Parity in the number of atoms in residue composition in proteins and contact preferences

Rudra Prasad Saha and Pinak Chakrabarti*

Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Kolkata 700 054, India

Two ways of looking into the amino acid composition in a database – one based on counting residues and another on the number of atoms each residue contributes to the total – give quite different results. In the former, the composition is inversely proportional to the residue size – the maximum deviation shown by Cys and Leu can be explained by their chemical property or the role in oligomeric structure, while the latter is size-invariant. Similarly, calculations of contact preferences between residues can be atom-based or residue-based, and the former method provides values that are size-independent, and thus should be preferred in modelling and docking studies.

Keywords: Amino acid composition, contact preferences, protein modelling, residue environment, residue size.

THE composition of DNA bases, like the GC content, has important implications in DNA sequence and genome analysis[1]. Some proteins, such as collagen[2,3], can be made up of only a few amino acids or have a skewed distribution of residues. Local stretches of polypeptide chain can be enriched in a residue of a particular type[4] or there could be internal duplications[5] giving rise to structural motifs, such as leucine-rich repeats. The amino acid composition of a protein has a bearing on the secondary structural content and the fold that it acquires[6,7]. Overall, however, in a non-redundant protein database created from a primary database – sequence[8] or structural[9] – the residue composition does not show much variation. Here we show that the amino acid composition in a protein database can be looked at from two perspectives – counting residues or atoms – the former giving values that show an approximate inverse relationship with the residue size, indicating the pressure that evolution has exerted on the size of individual residues and frequencies of their incorporation into protein molecules, while the latter does not exhibit such a dependence. This has implications in the derivation of different parameters on protein structures. For example, in the calculation of expected frequencies of interactions between residues, one needs to use the abundances of the amino acids, which should be atom-based to take into account the size difference between residues.

The dataset consisted of protein tertiary structures (a total of 555 chains in 531 files) selected using PDB_SELECT[10] (with conditions: R-factor ≤ 20%, resolution of ≤ 2.0 Å and sequence identity less than 25%) from the Protein Data Bank (PDB, as on April 2002) at the Research Collaboratory for Structural Bioinformatics (RCSB)[9]. The propensity[11] of a residue X to be in the environment of Y (the central residue) is

$$P_{XY} = \frac{N_{XY} / \sum_{x=1}^{20} N_{XY}}{N_x / \sum_{x=1}^{20} N_X}, \quad (1)$$

where N_{XY} is the number of atoms of residue X found close (within 4.5 Å) to atoms of residue Y, N_X is the total number of atoms contained in residue type X in the entire database and summations are over all the 20 residue types. The threshold value of distance was found suitable in earlier studies[12]. The numerator is the fractional occurrence of atoms of residue X around residue Y and the denominator represents the fraction of atoms belonging to residue type X in the whole dataset. P_{XY} 1.0 indicates a neutral preference of X to be in the environment of Y; a value > 1.0 implies preference to associate; < 1.0, to avoid.

Equation (1) involves counting of atoms. Another set of calculations was done in which instead of considering the constituent atoms, the counting was residue-based, i.e. N_{XY} is the number of residues of type X with at least one atom in contact with a residue of type Y.

The amino acid composition of protein structures is assumed to be such as to give rise to a stable native fold, with a rather hydrophobic core and a high proportion of polar residues in the exterior, with functionally important residues lining up the active site. To gain insight into the protein folding problem, the known structures are analysed from different angles[13] and for this a non-redundant dataset is normally used[10]. The percentage composition of residues in such a database, presented in Figure 1 *a*, indicates that there is a remarkable dependence on the size. The largest residue, Trp has a small occurrence and the two smallest residues Gly and Ala, the highest, the composition being inversely related to the accessible surface area of the residue (X) in a model tripeptide, Gly-X-Gly[14], representing its size. The correlation coefficient, which is –0.53, becomes –0.76 when Cys and Leu are excluded. It is interesting why these two residues do not follow the pattern. Because of the susceptibility to oxidation, nature has reasons to control the amount of Cys, usually limited to disulphide linkages and active sites of enzymes (such as iron–sulphur clusters). Leu, on the other hand, has a high propensity to be in α-helix[13], can be in the core because of its hydrophobic nature, but it also contributes the highest (10.4%) to the interface area of homodimeric proteins[15] – its 'stickiness' (the propensity of Leu residues to interact with each other) being responsible for its widespread use in motifs, such as leucine zippers[16] – and is

*For correspondence. (e-mail: pinak@boseinst.ernet.in)

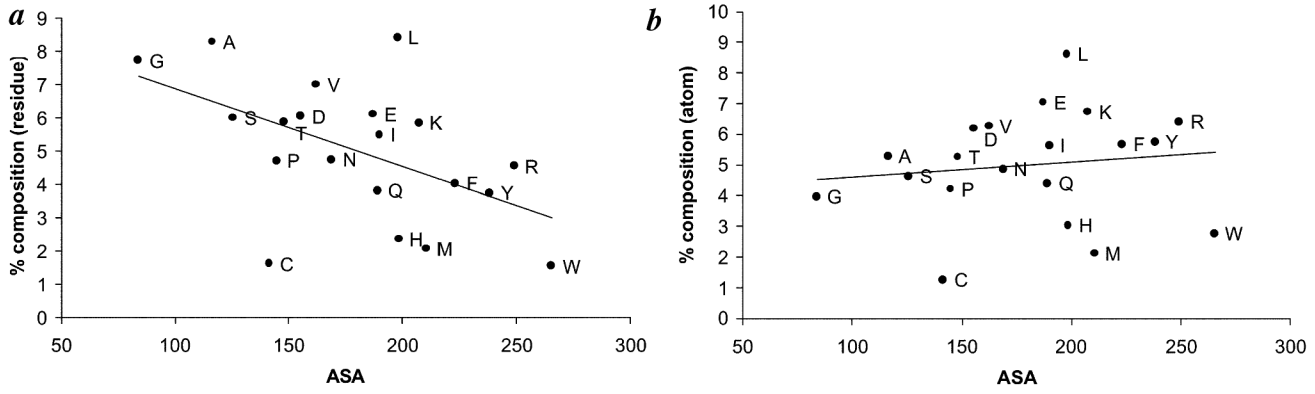

Figure 1. Plot of accessible surface area (ASA; Å^2) of residues (*X*, in a tripeptide Gly-*X*-Gly in an extended conformation) against amino acid composition based on the count of (***a***) residues and (***b***) atoms in each residue. One-letter amino acid code is shown against the points and equation of the least-squares line (excluding Cys and Leu) in (***a***) is $y = -0.03x + 10.35$.

therefore found in excess. Essentially the same relationship is obtained if the residue volume[17] or the number of constituent atoms in a residue replaces the surface area.

Compositions derived from a different dataset of PDB files, also calculated intra-molecularly and for residues located in inter-molecular interfaces[18], match with our values (with correlation coefficients of 0.75 and 0.84 respectively), showing consistency between datasets. There can be some differences in the amino acid composition depending on the intracellular or extracellular nature of the protein[19]. Nevertheless, the same trend is maintained against the size, correlation coefficients, excluding Cys and Leu, being –0.62 and –0.79 respectively, for the two categories of proteins. Likewise, the inverse relationship holds when the compositions of individual mesophilic organisms[20] are used; for example, *E. coli* giving a value of –0.70. However, for the two thermophilic organisms[20], because of the higher content of the charged residues and a discrimination against a residue like Gln, the correlation coefficient is only –0.28 (not considering Cys and Leu), which becomes –0.48 (when Glu and Lys are also excluded) for *Aquifex aeolicus*.

If the composition is based on the number of atoms each residue contributes to the total number of protein atoms, the result is quite different (Figure 1 *b*). The distribution of atom-based composition is narrower (the average and standard deviations are 5 and 1.8 respectively) than the residue-based composition (standard deviation of 2.1), indicating that the composition is such that each residue contributes nearly equally to the atom pool of protein structures. That this composition is independent of the size of the residues is indicated by a small correlation coefficient of 0.13. Cys and Leu, which showed maximum deviation from the overall trend in the residue-based composition, are also the ones with the minimum and maximum values respectively, in the atom-based composition.

As different residues are closer to each other in atom-based composition, an organism making all the twenty amino acids makes a more balanced distribution of atom sources amongst different residues.

Whether or not composition based on residue or atom is used may have important ramification in the calculation of residue–residue pair potentials[18,21-23] or the propensities of residues to be in the environment of another residue, for example Trp[24]. Assuming the residues to have no specific preference for each other, the likelihood of contact between any two residues would depend on the number of constituent atoms (i.e. the size). The composition of the residues is taken into account to get a normalized value for the propensity of interaction (eq. (1)). Two sets of propensity values have been calculated – atom- and residue-based – and are provided in Table 1. If one examines the propensities of the smallest (Gly) and the largest (Trp) of all the residues to be in the environment of all the twenty amino acid residues, the values are rather uniformly small (under the column 'Gly', Table 1 *b*) and large (under 'Trp') respectively. This is caused by both the factors, size and frequency of occurrence (from the residue-based composition), which are at the two extreme ends for the two residues, reinforcing each other and thereby diminishing any discriminatory capacity in the calculated values. To circumvent this, the number of residue–residue contacts needs to be normalized by residue volume/size[25]. The same result should also be achieved if the calculation is atom-based (counting the number of atoms of different residues involved in interaction and using atom-based composition), as this allows an implicit consideration of the size differences between residues. Thus in Figure 2, in which the propensities of residues to be in the environment of Trp are plotted, the residue-based values show an approximate linear dependence (which would improve further if Arg, Lys and Glu are excluded), while no such dependence can be seen for the atom-based values. Though in both the sets, the propensities of other aromatic residues (Phe, Tyr and His) to interact with Trp are high, indicating the preference of aromatic residues to associate among themselves, the effect

Table 1. Propensities of different residues (in columns) to be in the environment of any given residue (row) in pr otein structures

RES	Gly	Ala	Pro	Cys	Met	Leu	Val	Ile	Phe	Tyr	Trp	His	Arg	Lys	Asp	Glu	Asn	Gln	Ser	Thr
(*a*) Atom-based																				
Gly	1.15	1.28	0.98	1.32	0.97	0.93	1.11	1.03	1.00	1.04	1.05	0.99	0.85	0.78	0.97	0.72	1.03	0.87	1.23	1.19
Ala	1.07	1.61	0.81	1.10	1.13	1.16	1.27	1.18	0.99	0.89	0.88	0.85	0.82	0.75	0.83	0.82	0.86	0.90	1.08	1.03
Pro	1.01	1.00	0.89	1.16	1.03	0.98	1.10	0.99	1.14	1.28	1.43	1.16	0.89	0.66	0.92	0.80	1.04	0.99	1.04	1.05
Cys	1.06	1.05	0.92	8.76	1.05	0.98	1.04	0.95	1.19	0.99	0.94	1.08	0.70	0.59	0.70	0.58	0.84	0.81	1.07	1.00
Met	0.94	1.23	0.84	1.15	1.67	1.19	1.24	1.22	1.39	1.14	1.13	1.03	0.81	0.64	0.72	0.73	0.80	0.87	0.92	0.90
Leu	0.89	1.28	0.80	1.11	1.24	1.38	1.26	1.31	1.27	1.01	1.08	0.88	0.84	0.72	0.71	0.73	0.77	0.88	0.90	0.95
Val	0.96	1.31	0.86	1.15	1.23	1.26	1.47	1.35	1.28	1.00	0.98	0.93	0.72	0.68	0.71	0.69	0.80	0.77	0.94	1.01
Ile	0.90	1.24	0.79	1.06	1.28	1.32	1.38	1.41	1.24	1.06	0.98	0.90	0.78	0.68	0.71	0.71	0.80	0.78	0.94	1.02
Phe	0.94	1.13	0.89	1.28	1.38	1.23	1.25	1.20	1.50	1.15	1.27	1.00	0.78	0.67	0.69	0.66	0.79	0.80	0.97	0.95
Tyr	1.04	1.03	1.07	1.18	1.18	1.06	1.04	1.08	1.16	1.08	1.15	1.06	0.95	0.89	0.92	0.76	0.90	0.94	1.01	0.93
Trp	1.05	1.03	1.05	1.08	1.14	1.10	1.01	0.99	1.31	1.19	1.30	1.10	0.94	0.69	0.81	0.71	1.01	1.02	1.03	0.94
His	0.98	1.00	0.98	1.14	1.15	0.96	1.04	0.98	1.06	1.09	1.16	1.70	0.86	0.64	1.09	0.90	0.92	0.86	1.12	1.08
Arg	0.98	1.15	0.88	0.92	1.04	1.02	0.93	1.00	0.96	1.10	1.17	0.99	0.83	0.61	1.31	1.24	0.92	0.96	1.03	0.96
Lys	0.93	1.11	0.77	0.91	0.89	0.99	1.02	0.98	0.93	1.17	0.97	0.81	0.67	0.76	1.35	1.40	1.00	0.97	1.02	0.99
Asp	0.99	1.07	0.91	0.95	0.89	0.87	0.96	0.92	0.81	0.96	0.88	1.22	1.35	1.23	0.85	0.66	1.16	0.94	1.25	1.20
Glu	0.85	1.20	0.84	0.89	1.01	0.99	0.99	1.00	0.83	0.91	0.83	1.12	1.47	1.33	0.74	0.72	0.99	0.91	1.06	1.12
Asn	1.05	1.06	0.98	1.14	0.94	0.91	0.98	0.97	0.92	0.97	1.12	0.96	0.86	0.84	1.13	0.84	1.27	1.03	1.16	1.20
Gln	0.91	1.18	0.93	1.16	1.05	1.06	1.01	0.98	0.92	1.07	1.20	0.96	0.93	0.85	0.95	0.82	1.09	1.05	1.08	1.13
Ser	1.07	1.15	0.95	1.29	0.97	0.95	1.06	1.06	0.99	0.98	0.99	1.08	0.85	0.77	1.11	0.82	1.06	0.94	1.22	1.11
Thr	1.04	1.13	0.93	1.20	0.99	1.02	1.13	1.14	0.97	0.92	0.94	1.08	0.82	0.76	1.00	0.86	1.08	0.96	1.09	1.22
(*b*) Residue-based																				
Gly	0.69	1.00	0.93	1.19	1.09	1.04	1.03	1.08	1.22	1.29	1.39	1.10	1.08	0.91	0.92	0.82	1.00	0.96	1.01	1.05
Ala	0.63	1.21	0.76	0.95	1.23	1.27	1.21	1.25	1.25	1.14	1.25	0.95	1.04	0.87	0.78	0.89	0.83	0.96	0.85	0.89
Pro	0.65	0.86	0.89	1.15	1.15	1.12	1.10	1.08	1.31	1.49	1.61	1.19	1.11	0.80	0.89	0.88	1.00	1.03	0.90	0.98
Cys	0.64	0.82	0.90	5.95	1.09	1.08	1.04	1.06	1.35	1.28	1.33	1.11	0.94	0.72	0.69	0.68	0.84	0.91	0.90	0.91
Met	0.57	0.99	0.82	1.01	1.72	1.32	1.24	1.35	1.55	1.39	1.48	1.09	1.00	0.74	0.71	0.80	0.78	0.91	0.77	0.83
Leu	0.53	1.00	0.77	0.98	1.32	1.53	1.29	1.46	1.49	1.21	1.39	0.93	1.01	0.81	0.70	0.79	0.75	0.90	0.74	0.87
Val	0.56	1.02	0.81	1.02	1.31	1.39	1.43	1.48	1.49	1.19	1.30	0.98	0.89	0.79	0.72	0.75	0.78	0.83	0.77	0.91
Ile	0.53	0.98	0.75	0.95	1.32	1.45	1.38	1.55	1.47	1.26	1.28	0.94	0.96	0.77	0.71	0.76	0.77	0.80	0.78	0.93
Phe	0.56	0.91	0.84	1.11	1.42	1.38	1.27	1.36	1.73	1.38	1.61	1.05	0.92	0.76	0.71	0.73	0.78	0.84	0.82	0.89
Tyr	0.63	0.85	0.98	1.10	1.27	1.16	1.06	1.21	1.42	1.36	1.60	1.13	1.11	0.91	0.92	0.83	0.89	0.97	0.87	0.88
Trp	0.65	0.86	0.97	1.02	1.24	1.22	1.05	1.13	1.53	1.46	1.82	1.14	1.06	0.76	0.83	0.78	0.97	1.04	0.86	0.90
His	0.62	0.85	0.93	1.17	1.25	1.07	1.04	1.09	1.28	1.36	1.52	1.81	1.06	0.72	1.01	0.93	0.89	0.96	0.94	0.98
Arg	0.64	0.93	0.88	0.92	1.14	1.15	0.95	1.11	1.16	1.34	1.41	1.08	1.06	0.74	1.15	1.24	0.91	1.03	0.87	0.90
Lys	0.60	0.93	0.77	0.87	1.01	1.14	1.02	1.08	1.15	1.35	1.25	0.88	0.91	0.95	1.20	1.37	0.97	1.01	0.86	0.92
Asp	0.64	0.89	0.89	0.90	1.02	1.01	0.97	1.03	1.12	1.40	1.39	1.29	1.46	1.26	0.81	0.79	1.06	1.00	0.96	0.97
Glu	0.55	0.98	0.84	0.84	1.16	1.12	0.99	1.08	1.13	1.24	1.28	1.17	1.56	1.35	0.75	0.86	0.93	0.99	0.84	0.96
Asn	0.66	0.89	0.94	1.04	1.05	1.03	0.99	1.06	1.16	1.27	1.51	1.06	1.08	0.95	1.00	0.91	1.19	1.09	0.95	1.04
Gln	0.59	0.96	0.91	1.06	1.14	1.15	1.00	1.05	1.16	1.34	1.57	1.08	1.17	0.92	0.89	0.89	1.00	1.17	0.90	0.99
Ser	0.68	0.94	0.90	1.16	1.07	1.06	1.03	1.12	1.26	1.31	1.44	1.18	1.10	0.88	0.96	0.86	0.98	1.02	0.99	0.97
Thr	0.64	0.92	0.88	1.10	1.08	1.13	1.11	1.22	1.27	1.21	1.36	1.14	1.03	0.87	0.89	0.90	0.99	1.00	0.89	1.07

The ordering of residues is as follows: Gly (no side chain), Ala (no side -chain torsion angle), Pro (restricted torsion angles), Cys and Met (S -containing residues), L eu, Val and Ile (aliphatic branched side chains), Phe, Tyr, Trp and His (aromatics, the last one can also be charged), Arg, Lys, Asp and Glu (charged), and Asn, Gln, Ser and Thr (neutral polar).

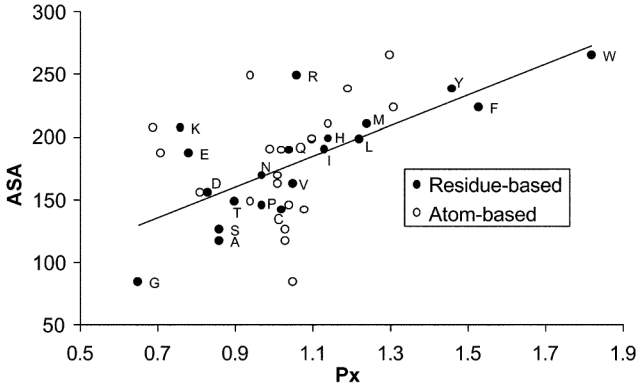

Figure 2. Plot of P_X, propensities of residues to be in the environment of Trp against ASA (defined in Figure 1). Filled circles represent values obtained from residue-based calculation and open circles are atom-based. For the former the equation of the least squares line is ASA = 122.24 P_X + 50.07 (with r^2 = 0.56).

of the residue size also gets reflected in the set of values given in Table 1 *b*. Thus atom-based calculations are likely to provide a more accurate estimate of the contact preferences – unbiased of the size – between residues.

In conclusion, conventionally defined composition of residues, quite routinely used in the discussion of protein structures, has an approximate inverse relationship with the residue size. This dependence on size, rather than on any other physico-chemical property, suggests that the size of residues has been an important factor in determining their level of incorporation into polypeptide chains during evolution. The atom-based composition does not show size dependence, and when used for calculation of interaction propensities between residues, provides values that reflect inherent preferences independent of size. Fold recognition programs[18] and protein–protein docking algorithms[22,25], which rely on the contact preferences, would benefit from such a consideration.

1. Bernardi, G., *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution*, Elsevier, Amsterdam, 2004.
2. Ramachandran, G. N., Structure of collagen at the molecular level. In *Treatise on Collagen* (ed. Ramachandran, G. N.), Academic Press, New York, 1967, pp. 103–183.
3. Bella, J., Eaton, M., Brodsky, B. and Berman, H. M., Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science*, 1994, **266**, 75–81.
4. Meyer, E. F. and Tollett Jr. W. J., WWWWhy does nature stutter? A survey of strands of repeated amino acids. *Acta Crystallogr. Sect. D*, 2001, **57**, 181–186.
5. Heringa, J., Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.*, 1998, **8**, 338–345.
6. Eisenhaber, F., Frömmel, C. and Argos, P., Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins*, 1996, **25**, 169–179.
7. Dubchak, I., Holbrook, S. R. and Kim, S. H., Prediction of protein folding class from amino acid composition. *Proteins*, 1993, **16**, 79–91.
8. Apweiler, R., Protein sequence databases. *Adv. Prot. Chem.*, 2000, **54**, 31–71.
9. Berman, H. M. *et al.*, The protein Data Bank. *Nucleic Acids Res.*, 2000, **28**, 235–242.
10. Hobohm, U. and Sander, C., Enlarged representative set protein structures. *Protein Sci.*, 1994, **3**, 522–524.
11. Narayana, S. V. L. and Argos, P., Residue contacts in protein structures and implications for protein folding. *Int. J. Pept. Protein Res.*, 1984, **24**, 25–39.
12. Bhattacharyya, R. and Chakrabarti, P., Stereospecific interactions of proline residues in protein structures and complexes. *J. Mol. Biol.*, 2003, **331**, 925–940.
13. Chakrabarti, P. and Pal, D., The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.*, 2001, **76**, 1–102.
14. Samanta, U., Bahadur, R. P. and Chakrabarti, P., Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng.*, 2002, **15**, 659–667.
15. Bahadur, R. P., Chakrabarti, P., Rodier, F. and Janin, J., Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 2003, **53**, 708–719.
16. Saha, R. P., Bahadur, R. P. and Chakrabarti, P., Inter-residue contacts in proteins and protein–protein interfaces and their use in characterizing the homodimeric interface. *J. Proteome Res.*, 2005, **4**, 1600–1609.
17. Pontius, J., Richelle, J. and Wodak, S. J., Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, 1996, **264**, 121–136.
18. Keskin, O., Bahar, I., Badretdinov, A. Y., Ptitsyn, O. B. and Jernigan, R. L., Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.*, 1998, **7**, 2578–2586.
19. Nakashima, H. and Nishikawa, K., Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, 1994, **238**, 54–61.
20. Deckert, G. *et al.*, The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, 1998, **392**, 353–358.
21. Russell, R. B. and Barton, G. J., Structural features can be uncovered in proteins with similar folds. *J. Mol. Biol.*, 1994, **244**, 332–350.
22. Moont, G., Gabb, H. A. and Sternberg, M. J. E., Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, 1999, **35**, 364–373.
23. Ofran, Y. and Rost, B., Analysing six types of protein–protein interfaces. *J. Mol. Biol.*, 2003, **325**, 377–387.
24. Samanta, U., Pal, D. and Chakrabarti, P., Environment of tryptophan side chains in proteins. *Proteins*, 2000, **38**, 288–300.
25. Glaser, F., Steinberg, D. M., Vakser, I. A. and Ben-Tal, N., Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, 2001, **43**, 89–102.

ACKNOWLEDGEMENTS. We thank the Council of Scientific and Industrial Research, New Delhi for providing fellowship to R.P.S. and a research grant to P.C.

Received 29 July 2005; accepted 18 October 2005