Does the validation tool in the Protein Data Bank accurately evaluate cryo-EM proteins?

Authors: Stefan C Atkinson, Kevin J Eppes, Department of Biology

Advisors: Lin Chen, Anurag Dasgupta, Department of Computer Science

Abstract

Background: Cryo-electron microscopy (cryo-EM) has become an essential technique for protein structure determination. In 2019, more than 1000 protein structures generated from cryo-EM technique were deposited into the Protein Data Bank (PDB). Though each of these structures has been validated by the validation system of the PDB at the time of deposition, many of them have been questioned for containing suspicious outlier conformations which might be introduced during the modeling process. *Methods:* We trained two unsupervised machine learning models, histogram-based outlier score (HBOS) and isolated forest (IF), with 9131 high-resolution (<=1.5 Å) protein structures generated from X-ray. The residues labelled as outliers by the two models are retrieved from the validation reports at PDB. *Results:* We noted that many residue conformations identified as reasonable in the validation reports were labeled as outliers by both HBOS and IF models. *Conclusion:* The current validation system designed for X-ray proteins at PDB does not work well on cryo-EM proteins.

Datasets

The training dataset, X-ray-1.5, contains 9131 protein structures that are derived from X-ray data with resolutions better than 1.5 Å. At the resolution 1.5 Å, major atoms in a protein are well identified. The protein structures in X-ray-1.5 were downloaded from wwPDB website [1] in March 2018 with a sequence similarity of less than 90%. RCSB PDB is a member of the wwPDB.

A cryo-EM protein dataset, EM-0-4-2019, was used as the test dataset. EM-0-4-2019 contains 1175 atomic structures derived from cryo-EM density maps with 0-4 Å resolution that are released before March 31, 2016 and those between April 1, 2018 and December 31, 2019. Due to the low quality of cryo-EM density maps, the structures in EM-0-4-2019 have the highest quality of the cryo-EM structures in PDB.

Table 1. X-ray atomic structure dataset and cryo-EM atomic structure dataset used for training and test.

Dataset	Resolution	Number of Proteins	Released Time
X-ray-1.5	< 1.5	9131	before Mar. 31, 2018
EM-0-4-2019	< 4.0	1175	from Apr. 1, 2018 to Dec. 31, 2018

Methodology

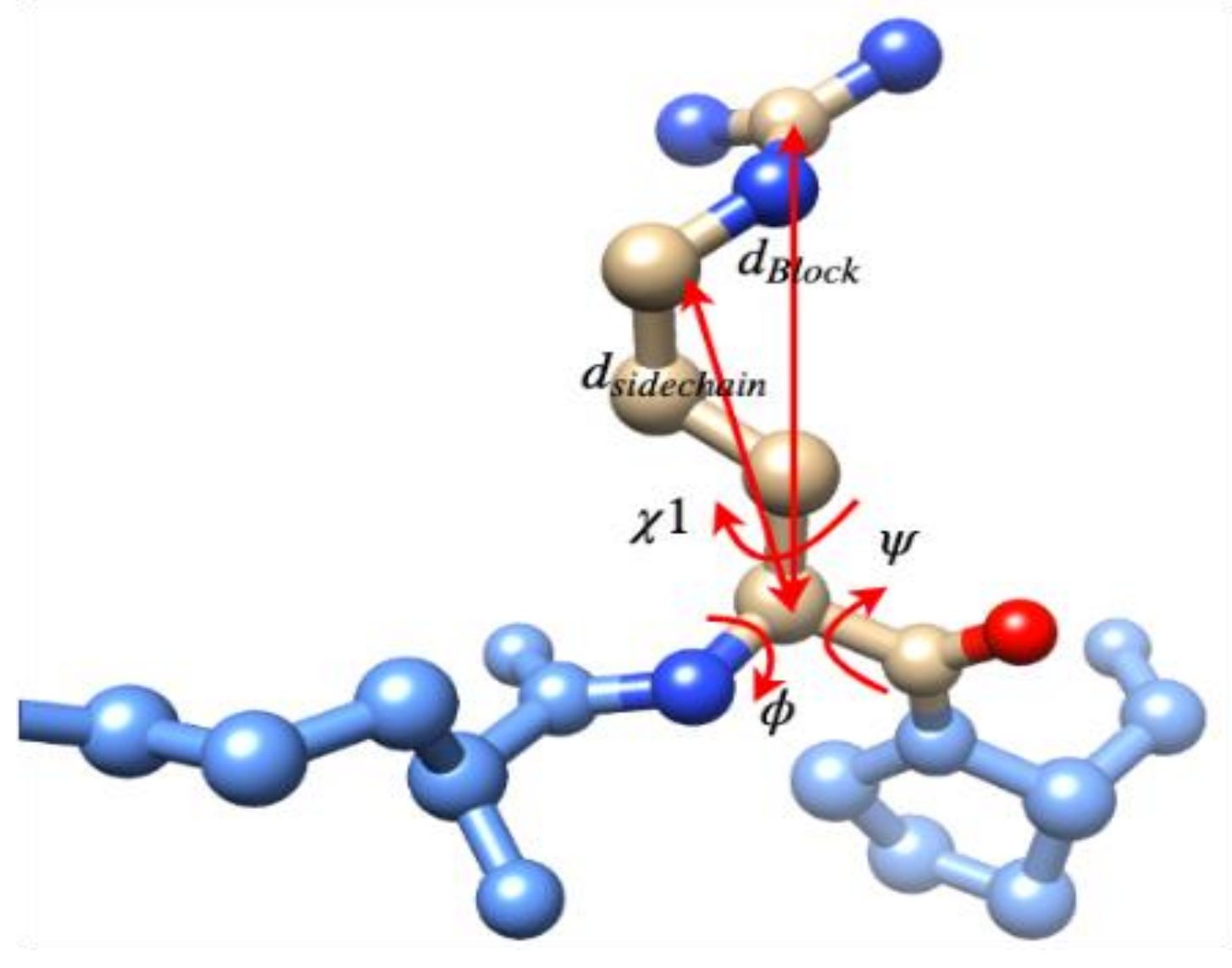

Figure 1. The five selected conformation features for protein residues, φ, ψ, χ_1, $d_{sidechain}$, and d_{block}.

$$HBOS = \sum_{i=1}^{5} HBOS(i) = \sum_{i=1}^{5} \log(\frac{1}{npdf_i(v_i)}) \qquad (1)$$

Five features, backbone torsion angle Phi (φ) and Psi (ψ), sidechain torsion angle (χ_1), sidechain size ($d_{sidechain}$), and block length (d_{block}), were selected to describe the conformation of a protein residue as shown in Figure 1. φ, ψ are torsional angles on the backbone of protein chains. χ_1 is the first torsion angle in the sidechain. 18 of 20 residues were used since glycine (GLY) and alanine (ALA) have no χ_1 due to their small size of sidechains. $d_{sidechain}$ is the distance between the CA atom on the backbone and the mass centroid of the sidechain atoms. d_{block} is the distance between the CA atom on the backbone and the mass centroid of the distal block of a specific residue. The blocks of the residues were defined in He's study [2].

Two unsupervised machine learning models, Histogram-Based Outlier Score (HBOS) [3] and Isolated Forest (IF) [4], were trained with the five features extracted from X-ray-1.5, then applied to test residues in EM-0-4-2019. HBOS model was proposed by Chen [3]. The normalized probability density functions (npdfs) for each feature of a residue type were used to calculate the HBOS score according to Equation 1. The residues with HBOS score above 10 are labelled as outliers in the current study. IF model uses the concept of isolation to explicitly isolate anomalies. IF is an algorithm which has a linear time complexity with a low constant and a low memory requirement. It is favorable in dealing with large datasets. X-ray-1.5 was used as the training data. The features were transformed with quantile transform [5], which transforms the features to follow a normal distribution. The model was trained with X-ray-1.5 by assuming 0.1% of the training data are outliers. Then the trained model was applied to EM-0-4-2019. The outlier residues are labeled as -1, labeled as 1 otherwise.

Results

The cryo-EM proteins in EM-0-4-2019 were evaluated by both HBOS model and IF model. The residues labelled as outliers by both models have the high preference of having error conformations. In PDB validation reports, a residue is color-coded as green if there is no outlier observed by the PDB validation system, yellow if there are outliers for one criterion, orange for two criteria, red for three or more criteria. The labels in the validation reports for the residues which are labelled by both HBOS and IF are retrieved by a python web crawler (https://github.com/lin-chen-VA/MDPI_Molecules_2020). Within those labelled residues by HBOS and IF, the number of residues that are identified as reasonable (green color) in EM-0-4-2019 are listed in Table 2.

Table 2. The number of residues labelled as outlier by HBOS and IF but not by the validation reports for 18 types of residues in EM-0-2019.

Residue	ARG	ASN	ASP	CYS	GLN	GLU	HIS	ILE	LEU
Outlier Number	2	0	1	0	7	3	0	11	483
Residue	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Outlier Number	1	0	2	1	0	0	0	6	24

Two residues that are labelled by HBOS and IF are color green in the validation reports in Figure 2. ARG 241 in Figure 2A has a long sidechain which tends to align with the nearby density cloud. Aligning the sidechain of ARG 241 to the cloud causes a φ angle of 38.44° which is highly unfavorable and barely observed in X-ray-1.5 since the nearby cloud maybe from the background noise. The validation report of 6ndy does not label it out. ILE 894 of 6sl1 is located at a turn of chain A. The d_{block} of ILE 894 is 2.81 Å which is abnormally short. It maybe caused by bending the sidechain into the cloud near the backbone turn which is more likely from the backbone instead of sidechain. The abnormal conformation of ILE 894 was not detected by the PDB validation tool and colored green.

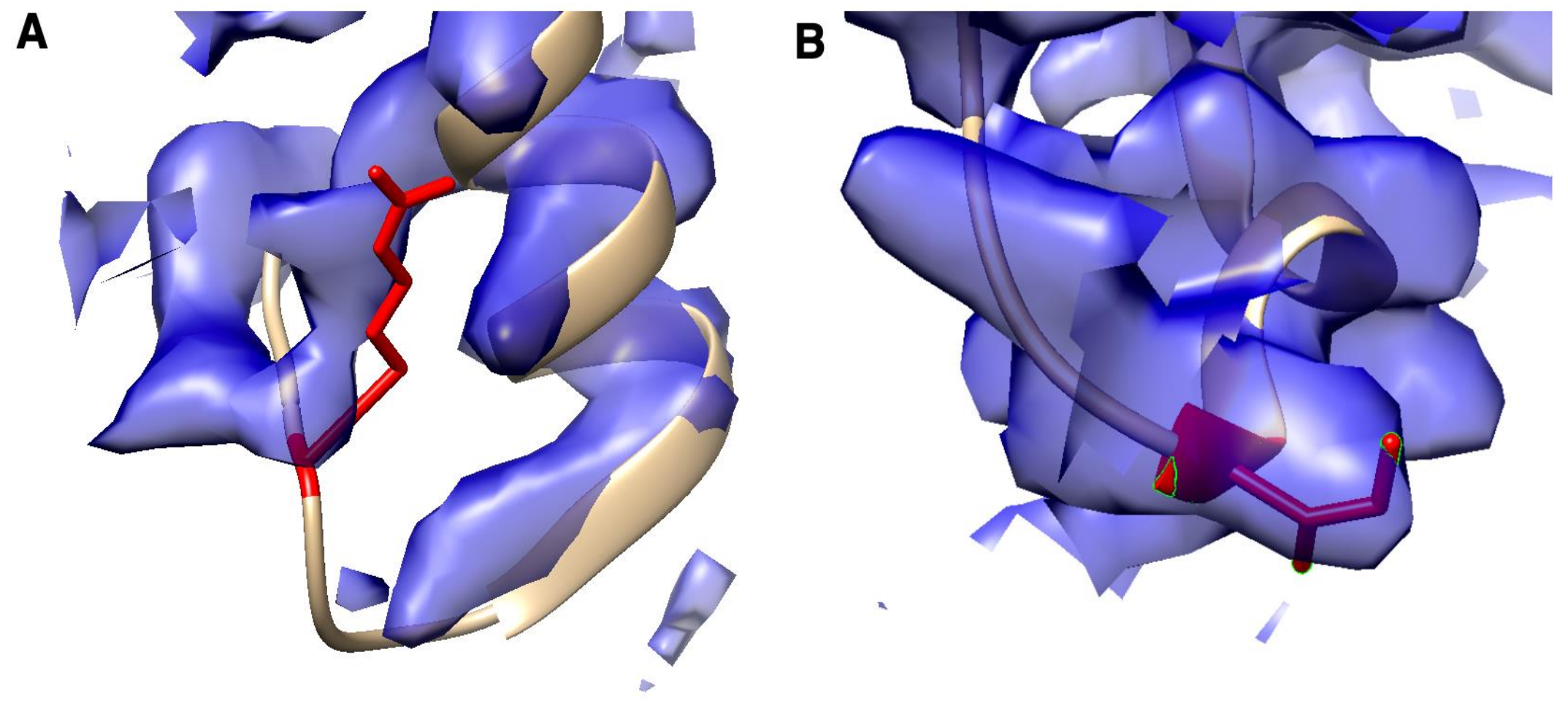

Figure 2. ARG 241 in B chain of 6ndy (A) and ILE 894 in chain A of 6sl1 (B) are colored green in the validation reports and are labelled as outliers by HBOS and IF.

Conclusion

There are many anomalous conformations identified by a probability-based model and a tree-based model. However, those potential outliers were not colored in the validation reports by the current pipeline validation tool which checks the metrics one by one. Though outliers may be genuine instead of errors, but they deserve attention. The current validation system at PDB may need to introduce more features or different validation methods as a complementary tool.

Reference

[1] wwPDB consortium et al., "Protein Data Bank: the single global archive for 3D macromolecular structure data," Nucleic Acids Research, vol. 47, no. D1, pp. D520–D528, Jan. 2019, doi: 10.1093/nar/gky949.

[2] L. Chen and J. He, "A distance- and orientation-dependent energy function of amino acid key blocks: Distance- and Orientation-Dependent Energy Function of Amino Acids," Biopolymers, vol. 101, no. 6, pp. 681–692, Jun. 2014, doi: 10.1002/bip.22440.

[3] L. Chen and J. He, "Using Combined Features to Analyze Atomic Structures derived from Cryo-EM Density Maps," in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18, Washington, DC, USA, 2018, pp. 651–655, doi: 10.1145/3233547.3233709.

[4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, Dec. 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.

[5] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," Bioinformatics, vol. 19, no. 2, pp. 185–193, Jan. 2003, doi: 10.1093/bioinformatics/19.2.185.

Acknowledgements

Lin Chen and Anurag Dasgupta participated in the method design. Stefan Atkinson wrote the poster. Stefan Atkinson and Kevin Eppes collected data and performed analysis.

The work in this paper is supported by Valdosta State University First-Year Tenure Track Faculty Fund and Valdosta State University Seed Grant.