3D protein structures similarity matching based on fractal features

Chenyang Cui* a, Donghui Wang b, Xin Yuan b
a State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China, 310012
b College of Computer Science, Zhejiang University, Hangzhou, China, 310012

ABSTRACT

In this paper, we propose a new method for finding similarity in 3-D protein structure comparison. Different from the other existing methods, our method is grounded in the theory of fractal geometry. The proposed feature vectors of protein structures are invariant to the rotation, translation, scaling of the protein molecule, and it is simple to implement. The method is very fast because it requires neither alignment of the chains nor any chain-chain comparison. We calculate the fractal features of a set of 200 protein structures selected from PDB (Protein Data Bank). The experimental result shows that our method is very effective in classification of 3-D protein structures.

Keywords: 3-D protein structure comparison, fractal features, fractal dimension

1. INTRODUCTION

In protein database (e.g. PDB), each year a high number of new 3-D structures of protein molecules is determined either by x-ray crystallography or by NMR techniques or by theoretical prediction. 3-D protein structures comparison is becoming a necessity for the enormous growth in the number of known protein structures. To avoid potential exponential explosion of structures, a basic problem is to which class the new protein molecule belong. From a biologist’s point of view, sequence comparison alone can’t provide some required information [8]. Structure of biological molecules is a very important clue to understanding and manipulating biological function. Proteins belonging to the same functional family often share common structural features even if there is no evolutionary dependence or sequence similarity between them. So, the protein molecules can be classified by comparing its 3-D structures similarity [21, 22].

In recent years, many different approaches for extracting protein structural features are developed. Chen [1] extracts 3D structure features such as n order moments, and the ratios defined as height and depth divided by width and so on. Geometric hashing based algorithms [2] and index based algorithm [7] choose a set of reference frames from each target protein and place the other elements of the protein in a hash table, based on each reference frame, the space complexity of this approach is relative to the number of elements considered for each target protein. Tolga Can [3] uses three independent smoothing splines parameterized with respect to the polygonal arc length t to infer the global structure. Similarly, Peter Rogen [4] uses Gauss integrals to analyze and compare protein structure, this method is very fast and used in CATH2.4 database. A popular method is distance matrix [5] and used in FSSP database, in which the distance is defined as the one between two atoms. Mihael Ankerst [6,13] used a net-shaped to segment 3D protein and construct shape histogram. The CE algorithm [15] performs a combinational extension of aligned fragment pairs, it builds an alignment between two protein structures through a combinatorial extension of an alignment path defines by aligned fragment pairs. Alexandre [10] proposes a method HPM (Hybrid Protein Model) to compare protein 3D structure information and physicochemical properties. RMSD (Root Mean Square Distance) is an excellent measure of similarity for nearly identical structures, but once the shape of two proteins begins to diverge, RMSD looses it effectiveness. Two completely unrelated proteins maybe have a very large RMSD. In [17], an octree spatial decomposition of 3-D protein structures and a node graph is used to analyze the features of different regions of protein structures, but this method in sensitive to the rotation of the protein. The key problem in structural alignment of proteins is to find the optimal mapping between the atoms in two molecular structures.

The most of methods above mentioned is based on high dimension features, when performing a database search, these methods need exhaustive searching. They compare the query structure feature against the structures in the
database one by one and then report the ones most similar to the query structure, this cost very much time. Now, there is no universal agreement of the similarity of proteins, it is not easy to assess the results of the similarity retrieval systems to tell which one is the best [12]. The challenges outlines in the preceding section motivated us to step away from high dimension features and to instead compare and classify proteins on the basis of their fractal properties.

2. PROPOSED METHOD

2.1. Fractal Background

Fractals are not relegated exclusively to the realm of mathematics. It can be found virtually everywhere in the natural world, e.g. protein molecules. Proteins are heteropolymers with a variable composition of twenty different amino acids. The amino acid sequence shows that 3-D structure of the protein for the varied composition and nature of their side groups result in a range of possible interactions within the protein. These interactions determine the final structure of protein. So, proteins have an intrinsic self-similarity in the compactness and the packing of their structure. This is a simple form of fractal behavior but it has important consequences for the morphology of the protein and for thermodynamics of protein folding [18]. Whether natural or mathematical, all fractals have particular fractal dimensions. These are not the same as the familiar Euclidean dimensions, measured in discrete whole integers such as 1, 2, or 3, but a different kind of quantity. Usually non-integer, a fractal dimension indicates the extent to which the fractal object fills the Euclidean dimension in which it is embedded. A natural fractal of fractal dimension 2.8, for example, would be a sponge-like shape nearly 3-D in appearance. A natural fractal of fractal dimension 2.2 would be a much smoother object that just misses being flat [19]. A protein molecule is made up of one or several polypeptide chains and the protein backbone is a space curve composed of $C^\alpha$ atoms. In Figure 1, the backbones of protein molecule 1CMK and 2HHB are shown. This motivates us to consider the fractal features of the 3-D protein structures. The following section describes how to extract the fractal features of the protein in our methods.

![Fig. 1. The backbone of the protein molecules 1CMK and 2HHB](image)

2.2. Fractal Feature Extraction

The fractal dimension of a curve may be defined by measuring the length $L$, with rulers of fixed length $\varepsilon$. The relationship between the length $L$ and length $\varepsilon$ is:

$$L(\varepsilon) \propto \varepsilon^{-(1-D(\varepsilon))}$$ (1)

Here, $D(\varepsilon)$ is the fractal dimension. $L(\varepsilon)$ is defined as the length of the protein molecule chain from the $C^\alpha$ of $N$ polar in polypeptide chain, in which the $1^{\text{st}} C^\alpha$, $(\varepsilon+1)^{\text{th}} C^\alpha$, $(2\varepsilon+1)^{\text{th}} C^\alpha$, ....., and $(K\varepsilon+1)^{\text{th}} C^\alpha$ are connected to a zigzag, shown in Figure 2. $C^\alpha$ is represented as amino acid residue, the number of the amino acid...
residues is \(N\), if the number of residues \(n\) between \((Ke + 1)^{th}\) \(C^\alpha\) and \(C\) polar in polypeptide chain is less than \(\epsilon\), the connected zigzag stopped at the \((Ke + 1)^{th}\) \(C^\alpha\).

Then, the length of the protein molecule chain \(L(\epsilon)\) is defined as following:

\[
L(\epsilon) = L_\epsilon(\epsilon) + \frac{n}{\epsilon} \cdot \frac{L_\epsilon(\epsilon)}{K}
\]

(2)

where, \(1 \leq \epsilon \leq N - 1\), and \(L_\epsilon(\epsilon)\) is the length of the connected zigzag, and \(\frac{n}{\epsilon} \cdot \frac{L_\epsilon(\epsilon)}{K}\) is the length of the remainder whose \(C^\alpha\) is not connected. By the curve approximation, the slope of the fractal curve \(S(\epsilon)\) can be computed, and then

\[
D(\epsilon) = 1 - S(\epsilon)
\]

(3)

Obviously, \(D(\epsilon)\) is sensitive to the fixed length \(\epsilon\), if \(\epsilon\) is in certain region, the curve is linear. To improve the robust of \(D(\epsilon)\), the following fractal dimensions \(D_M(\epsilon)\) and \(D_F(\epsilon)\) can be defined:

\[
D_m(\epsilon) = \frac{1}{i} \sum_{\epsilon=1}^{i} D(\epsilon)
\]

(4)

where \(i\) is the number of the fractal dimension.

\[
D_F(\epsilon) = \left\{ \begin{array}{ll}
\frac{2}{N} \sum_{\epsilon=1}^{N/2} D_s(\epsilon), & N = 2n(n = 1, 2, \ldots) \\
\frac{2}{(N + 1)} \sum_{\epsilon=1}^{(N+1)/2} D_s(\epsilon), & N = 2n - 1(n = 1, 2, \ldots)
\end{array} \right.
\]

(5)

Here, \(D_s(\epsilon)\) is defined as:

\[
\left( \langle \epsilon^2 \rangle_N \right)^{D_s(\epsilon)} = A\epsilon, 1 \leq \epsilon \leq N - 1
\]

(6)

\[
\langle \epsilon^2 \rangle_N = \frac{1}{N - \epsilon + 1} \sum_{j=1}^{N-\epsilon} \langle \epsilon_{j, j+\epsilon}^2 \rangle_N
\]

(7)

\(\langle \epsilon_{j, j+\epsilon}^2 \rangle_N\) is the mean square distance between the \(i^{th}\) \(C^\alpha\) and the \(l+i^{th}\) \(C^\alpha\). So, we can compute three kinds of the fractal dimensions of the protein molecule \(D(\epsilon), D_m(\epsilon)\) and \(D_F(\epsilon)\), and use them as the features of 3-D protein structures. So, three features are extracted at most for each protein molecule.

Fig.2 The backbone and the zigzag of the protein molecule 159D
So, we can compute the fractal dimensions of the protein molecule $D_M(\varepsilon)$, and use it as the feature of 3-D protein structures.

3. EXPERIMENTS

In our experiments, we download about two hundred protein structures from PDB. Experiment results show that the fractal dimension features perform well in comparing 3-D protein structure similarity. See Figure 3, the backbones of several protein molecules are shown and the corresponding fractal dimensions are given in Table 1. Simultaneously, some protein molecules in [1, 20] are also given in Figure 2 and Table 1. Here, we only consider those protein molecules that have the single chains. The similarity between two protein molecules $i$ and $j$ can be computed by using the following equation:

$$\delta = |D_i(\varepsilon) - D_j(\varepsilon)|$$

(8)

![Image of protein structures](image_url)

Fig. 3. The backbone of some protein molecules

Where, $D_M(\varepsilon)_i$ and $D_M(\varepsilon)_j$ are the fractal dimension of protein molecule $i$ and $j$ respectively. $\delta$ is the difference between the fractal dimensions of any two protein molecules. Several $\delta$ are shown in Table 2. Experiment results show that if $\delta < 0.002$, two compared protein structures and function are similar, the fractal dimension is independent of the number of residues. In our experiment, two hundred proteins are classified to forty groups.

4. DISCUSSION

In our experiments, all protein molecules have only one polypeptide chain. Only the backbone composed of $C^\alpha$ is used to extract the fractal dimensions. Compared with the existing methods, the fractal dimension is easy to compute and low dimensional, and experiment result shows that the fractal dimension performs well in classifying 3-D protein molecules. To improve the accuracy of the classification, more fractal dimensions and better fractal dimensions should be defined and computed. In the future, all atoms of protein molecule will be used in extracting the fractal dimensions, furthermore, several polypeptide chains will be included in our research.
Table 1. The fractal dimensions of partial protein molecule

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue number</th>
<th>$D_M(\varepsilon)$</th>
<th>Ref.</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1lqn</td>
<td>207</td>
<td>1.948416</td>
<td>-</td>
<td>PROTEIN BINDING</td>
</tr>
<tr>
<td>1loq8</td>
<td>122</td>
<td>1.948392</td>
<td>-</td>
<td>RNA BINDING PROTEIN</td>
</tr>
<tr>
<td>2che</td>
<td>128</td>
<td>2.246084</td>
<td>[20]</td>
<td>SIGNAL TRANSDUCTION PROTEIN</td>
</tr>
<tr>
<td>1e6m</td>
<td>128</td>
<td>2.245989</td>
<td>[20]</td>
<td>SIGNALING PROTEIN</td>
</tr>
<tr>
<td>1rnl</td>
<td>200</td>
<td>2.246022</td>
<td>[20]</td>
<td>SIGNAL TRANSDUCTION PROTEIN</td>
</tr>
<tr>
<td>1e6k</td>
<td>130</td>
<td>2.245890</td>
<td>[20]</td>
<td>SIGNALING PROTEIN</td>
</tr>
<tr>
<td>1at8</td>
<td>105</td>
<td>2.924617</td>
<td>-</td>
<td>RIBONUCLEIC ACID</td>
</tr>
<tr>
<td>1auf</td>
<td>105</td>
<td>2.924620</td>
<td>-</td>
<td>RIBONUCLEIC ACID</td>
</tr>
<tr>
<td>1sun</td>
<td>765</td>
<td>2.937580</td>
<td>-</td>
<td>RIBONUCLEIC ACID</td>
</tr>
<tr>
<td>305d</td>
<td>23</td>
<td>2.712343</td>
<td>-</td>
<td>DEOXYRIBONUCLEIC ACID</td>
</tr>
<tr>
<td>267d</td>
<td>35</td>
<td>2.713475</td>
<td>-</td>
<td>DEOXYRIBONUCLEIC ACID</td>
</tr>
<tr>
<td>4mbn</td>
<td>153</td>
<td>2.468260</td>
<td>[1]</td>
<td>MYOGLOBIN (MET)</td>
</tr>
<tr>
<td>101m</td>
<td>154</td>
<td>2.467826</td>
<td>[1]</td>
<td>MYOGLOBIN MUTANT</td>
</tr>
<tr>
<td>5mbn</td>
<td>153</td>
<td>2.467734</td>
<td>[1]</td>
<td>MYOGLOBIN (DEOXY)</td>
</tr>
</tbody>
</table>

Table 2. The difference of fractal dimension between any protein molecule

<table>
<thead>
<tr>
<th>1&lt;sup&gt;st&lt;/sup&gt; Protein ID</th>
<th>$D_1(\varepsilon)$</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt; Protein</th>
<th>$D_2(\varepsilon)$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1lqn</td>
<td>1.948416</td>
<td>1.948392</td>
<td>0.000024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2che</td>
<td>2.246084</td>
<td>0.297668</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1at8</td>
<td>2.924617</td>
<td>0.976201</td>
<td></td>
</tr>
<tr>
<td></td>
<td>305d</td>
<td>2.712343</td>
<td>0.763927</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4mbn</td>
<td>2.468260</td>
<td>0.519844</td>
<td></td>
</tr>
<tr>
<td>1auf</td>
<td>2.924620</td>
<td>1.948416</td>
<td>0.976204</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2che</td>
<td>2.246084</td>
<td>0.678536</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1at8</td>
<td>2.924617</td>
<td>0.000003</td>
<td></td>
</tr>
<tr>
<td></td>
<td>305d</td>
<td>2.712343</td>
<td>0.212277</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4mbn</td>
<td>2.468260</td>
<td>0.45636</td>
<td></td>
</tr>
</tbody>
</table>

**REFERENCE**

3. Tolga Can, Yuan-Fang Wang. CTSS: A Robust and Efficient Method for Protein Structure Alignment Based on Local Geometrical and Biological Features. Proceedings of the computational Systems Bioinformatics(CSB’03)
10. Alexandre G.de Brevern, Serge A.Hazout. Hybrid Protein Model (HPM): a method to compact protein 3D-structure information and physicochemical properties

*ccy@cad.zju.edu.cn; phone 86 571 8795-1045; fax 86 571 8795-1670; State Key Lab. of CAD&CG, Zhejiang University.