Computational and Structural Biotechnology Journal xxx (2017) xxx–xxx

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Protein Structure Classification and Loop Modeling Using Multiple Ramachandran Distributions☆

Seyed Morteza Najibi[a], Mehdi Maadooliat[b], Lan Zhou[c], Jianhua Z. Huang[c], Xin Gao[d,*]

[a]Department of Statistics, Persian Gulf University, Bushehr 75169, Iran
[b]Department of Mathematics, Statistics and Computer Science, Marquette University, WI 53201-1881, USA
[c]Department of Statistics, Texas A&M University, TX 77843-3143, USA
[d]Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Article history:
Received 30 October 2016
Received in revised form 26 January 2017
Accepted 28 January 2017
Available online xxxx

Keywords:
Bivariate splines
Log-spline density estimation
Protein structure
Ramachandran distribution
Roughness penalty
Trigonometric B-spline
Protein classification
SCOP

ABSTRACT

Recently, the study of protein structures using angular representations has attracted much attention among structural biologists. The main challenge is how to efficiently model the continuous conformational space of the protein structures based on the differences and similarities between different Ramachandran plots. Despite the presence of statistical methods for modeling angular data of proteins, there is still a substantial need for more sophisticated and faster statistical tools to model the large-scale circular datasets. To address this need, we have developed a nonparametric method for collective estimation of multiple bivariate density functions for a collection of populations of protein backbone angles. The proposed method takes into account the circular nature of the angular data using trigonometric spline which is more efficient compared to existing methods. This collective density estimation approach is widely applicable when there is a need to estimate multiple density functions from different populations with common features. Moreover, the coefficients of adaptive basis expansion for the fitted densities provide a low-dimensional representation that is useful for visualization, clustering, and classification of the densities. The proposed method provides a novel and unique perspective to two important and challenging problems in protein structure research: structure-based protein classification and angular-sampling-based protein loop structure prediction.

1. Introduction

Proteins are large biomolecules or macromolecules that perform a vast array of functions for the biological processes within the cell of organisms. A protein is a linear chain of amino acids, each of which is composed of an amino group ($-NH_2$), a central carbon atom (C_α), a carboxyl group (–COOH), and a side-chain group that is attached to C_α and is specific to each amino acid. Depending on the amino acid sequence (different amino acids have different biochemical properties) and interactions with their environment, proteins fold into a three-dimensional structure, which allows them to interact with other proteins and molecules to perform their function. Hence, an important topic in the field of structural biology is the determination of the three-dimensional (3D) structure of a protein. In a protein, each amino acid is called a residue and the chain of carbon, nitrogen and oxygen atoms are referred to as the backbone. While the side-chain structures determine local structures and interactions of the amino acids of the protein, the backbone structure determines the overall shape of the protein and is the focus of much research.

The backbone conformation of proteins can be represented equivalently by Cartesian coordinates of carbon, nitrogen and oxygen atoms, or the backbone dihedral angles (ϕ,ψ), and ω, with the assumption of standard bond lengths and angles. Moreover, the global folds of proteins can be equivalently represented by either the Cartesian coordinates of C_α traces or the 2 pseudo-angles (θ,τ) between the two consecutive planes formed by 4 successive C_α. The Ramachandran plot, a scatter plot of ϕ vs. ψ, can reflect

☆ The first two authors, Najibi and Maadooliat, made equal contributions to the paper.

* Corresponding author.
E-mail addresses: najibi@pgu.ac.ir (S.M. Najibi), mehdi@mscs.mu.edu (M. Maadooliat), lzhou@stat.tamu.edu (L. Zhou), jianhua@stat.tamu.edu (J.Z. Huang), xin.gao@kaust.edu.sa (X. Gao).

http://dx.doi.org/10.1016/j.csbj.2017.01.011

the allowed regions of conformational space available to protein chains. By analogy to Ramachandran's concept of dihedral angles, the pseudo-Ramachandran plot, a scatter plot of θ vs. τ, can provide a distinctive classification of protein structures and largely contribute to different applications [1].

In the development of protein tools over the last two decades, the angular representation of proteins and Ramachandran plots have been applied in various protein structure-related problems, such as protein structural model checking [2–4], structure prediction [5–9], model quality assessment [10–12], prediction server ranking [13, 14], protein structure alignment [15, 16], free energy function learning [17–19], molecular dynamics simulation [20], empirical energy functions [21] and classification functions such as backbone-dependent rotamer library [22, 23].

Since the seminal work of Ramachandran et al. [24], the two-dimensional histogram of Ramachandran plot has been commonly used to determine accessible regions and validate new protein structures [2, 3]. The histogram is a rough non-parametric density estimation where the number of parameters is equal to the number of data points. Furthermore, because of the circular nature of the protein angles, the traditional parametric or non-parametric density estimation methods cannot be used for estimating Ramachandran distributions. In the last decade, novel parametric and non-parametric methods have been introduced to address this problem. The parametric methods propose to use directional distributions such as von Mises distribution or short Fourier series that are naturally designed for periodic data [25–29]. On the other hand, the non-parametric techniques use kernel density estimates with periodic kernels, Dirichlet process with boundary modification, or a mixture of directional distributions [30–32].

Depending on the purpose of the study, one may produce Ramachandran plots based on residues associated with some specific amino acids, and/or some specific structural elements. In some cases, the number of residues (data points) is too small, and that makes it challenging to obtain reliable bivariate densities using techniques that estimate each Ramachandran distribution separately. An intuitive solution to this problem is to borrow information from a group of Ramachandran plots that has some common features. To this end, Lennox et al. [33] proposed a hierarchical Dirichlet process technique based on bivariate von Mises distributions that can simultaneously model angle pairs at multiple sequence positions. This method is typically used for predicting highly variable loop and turn regions. Ting et al. [34] and Joo et al. [35] also used this technique with further modification to produce near-native loop structures. In another approach, Maadooliat et al. [36] proposed a penalized spline collective density estimator (PSCDE) to represent the log-densities based on some shared basis functions. This method showed some significant improvements for loop modeling of the hard cases in a benchmark dataset where existing methods do not work well [36].

Comparing to other competitive approaches, PSCDE is more efficient in estimating the densities in the sparse regions by incorporating the shared information among the distributions. In this technique, the bivariate log-densities are represented using a common set of basis functions. Each log-density has its own coefficient vector in the basis expansion, and it can be used for clustering and classification of the densities. Furthermore, using a common set of basis functions significantly reduces the number of parameters to be estimated. This method has been applied to estimate the neighbor-dependent Ramachandran distributions to make the angular-sampling-based protein structure prediction more accurate. In this paper, we make an innovative and constructive development over the PSCDE method.

The PSCDE method is constructed based on Bernstein-Bézier spline basis functions defined over triangles to estimate the log-densities in a complex domain [36]. In simple words, in PSCDE, we artificially extended the constraints of the adjacent triangles to the triangles in boundaries in order to estimate the densities in a two-dimensional circular domain. Here, we propose an alternative approach that uses the tensor product of trigonometric B-spline basis to handle the angular nature of the data. The main advantage of the proposed method is that there is no need to implement any further constraints to take into account the continuity and circularity of the data since the new bases are trigonometric functions that are smooth and intrinsically periodic. Another improvement in the proposed procedure is on selecting the smoothing parameter. In the existing PSCDE procedure, the tuning parameter is selected using the Akaike Information Criterion. Therefore a grid search is needed to choose the optimal tuning parameter and that could become time-consuming, especially if different tuning parameters are used for different basis functions. Following Schellhase and Kauermann [37], we propose to update the smoothing parameter within the Newton–Raphson iterative procedure that is used for the density estimation.

The PSCDE method is originally applied to the protein loop modeling problem. Here, we focus on a new application and use an extension of PSCDE to the protein structure classification problem. There is a large literature on the classification of the protein structures in the Protein Data Bank (PDB) [38–40]; because a good classification can reveal the evolutionary relationship between the proteins and step toward understanding the protein functions. While a vast majority of the literature deals with the protein classification in a pairwise structural comparison framework, the proposed estimated densities can be used as an alternative technique based on angular representation for the structural classification.

Specifically, the estimated angular density corresponding to a protein structure has a basis expansion whose coefficients can be used as an input to a clustering algorithm. Furthermore, most of the existing techniques for protein classification are using sequence and/or 3D structure comparison to classify the proteins based on some (dis)similarity scores obtained after pairwise alignments. The proposed method is an alignment-free procedure that provides a vector of coefficients (i.e. features), associated with each structure (density), that can be directly used to classify the proteins.

We also applied the proposed method to the loop modeling problem and compared the result with the other methods in the online supplementary. In this application, we trained the neighbor-dependent distributions of the backbone dihedral angles (i.e., neighbor-dependent Ramachandran distributions) using the new collective density estimation approach and fed the results into the Rosetta loop modeling procedure to study the accuracy and efficiency of the Rosetta server in predicting the loop regions. The main concern of using the neighbor-dependent Ramachandran distributions is that we are partitioning the data into smaller groups, some partitions may end up with a limited number of observations, and therefore we may lose accuracy in estimating the Ramachandran distributions due to the data sparsity. The proposed collective estimation procedure can overcome this difficulty and thereby improve the accuracy of the estimated densities. We encourage the interested readers to read the online supplementary materials for the implementation of the proposed method on loop-modeling application.

The rest of the paper is organized as follows. Section 2 introduces the penalized spline collectively density estimator procedure based on the new trigonometric basis functions to incorporate the circular nature of data. Section 3 presents the protein structure classification problem and the implementation of the new procedure for this application. Section 4 concludes the paper with a discussion. A web-based toolbox is also introduced in the Appendix to illustrate the advantages of the proposed technique. This toolbox can be used further by the research community to obtain the collective estimation of Ramachandran distributions for any other related application (e.g. backbone-dependent rotamer library [22, 23]).

Please cite this article as: S.M. Najibi et al., Protein Structure Classification and Loop Modeling Using Multiple Ramachandran Distributions, Computational and Structural Biotechnology Journal (2017), http://dx.doi.org/10.1016/j.csbj.2017.01.011