Computational Biology and Chemistry 73 (2018) 105–119

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Guiding exploration in conformational feature space with Lipschitz underestimation for ab-initio protein structure prediction

Xiaohu Hao, Guijun Zhang*, Xiaogen Zhou

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

ARTICLE INFO

Article history:
Received 13 April 2017
Received in revised form 25 January 2018
Accepted 3 February 2018
Available online 6 February 2018

Keywords:
Ab-initio
Lipschitz underestimation
Rosetta
Differential evolution
Fragment assembly

ABSTRACT

Computing conformations which are essential to associate structural and functional information with gene sequences, is challenging due to the high dimensionality and rugged energy surface of the protein conformational space. Consequently, the dimension of the protein conformational space should be reduced to a proper level, and an effective exploring algorithm should be proposed. In this paper, a plug-in method for guiding exploration in conformational feature space with Lipschitz underestimation (LUE) for ab-initio protein structure prediction is proposed. The conformational space is converted into ultrafast shape recognition (USR) feature space firstly. Based on the USR feature space, the conformational space can be further converted into Underestimation space according to Lipschitz estimation theory for guiding exploration. As a consequence of the use of underestimation model, the tight lower bound estimate information can be used for exploration guidance, the invalid sampling areas can be eliminated in advance, and the number of energy function evaluations can be reduced. The proposed method provides a novel technique to solve the exploring problem of protein conformational space. LUE is applied to differential evolution (DE) algorithm, and metropolis Monte Carlo(MMC) algorithm which is available in the Rosetta; When LUE is applied to DE and MMC, it will be screened by the underestimation method prior to energy calculation and selection. Further, LUE is compared with DE and MMC by testing on 15 small-to-medium structurally diverse proteins. Test results show that near-native protein structures with higher accuracy can be obtained more rapidly and efficiently with the use of LUE.

1. Introduction

Proteins play a vital role as a class of macromolecules present in all biological organisms. They form the basis of cellular and molecular life and significantly affect the structural and functional characteristics of cells and genes (Gu and Bourne, 2009). Determining the structure and biological activity state of a protein molecule is significant to facilitate further understanding and curing of many diseases caused by changes in protein structure. Predicting the 3D structure of protein molecules has become a major research topic and an important task in the current bioinformatics, which directly predicts the protein tertiary structure from amino acid according to the principle of Anfinsen (Anfinsen, 1973). With the complexity and high dimensionality of protein conformational space, the problem of gaining near-native protein conformations by computing becomes NP-Hard. To solve the bottleneck problem in ab-initio protein structure prediction, an algorithm efficiently exploring for the conformational space needs to be developed (Kim et al., 2009).

As it predicts the structure with the use of only sequence information, ab-initio protein structure prediction is of great significance for protein molecular design and protein folding research. A variety of methods have been developed for ab-initio protein structure construction (Xu and Zhang, 2013), ranging from atomic-level molecular dynamic simulation (Brooks et al., 1983; Case et al., 1997) to reduced-level physics-based (Liwo et al., 1993; Klepeis and Floudas, 2003) and knowledge-based (Simons et al., 1997; Xu and Zhang, 2012) Monte Carlo (MC) assembly, to topology-level fold enumeration (Taylor et al., 2008) and to residue-contacts constrained conformational reconstruction (Wu et al., 2011; Marks et al., 2011). The Rosetta (Leaver et al., 2011) developed by the research group of Baker, and QUARK (Xu and

* Corresponding author.
E-mail address: zgj@zjut.edu.cn (G. Zhang).

https://doi.org/10.1016/j.compbiolchem.2018.02.003
1476-9271/

Zhang, 2012) by Zhang are ab-initio methods that performed well in previous CASP versions and have become internationally advanced ab-initio prediction methods.

Owing to the Lennard-Jones energy term in protein molecule, the local minima solutions in the conformational space quadratically grow with an increase in sequence length. Thus the energy surface becomes too rugged to explore, due to the increased energy barriers in conformational space. As a result, the time for computing a conformation also quadratically grows with chain length increase. Thus, the exploring process needs a large consumption of computing resource in evaluating the energy of generated conformations. However, not every conformation generated in the exploring process is helpful for finding the near-native conformations. How to exclude the invalid exploring regions and meaningless conformations in advance, save the evaluation times and lead the exploring to a more potential area, seems to be an important work in protein structure prediction.

ACUE (Hao et al., 2016) proposed in our previous work is an independent algorithm that could help save the function evaluations. Abstract convex underestimate technique (Zhou et al., 2016 a) is mainly used, which is a deterministic global optimization method originally mainly used for solving the integer programming problem. It achieves the asymptotic convergence of the upper and lower estimate bounds of the original optimization problem by constructing a series of convex relaxed models, and the upper and lower estimate bounds eventually converge to the global optimal solution of the original optimization problem. The abstract convex theory (Andramonov et al., 1999) provides strong theoretical support to deterministic global optimization (Rubinov, 2000).

LUE, which is a new plug-in method for guiding exploration in conformational feature space with Lipschitz underestimation for ab-initio protein structure prediction, is proposed on the basis of the Lipschitz estimation theory. Unlike ACUE (Hao et al., 2016), more accurate underestimation of the original problem can be obtained with neither model transformation nor sample training needs with the use of Lipschitz method (Rubinov and Andramonov, 1999) in LUE, and it can be ported to other prediction algorithms. The most significant contributions of the proposed LUE are as follows: the use of tight lower bound estimate information for exploration guidance, the advance elimination of invalid sampling area, and the reduction of the number of evaluations as a consequence of using underestimation model. LUE provides a novel technique to solve the exploring problem of protein conformational space. In this study, LUE is applied to DE (Storn and Price, 1997) algorithm, and MMC (Leaver et al., 2011) algorithm which is available in the Rosetta; When LUE is applied to DE and MMC, it will be screened by the underestimation method prior to energy calculation and selection. Further, LUE is compared with DE and MMC by testing on 15 small-to-medium structurally diverse proteins. Test results show that near-native protein structures with higher accuracy can be obtained more rapidly and efficiently with LUE.

2. Preliminary

2.1. Differential evolution

Differential evolution (Storn and Price, 1997) is a population-based stochastic search algorithm for global optimization, which has been successfully used for protein structure prediction problems (Hao et al., 2016, 2017; Xin et al., 2010; Zhou et al., 2016b; Zhang et al., 2017). Basic DE is conducted as follows. An initial population with N individuals is randomly sampled from the feasible solution space Ω. During each generation, mutation operation is performed for each target individual, followed by crossover operation. Then, trial individuals are generated. Finally, an updating operation is performed to determine whether a target individual or a trial individual will survive the next generation. DE outputs the last population until the terminating condition is achieved.

2.2. Metropolis Monte Carlo

Metropolis Monte Carlo (MMC) algorithm is very widely used in the field of protein structure prediction. In general MMC algorithm, in obtaining a new conformation through fragment assembly, the new conformation is accepted according to the Boltzmann probability density function at a given temperature, T, by calculating the energy difference value, ΔE. The process runs iteratively until the lowest energy conformation is obtained. Specifically, as expressed by (1):

$$P_{MMC}(new \leftarrow old) = \begin{cases} 1, & \Delta E < 0 \\ e^{-\beta \cdot \Delta E}, & \Delta E > 0 \end{cases} \quad (1)$$

If the energy of new conformation is lower than that of the former conformation, the new one will be accepted directly. Otherwise, it is accepted with probability $e^{-\beta \cdot \Delta E}$, where ΔE is the difference in energy after fragment assembly and β is a temperature-scaling factor, $\beta = 1/k_B T$, where k_B is the Boltzmann constant. In this work, β is chosen according to the PyRosetta application program interface that the temperature is close to room temperature.

2.3. Fragment assembly

Fragment assembly technique is an important measure for protein structure prediction owing to the high-dimensionality feature of protein conformational space (Xu and Zhang, 2013). The introducing of fragment assembly can greatly reduce the exploring space of protein conformations and significantly improve computing speed, and more reasonable three-dimensional protein structure can be formed. Generally, when purely optimization method is used, only the considered stable conformation of polypeptide chain with sequence length is about 5–20 can be obtained, whereas with fragment assembly technique, a high-accuracy prediction structure with a sequence length of 150 can be obtained (Xu and Zhang, 2013, 2012). Thus, fragment assembly is also adopted in our proposed method in this manuscript. The process of fragment assembly can be briefly described as follows: target sequence is divided into a number of continuous sections, called fragments, then, fragments that fit local known protein structures are searched through sequence alignments and are further assembled into a complete target structure.

2.4. Ultrafast shape recognition

Ultrafast shape recognition (USR) is originally presented for molecular shape comparison, capable of screening billions of compounds for similar shapes using a single computer and without the need of aligning the molecules before testing for similarity (Ballester and Richards, 2007). Shehu introduced USR for simplifying a protein conformation for the first time in FeLTr (Shehu, 2009). The USR method is used in following way in this manuscript: Based on the protein coarse-grained representation model, USR first extracts four key Cα-atoms of the protein coarse-grained representation model, the centroid Cα-atom (CTD), the nearest Cα-atom from CTD (CST), the farthest Cα-atom from CTD (FCT), and the farthest Cα-atom from FCT (FTF); The CTD, CST, FCT, and FTF Cα-atoms can express the structural feature of a conformation; Then, the distance value between CTD and origin