Gene 663 (2018) 138–147

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Three-dimensional protein shape similarity analysis based on hybrid features

Zhong Li[a,*], Jiangjiang Yu[a], Hailong Hu[a], Shuiwang Ji[b]

[a] School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China
[b] School of Electrical Engineering and Computer Science, Washington State University, Pullman 99163, USA

ARTICLE INFO

Keywords:
3D structure
Similarity analysis
Skeleton
Salient geometric feature
Heat kernel signature
Tensor

ABSTRACT

The analysis of protein similarity is a matter of concern in the bioinformatics field, since studying the protein similarity can help understand the protein structure-function relationship. To this aim, several methods have been proposed, but currently, protein similarity results are still not satisfactory. Here we presented a novel method for evaluating the similarity of 3D protein models based on hybrid features, including the local diameter (LD), the salient geometric feature (SGF) and the heat kernel signature (HKS). LD is suitable to the topological deformation of 3D models, SGF is an important local feature on the protein model surface, and HKS is invariant under isometric deformations. Our method provides the improved feature extraction procedure to calculate LD, SGF and HKS of a protein model, and then uses these features to construct a tensor based feature descriptor for 3D protein models. The method finally analyzes the similarity of 3D protein models by using this tensor descriptor and the extended grey relation analysis. Experimental data indicated that our method is effective and can outperform the existing similarity analysis results obtained by previously reported methods.

1. Introduction

During the last years, protein similarity analysis has become a topic of growing interest in bioinformatics and computational biology, being helpful in understanding and revealing the structure and functions of proteins (Liu et al., 2009). With the development of cryoelectron microscopy and related techniques, we can easily convert 3D images or voxel data for the construction of 3D protein models. The analysis of the shape of these proteins plays an important role in medical research, computer-based molecular design, and protein structure retrieval and prediction.

Many methods for shape similarity analysis of 3D proteins have been proposed (Ballester and Richards, 2007). Some of these methods are based on topological features of 3D protein models. For example, Hilaga et al. (2001) compared the shape similarity on the basis of the multi-resolution Reeb graphs (MRGs) of 3D models. In this method, MRG is constructed as a continuous function based on the geodesic distance of 3D models. Although it can represent the skeleton and topology of 3D models at different resolution levels, the shape analysis results are not satisfactory when 3D models include some holes. Li et al. (2014) proposed a similarity analysis method applicable for proteins by improving the MRG algorithm. They first extracted the skeleton and calculated the local diameter (LD) of proteins, and then analyzed the protein similarity based on LD. In order to compare the similarity of protein models, which may be deformed by another protein model, Liu et al. (2009) used the Inner Distance Shape Signature (IDSS) to represent the 3D protein shape. The limitation is that the calculation of the IDSS is influenced by the topology changement in shape and the proteins with similar descriptors perhaps have no evolutionary relationship. Fang et al. (2009) proposed a method for comparing 3D protein shape similarity based on LD. However the LD descriptor is not sensitive to the topological variance of protein models and it cannot be used to compare the similarity of proteins with evident deformation.

Another group of methods for similarity analyses is based on the geometric feature on the surface of 3D models. For example, Yao et al. (2016) proposed a method based on the salient geometric feature. Osada et al. (2002) compared the shape similarity according to the shape distribution. They constructed a shape function by using the Euclidean distance for each pair of sampled vertices, and then formed a statistical histogram that measures the vertex distribution of the whole model surface, finally calculated the similarity by comparing two similar distances. Sael et al. (2008) introduced a global surface feature

Abbreviations: SGF, salient geometric feature; HKS, heat kernel signature; MRGs, multi-resolution Reeb graphs; LD, local diameter; EGI, Extended Gaussian Images; IDSS, Inner Distance Shape Signature; SI, shape index

* Corresponding author.
E-mail address: lizhong@zstu.edu.cn (Z. Li).

https://doi.org/10.1016/j.gene.2018.04.049
Received 3 October 2017; Received in revised form 25 March 2018; Accepted 18 April 2018
Available online 21 April 2018

representation of protein models based on 3D Zernike, and then provided a fast comparison method for the similarity of 3D proteins. This descriptor can conveniently represent the protein structure and can be used for the protein database retrieval. However, it is not applicable for 3D proteins with holes or when the genus of a 3D protein model is not 0.

The final group of methods for shape comparison is based on other features of 3D models. By using the Laplace-Beltrami operator, Reuter et al. (2005) proposed a spectral method, which possessed an isometry-invariant global geometric property, ignoring the local features of 3D models. Horn (1984) provided the concept of Extended Gaussian Images (EGIs) that could represent the features on the model surface. They calculated the extended Gauss vector by mapping the mesh of a 3D model surface onto the unit sphere, and then obtained the directional histogram. Min et al. (2002) introduced a 3D shape retrieval method, obtaining the shape signature by mapping the 3D shape to 2D images from multiple viewpoints. But this method could lack feature information, not effectively reflecting topological and local features of 3D models. Recently, the heat kernel signature (HKS) has been proposed and used in several applications involving deformable models (Au et al., 2008) because it is invariant under isometric deformations and thus reflects the intrinsic geometry of a model.

Considering the shape flexibility of protein models, only topological or local features of a protein shape are not suitable to capture all the detailed information of a protein shape. Indeed, here we proposed a protein shape comparison method based on hybrid features including LD, SGF and HKS. Furthermore, we constructed a tensor-based grey relation analysis descriptor for the similarity analysis of 3D proteins. Experimental comparison indicated that our method could obtain satisfactory results in comparing protein shape similarity.

2. Methods

Firstly, we extracted the skeleton of 3D protein models, and then calculated as reported below the local diameter (LD), which is used to reflect the topological feature of 3D proteins. Secondly, we calculated the salient geometric feature (SGF) to capture the local surface features of 3D proteins. Being the heat kernel signature (HKS) stable under perturbations of the shape and invariant under isometric deformations, we combined this feature with LD and SGF to obtain the hybrid feature of 3D proteins. We furthermore constructed a tensor-based descriptor of 3D proteins by using the above-mentioned three features. Finally, we extended the grey relation analysis method and combined our tensor based descriptor for the similarity analysis of 3D proteins. The global process on the basis of our method is schematized in Fig. 1.

2.1. Skeleton extraction and LD description of 3D protein models

2.1.1. Skeleton extraction from protein models

The skeleton of 3D models can represent their topological properties. It is useful in many shape analysis applications, such as shape retrieval, feature matching, and others (Tierny et al., 2006). Au et al. (2008) proposed a robust skeleton extraction method based on the mesh contraction, which directly works on the mesh domain. However, some results would be unsatisfactory when extracting the skeleton of protein models containing local surface features and blurred topological features. Hilaga et al. (2001) proposed a method based on multi-resolution Reeb graphs (MRGs) that, although representing the skeleton and topology of a 3D model at different resolution levels, resulted helpless in the case of objects with holes or whose genus is not equal to 0. Li et al. (2014) provided a method by improving the MRG algorithm. But it needs to calculate the geodesic distance and is time consuming for large protein models.

For a rapid analysis of the shape similarity among protein models, we provided a simple and effective method for extracting the skeleton of 3D proteins. Given a protein mesh S with n vertices, P is a matrix of size $n \times 3$, which represents the coordinates of vertices. The inner normal vectors of vertices can be calculated and the contraction entry is defined as

$$\overline{P} = P + WN \tag{1}$$

where N is a matrix of size $n \times 3$, each row of it represents an inner normal vector. W is a diagonal matrix, $W(i,i)$ is the weight of contraction that the vertex i moves along the inner normal which is calculated by

$$W(i,i) = \lambda \cdot \frac{\sum_{i=1}^{n} \sum_{p=1}^{k(i)} A(i,p)}{n} \cdot \sum_{j=1}^{m(i)} d(i,j)/m(i) \tag{2}$$

where λ is a constant and we usually set $\lambda = 2$, $A(i, p)$ is the area of the triangle p associated with vertex i, $k(i)$ represents the number of triangles associated with vertex i, $m(i)$ represents the number of vertex associated with vertex i, $d(i, j)$ is defined as the Euclidean distance from vertex i to vertex j. The process of the mesh contraction produce the corresponding rough mesh, we then smooth the obtained mesh by

$$\overline{P_i} = \sum_{j=1}^{m(i)} \overline{P_i(j)}/m(i) \tag{3}$$

where $\overline{P_i(j)}$ represents the coordinate of vertex j associated with the vertex i.

The process reported above can be regarded as one iteration of skeleton extraction. Normally after implementing three iterations, the protein mesh model will be contracted to be a new skeleton preserving the original topological connection. We then cluster the points on the mesh skeleton with k-means algorithm (Au et al., 2008) to get the linear skeleton. Our method is simple and convenient for the skeleton extraction of protein models. In Fig. 2a, it is reported the linear skeleton of protein 2WRP obtained by using our method. The new skeleton contains more details compared to the skeleton obtained by the method from Li et al. (2014) (Fig. 2b). For more skeleton extraction and comparison, we can see Fig. 7 in the following experimental section.

2.1.2. Local diameter (LD)

The local diameter (LD) of a 3D shape is defined as the distance from one vertex to the opposite vertex on the surface of an object (Fang et al., 2009). LD is approximately equal to two times of the shape radius (Choi et al., 1997), which can be defined as the distance between a point on the surface and the axis or skeleton of the model. Here, we used the above-indicated skeleton information to calculate LD of each vertex on the surface of a protein model, reflecting the topological feature of a 3D protein model.

2.2. Local surface feature of 3D protein shape

With the increasing of protein database, there exist some protein models with similar topological features while containing different function (Fang et al., 2009). Skeleton feature and local diameter are not enough to collect complete information about protein models. Other local geometric features are needed to analyze the similarity of protein surfaces. There are many methods based on different local surface feature descriptions, such as various representations of curvature, the shape index (Bradford and Westhead, 2008), the salient geometric feature (Yao et al., 2016; Gal and CohenOr, 2006), the shape diameter function (Shapira et al., 2008) and others.

The curvature of 3D models can effectively reflect the local geometric feature on the model surface. To strengthen the local feature of models, Hoffman et al. (Gal and CohenOr, 2006) constructed the salient geometric feature (SGF) based on the curvature feature. Compared to the curvature feature, the shape index (SI) is introduced to express the concave or convex information of models (Bradford and Westhead, 2008), which is described as