C-QUARK: contact-assisted *de novo* protein structure prediction on XSEDE Comet

Chengxin Zhang (zcx@umich.edu), Robin Pearce (robpearc@umich.edu), Yang Zhang (zhng@umich.edu)
Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

INTRODUCTION

Protein structure prediction, which infers the three-dimensional (3D) structures of proteins from their amino acid sequences, is an unsolved problem in structural bioinformatics. *De novo* structure prediction, which aims to fold proteins without using homologous structure templates, is the most challenging, yet generalizable, approach to solving the protein folding problem. Due to the huge conformational space that needs to be explored [1], *de novo* structure prediction, which is implemented using Simulated Annealing (SA) or Replica-Exchange Monte Carlo (REMC) simulations, typically takes hundreds of CPU core hours for a medium size protein with up to several hundred amino acid residues [2-4]. To address the challenging problem of *de novo* structure prediction, we developed a new server, C-QUARK, as an extension of our earlier QUARK [3] structure prediction pipeline. C-QUARK significantly reduced the conformational search space and improved modeling accuracy by incorporating predicted residue-residue contacts, i.e., residue pairs predicted to be in close proximity in 3D space. As an Extreme Science and Engineering Discovery Environment (XSEDE) science gateway, the extensive REMC simulations employed by C-QUARK are parallelized on shared CPU nodes on the XSEDE Comet supercomputer cluster for rapid job processing. C-QUARK is freely available to academic users at https://zhanglab.ccmb.med.umich.edu/C-QUARK/.

METHODS

The C-QUARK pipeline consists of four stages: (i) contact prediction, (ii) local structure feature and fragment generation, (iii) REMC simulation, and (iv) structure clustering and refinement. Stage 1 first constructs a profile from sequences that are homologous to the query. From the sequence profile, direct coupling analysis is used to estimate the evolutionary coupling parameters between every pair of residue positions in the query, and these parameters are used as input features by a deep convolutional neural network to predict the residue-residue contacts [5]. Next, in Stage 2, local structural features, including secondary structure and backbone torsion angles, are predicted by traditional (shallow and fully connected) neural networks from the sequence profile [6]. Then, structure fragments ranging in size from 1 to 20 residues are generated for each query residue position based on the predicted local structure similarity to experimental structures in the PDB database. Stage 3 assembles these fragments into full-length structures using REMC simulations guided by a composite force field that consists of knowledge-based energy terms derived from PDB statistics, together with the residue-residue contacts and other local structural restraints predicted in Stage 1 and 2. Stage 3 is the most time-consuming stage and needs 5 parallel simulation runs with different random seeds. Since communication is not required among different runs, these runs can be easily parallelized by submitting 5 individual jobs to "shared" computing nodes on COMET. Finally, Stage 4 clusters the thousands of conformations generated from the Stage 3 simulation trajectories based on their structural similarity [7]. The centroid of the largest cluster is refined by a short molecular dynamics simulation [8] to generate the final structure model.

RESULTS

C-QUARK was tested in the CASP13 community-wide protein structure prediction challenge as "QUARK" server. Compared to an earlier version of QUARK that did not utilize contacts predicted by deep learning, C-QUARK increased the number of targets with correct topologies (TM-score>=0.5 [9]) from 68 to 97, even though the number of parallel simulation runs was decreased from 10 to 5. Thanks to Comet parallelization, server jobs can typically be completed within 48 hours.

ACKNOWLEDGMENTS

We thank Dr. Baoji He and Dr. S. M. Mortuza for assistance in the early stage of C-QUARK development. This work used XSEDE, which is supported by National Science Foundation grant number ACI1548562.

REFERENCES

[1] C. Levinthal, J Chim Phys Pcb **65**, 44 (1968).
[2] S. Ovchinnikov, et al., Proteins **86**, 113 (2018).
[3] D. Xu and Y. Zhang, Proteins **80**, 1715 (2012).
[4] W. Zheng, et al., Future Gener Comput Syst **99**, 73 (2019).
[5] Y. Li, et al., Bioinformatics, btz291 (2019).
[6] S. Wu and Y. Zhang, Plos One **3** (2008).
[7] Y. Zhang and J. Skolnick, J Comput Chem **25**, 865 (2004).
[8] J. Zhang, et al., Structure **19**, 1784 (2011).
[9] Y. Zhang and J. Skolnick, Proteins **57**, 702 (2004).