Protein Engineering vol. 16 no. 10 pp. 713–715, 2003
DOI: 10.1093/protein/gzg100

COMMUNICATION

A novel fingerprint for the characterization of protein folds

Mihaly Mezei

Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York University, New York, NY 10029, USA

E-mail: mezei@inka.mssm.edu

A novel fingerprint, defined without the use of distances, is introduced to characterize protein folds. It is of the form of binary matrices whose elements are defined by angles between the C=O direction, the backbone axis and the line connecting the α-carbons of the various residues. It is shown that matches in the fingerprint matrices correspond to low r.m.s.d.

Keywords: fold recognition/molecular similarity/protein fingerprint

Introduction

The tertiary structures of proteins appear to fall into a limited number of classes. The characterization of these classes, however, is not a trivial matter—even the uniqueness of structural alignments may be questionable (Godzik, 1996). As summarized recently (Shindyalov and Bourne, 1998), measures of structural similarities fall into the following classes: '(i) structure superposition as rigid bodies; (ii) inter-residue distances; (iii) environmental properties (for example, exposure, secondary structure); and (iv) conformational properties (for example, bond angles dihedral angles and orientation with respect to the protein center of mass)'.

One reason for this difficulty is that rather large deviations in the positions of atoms can occur within the same fold. This means that the set of internal coordinates of a protein domain with a given fold contains too much information with respect to the fold. It is therefore of interest to establish a characterization of the tertiary structure of proteins that contains just enough information about it so that it can be distinguished from other folds. The aim of this communication is the introduction of such a novel protein fingerprint with reduced information content and the exploration of its ability to characterize the secondary structure. This fingerprint forms a new category of conformational properties [class (iv) above], as it relies solely on angles between inter-residue lines and locally defined directions.

Methods

For a protein of n residues, its primary fingerprint is defined as an $n \times n$ binary matrix $\mathbf{FP^0}$ whose elements are defined by the angles that the line connecting the backbone carbonyl carbons of residues i and j forms with the C=O bond's direction on residue i:

$$FP^0_{ij} = \text{sign}\{[\mathbf{r}(O_i) - \mathbf{r}(C_i)]\cdot[\mathbf{r}(C_j) - \mathbf{r}(C_i)]\}$$

i.e. $FP^0_{ij} = -1$ if the angle $\phi(O_iC_iC_j)$ is >90° and 1 otherwise. Such a matrix can be conveniently visualized by drawing black and white squares at the places corresponding to –1 and 1, respectively. Since the C=O directions essentially alternate by 180° in sheets, $\mathbf{FP^0}$ will be dominated by alternating white and black bars in such regions. On the other hand, the C=O directions are essentially parallel in helices, resulting in black equilateral right-angle triangles located above the diagonal. Note also that an essentially equivalent definition for $\mathbf{FP^0}$ can be obtained using the N–H bond and N–N distances since the C=O and N–H bond directions of a peptide bond are essentially antiparallel. The choice of using the C=O direction was dictated by the frequent lack of hydrogen coordinates in experimentally determined structures.

It is easy to see, however, that $\mathbf{FP^0}$ is insensitive to the backbone direction for β-sheets and to the helix packing arrangements for parallel helix bundles. This means that for some folds using the primary fingerprint only will result in partitioning of the proteins into superfamilies. The separation of the members in these superfamilies, however, can be achieved with the use of two secondary fingerprints, defined as

$$FP^1_{ij} = \text{sign}\{[\mathbf{r}(N_i) - \mathbf{r}(C_i)]\cdot[\mathbf{r}(C_j) - \mathbf{r}(C_i)]\}$$

$$FP^2_{ij} = \text{sign}\{[(\mathbf{r}(O_i) - \mathbf{r}(C_i))\times(\mathbf{r}(N_{i+1}) - \mathbf{r}(C_i))]\cdot[\mathbf{r}(C_j) - \mathbf{r}(C_i)]\}$$

where $FP^1_{ij} = -1$ if the angle $\phi(N_iC_iC_j)$ is >90° and 1 otherwise and $FP^2_{ij} = -1$ if the angle between the line C_i–C_j and the normal to the plane of O_i, C_i and N_{i+1} is >90° and 1 otherwise. Again, neither FP^1 nor FP^2 is symmetric. The line N_i–C_j is largely parallel to the backbone, hence the information in $\mathbf{FP^1}$ encodes the direction that the backbone path takes, allowing one to differentiate between parallel and antiparallel sheets. On the other hand, the vector normal to the plane of O_i, C_i and $C\alpha_i$ is largely perpendicular to the backbone direction, thus the information in $\mathbf{FP^2}$ encodes the relative arrangements of these backbone segments, allowing one to differentiate between different packing of helices.

The utility of these fingerprint matrices rests on the question of whether they really contain enough information to tell different folds apart. To answer this question, an alignment procedure has been implemented, in which the fingerprint matrix of protein A (representing the query structural element) is compared with protein B's diagonal submatrices of identical size. Positions with good agreement between the submatrices indicate substructures of protein B with similar fold to that of A. Concurrently with the comparison of the fingerprint of protein A with successive diagonal submatrices of protein B, the r.m.s.d. between protein A and the substructure corresponding to the selected submatrix was also calculated using the formula of Kabsch (Kabsch, 1976). Comparison of the plots of the r.m.s.d. and the matrix differences as a function of

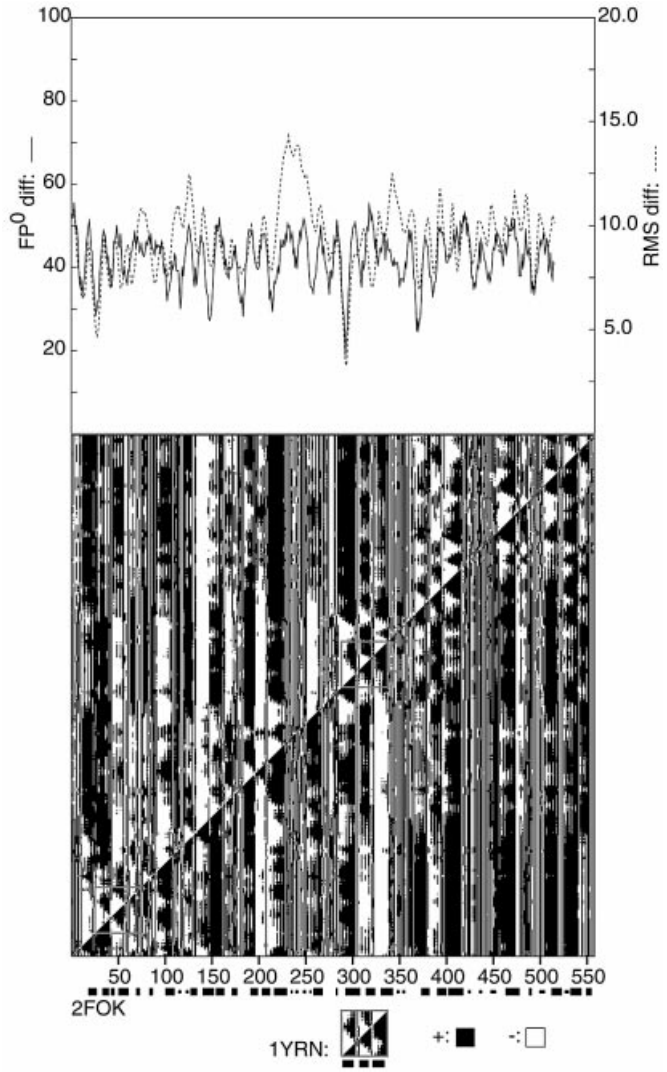

Fig. 1. The primary fingerprint matrices of chain A of the protein 1yrn and of the protein 1fok. The graphs above the fingerprints show both the fingerprint matrix percentage differences (full line) and the r.m.s.d. (in Å) of the overlay between the corresponding backbone segments (dotted line) as 1yrn is aligned with 1qav. The two best three-helix bundles detected in 1fok are enclosed in gray squares. The helix and sheet segments of both proteins are marked below the matrices by thick and thin black lines, respectively.

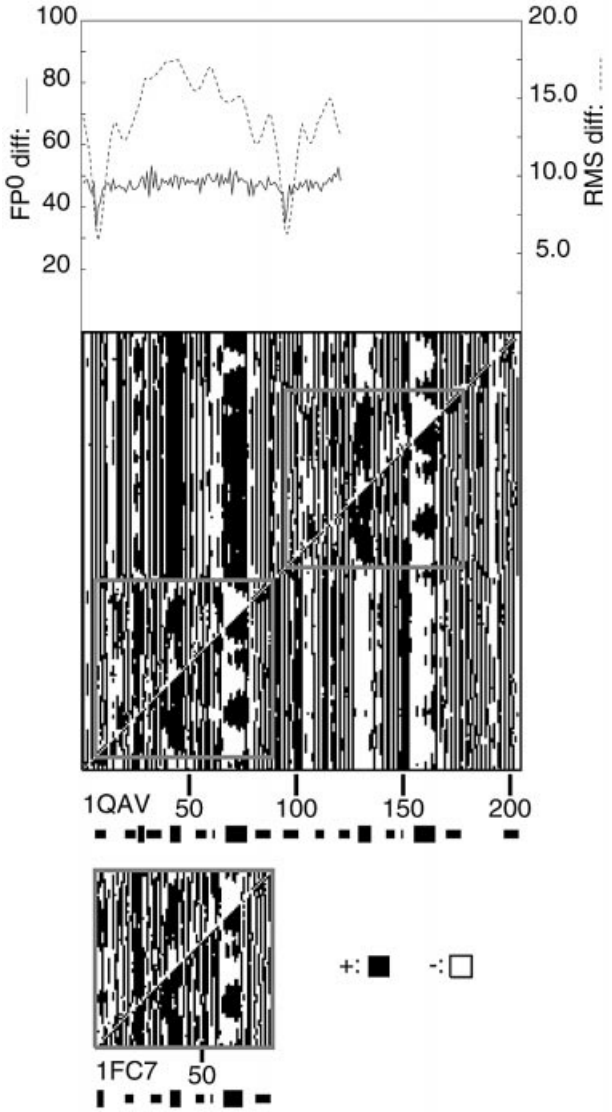

Fig. 2. The primary fingerprint matrices of the protein 1qav containing two PDZ domains and of the PDZ domain of the protein 1fc7. The graphs show both the fingerprint matrix percentage differences (full line) and the r.m.s.d. (in Å) of the overlay between the corresponding backbone segments (dotted line) as the PDZ domain of 2fc7 is aligned with 1qav. The PDZ domains detected in 1qav are enclosed in gray squares. The helix and sheet segments of both proteins are marked below the matrices by thick and thin black lines, respectively.

alignment position characterizes the ability of our fingerprint matrices to identify various folds reliably.

The calculations demonstrating these new concepts were performed with the Fortran-77 program PFP. The program is available at the URL http://inka.mssm.edu/~mezei/pfp.

Results

The fingerprint matrices were tested on proteins with know structures from the Protein Data Bank (PDB) (Berman *et al*., 2000). The first set of tests used a 49-residue long segment (chain A) of the mating type protein A-1 (PDB i.d.: 1yrn) that forms a three-helix bundle as the representative of that fold and tested several DNA-binding proteins that contained this motif. Invariably, the matchups with the lowest (backbone) r.m.s.d. were also the matchups with the smallest deviation between the fingerprint matrices $\mathbf{FP^0}$ of 1yrn and the respective submatrix of the larger protein. Figure 1 shows a comparison of the primary fingerprint matrices for the largest protein considered here, the nucleic acid recognition structure of restriction endonuclease foki (PDB i.d.: 2fok). The variations in the fingerprint match track the variation in the r.m.s.d.s. The lower the r.m.s.d., the better is the correlation and it is particularly good when the fingerprint deviation is <30%. Note that random alignment would result, on average, in 50% agreement between the fingerprint matrices.

A larger and more complex motif (including both helices and sheets) tested involved the PDZ domain. Figure 2 shows the comparison of the PDZ domain of the photosystem II D1 C-terminal processing protease (PDB i.d.: 1fc7) with the protein complex of α-1 synthropin and neuronal nitric oxide synthetase (PDB i.d.: 1qav). each member of which contains a PDZ domain. Again, the deviations in the primary fingerprint

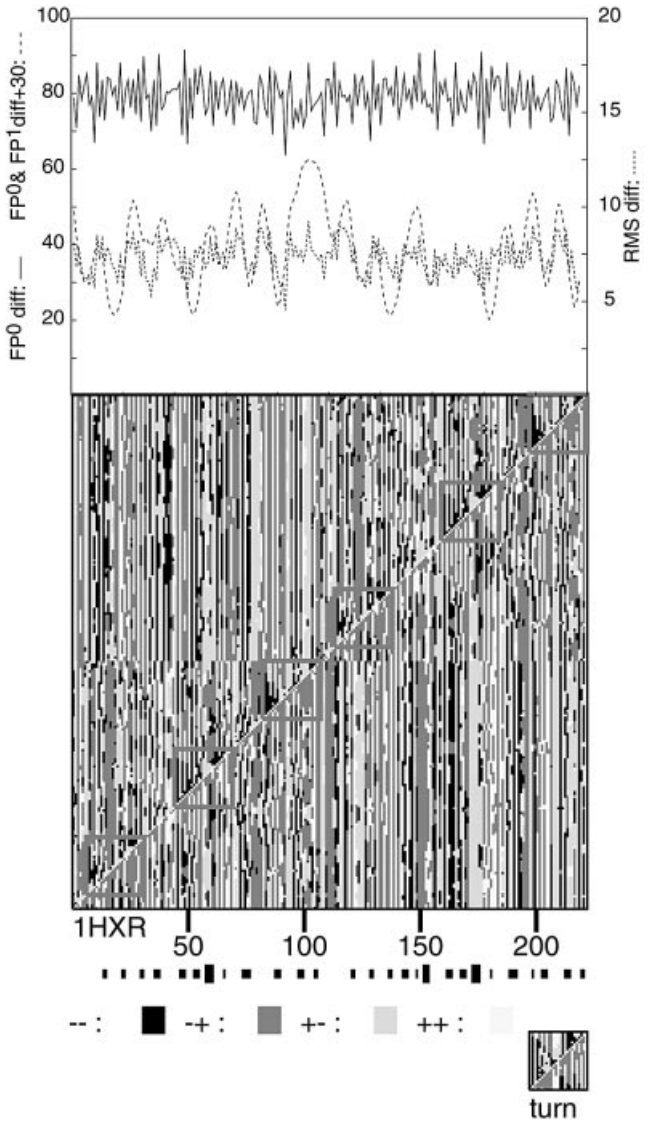

Fig. 3. The combined fingerprint matrices $\mathbf{FP^0}$ and $\mathbf{FP^1}$ of the simple turn and the protein 1hxr. The graphs above the fingerprints show the primary fingerprint matrix percentage differences (full line, shifted up by 30%), the combined fingerprint matrix differences (broken line) and the r.m.s.d. (in Å) of the overlay between the corresponding backbone segments (dotted line).

matrices track the r.m.s.d. well when the r.m.s.d. is not too high.

A test involving the recognition of a simple antiparallel β-sheet [residues 49–73 of chain A of the murine olfactory marker protein (PDB i.d.: 1f35)] in the protein guanine nucleotide exchange factor MSS4 (PDB i.d.: 1hxr) was also performed. The matches picked by the comparison of the fingerprint matrices $\mathbf{FP^0}$ all corresponded to the same motif, but the scores displayed an oscillating pattern, corresponding to the alternating black and white bars on the fingerprint matrix. Combination of $\mathbf{FP^0}$ and $\mathbf{FP^1}$ resulted similarly in picking the right motifs, but the oscillations in the combination score comparison were markedly reduced. Figure 3 shows the comparison scores for $\mathbf{FP^0}$ (full line, shifted upwards by 30%) and for the combination of $\mathbf{FP^0}$ and $\mathbf{FP^1}$ (dotted line) and the combined fingerprint matrices. The matches found by the minima of the fingerprint matches were also examined visually, and all matches corresponded to the same fold as the reference conformation. For the evaluation of Figure 3, it is important to point out that the relevance of the r.m.s.d. to describe similarities deteriorates as it becomes larger.

The possibility of obtaining a false positive was examined by comparing a mutant of apoliprotein-E2 (PDB i.d.: 1le2), a four-helix bundle with the chain A of a DNA-binding ferretin homolog (PDB i.d.: 1dps), another four-helix bundle with a crossover loop separating the first two helices from the rest, resulting in reversed directions of the last two helices. The corresponding two fingerprint matrices (not shown) displayed the dark triangles diagnostic of helices but in a markedly different pattern, resulting in comparison scores of the matrices in the 50% range (i.e. only as good as random).

Discussion

The results show that the novel fingerprint matrices defined in this communication are capable of recognizing different protein folds. This means that they can form the basis of structural comparison and fold detection algorithms. The major difficulty in this is the efficient handling of gaps. Work in this direction is in progress.

In addition to the inherent interest in being able to represent a fold of an n-residue domain with just $2n^2$ bits, there are several other attractive features of these fingerprints. First, it presents a description of the fold with very little redundancy, thereby increasing the robustness of comparisons based on it since redundancy allows different descriptions for identical objects. Secondly, they avoid the use of distances, allowing for the significant variations in the internal coordinates between members of a given fold. Thirdly, there is no cutoff involved, so the arbitrariness is reduced. Fourthly, it provides a two-dimensional representation of the fold that can be recognized visually or with the appropriate software – it is planned that a library of these representations will be generated and software for the recognition of these motifs is developed. Finally, even if the fold descriptor proposed here did not have any specific advantage over currently used ones, it can be used to complement existing ones in cases where there is some ambiguity in the comparison.

References

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Godzik,A. (1996) *Protein Sci.*, **5**, 1325–1328.
Kabsch,W. (1976) *Acta Crystallogr.*, **A32**, 922–923.
Shindyalov,I.N. and Bourne,P.E. (1998) *Protein Eng.*, **11**, 739–747.

Received November 15, 2002; revised July 25, 2003; accepted August 31, 2003