Structure-based barcoding of proteins

Rahul Metri,¹ Gaurav Jerath,² Govind Kailas,¹ Nitin Gacche,² Adityabarna Pal,² and Vibin Ramakrishnan^{1,2*}

¹Institute of Bioinformatics & Applied Biotechnology, Bangalore 560100, India

²Department of Biotechnology, Indian Institute of Technology, Guwahati 781039, India

Received 6 August 2013; Revised 15 October 2013; Accepted 21 October 2013

DOI: 10.1002/pro.2392

Published online 29 October 2013 proteinscience.org

Abstract: A reduced representation in the format of a barcode has been developed to provide an overview of the topological nature of a given protein structure from 3D coordinate file. The molecular structure of a protein coordinate file from Protein Data Bank is first expressed in terms of an alpha-numero code and further converted to a barcode image. The barcode representation can be used to compare and contrast different proteins based on their structure. The utility of this method has been exemplified by comparing structural barcodes of proteins that belong to same fold family, and across different folds. In addition to this, we have attempted to provide an illustration to (i) the structural changes often seen in a given protein molecule upon interaction with ligands and (ii) Modifications in overall topology of a given protein during evolution. The program is fully downloadable from the website <http://www.iitg.ac.in/probar/>.

Keywords: barcode; protein structure comparison; fold classification

INTRODUCTION

The strength of protein data bank (PDB) has been growing exponentially over last 3 decades.¹ As structural genomics initiatives gain momentum, this trend is expected to continue in the following years as well, principally because of the rapid advance-

ment in high throughput structure determination techniques.^{2,3} Total number of structures reported in PDB is inching closer to the milestone of 1 lakh structures. Total number of folds identified so far is 1392 and 1282 as per SCOP^{4,5} and CATH⁶ classification, respectively, and no additions to this number have been reported since 2009. Nevertheless proteins belong to the same fold family do exhibit variations at sequential, structural (to some extent) as well as functional levels.^{7,8} Numerous tools are available as open source programs for protein visualization⁹ and structure prediction.^{10,11} There have also been attempts to present reduced representations to three-dimensional⁶ protein structures in 2D and 1D. TOPS diagrams¹² and contact maps¹³ show protein secondary structure and topology in two dimensions, while DSSP presents secondary structure information of a protein molecule sequentially from N terminus to C terminus as a 1D string.¹⁴ We present here a new representation of protein structure in the form of a "barcode." The advantage of

Abbreviations: CATH, class architecture topology homology; CBIR, content-based image retrieval; DHFR, dihydrofolate reductase; DSSP, dictionary of protein secondary structure; PDB, protein data bank; SSE, secondary structure elements; TOPS, topology of protein structure.

Additional Supporting Information may be found in the online version of this article.

Grant sponsors: Department of Biotechnology, Govt. of India (Innovative Young Biotechnologist Award [IYBA] Scheme) and Department of Information Technology, Government of India (DIT-CoE scheme, to G.K.).

*Correspondence to: Vibin Ramakrishnan, Department of Biotechnology, Indian Institute of Technology, Guwahati 781039, India. E-mail: vibin@iitg.ernet.in

Figure 1. Generation of protein barcode from 3D representation of protein G (1PGB.pdb) (A). TOPS diagram of protein G showing secondary structure and their relative orientation (B). SSEs with the previous and successive ones are assigned based on a tableaux representation with space width assigned in parenthesis (C). ANCODE generated for protein G as explained in validation Section (D) and its corresponding barcode format (E).

this type of representation is that, it can encode secondary structure as well as their relative orientation in space. We can align different “barcodes” to compare and contrast structural and topological information of a given structure. Inspiration to this type of a representation was drawn from the pioneering contribution in encoding information as “barcodes” by Bernard Silver and Norman Woodland in 1949.¹⁵ It took 3–4 decades to completely operationalize the technology using barcodes for cataloguing articles across a wide variety of applications. We present in this article, the design and utility of this computational tool in cataloguing proteins according to their structure. The program is fully downloadable from the website <http://www.iitg.ac.in/probar/>; we also provide a webserver that can display barcode images of close to about 70,000 protein molecules in PDB.

VALIDATION OF COMPUTATIONAL METHODS

Crystal structure of B1 immunoglobulin-binding domain of streptococcal protein G¹ (1PGB.pdb) is used as a model structure to illustrate the design of protein barcode representation. The 56 residue protein molecule with one alpha helix and one beta sheet consisting of four beta strand has a well-defined hydrophobic core. Total number of secondary structure elements is five, with first and second strands forming an antiparallel beta sheet followed by a helix. Another antiparallel beta sheet follows the helix, coplanar with the first sheet with final beta strand being parallel to the first strand. As all four strands form one continuous sheet, all four strands are colored same (blue in this case). Secondary structure elements (SSEs) not part of the same sheet are colored differently as illustrated in Figures 3 and 4. All successive secondary structures in protein G are antiparallel in their relative orientation and hence having an identical space width of three units. Space width is customizable by appropriately modifying the code. Space width may change according to the relative topology of successive SSEs. Therefore, protein barcode provides information about SSEs and their relative topology with necessary clarity. Furthermore, it is possible to derive TOPS representation from barcode with reasonable accuracy and vice versa (Figs. 1 and 2).

Structure comparison using barcode identity index (BII): analyzing the spatial orientations of proteins is significant for their functional and evolutionary studies¹⁶ and such an objective may be achieved by comparison of barcodes. To indicate the utility of protein barcode, we further examined the barcode images generated from structure files of all PDB structures of DHFR (dihydrofolate reductase) across different species.¹ Although the barcode images look more or less identical, subtle differences can be observed in structures adapted during evolution from left to right (Fig. 3). A barcode identity index (BII) has also been formulated to compare structures quantitatively (Fig. 4) and structural adaptations at specific loci can be identified by carefully comparing two barcode images. Barcode identity index (BII) is calculated from a metadata of barcode image, consisting of numbers that correspond to the “barcode” and aligning them. In a typical case, Helix is represented as 0, Strand as 1, and the orientation between secondary structures as 3, 4, 5, and 6 based on space width between 2 bars in the barcode

Figure 2. Barcode images of representative protein structures corresponding to all beta, all alpha, and alpha/beta folds in the SCOP database. The respective TOPS diagram and “Barcodes” present the utility of “barcode” representation in encoding the structure and topology of any given protein structure.

Figure 3. Barcodes corresponding to dihydrofolate reductase enzyme in different species. Only those species with structures available in PDB were shown in this figure. The differences in barcode can be attributed to the differences in the secondary structures that are altered during the course of evolution. However, there is a common string of bars in the barcode depicting the structural conservation for DHFR in the bacterial species. Similarly, the barcodes for the vertebrates and fungi are somewhat identical within their respective sets.

representation. For example, 1A41.pdb may be represented as 03030413140304030303030. The number that represents a barcode (query) is aligned with another number (subject) using Needleman Wunsch algorithm.¹⁷ Further details may be found in Supporting Information and BII code may be downloaded from Barcode webpage.

Protein barcode is presented as a TIFF image. If this representation is widely accepted by the scientific community, then it will help in locating proteins in a “protein-barcode” database by making use of Content-based image retrieval (CBIR) tools.^{18,19} This method is basically meant for addressing the problem of searching digital images in large databases. It analyzes the content of the image rather than the meta-data or descriptions or tags associated with the image. Barcode representation foresees this opportunity in subsequent phases of its development, although it is beyond the scope of this manuscript. Furthermore, we tested barcode image comparison to study the possible structural alterations during ligand binding on the same DHFR structure. The number and type of ligands bound to DHFR receptor were given in Table S1 (Supporting Information). The disparities in structures are pictorially represented as barcodes and their relative similarities in overall topology may be quantified from calculating BII. For illustrative purpose, topologically similar structures are clubbed together and structurally dissimilar molecules are separated in a VIBGYOR color scheme.

COMPUTATIONAL METHODS

Protein barcode is the representation of secondary structures, and their orientations as barcode images. The colored bars in the barcode image correspond to the SSEs and white spaces between the secondary structures represent the orientation between the

two SSEs. Three-dimensional co-ordinate file from PDB is used to generate these barcodes. DSSP program is used to obtain secondary structure information. The information about strands and the sheet they belong to is also obtained from DSSP file.¹⁴ The orientation between secondary structures is the angle in radians calculated by atan2 method. The first step in generating a “barcode” is the generation of an alpha-numero code (ANCODE). ANCODE is a combination of alphabets, H (for helix), and S (for strand/sheet) followed by a four-digit number

Figure 4. Differences in protein structures illustrated using “barcode” representation when the same DHFR molecule is bound with different ligands. All structures are obtained from PDB.³

divided into two pairs. First pair represents overall SSE count and second pair represents the count of secondary structure each SSE belongs to. For example, S0401 in Figure 1(D) signifies that the given strand is the fourth SSE in the overall structure, but belongs to the first sheet. Similarly, H0301 in Figure 1(D) signifies that Helix (H) is the third SSE⁵ but is first (01) helix in the overall structure.

The orientation of each SSEs with the previous and successive ones is assigned based on a tableaux representation [Figure 1(C)]. If both secondary structures are pointing within 90° against each other, they are considered parallel (P) and if they are between -135° and +135°, antiparallel. The relative orientations in between are designated as L and R in either directions as shown in Figure 1(C).

BARCODE is derived from ANCODE generated using pdb file. H is always colored black, S is colored based on the corresponding sheet id. Each sheet id is colored unique. For example, Figure 2(A) has seven strands with four strands forming one sheet (green) and the remaining three forms second sheet (blue). Orientations of successive SSEs are represented by the “width” of white space between the bars in barcode image. Orientation and pixel width is as follows, P = 6 units, A = 3 units, R = 4 units, and L = 5 units. Representations of successive SSEs are denoted in ANCODE in the sixth and seventh spaces after a colon. The first letter shows orientation between previous SSE and second letter shows the succeeding one. If the previous SSE and succeeding SSE is missing (as in the case of N terminus and C terminus) it is denoted as “O” [Fig. 1(C,D)]. Thus, secondary structures and topology are encoded in the ANCODE string and further translated to barcode image in TIFF format in MATLAB.²⁰

CONCLUSION

In this methodology article, we attempted to present a new reduced representation of protein structures so as to compare and contrast two structures based on their secondary structure and topology. Apart from the structural and topological information conveyed, we can also quantify the overall comparison by way of a barcode identity index (BII). The two experiments described above are indicative of the utility of the tool. Addressing a scientific problem and comparison with other tools are not within the scope of this article, yet the value of the method for qualitative and quantitative comparison of protein structures may not be discounted. The program is fully downloadable from the webpage <http://www.iitg.ac.in/probar/>.

Acknowledgments

Authors acknowledge the contributions of Prof. P. K. Bora of Electrical Engineering at IIT Guwahati for useful suggestions and Rakesh Kumar of Biotechnology, IIT Guwahati in the final formulation of this manuscript and creation of webpage.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
2. Pieper U, Schlessinger A, Kloppmann E, Chang GA, Chou JJ, Dumont ME, Fox BG, Fromme P, Hendrickson WA, Malkowski MG, Rees DC, Stokes DL, Stowell MHB, Wiener MC, Rost B, Stroud RM, Stevens RC, Sali A (2013) Coordinating the impact of structural genomics on the human [alpha]-helical transmembrane proteome. *Nat Struct Mol Biol* 20:135–138.
3. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J (2000) The Protein Data Bank and the challenge of structural genomics. *Nat Struct Mol Biol* 7:957–959.
4. Day R, Beck DAC, Armen RS, Daggett V (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* 12:2150–2160.
5. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36:D419–D425.
6. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res* 41:D490–D498.
7. Krissinel E (2007) On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* 23:717–723.
8. Eidhammer I, Jonassen I, Taylor WR (2000) Structure comparison and structure patterns. *J Comp Biol* 7:685–716.
9. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14:33–38.
10. Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294:93–96.
11. Zhang Y (2009) Protein structure prediction: when is it useful? *Curr Opin Struct Biol* 19:145–155.
12. Michalopoulos I, Torrance GM, Gilbert DR, Westhead DR (2004) TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res* 32:D251–D254.
13. Yuan X, Bystroff C, Protein contact map prediction. In: Xu Y, Xu D, Liang J, Ed. (2007) Computational methods for protein structure prediction and modeling. New York: Springer, pp 255–277.
14. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.
15. Woodland NJ, Silver B (1952) Classifying apparatus and method. US Patent no. 2612994.
16. Shi S, Chitturi B, Grishin NV (2009) ProSMoS server: a pattern-based search using interaction matrix representation of protein structures. *Nucleic Acids Res* 37: W526–W531.
17. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443–453.
18. Lew MS, Nicu S, Chabane D, Ramesh J (2006) Content-based multimedia information retrieval: state of the art and challenges. *ACM Trans Multimedia Comp Commun Appl* 2:1–19.
19. Ritendra D, Dhiraj J, Jia L, James ZW (2008) Image retrieval: ideas, influences, and trends of the new age. *ACM Comput Surv* 40:1–60.
20. MATLAB version 7.10.0. (2010) Natick, Massachusetts: The MathWorks Inc.