Chaos game representation of proteins

Soumalee Basu, Archana Pan, Chitra Dutta, and Jyotirmoy Das

Biophysics Division, Indian Institute of Chemical Biology, Calcutta, India

The present report proposes a new method for the chaos game representation (CGR) of different families of proteins. Using concatenated amino acid sequences of proteins belonging to a particular family and a 12-sided regular polygon, each vertex of which represents a group of amino acid residues leading to conservative substitutions, the method can generate the CGR of the family and allows pictorial representation of the pattern characterizing the family. An estimation of the percentages of points plotted in different segments of the CGR (grid points) allows quantification of the nonrandomness of the CGR patterns generated. The CGRs of different protein families exhibited distinct visually identifiable patterns. This implies that different functional classes of proteins follow specific statistical biases in the distribution of different mono-, di-, tri-, or higher order peptides along their primary sequences. The potential of grid counts as the discriminative and diagnostic signature of a family of proteins is discussed. © 1998 by Elsevier Science Inc.

Keywords: chaos game, pictorial representation, protein family, conservative substitution, visual patterns

INTRODUCTION

Using the chaos game algorithm of nonlinear dynamics,[1] it has been possible to represent nucleotide sequences pictorially,[2,3] by considering the sequences as strings composed of the four units G, A, C, and T/U. This technique, known as chaos game representation (CGR) of DNA/RNA sequences, can generate fractal structures[3] and has shown promise in recognizing the underlying patterns or bias in the selection of nucleotides in DNA sequences of genes, the products of which perform similar or identical functions. The usefulness or limits of CGR were further established by the mathematical characterization of the technique[4] and it has been shown that variations in the bias of distribution of di-, tri-, or higher order nucleotides along the DNA/RNA sequences can generate distinct patterns in the CGR, which can be used as diagnostic patterns for different families of genes.[4,5] The major advantage of this technique is that it does not require prior knowledge of 'consensus' sequences, nor does it involve exhaustive searches for sequences in databases. This approach takes advantage of the uneven distribution of subsequences of length k ($k = 1, 2, 3\ldots$) along the sequence.

The technique of the CGR has been generalized and applied to analyze both the primary and secondary structures of protein sequences.[6] Using a regular 20-sided polygon, the primary sequences of all the proteins in the database have been plotted together irrespective of their functions or origins, to look for the frequencies of occurrences of special sequence motifs in the whole protein database. Such a generalized application, however, suffers from two serious limitations. First, to visualise the characteristic CGR patterns of different protein families, one should have sorted out the amino acid sequences of the members of individual families and plotted them separately in different polygons instead of plotting together a random sampling of proteins of different functions and origins in a single CGR.[6] Second, in the homologous protein sequences belonging to a particular family, the amino acid residues in different positions are often replaced by their conservative substitutions, keeping their functions invariant. A 20-vertex CGR, in which each of the amino acid residues is plotted separately, cannot be used to differentiate between similar and dissimilar residues and hence it would exhibit similar patterns only for the sequences with identical residues in most of their respective positions, but not for sequences having conservative substitutions at the key positions.

To exploit the immense potential of the technique of CGR in generating visually identifiable distinct patterns for the amino acid sequences of proteins belonging to different functional classes, it was, therefore, necessary to modify it in a way to exhibit similar patterns for homologous amino acid sequences having appreciable conservative substitutions between them. It is in this context that the present article proposes a new algorithm for generating the CGRs of different families of proteins using a 12-sided regular polygon, each vertex of which represents a group of amino acid residues leading to conservative substitutions. An attempt has also been made to quantify the distinct bias in the CGR patterns of different protein families.

PROBLEMS IN EXTENDING THE TECHNIQUE OF CGR TO INDIVIDUAL PROTEIN FAMILIES, AND THEIR SOLUTIONS

A major difficulty in extending the CGR technique to individual protein sequences concerns the minimum number of resi-

Address reprint requests to: Chitra Dutta, Biophysics Division, Indian Institute of Chemical Biology, 4 Raja S.C. Mullick Road, Calcutta 700032, India; e-mail: ⟨iichbio@giasc101.vsnl.net.in⟩; ⟨biophy@cal.vsnl.net.in⟩.

Journal of Molecular Graphics and Modelling 15, 279–289, 1997
© 1998 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/97/$17.00
PII S1093-3263(97)00106-X

dues that might be required to generate identifiable patterns. It has been reported that for nucleotide sequences at least 2 000 bases are required to generate patterns.[2] Because the number of amino acids in the majority of proteins is much less than this threshold value, in the present study the amino acid sequences of proteins belonging to a particular family have been concatenated so that the number of residues available for CGR analysis exceeds 2 000. It has been verified that the characteristic pattern of a family of proteins does not depend on the number of sequences added together, or on the order in which they are concatenated (data not shown). Furthermore, once the pattern is generated, adding more protein sequences of the family does not alter the pattern.

The second problem of applying the present formalism of CGR to study the characteristic patterns of different families was that it could not be used to distinguish between groups of similar and dissimilar amino acid residues. For example, most of the existing protein sequence alignment programs will recognize the peptide IDEAL as being similar to MEEGL, but not to SREYL. In the former, isoleucine, aspartic acid, and alanine residues of the peptide have been replaced by their conservative substitutions[7] methionine, glutamic acid, and glycine, respectively, but in the latter, they were replaced by the dissimilar residues serine, arginine, and tyrosine. Following the same logic, SREYL is similar to TKEFL. If the technique of CGR is to be applied to differentiate between homologous and nonhomologous amino acid sequences, it must be modified in such a way that the two pentapeptides IDEAL and MEEGL would be plotted in an identical point, which would be different from the point representing SREYL or TKEFL. To achieve this, in the present study the number of vertices in protein CGR has been reduced to 12 (Figure 1) by grouping together similar amino acid residues, which can be conservatively substituted.[7] For example, alanine (A) and glycine (G), being conservative substitutions, are considered as one vertex; serine (S) and threonine (T) represent a vertex; isoleucine (I), leucine (L), valine (V), and methionine (M) represent one vertex; and so on. Using this 12-vertex CGR, we have been able to generate visually identifiable patterns for amino acid sequences belonging to different protein families. The number of vertices, however, could be reduced further by grouping more residues together. For example, proline (P) could be grouped together with S and T or histidine (H) and glutamine (Q) could be represented together. But the 12-vertex representation of protein CGR has been found to be optimum for generation of distinct patterns for different protein families. Any further reduction in the number of vertices reduces the resolution of CGR patterns to a great extent, whereas a CGR with an increased number of vertices often fails to generate similar patterns for proteins having significant sequence homology.

ALGORITHM FOR GENERATION OF THE CGR OF A PROTEIN FAMILY

Using a 12-sided regular polygon and concatenated amino acid sequences of the members of any particular protein family, an algorithm for generation of the CGR pattern of the family has been developed. Each vertex of the polygon is assigned to a particular group of conservatively substitutable amino acid residues (Figure 1). Following the chaos game algorithm,[2] the first amino acid residue of the concatenated protein sequence is plotted halfway between the center of the polygon and the vertex labelled with the code of the first residue. For example, to obtain the CGR of the first seven residues MASETFE[8] in the sequence of yeast heat shock protein 90 (Hsp90), the first residue must be plotted halfway between the center of the polygon and the vertex labelled ILVM (Figure 1a). The second residue (A) in the sequence is then plotted halfway between the first point and the vertex labelled with the code of the second residue (Fig. 1b). The process must be repeated until the last residue in the sequence is plotted (Fig. 1c).

Figure 1d represents the CGR of the concatenated sequence of the Hsp90 family of proteins. The distribution of points in the CGR, as seen in Figure 1d, is not random. It exhibits clustering of points along the vertices joining ILVM–RK and RK–DE vertices. This implies that in this class of proteins, the frequencies of occurrence of dipeptides *mn* (where $m, n =$ I, L, V, M, R, K, D, and E) are much higher than those of other dipeptides. This is in agreement with the fact that these Hsps contain clusters of charged residues along their sequences.[9] The protein CGR algorithm thus allows pictorial representation of patterns in amino acid sequences of proteins belonging to any family. However, for proper interpretation of such patterns, a mathematical characterization of the CGRs of proteins will be more useful.

GRID-COUNTING ALGORITHM

As demonstrated in the case of nucleotide CGR, the visually identifiable patterns exhibited in protein CGR also reveal the bias in the distribution of di-, tri-, or higher order peptides in the sequence. With a view to creating a quantitative estimation of such bias, a grid-counting algorithm, which determines the density of points in any region of the CGR, has been developed as follows.

The 12-sided polygon is divided into 24 segments (grid) as shown in Figure 1e and the segments are labelled serially with numbers 1–24 (not shown in Figure 1). For each segment, a counter (C_j for the *j*th segment) is set and initialized to zero. As each point is plotted within the CGR, the countervalue of the segment, in which the point lies, will increase by 1, keeping all other countervalues unchanged. For example, if a point is plotted in the *k*th segment, then $C_k = C_k + 1$ and $C_j = C_j$ ($j \neq k$). The counts for the points falling on the boundaries of different segments must be included in the countervalues of any one of the neighboring segments. The process must be continued until the last point in the sequence is plotted (Figure 1e). The percentages of points falling in different segments of the grid are calculated such that

$$G_j = (C_j/N) \times 100 \ (j = 1, 2, 3. . .24)$$

where N is the total number of residues in the sequence plotted. G_j, which represents the percentage of points plotted within or on any boundary of the *j*th segment, will be referred to henceforth as the grid count of the *j*th segment. Figure 1f shows the grid counts for the sequences of the members of the Hsp90 family.

It has been observed that the grid counts remain invariant (with standard deviations of ~1) for a particular family of proteins and for a particular orientation of the residue groups along the vertices of the CGR irrespective of the number of protein sequences concatenated or the order in which they were

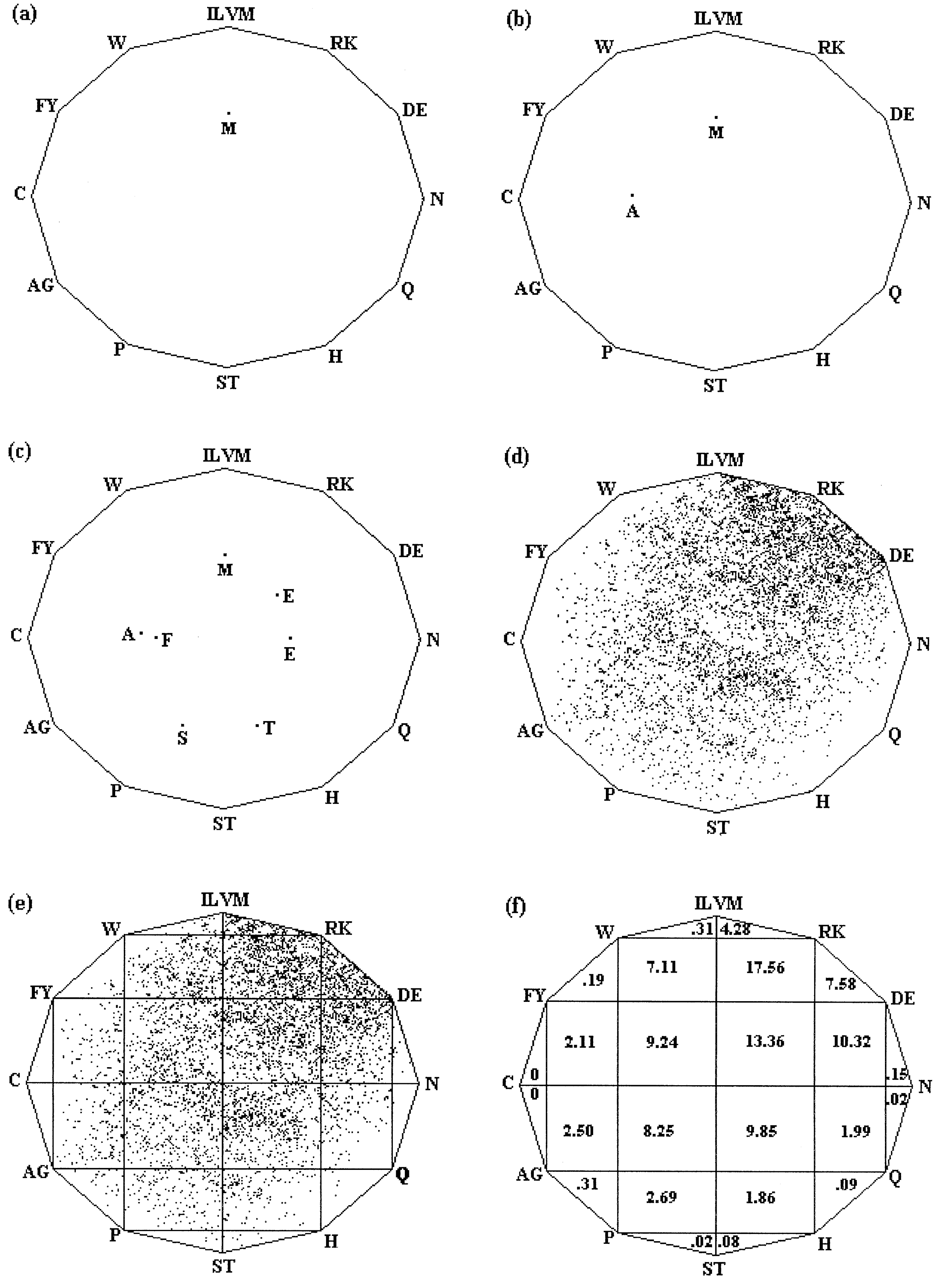

Figure 1. Generation of patterns in the CGR of proteins, following the algorithms described in text. (a–c) CGR of the first seven residues of yeast Hsp90. (d) CGR of the concatenated sequence of the Hsp90 family of proteins, having 7 847 residues. (e) Same as (d), except that the polygon has been divided into 24 grids. (f) Grid counts of the patterns shown in (d) and (e).

joined together. It is true not only for the CGR of Hsp90 family of proteins, but also for CGRs of all other families of proteins examined so far. It has been shown that the grid counts for the concatenated sequence of any family of proteins are similar (within the limit of standard deviations) to that obtained for the individual protein sequences belonging to that family.

APPLICATIONS OF THE ALGORITHMS TO DIFFERENT PROTEIN FAMILIES

Using these two algorithms, CGRs of several protein families were generated and the corresponding grid counts were determined (Figures 2 and 3). In each case, the amino acid sequences of different members of the family have been concatenated together to obtain a sufficient number of residues to exhibit visually identifiable patterns in the CGR. There is, however, no upper limit to the number of residues that can be used. The limit of resolution of the 12-vertex CGR has been determined and it has been observed that sequences in which the last eight residues are identical cannot be resolved under the normal viewing condition. The resolving power of the CGR can be increased by zooming in on the area of interest.

For most of the protein families examined in the present study, the distributions of points along the 12-sided polygon follow distinct, nonrandom, family-specific biases. The CGR of the Hsp70 family of proteins (Figure $2a_1$) showed a clustering of points near ILVM, RK, and DE vertices, although the clustering is less prominent than that observed in the CGR of the Hsp90 family of proteins (Figure 1d and e). The grid counts of the CGRs of the Hsp70 (Figure $2a_2$) and Hsp90 (Figure 1f) proteins show that the number of points lying in the lower half of the CGR of Hsp70 is significantly higher than is the case for Hsp90. This indicates that the frequencies of occurrence of residues such as A, G, P, S, and T are greater in the sequences of Hsp70 proteins than in the sequences of Hsp90 proteins. The CGR of Hsp60 proteins (Figure $2b_1$) apparently looks similar to that of Hsp70 proteins, although the grid counts of the two CGRs differ appreciably (Figures $2a_2$ and b_2). One common feature of all high molecular weight Hsp families is the paucity of points near W, FY, and C vertices (Figures 1d, $2a_1$, and $2b_1$), reflecting the rare occurrences of dipeptides *pq* (p, q = W, F, Y).

The CGR of the Hsp20 family of proteins (Figure $2c_1$) differs from those of high molecular weight Hsps in two respects. First, it does not exhibit any apparent cluster of points in any region of the 12-vertex polygon, although the grid counts (Figure $2c_2$) show a positive bias in the distribution of points towards the upper right quadrant of the polygon. Second, it required a larger number of residues to generate the pattern compared with other Hsps. Even for a concatenated sequence having about 10 000 residues (Figure $2c_1$), very few points are displayed in the CGR, indicating overlapping of large numbers of points. Thus, the extensive amino acid sequence homology between the members of Hsp20 family is reflected in the requirement for a large number of residues.

The CGR pattern as well as grid counts of rhodopsins (Figure $3a_{1,2}$) exhibit a clustering of points in the left half of the polygon, indicating that members of this family of proteins are rich in hydrophobic residues. Parenthetically, very few protein families examined so far showed abundance of phenylalanine(F) and tyrosine(Y).

The CGRs and grid counts of the Myc family (Figure $3b_{1,2}$), collagens (Figure $3c_{1,2}$), cytosol keratins (Figure $4a_{1,2}$), periodic proteins (related to circadian rhythms[10,11]) (Figure $4b_{1,2}$), and metalloproteases (Figure $4c_{1,2}$) all exhibit characteristic patterns indicating family-specific biases in the distribution of different residue groups along their sequences. For example, collagens (Figure $3c_{1,2}$) display a clustering of points along the side AG–P of the CGR, which implies that such proteins have high frequencies of the dipeptides *mn* (*m*, *n* = A, G, or P) along their sequences, whereas keratins (Figure $4a_{1,2}$) and periodic proteins (Figure $4b_{1,2}$) exhibit dense lines joining the AG and ST vertices of the CGR, suggesting an overrepresentation of the dipeptides *rs*, where *r* or *s* = A, G, S, or T in their primary structure.

SIMILARITY OF CGR PATTERNS OF TWO PROTEINS MAY NOT DEPEND ON THE EXTENT OF THEIR SEQUENCE HOMOLOGY

Voltage-dependent sodium channel (NaCh) proteins comprise a multigene family with at least six distinct isoforms known to exist in mammalian heart, brain, and skeletal muscle.[12–14] CGRs of all these protein sequences have been generated and their grid counts were calculated. Because there are about 2 000 residues in the deduced amino acid sequence of each of these NaCh proteins, the CGR patterns were obtained for each protein individually rather than for concatenated proteins. All the sodium channel proteins examined exhibit similar patterns as well as grid counts in their CGRs (Figure $5a_{1,2}$). This is in agreement with the fact that most of the deduced amino acid sequences from cloned NaCh cDNAs exhibit a very high degree of homology (about 90% in some cases) with one another.[12,13] However, the cDNA of the NaCh protein expressed in human heart and uterus (hNaV2.1) exhibits less than 50% overall amino acid sequence homology with other NaCh proteins and appears to represent a distinct subfamily of NaCh genes.[15] Surprisingly, the CGR of the deduced amino acid sequence of this atypical NaCh hNaV2.1 (Figure $5b_{1,2}$) is strikingly similar to the CGRs of other NaCh proteins (Figure $5a_1$) and their grid counts are almost identical within the limit of the standard deviation (1.5) (Figure $5a_2$ and b_2.). Thus, although hNaV2.1 exhibits about 50% overall primary sequence homology with other NaCh proteins, the frequency distributions of different conservative-substitution residue groups are similar to those of other sodium channel proteins. In sodium channel proteins, the frequencies of occurrence of both hydrophobic (isoleucine/leucine/valine/methionine, tryptophan, and phenylalanine/tyrosine) and hydrophilic (arginine/lysine and aspartate/glutamate) residues are much higher than for the relatively neutral residues. This is not surprising in view of the fact that NaCh proteins are transmembrane proteins.

DEPENDENCE OF CGR PATTERNS ON THE RELATIVE ORDER OF DIFFERENT RESIDUE GROUPS ALONG THE VERTICES

In a 12-vertex CGR, the vertices can be labelled in 11! different ways. It has been observed that the CGR pattern of a protein

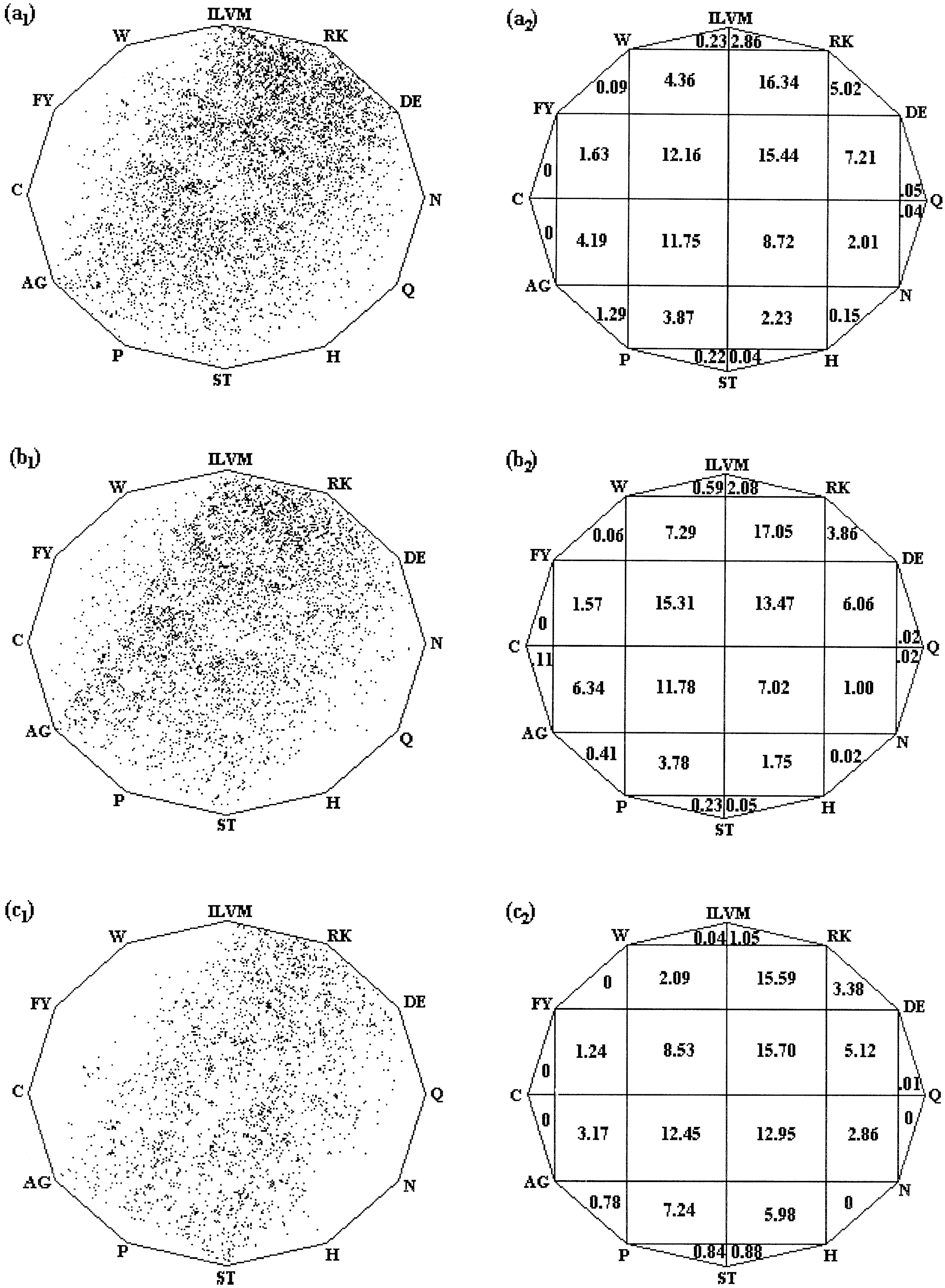

Figure 2. CGRs and grid counts of different protein families. (a_1 and a_2) Hsp70 (9 867 residues); (b_1 and b_2) Hsp60 (7 236 residues); (c_1 and c_2) Hsp20 (6 575 residues).

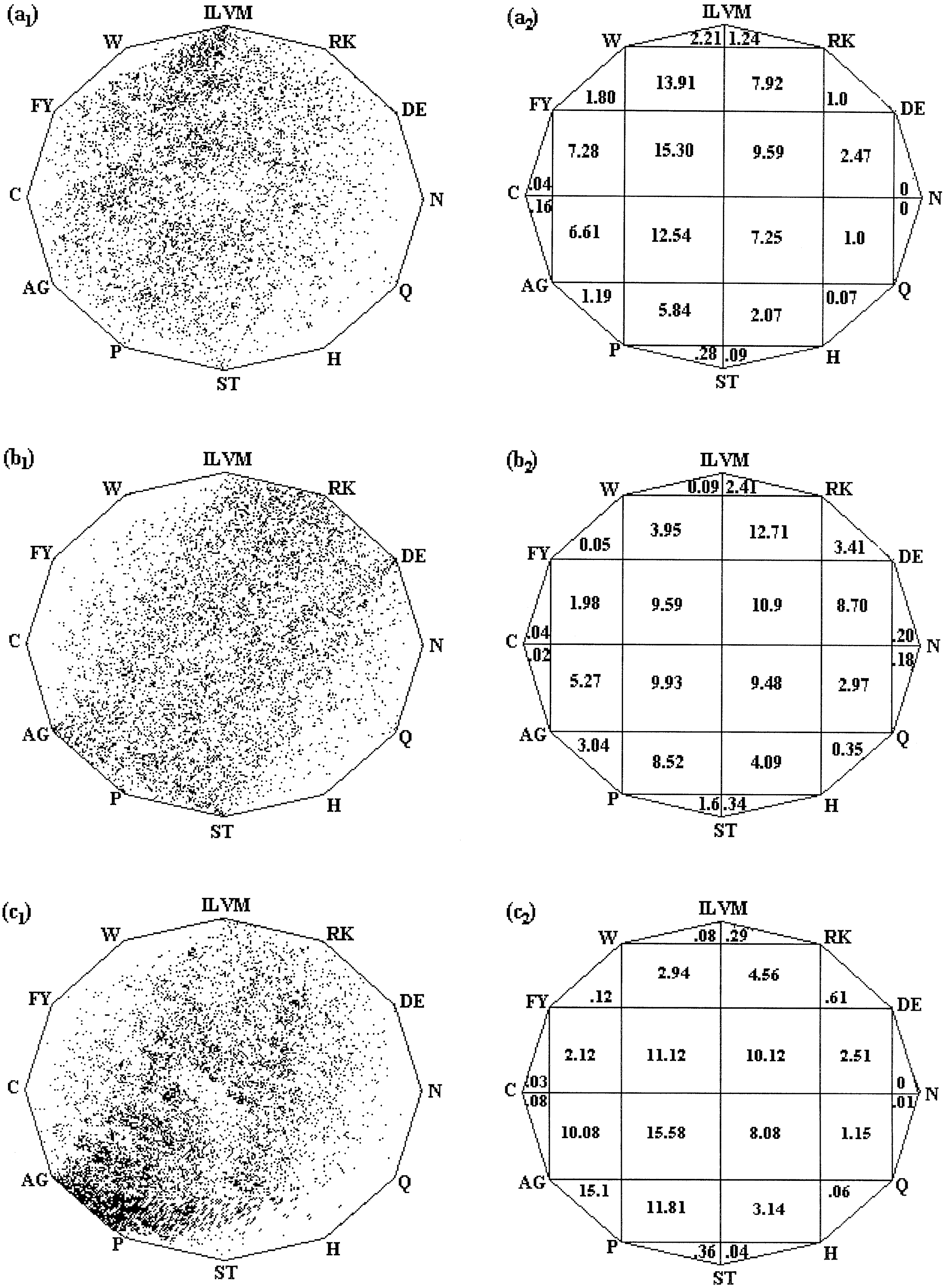

Figure 3. CGRs and grid counts of different protein families. (a_1 and a_2) Rhodopsins (12 893 residues); (b_1 and b_2) Myc (12 119 residues); (c_1 and c_2) collagen (14 424 residues).

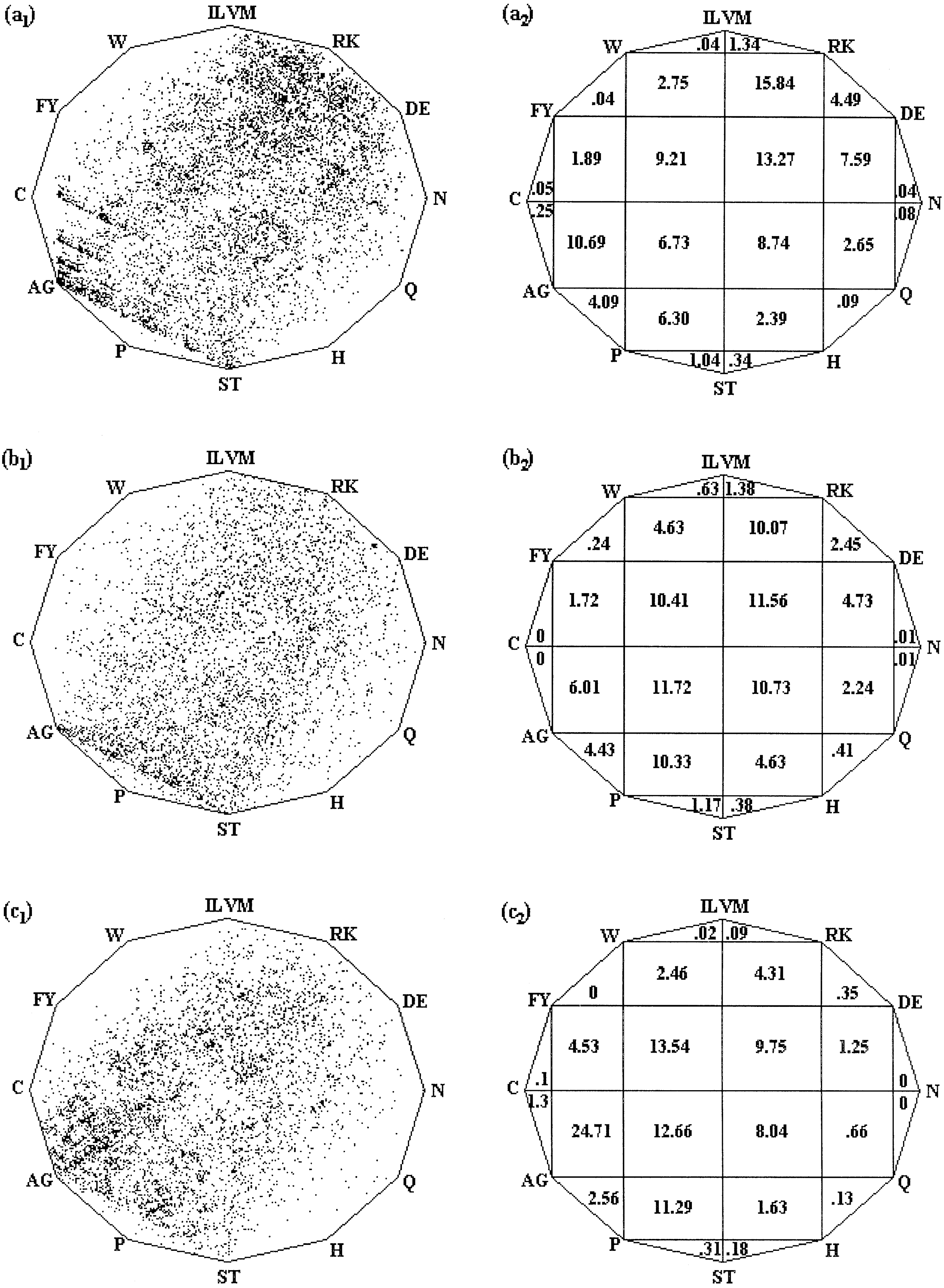

Figure 4. CGRs and grid counts of different protein families. (a_1 and a_2) Keratin (cytosol) (17 858 residues); (b_1 and b_2) periodic proteins (15 357 residues); (c_1 and c_2) metalloproteases (12 694 residues).

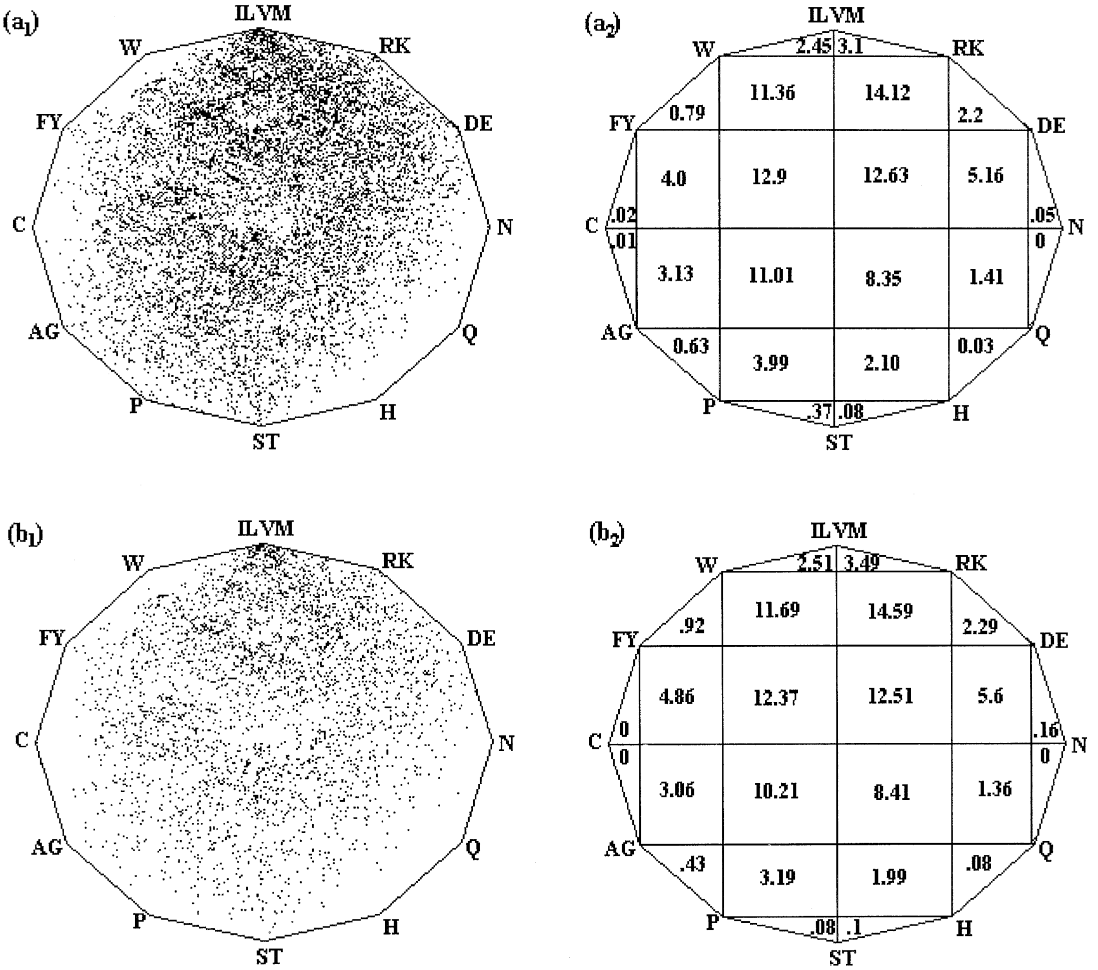

Figure 5. CGRs and grid counts of (a_1 and a_2) concatenated sequences of the sodium channel II proteins (8 132 residues); (b_1 and b_2) the human sodium channel protein hNaV2.1 (1 682 residues).

family is dependent upon the order in which the amino acid residue groups are arranged along the vertices. The patterns generated are insensitive to alteration with respect to less abundant residues such as cysteine (C), trytophan (W), asparagine (N), and histidine (H), but reorientation of frequently occurring residue groups such as ILVM, DE, RK, and AG along the vertices greatly affects the CGR patterns. Among various possibilities of labelling the vertices of the polygon, in the present study, the vertices are labelled in order of decreasing normalized hydrophobicity[16] of the residue groups in the counterclockwise direction starting from ILVM (Figures 1–5). This orientation will allow the derivation of information about the relative densities of hydrophobic and hydrophilic residues in the primary sequences of proteins from the CGR. Labelling of vertices can also be done according to hydropathicity, bulkiness, polarity, helix, or β sheet-forming propensities and other physicochemical properties of the amino acid residues.

In 12-vertex CGRs of proteins, a point in the CGR does not represent a unique amino acid sequence as do points in the 4-vertex DNA CGRs. In such protein CGRs, even within the resolution limit of the monitor, two entirely different amino acid sequences may be plotted at the same point within the polygon for a particular orientation of vertices. To resolve these degeneracies and for proper interpretation of patterns the relative positions of residue groups along the vertices of the CGR have been changed and the patterns compared (Figure 6). CGRs of Hsp70, Hsp60, and Myc (Figures 2a and b and 3b) visually appear to be similar (except the higher density of points along the sides AG–P and P–ST in the CGR of Myc), when plotted using the orientation of residues as in Figures 1–5. However, when the relative orientations of different residue groups along the vertices were altered to generate the CGRs described in Figure 6, the CGRs of Hsp70 and Hsp60 proteins (Figure 6a and b), although looking similar to one another, are both visibly different from that of Myc proteins (Figure 6c). The CGRs of the Hsp classes of proteins have been plotted in many different orientations of the residues along the vertices and it has been observed that for any particular orientation, they visually resemble one another. Similarly, some related oncogene families such as the Myc, Jun, and Fos families of proteins also exhibit CGR patterns similar to one another for any particular orientation of the vertices (data not shown). This implies that the apparent similarity between oncogene and Hsp classes of proteins in Figures 1 and 2 was an artifact whereas the intrasuperfamily[17] similarities between the different classes of Hsps or oncogenes are genuine. The members of the rhodopsin, keratin, and collagen families also ex-

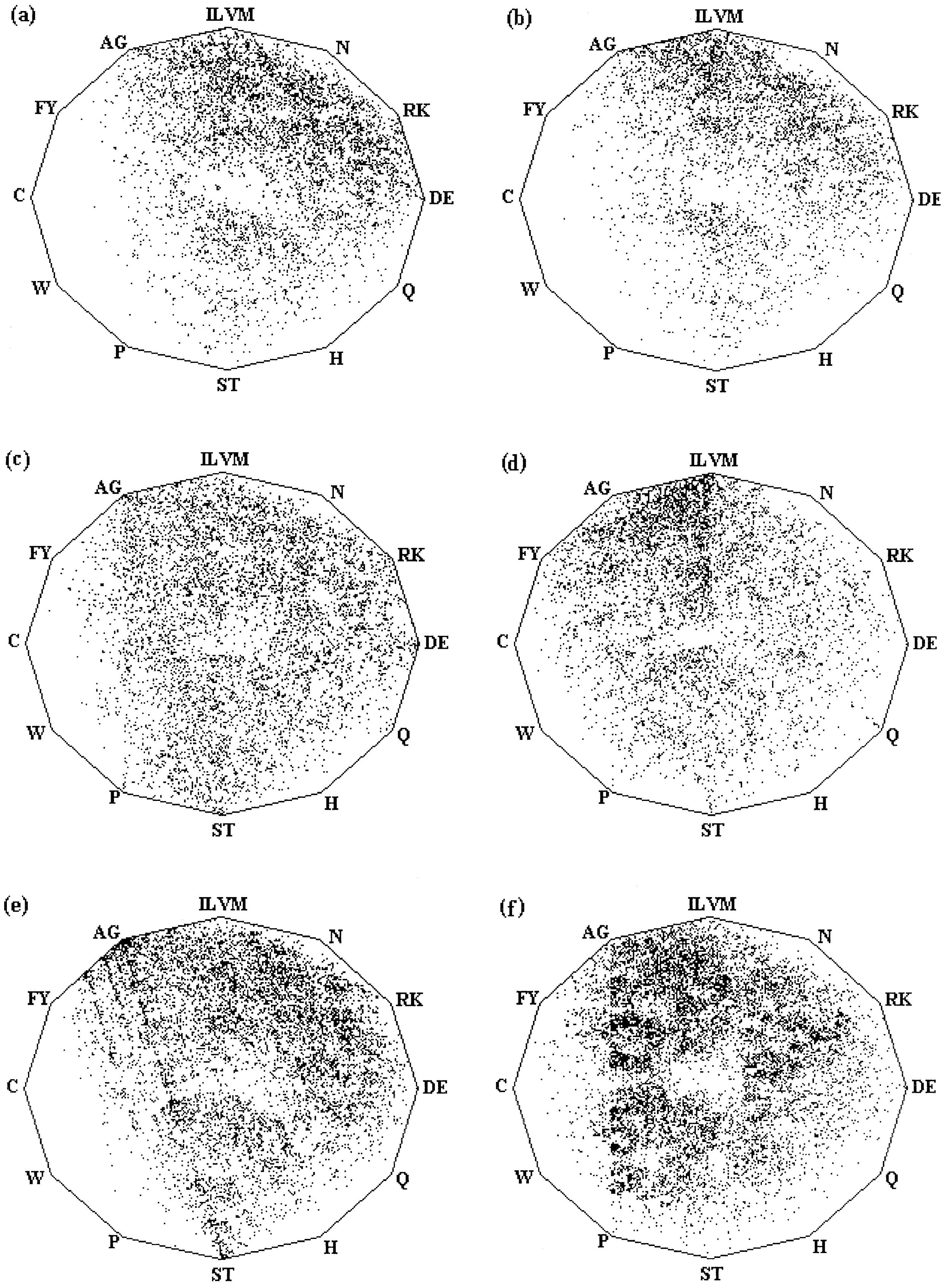

Figure 6. CGRs of (a) Hsp70, (b) Hsp60, (c) Myc, (d) rhodopsin, (e) keratin (cytosol), and (f) collagen protein families, using a relative orientation of vertices different from that used in Figures 1–3. The numbers of residues used to generate the patterns are the same as in Figures 1–4 for respective protein families.

hibit different characteristic patterns in the CGRs having a new orientation of vertices (Figure 6d–f).

POTENTIAL OF GRID COUNTS AS DIAGNOSTIC SIGNATURE OF A PROTEIN FAMILY

The grid counts of the CGR of any particular protein family depend only on the relative positions of different residue groups along its vertices and not on the number or order of proteins concatenated. To verify this, for each of the protein families examined in the present study, 70% of total protein sequences available in the database were taken as the "data set", while the remaining proteins were left as the test set. The proteins in the data set were then divided into different subsets. The members of any two subsets may or may not be mutually exclusive and even all of them may be identical but joined together in different orders. The grid counts were calculated for each of these subgroups for different orientations of the residue groups along the vertices of the CGR. For any particular orientation of residue groups, the grid counts remained the same (with small standard deviations) for all subsets belonging to a particular protein family. This implied that the grid counts represent a set of characteristic parameters of any protein family. For each of the protein families, average grid counts and standard deviations were calculated for three to five different orientations of vertices. To examine whether these grid counts can be used as discriminative and diagnostic signatures for characterization of new sequences, different protein sequences were picked from the test sets belonging to different families, CGRs were generated for each of these proteins individually, grid counts were determined, and the following calculations were done.

1. For any particular protein family, let $(G)_j$ be the average grid count and $(\mathrm{SD})_j$ be the standard deviation of the *j*th segment of the grid for a particular orientation of the vertices [in most cases, $0.2 \leq (\mathrm{SD})_j \leq 1$].
2. Let $(T)_j$ be the grid count of the test protein for the same orientation of the vertices.
3. Let $(\mathrm{DEV})_j$ be a parameter such that if

$$|(G)_j - (T)_j| > (\mathrm{SD})_j \quad \text{then} \quad (\mathrm{DEV})_j = 1$$

$$\text{otherwise } (\mathrm{DEV})_j = 0$$

4. $$\text{TOTAL DEV} = \sum_{j=1}^{24} (\mathrm{DEV})_j$$

5. Repeat steps 1–4 for two different orientations of the vertices.
6. If for each of the three different orientations, TOTAL DEV is less than the cutoff value, then the test sequence has a finite probability of being a member of the protein family under consideration. It has been seen that in general, for most of the protein families studied so far, if the cutoff value is set to be 4, the predictability of the preceding method is nearly 90%. However, in some cases, when the CGRs of two or more different protein families resemble one another even for different orientation of vertices, it becomes difficult to differentiate between their members. For example, the method described here is not very efficient in differentiating a member of the Myc family from that of the Jun/Fos family or in discriminating between the members of the Hsp60 and Hsp70 families of proteins.

CONCLUDING REMARKS

The present article describes the algorithms for generation of CGRs of the primary sequences of the members of different protein families and quantitative measurement of the bias, if any, in such CGR patterns. It has been demonstrated that different protein families exhibit distinct patterns in their CGRs with characteristic grid counts. It is, therefore, possible to use the grid counts as diagnostic features of such protein families for identification of new members of the families. In 12-vertex CGRs of proteins, the patterns are dictated not only by the frequencies of occurrence of different residues but also by the statistical bias in the occurrence of di-, tri-, or higher order peptides in the amino acid sequences of proteins.

A major advantage of the technique of protein CGR over the usual homology-searching sequence alignment programs is that it has the potential to reveal the evolutionary and/or functional relationships even between the proteins having no significant sequence homology, provided the sequences follow similar statistical distributions of di- or tripeptides, as seen in the case of the sodium channel proteins in the present analysis.

Nonrandom genomic sequences often generate fractal patterns in their corresponding CGRs. Surprisingly, no fractal pattern was detected in the CGRs of any of the protein families examined. In 12-vertex CGRs, intrinsic degeneracy is associated with the generated patterns in which points are not representing unique amino acid sequences. Hence, the fractal structure, normally originating from the paucity or abundance of any particular symbol in the sequence,[4] might be suppressed owing to the random occurrence of other residues in protein sequences. The degeneracy of points in the 12-sided CGR could be avoided, if the *m*th point were plotted following the formula provided in Fiser et al.,[6] instead of plotting it exactly halfway between the *m*–*l*th point and the vertex representing the *m*th residue. But in that case, the points would have appeared only in small regions around the vertices, leaving most of the central portion of the CGR empty. It has been observed that the CGRs generated in this way, in general, do not exhibit any visually identifiable family-specific pattern as observed in the CGRs plotted following the technique discussed in the present article.

It is worth mentioning at this point that in any CGR having more than four vertices, if the successive points are plotted at a specified fraction of the distance between the previous point and the new vertex, a fractal structure may emerge in the case of some special types of sequences, but using Fiser's formula[6] it can be shown that in such cases there may be a complex relationship between the visible CGR patterns and the fraction. It will be intriguing to study if any such complex relationship does exist in the case of 12-vertex protein CGRs.

ACKNOWLEDGMENTS

We thank Dr. P.M. Bhargava (Anveshna, Hyderabad, India) for encouragement and helpful suggestions. This investigation was supported by the Department of Biotechnology, Government of India. C.D. is an Associate of the CSIR Centre for Mathematical Modelling and Computer Simulation (NAL, Bangalore, India). S.B. and A.P. are grateful to the Council of Scientific and Industrial Research, Government of India, for their predoctoral fellowships.

REFERENCES

1 Barnsley, M.F. In: *Fractals Everywhere.* Springer-Verlag, New York, 1988, pp. 118–171

2 Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.* 1990, **18,** 2163–2170

3 Jeffrey, H.J. Chaos game visualization of sequences. *Comput. Graphics* 1992, **16,** 25–34

4 Dutta, C. and Das, J. Mathematical characterization of chaos game representation: New algorithms for nucleotide sequence analysis. *J. Mol. Biol.* 1992, **228,** 715–729

5 Solovyev, V.V., Korolev, S.V., Tumanjan, V.G., and Lim, H.A. Novij Podhod k klaccifikacii ysactkob DNK, ocnovannij ha fraktalnom predctavlenii nabora funktionalnovo chodhih pocledovatelnoctej. *Dockladi Akademii Nauk SSSR.* 1991, **319,** 1496–1500

6 Fiser, A., Tusnady, G.E., and Simon, I. Chaos game representation of protein structures. *J. Mol. Graphics.* 1994, **12,** 302–304

7 Dayhoff, M. *Atlas of Protein Sequence and Structure,* Vol. 5, Suppl. 3. National Biomedical Research Foundation, Silver Spring, Maryland, 1978

8 Farelly, F.W. and Finkelstein, D.B. Complete sequence of the heat shock inducible Hsp90 gene of *Saccharomyces cerevisiae. J. Biol. Chem.* 1984, **259,** 5745–5751

9 Karlin, S., Blaisdell, B.E., and Brendel, V. Identification of significant sequence patterns in proteins. *Methods Enzymol.* 1990, **183,** 388–402

10 Jackson, F.R., Bargiello, T.A., Yun, S.H., and Young, M.W. Product of *per* locus of *Drosophila* shares homology with proteoglycans. *Nature (London)* 1986, **320,** 185–188

11 Peixoto, A.A., Campesan, S., Costa, R., and Kyriacon, C.P. Molecular evolution of a repetitive region within the *per* gene of *Drosophila. Mol. Biol. Evol.* 1993, **10,** 127–139

12 Gellens, M.F., George, A.L., Chen, L., Chahine, M., Horn, R., Barchi, R.L., and Kallen, R. Primary structure and functional expression of the human cardiac tetrodotoxin-insensitive voltage-dependent sodium channel. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89,** 554–558

13 Noda, M., Ikeda, T., Kayano, T., Suzuki, H., Takeshima, H., Kurasaki, M., Takahashi, H., and Numa, S. Existence of distinct sodium channel messenger RNAs in rat brain. *Nature (London)* 1986, **320,** 188–192

14 Rogert, R.B., Gibbs, L.L., Muglia, L.K., Kephart, D.D., and Kaiser, M.W. Molecular cloning of a putative tetrodotoxin-resistant rat heart Na channel isoform. *Proc. Natl. Acad. Sci. U.S.A.* 1989, **86,** 8170–8174

15 George, A.L., Jr., Knittle, J.T., and Tamkun, M.M. Molecular cloning of an atypical voltage-gated sodium channel expressed in human heart and uterus. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89,** 4893–4897

16 Eisenberg, D., Schwarz, E., Komarony, M., and Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 1984, **179,** 125–142

17 Doolittle, R.F. Searching through sequence databases. *Methods Enzymol.* 1990, **183,** 99–110