Chaos game representation of protein structures

András Fiser, Gábor E. Tusnády, and István Simon

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary

Chaos game representation (CGR) was proposed recently to visualize nucleotide sequences as one of the first applications of this technique in the field of biochemistry.[1] In this paper we would like to demonstrate that representations similar to CGR can be generalized and applied for visualizing and analyzing protein databases. Examples of applications will be presented for investigating regularities, and motifs in the primary structure of proteins, and for analyzing possible structural attachments on the super-secondary structure level of proteins. A further application will be presented for testing structure prediction methods using CGR.

Keywords: chaos game representation, protein structure, fractal geometry

INTRODUCTION

Fractals and fractal-based mathematical procedures emerge in all areas of science. Chaos Game Representation (CGR) was recently proposed to visualize nucleotide sequences in the following way: in a square, where the four vertices correspond to the four nucleic acid bases A, C, G, T (or A, C, G, U), the first point of the plot is placed halfway between the center of the square and the vertex corresponding to the first nucleic acid base of the sequence.[1-3] The *m*th point of the plot (also called an *attractor*) is placed halfway between the (*m*−1)th point and the vertex corresponding to the *m*th base. An infinitely long random sequence results in an evenly filled-in square, while any nonrandom sequence corresponds to a characteristic attractor. Theoretically, the location of the *m*th point unambiguously represents the entire sequence that precedes it. To "decode" the sequence one needs only to identify the quarter of the square in which the point is located, as the vertex that belongs to this quarter indicates the last residue of the sequence. This quarter should then be further divided into four parts, and the procedure repeated as many times as needed to find out all preceding residues. Of course, practically, only the last 8–10 residues can be identified due to the finite resolution of computer displays.

This representation was further developed and applied to classify DNA sequences and to investigate possible homologies between the functionally important parts of DNA.[4] There have been other attempts at the graphical representation of nucleotide sequences, all of them sharing the common feature that they can handle only four different elements (e.g., the nucleic acid bases).[5-7] While this method can visualize gene structures, there has been no attempt to apply it to proteins.

This paper demonstrates that CGR can be extended in such a way that it becomes applicable for visualizing and analyzing both the primary and the three-dimensional (3D) structures of proteins.

PROGRAMMING

The computer programs were written in C, and were developed on a Silicon Graphics Personal Iris workstation under a UNIX operating system. Every attractor was visualized as a 1200 * 1000 pixel array, and we used the normal GL function. Every type of attractor required its own program. The executable versions on SGI are available on request (tusi@enzim.hu).

GENERALIZATION OF CGR

When an attractor is created in a square[1] for DNA and RNA sequences containing four elements, it is particularly important that the distance between the proper corner and the (*m*−1)th point be divided at an equal rate ($S = 0.5:0.5$), because at a higher dividing rate (i.e., when the distance between the proper corner and the *m*th point is greater than between the *m*th and the (*m*−1)th points), several parts of the attractor would overlap and the attractor would become ambiguous and undecodeable. At a lower dividing ratio the procedure results in a picture of a fractal (Color Plate 1) which is theoretically as decodeable and unambiguous as the ones created by using an equal dividing rate. Fractals may be less convenient to represent sequences containing four kinds of elements than a simple square, but the fractal representation can be generalized to become applicable for sequences of any number of elements, e.g., for the 20 residues of proteins.

Generalization may take place in several ways. In one of the simplest cases, the square is replaced by an *n*-sided regular polygon (*n*-gon), where *n* is the number of different

Color Plates for this article are on page 295.

Address reprint requests to Dr. Simon at the Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, H-1518 P.O. Box 7, Hungary.

Received 2 February 1994; revised 3 April 1994; accepted 12 April 1994

elements in the sequence which should be represented. In this case the attractor can be an unambiguous and consequently decodeable fractal showing an n-gon in which there are small, separated n-gons at every vertex. The small n-gons contain even smaller n-gons in their vertices, etc. When the circles around the inner polygons touch each other, the polygons do not overlap. In this case the dividing rate is:

$$S = S_1 : S_2 \qquad (1)$$

where $S_1 = \sin(2\pi * i/n)/(1 + \sin(2\pi * i/n))$ and $S_2 = 1/(1 + \sin(2\pi * i/n))$.

The (x,y) coordinates of a certain vertex i (if the circle around the main polygon is a unit circle) are:

$$v_{i,x} = \cos(2\pi * i/n) \qquad (2)$$

$$v_{i,y} = \sin(2\pi * i/n) \qquad (3)$$

The coordinates of the 0th point is [0,0] and the mth point are:

$$p_{m,x} = (v_{m,x} - p_{m-1,x}) * S_2 + p_{m-1,x} \qquad (4)$$

$$p_{m,y} = (v_{m,y} - p_{m-1,y}) * S_2 + p_{m-1,y} \qquad (5)$$

As proteins consist of 20 kinds of amino acids, a 20-sided regular polygon (regular 20-gon) and $S = 0.135 : 0.865$ dividing ratio (calculated from Equation (1)) is the most adequate for sequence representation, because in that case we get 20 separated but almost touching smaller 20-gons. Color Plate 2a shows the representation of a pentapeptide having the amino acid sequence IDEAL. A few thousand points result in an ''attractor'' which gives a visible impression of the rare or frequent residues and sequence motifs. In areas corresponding to the rare or never occurring pairs, triplets, etc. we will find poorly dotted or empty 20-gons. The new small twenty 20-gons represent the preceding amino acids in the sequences. Theoretically, each point represents the whole preceding sequence motif, but practically, we cannot recognize more than two or three preceding neighbors due to the finite resolution of the monitor (Color Plate 2b).

When the number of residues represented exceeds 10^5, all polygons look equally filled in, and the abundance of the various sequence motifs become indistinguishable. The efficiency of this representation can be improved significantly by putting any of the inner polygons into the empty space in the center of the original n-gon and zooming it out to the border of the empty space; then the next one from the resulting smaller polygons can be put in the center of the concentric polygon and zoomed out again, etc.

Longer sequence motifs can be displayed by putting the zoomed smaller 20-gons inside the bigger one. To avoid overlapping of the concentric 20-gons at every step the radius of the kth zoomed circle should be:

$$r_k = r_{k-1} - 2 * S_1 * r_{k-1} \qquad (6)$$

where $k_0 = 1.0$. For a 20-gon:

$$r_k = 0.72945 * r_{k-1}$$

as has been described above. As a result of this procedure, the sequence information becomes visible for various segments along the radii of the original 20-gon, instead of the 20-gons being fused into each other.

One can look for special sequence motifs like IDEAL in a database[12] (Color Plate 2c). In this case, the 20-gon L should be zoomed in the center of the main polygon. Then the 20-gon A of the L polygon should be zoomed in the center. In the next step, the 20-gon E of the A polygon should be zoomed in the center. The 20-gon D of the E polygon is scarcely dotted, indicating the limited number of tetrapeptide DEAL in the database. When this 20-gon is zoomed in the center, one should note that there are only two points of the 20-gon I of the D polygon. Finally this 20-gon I can be zoomed in the center (and its points made bigger than the others).

The resulting two bright green points indicate that there are two IDEAL pentapeptides in the database. The position of these points indicates that one IDEAL pentapeptide follows residue G, and the other follows residue A. In same way, any other sequence fragment, including repetitive sequences, can be studied. On an average-sized monitor (1000*1000 point) one can analyze sequence motifs up to 22 residues long ($0.72954^{22} \sim 10^{-3}$).

SECONDARY STRUCTURE REPRESENTATION

Chaos Game Representation can also be used to study 3D structures of proteins. Protein conformations can be characterized by a sequence of dihedral angles (ϕ, ψ) of the single bonds of C_α atoms in the polypeptide chain. Due to steric restrictions that stem from high-energy atomic overlap, there are only 16 areas on the ϕ,ψ map—Ramachandran plot[8]—available for a low-energy structure. The conformation of a polypeptide chain can be characterized by the sequence of these low-energy areas along the polypeptide chain.[9] Thus, protein structures can be visualized in a way analogous to that put forward for sequences by using 16-gons instead of 20-gons.

In most cases, a less detailed structure description, with reference to helix, sheet, turn, and ''random coil'' structures are used for characterizing the polypeptide structure. When one deals with exactly four kinds of elements, the original CGR suggested[1] can be used by replacing the four nucleotides with the four secondary structure elements at the vertices of the square (Color Plate 3a). A random sequence would result in a filled in square, while this attractor indicates the nonrandomness of the structural elements in proteins.

Up to this point, we have discussed only sequences of similarly ranking elements. However, one of the four structural elements, the random coil, is not a regular one, so it is not in the same rank as the other three. Therefore, instead of a square, one of the regular structure elements (helix, sheet, or turn) can be selected and placed at the vertices of a regular triangle, while the random coil structure is represented by the center of this triangle (Color Plate 3b). If the distribution of the secondary structure elements were random, it would result in a Sierpinski-like triangle,[1,3] but as the central point is also used for reference point, a new Sierpinski triangle appears without overlapping the original one. (Note that it is simply a projection picture of the space points of a tetrahedron projected to its favored side, which contains the three vertices of the tetrahedron as reference to the three regular structural elements.) In this way, almost all the space is covered in a random case. Using protein structure information (Protein Data Bank[10]) the attractor is significantly modified and dotted areas are much more scarce, reflecting the 3D structure restrictions. The attractor loses its threefold symmetry, reflecting the secondary structure regularities. This may be a useful tool in studying the frequency of the attachment of various secondary structure elements, or to test the several prediction methods by comparing the predicted attractors with the ones based on the native structure. Color Plate 3c shows the CGR representation of the test of the Chou-Fasman prediction methods.[11] The yellow dots represent native protein structures (Protein Data Bank[10]), while the Chou-Fasman prediction on the same data set results in the blue and red points. If the predicted structure matches the native one, the yellow point turns to red; the blue dots are the mispredicted ones. We can check, using the attractor, which structures are the well-predicted ones and in which cases the method provides false results. The red points around the corners and along the lines between the center and corners indicate that the prediction is efficient in recognizing long regular secondary elements. The vertex marked *turn* is full of yellow dots because the prediction hardly recognizes consecutive turns, while in the native structures not only consecutive but overlapping turns (e.g., a 15-membered continuous turn) also occur. An abundance of blue points inside the inner center-sheet-helix and center-helix-turn triangles shows that the prediction and method cannot handle the immediate transition between consecutive secondary elements precisely. For example, the helix-former potential is already low at the end of a helix, and the sheet-forming potential is still not high enough, but in the native structure these structural elements are consecutive. However, the prediction often calculates a short random segment between them which "fills" the inner triangle with blue dots. The center-turn-sheet triangle is almost clear, and is much more regular because of the high occurrence of 4-residue long turns. This also causes the empty spaces within the other two inner triangles.

CONCLUDING REMARKS

All informational macromolecules of biological interest (DNA, RNA, peptides and proteins) are linear polymers. Not only their chemical structures, but also their 3D structures, can be represented as a linear array; i.e., the conformation of a macromolecule may be given as the sequence of dihedral angles around the single bonds along the polymer chain. As shown in this paper, CGR is an effective method for visualizing any structural features if it is given as a sequence of elements. The examples discussed in this paper demonstrate that CGR can be applied for questions related to both the primary and the 3D structures of proteins.

ACKNOWLEDGMENTS

This work was started in the frame of a scientific collaboration project between the Hungarian Academy of Sciences and the Research Council of Norway, and was supported by research grant OTKA 1361 and 12890. Fiser and Tusnády appreciate support from the foundation "For Hungarian Science." The authors are grateful for helpful discussions with Professor Dietmar Kamp, Professor Ferenc Vonderviszt, and Dr. Péter Tompa.

REFERENCES

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

2 Jeffrey, H.J. Chaos game visualization of sequences. *Comput. Graphics* 1992, **16,** 25–33

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

4 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

5 Pickover, C.A. DNA and protein tetragrams: biological sequences as tetrahedral movements. *J. Mol. Graphics.* 1992, **10,** 2–17

6 Dutta, C. and Das, J. Mathematical characterization of chaos game representation; new algorithms for nucleotide sequence analysis. *J. Mol. Biol.* 1992, **228,** 715–719

7 Goldman, N. Nucleotid, dinucleotid and trinukleotid frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acid Res.* 1993, **21,** 2487–2491.

8 Ramachandran G.N. and Sasisekharan, V. Allowed conformations of polypeptide chain. *Adv. Prot. Chem.* 1968, **23,** 325–347

9 Simon, I., Glasser, L., and Scheraga, H.A. Calculation of protein conformation as an assembly of stable overlapping segments: application to bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA.* 1991, **88,** 3661–3665

10 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542

11 Fasman, G.D. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Plenum Press, New York, 1989, pp. 391–417

12 George, D.G., Barker, W.C., and Hunt, L.T. The protein identification resource (PIR). *Nucleic Acid Res.* 1986, **14,** 11–15