most organisms. Some amino acids do not appear as higher-order homorepeats. In the database scanned, tryptophan repeats beyond trimers were not found. Tyrosine and cysteine were not found beyond pentameric repeats.

A detailed analysis of the *E. coli* proteins with higher-order homorepeats (higher than trimers) reveals some interesting features. The higher-order repeats of histidine, glutamine, glutamate occur as multimeric repeats in only one protein each. A histidine septamer is found in the His operon leader peptide (attenuator). A stretch of penta-glutamine is found in the sbcc protein, which is involved in recombinogenic activity, and a hexamer of glutamate is present in the 30S ribosomal protein at the C-terminal end. The analysis of multiplets of leucine reveals that only nine proteins have pentamers or higher orders of leucine in a tandem array. Of these nine proteins, seven are membrane proteins revealing the contribution of leucine repeats to membrane anchorage. Three proteins were found to have valine higher-order repeats, two of these being membrane proteins and one a regulator of valine and other branched chain amino acid biosynthetic pathways. Though glycine, alanine, arginine and threonine homorepeats are found as higher orders in less than ten proteins each, their distribution with respect to function of proteins is random. It is not possible to correlate the lower-order repeats to functions assumed by proteins due to their abundant occurrence.

Occurrence of significant numbers of homorepeats in proteins indicates that they should have important roles in protein structure and function. The frequency of higher-order homorepeats in *E. coli* proteins is markedly lower

compared to the other higher evolved species. The preferential selection of homorepeats of certain amino acids in proteins of different species is a reflection of the mechanisms to achieve functional and structural versatility by proteins in the course of evolution. As the database analysed is large, we believe that the percentage frequency results will not be biased by multiple entries of single proteins. A similar trend of homorepeat frequency can be expected of more exhaustive forthcoming data as the number of sampling points analysed is quite large.

1. Cantor and Schummel, in *Biophysical Chemistry*, W. H. Freeman and Co., 1980, Part I, pp. 41–154.
2. Doolittle, R. F., *Sci. Am.*, 1985, **253**, 88–99.
3. Doolittle, R. F., *Methods Enzymol.*, 1990, **183**, 99–110.
4. Borivoj Keil, *Methods Enzymol.*, 1990, **183**, 50–60.
5. Zamyatnin, A. A., *Protein Seq. Data Anal.*, 1991, **4**, 57–60.
6. Kolaskar, A. S. and Samuel, S. L., *Protein Seq. Data Anal.*, 1991, **4**, 105–110.
7. Argos, P., *Nucleic Acids Res.*, 1988, **16**, 9909–9916.
8. Kolaskar, A. S. and Kulkarni-Kale, U., *J. Mol. Biol.*, 1992, **223**, 1053–1061.
9. Jones, T. A. and Thirup, S., *EMBO J.*, 1986, **5**, 819–822.
10. Gibrat, G. J., *J. Mol. Biol.*, 1987, **198**, 425–443.
11. Blundell, T. M., *Trends Biochem Sci.*, 1990, **15**, 425–430.
12. McLachlan, A. D., *Symp Quant. Biol.*, 1987, **II**, 411.
13. Shin, M. S., Bargiello, T. A., Clark, B. T., Jackson, F. R. and Young, M. W., *Nature*, 1985, **317**, 445–448.
14. Bairoche, A. and Boeckmann, B., *Nucleic Acids Res.*, **20**, 2019–2022.
15. Pain, P., *Trends Biochem. Sci.*, 1987, **12**, 309–312.
16. Higgins, T. J. V., *Annu. Rev. Plant Physiol.*, 1984, **35**, 191–221.

Received 9 March 1993; revised accepted 28 May 1993

γ-Rays- and EMS-induced leaf mutants in mung bean (*Vigna radiata* (L) Wilczek)

V. P. Singh and Rashmi Sharma

Department of Plant Sciences, Rohilkhand University, Bareilly 243 005, India

A few pentafoliate and tetrafoliate mutants were isolated from the γ-rays and EMS-treated M_2 population. These mutants showed a significant increase in dry matter production, total chlorophyll contents and yield compared to their parents in M_2 and M_3 generations.

THE role of mutation breeding in the induction of leaf mutants of agronomic interest is well established[1]. Being easily discernible and stable phenotypes, the leaf mutants offer interesting experimental material. In a short-duration crop such as mung bean, it is important that the leaf area should expand and reach its optimal level as rapidly as possible for maximum interception of the incident light. Earlier studies indicate that the number of pods per plant, reduction in leaf number per plant and leaf area, and insufficient dry-matter production are the principal factors limiting the yield[1,2]. The present report describes an attempt to study the morphological and physiological components of yield in the tetrafoliate and pentafoliate mutants induced in mung bean cv. PDM-116 and PDM-11.

Dry seeds of mung bean (*Vigna radiata* (L) Wilczek) cv. PDM-116 and PDM-11 obtained from Pulse Directorate, Kalayanpur, Kanpur, were irradiated with different doses of γ-rays (15, 30 and 45 kR) at the Indian Agricultural Research Institute, New Delhi, delivered from a source of ^{60}Co and sown in the field. In another experiment presoaked seeds (12 h in distilled water) were treated with an aqueous solution of chemical mutagen (0.1%, 0.2%, 0.3% ethylmethanesulphonate) for 6 h with intermittent shaking of the mutagenic solution. After the termination of chemical treatments, the seeds were washed in running water and directly sown in the field. Seeds from each M_1 plant were collected and sown in the field in randomized-block single-row design to raise M_2 generation. The mutants isolated from M_2 generation were carried over to M_3 generation to study their breeding behaviour and productivity. The protein content was estimated following the modified Kjeldhal's method[3] and the chlorophyll contents were estimated following Arnon's method[4].

In M_2 generation, 1.66% and 2.50% tetrafoliate mutants in variety PDM-116 and PDM-11, respectively, and 0.83% pentafoliate mutants in both these varieties were isolated from the mutagen-treated population. The highest frequency of induced mutants was reported in γ-rays-treated population of both the varieties. It was interesting to note that the tetrafoliate mutants were recorded from γ-rays-treated population and pentafoliate mutants were recorded from EMS-treated population only.

The leaf characteristics of the mutants and their productivity are given in Tables 1 and 2, respectively. In M_2 and M_3 generation, the leaf area increases significantly in both the mutants along with the dry-matter production and total chlorophyll contents per plant. The total yield in the induced mutants was significantly higher, the high-yield contributing factor being the number of pods per plant. The protein contents remain unaltered in tetrafoliate and pentafoliate mutants except in the pentafoliate mutant isolated from variety PDM-116, where a significant increase was observed in the protein content and the yield. A similar result of the induction of desirable leaf mutants by the use of various physical/chemical mutagen in pulses has been reported earlier[2,5–7]. The M_3 segregation population of the mutants showed a 3:1 segregating ratio, confirming that the mutant character is controlled by a single recessive gene (Table 3).