molecules

MDPI

Article

An Investigation of Atomic Structures Derived from X-ray Crystallography and Cryo-Electron Microscopy Using Distal Blocks of Side-Chains

Lin Chen [1,*], Jing He [2], Salim Sazzed [2] and Rayshawn Walker [1]

[1] Department of Mathematics and Computer Science, Elizabeth City State University, Elizabeth City, NC 27909, USA; rrwalker135@students.ecsu.edu

[2] Department of Computer Science, Old Dominion University; Norfolk, VA 23529, USA; jhe@cs.odu.edu (J.H.); ssazz001@odu.edu (S.S.)

* Correspondence: lchen@ecsu.edu; Tel.: +1-(252)-335-3634

Received: 22 December 2017; Accepted: 27 February 2018; Published: 8 March 2018

Abstract: Cryo-electron microscopy (cryo-EM) is a structure determination method for large molecular complexes. As more and more atomic structures are determined using this technique, it is becoming possible to perform statistical characterization of side-chain conformations. Two data sets were involved to characterize block lengths for each of the 18 types of amino acids. One set contains 9131 structures resolved using X-ray crystallography from density maps with better than or equal to 1.5 Å resolutions, and the other contains 237 protein structures derived from cryo-EM density maps with 2–4 Å resolutions. The results show that the normalized probability density function of block lengths is similar between the X-ray data set and the cryo-EM data set for most of the residue types, but differences were observed for ARG, GLU, ILE, LYS, PHE, TRP, and TYR for which conformations with certain shorter block lengths are more likely to be observed in the cryo-EM set with 2–4 Å resolutions.

Keywords: protein; structure; cryo-electron microscopy; validation; statistics; X-ray; crystallography; side chain

1. Introduction

Cryo-electron microscopy (cryo-EM) is an emerging structure determination technique in addition to two other techniques: X-ray crystallography (X-ray) and Nuclear Magnetic Resonance (NMR). Currently, over 1900 atomic structures have been derived from electron density maps produced using cryo-EM technique, and they are deposited in Protein Data Bank (PDB) [1]. Some of the atomic structures are derived from cryo-EM density maps with 2–4 Å resolutions. Others are obtained from density maps of much lower resolutions. It is generally expected that atomic structures that are derived from high-quality density maps are more accurate than those derived from lower-quality maps. Although there are more than 5700 EM density maps in the EM Data Bank (EMDB) as of January 2018, about 54% of them have resolution lower than 10 Å and 24% with 5–10 Å resolution [2]. The set of highest-quality cryo-EM density maps contains those with 2–4 Å resolutions. The number of such high-quality maps has increased rapidly since 2012. As of December 2016 when the data were downloaded for the study, there were 416 cryo-EM density maps with better than or equal to 4 Å resolutions. Some of them are protein structures and others are nucleic acid structures. Atomic structures derived from these maps represent a set of most accurate structures derived from cryo-EM density maps.

Although both X-ray crystallography and cryo-EM produce electron density maps from which atomic structures are derived, they have different sample preparation, different data collection, and different data processing. In addition, structure changes have been observed at different temperatures [3]. The concept of resolution in X-ray depends on the highest resolvable diffraction spots. However, a commonly used method to evaluate resolution for a cryo-EM density map, Fourier Shell Correlation [4], quantifies the correlation of two halves of the data set in Fourier space. Regardless of the differences between the two structure determination methods, side chains are generally distinguishable in a cryo-EM density map with about 3 Å resolution [5–9]. It is an open question how accurate cryo-EM structures are, particularly for those derived from the highest-quality cryo-EM density maps.

Validation of protein structures is an important step to maintain an accurate archive of atomic structures [10–14]. Various analysis tools have been developed since 1990 to identify outliers [12,15–25]. In order to create comprehensive assessment criteria, Worldwide Protein Data Bank (wwPDB) convened a Validation Task Force (VTF) to establish policies, standards, and specifications of formats. The VTF includes X-ray VTF, NMR VTF, and 3DEM VTF. In 2011, X-ray VTF published its recommendation report on how to validate protein structures derived from X-ray density maps [26]. After that, 3DEM VTF and NMR VTF released their recommendation reports in 2012 [27] and 2013 [28]. The wwPDB accepts and curates depositions using wwPDB Deposition & Annotation (D&A) system that implements the recommendations from wwPDB VTF [29]. wwPDB generates validation reports for new depositions containing the results from rigorous tests of structure model quality. The validation reports for X-ray structures have been updated with the D&A system in March 2017 with 2016 statistics. The validation reports for NMR and 3DEM structures in PDB were available since May 2016.

3DEM VTF recommended to validate both EM density map and models. Several methods have been proposed to prevent overfitting of data: cross-validation [27], limit of the reconstruction with low signal-to-noise-ratios [30], and regulation using a Gaussian [31]. Practices implemented to validate density maps include absolute hand determination [32,33], data coverage and agreement between images and class averages [34], and statistical assessment of maps [35]. Rfree [36] factor is used to check the correlation coefficient between a model and its cryo-EM map. According to wwPDB EM validation reports, EM models are often validated using existing tools for X-ray data and the correlation coefficient between an EM density map and its atomic model [37].

As more atomic structures are determined using high-quality cryo-EM density maps, it is becoming possible to characterize statistical behaviors of such structures. In this paper, we present a statistical analysis of structures derived from ultra-high-resolution density maps of X-ray and those derived from EM density maps with 2–4 Å resolutions. The analysis uses the length of the distal block of a side-chain in addition to backbone and side-chain torsion angles to characterize the conformational distribution of each amino acid.

2. Results

Two X-ray data sets and two EM data sets were analyzed in this study. An X-ray data set, referred as X-ray-1.5, contains 9131 PDB protein structures that are solved using X-ray crystallography and have a resolution better than or equal to 1.5 Å. The protein structures were extracted from the PDB website using the default sequence similarity value of 90%. X-ray-1.5 was used as the reference in the study because side-chain positions are more precisely defined at such resolutions than lower resolutions. In order to characterize the statistics from structures that are derived from high quality EM density maps, we downloaded an EM data set in December 2016, referred as EM-2-4, containing 237 PDB protein structures that are derived from EM density maps with resolutions between 2 and 4 Å (including 4 Å). These maps represent the most accurate structures derived from cryo-EM density maps. The third data set, referred as EM-4-6, contains 168 protein structures derived from EM density maps with resolutions greater than 4 Å and less than or equal to 6 Å resolutions. The fourth data set, referred as X-ray-237, contains 237 protein structures that are randomly picked from the 9131 X-ray structures that are used in the first data set.

2.1. 3D Scatter Plots for Backbone and Side-Chain Torsion Angles

The conformation of a residues is widely used to validate the quality of a model. One of the practices in the validation pipeline uses combination of torsion angles from either the backbone or the side-chain. Most of modeling methods have considered Ramachandran criteria [38] and side-chain rotamer library [39] to assign favorable combination values to residues.

We investigated the distribution of combined torsion angles (ϕ, ψ, χ_1) of arginine (ARG) using a 3D scatter plot (Figure 1). Each conformation of ARG is represented as a point in the plot. For clear visualization, points are colored according to χ_1 angle. Although the X-ray-1.5 data set contains the most number of proteins as compared to EM-2.4 and EM-4-6 data sets, the points are highly clustered within several areas. ARG conformations are rarely observed outside those areas (Figure 1A). Due to EM data set has much lower resolution, EM points are less concentrated in those clusters and more conformations are seen in between clusters as expected (Figure 1B,C). As an example for the top layer (yellow), there are two large clusters at (250°, 330°, 300°), (250°, 150°, 300°), one small cluster at (50°, 50°, 300°), and two weak clusters at (80°, 350°, 300°) and (50°, 230°, 300°). In the same layer for the two EM data sets (Figure 1B,C), there is a belt spanning from about (50°, 0°, 300°) to (50°, 360°, 300°). The two clusters at (250°, 330°, 300°) and (250°, 150°, 300°) are broader (Figure 1B,C) suggesting that many ARG conformations are unfavorable according to the distribution of X-ray data. The middle layer (cyan in Figure 1A) contains a cluster at (290°, 330°, 180°) and a larger cluster spanning from about (200°, 140°, 180°) to (300°, 140°, 180°). The middle layer of EM data (Figure 1B,C) has the cluster spanning from (200°, 330°, 180°) to (330°, 330°, 180°) and a large cluster spanning (200°, 140°, 180°) to (320°, 140°, 180°). EM data show a "cyan belt" in the middle layer from (50°, 0°, 180°) to (50°, 360°, 180°) (Figure 1B) which does not exist in the corresponding space in Figure 1A. In the bottom layer (blue), X-ray points have three clusters at about (300°, 330°, 50°), (200°, 150°, 50°) and (300°, 150°, 50°). EM points show broader blue clusters compared to those of X-ray data.

Although current modeling and validation methods have considered favorable residue conformations using Ramachandran criteria and rotamer libraries, more combinations of torsion angles should be considered as more data are available. Using the Ramachandran plot that is colored based on χ_1, we observed that the X-ray data set and the EM data sets generally have a similar distribution of (ϕ, ψ) (Figure 1D–F). However, EM-2-4 set has many points outside the clusters of X-ray-1.5 set. The broader clusters of EM data sets shown in the 3D scatter plot (Figure 1A–C) and the colored Ramachandran plot (Figure 1D–F) suggest that some conformations of ARG are unfavorable if X-ray high-resolution data are used as a reference. EM structures may benefit from a refinement of side-chain conformations if more dependency of torsion angles is incorporated in model building.

2.2. Normalized Probability Density Function for Block Lengths

We calculated block length (d_{Block}) using the distal block of a side-chain for each of 18 types of residues (see Materials and Methods), since the position and orientation of the distal block are sensitive in distinguishing side-chain conformations [40]. The block lengths of lysine (LYS) range from 4 Å to 6 Å approximately. The side-chains of LYS show two peaks at 5.2 Å and 5.7 Å respectively in the normalized probability density function (npdf) from both the EM-2-4 data set and the X-ray-1.5 data set, suggesting two preferred block lengths (Figure 2A). The height of the peak at 5.7 Å is lower in the npdf of the EM-2-4 set than that of the X-ray-1.5 set. This suggests that it is less likely to find a conformation of the side-chain with block length of 5.7 Å in the EM-2-4 data set than in the X-ray-1.5 set. However, the difference in the height of the highest peak only reflects the situation at one point (the peak). When Figure 2A and 2B are both considered that the most preferred conformation of LYS has (ϕ, d_{Block}) as (298°, 5.7 Å). The EM-2-4 set has lower probability of having the most preferred conformation than the X-ray-1.5 set. In the region where $4 \text{ Å} < d_{Block} \leq 5.2 \text{ Å}$, the EM curve appears to be mostly above the X-ray curve (Figure 2A and Figure S1) indicating that most points in this region (that corresponds to shorter block lengths) show higher probability in the EM-2-4 set than in the X-ray-1.5 and X-ray-237 sets. In the range between 4 Å and 4.8 Å, npdf value of X-ray-1.5 is smaller than that of npdf of EM-2-4 (Figure 2A). This suggests that it is more likely to observe a shorter side-chain for LYS, within the range of 4 Å to 4.8 Å, in the EM-2-4 set than in the X-ray-1.5 set. This may suggest that the visibility of the density map at the side chain region may have given preference to a shorter LYS side-chain when modeling some EM structures.

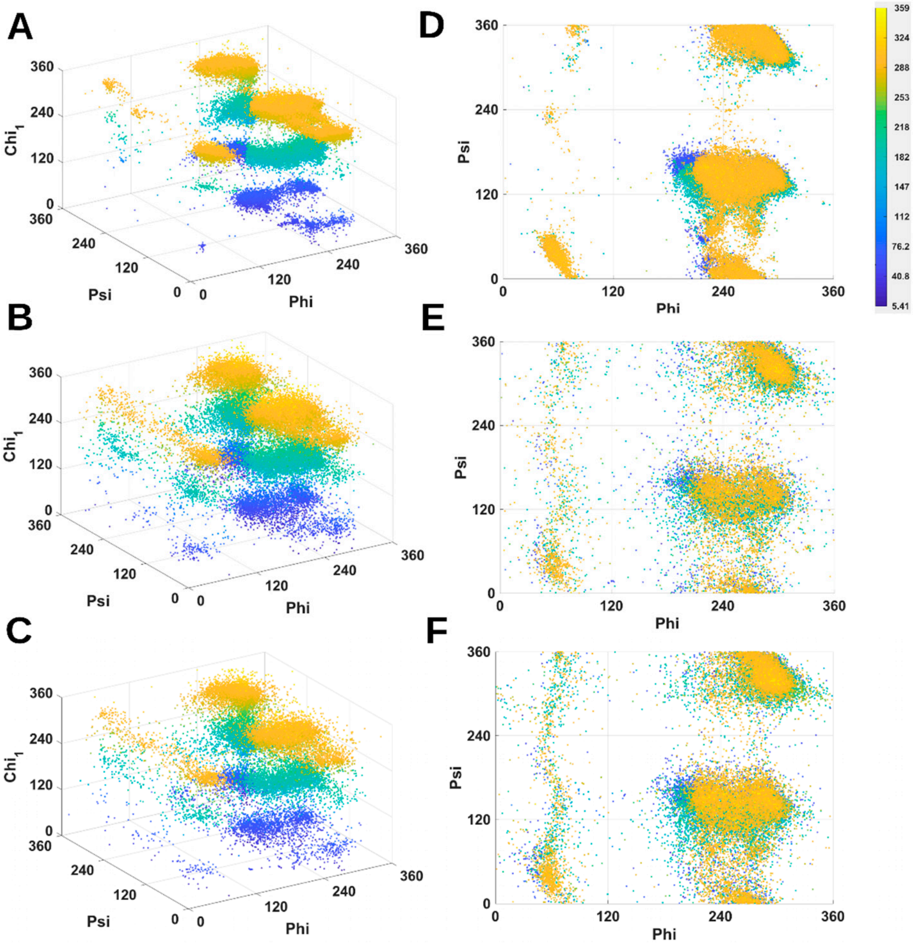

Figure 1. 3D scatter plots and Ramachandran plots of Arginine (ARG) residue for three data sets. 3D scatter plots of (ϕ, ψ, χ_1) are shown for X-ray-1.5 data set (**A**); EM-2-4 data set (**B**); and EM-4-6 data set (**C**); Ramachandran plots of (ϕ, ψ) are shown for X-ray-1.5 data set (**D**); EM-2-4 data set (**E**); and EM-4-6 data set (**F**). The color code (shown) is based on χ_1, and it is the same for each panel.

The block length of isoleucine (ILE) is about 2.8 Å-4 Å (Figure 2C). The height of the peak at 3.88 Å of block length is about the same for both X-ray-1.5 set and the EM-2-4 set. The right side of the curve falls faster in the EM-2-4 set (blue) than in the X-ray-1.5 set (red) (Figure 2C). This suggests that it is less likely to observe those side-chain conformations with block length between 3.88 and 4 Å in the EM-2-4 data set. There is a small peak at about 3.4 Å block length for EM structures that is not obvious for X-ray structures. The peak of the X-ray npdf at 3.08 Å of block length shifts slightly left towards shorter block lengths. This suggests that it is more likely to assign a side-chain conformation of ILE with a slightly shorter block length at about 3 Å in the EM-2-4 data set. The npdf of backbone ϕ angle also shows some differences between the X-ray structures and EM structures for ILE, such as at 62.5° of ϕ angle, the unfavorable angle if judged by the X-ray curve (red) (Figure 2D).

The npdf of block length shows that both X-ray-1.5 set and the EM-2-4 set have the same peaks located at 3.78 Å for phenylalanine (PHE) and 6.43 Å for Tyrosine (TYR) respectively (Figure 3E,F). However, the EM-2-4 data set has a higher peak than the X-ray-1.5 set for both PHE and TYR. This suggests that it is more likely to assign a side-chain with the most popular conformation in the EM structures than in the X-ray structures. The curves at the right side of the peak fall faster in the EM-2-4 set than in the X-ray-1.5 set. This suggests that the probability of finding side-chain with a block length longer than the most popular length is smaller in the EM-2-4 set than in the X-ray-1.5 set.

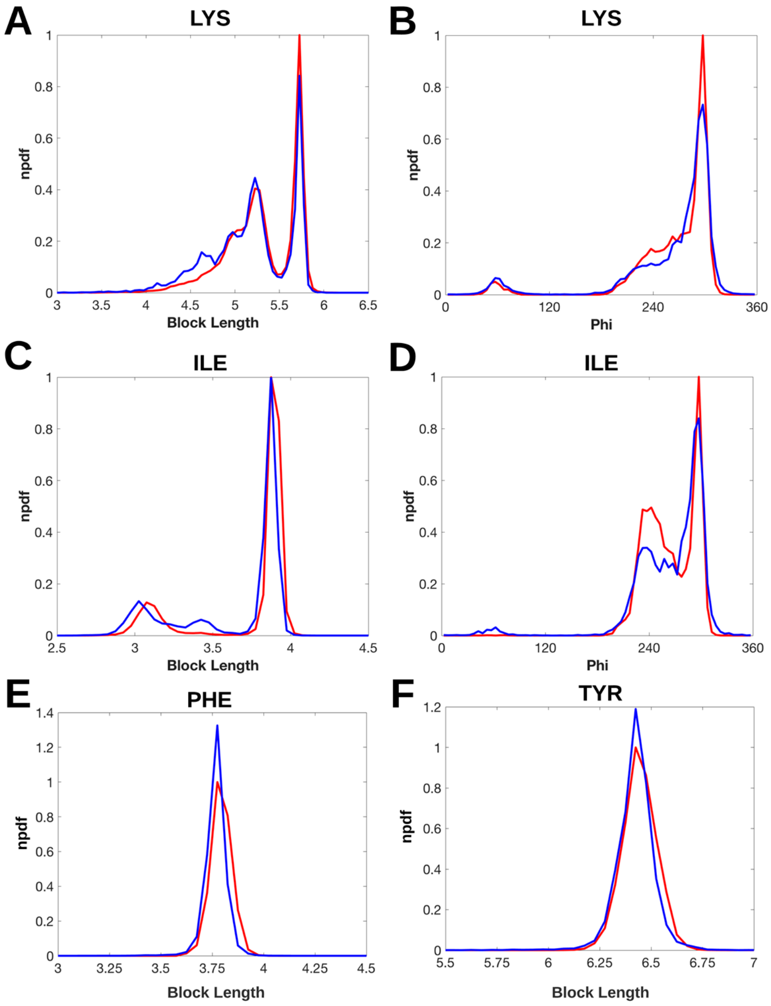

Figure 2. Normalized probability density function of block length and torsion angle ϕ. The distribution of block length and ϕ is shown respectively as a probability density function (npdf) normalized by the highest peak of the pdf of the X-ray-1.5 data set. The npdf of X-ray-1.5 data set (red line) and the npdf of EM-2-4 data set (blue line) are shown for lysine (LYS) (**A**,**B**); isoleucine (ILE) (**C**,**D**); phenylalanine (PHE) (**E**); and tyrosine (TYR) (**F**) respectively.

In general, the lower the resolution of the density map, the more challenging to determine precisely the position of the distal block in electron density maps. The lengths of distal blocks were characterized from X-ray-1.5 set containing 9131 proteins and EM-2-4 set containing 237 proteins. We observed that the EM-2-4 set has bias towards those side-chain conformations with slightly shorter block lengths for LYS, ILE, PHE, and TYR. For example in the case of LYS, EM-2-4 has lower probability at 5.7 Å block length but higher probability at 4-5.2 Å when it is compared to X-ray-1.5 set. It is possible that some side-chain conformations with long block length of about 5.7 Å were modeled to shorter side-chains. The bias is less likely due to the difference in the number of samples for the two data sets (see Supplementary Figure S1). We randomly selected 237 proteins from the X-ray-1.5 set to create X-ray-237 set that is expected to have the same magnitude of residues as for the EM-2-4 set. Yet the same bias towards shorter block length was observed when the X-ray-237 and EM-2-4 were compared (Supplementary Figure S1). In addition to those four residues, we also observed higher chance of finding shorter side-chain for ARG, Glutamic acid (GLU) and Tryptophan (TRP) (see Supplementary Figure S3). No obvious difference was observed in the length distributions of the other 11 residues (see Supplementary Figure S2).

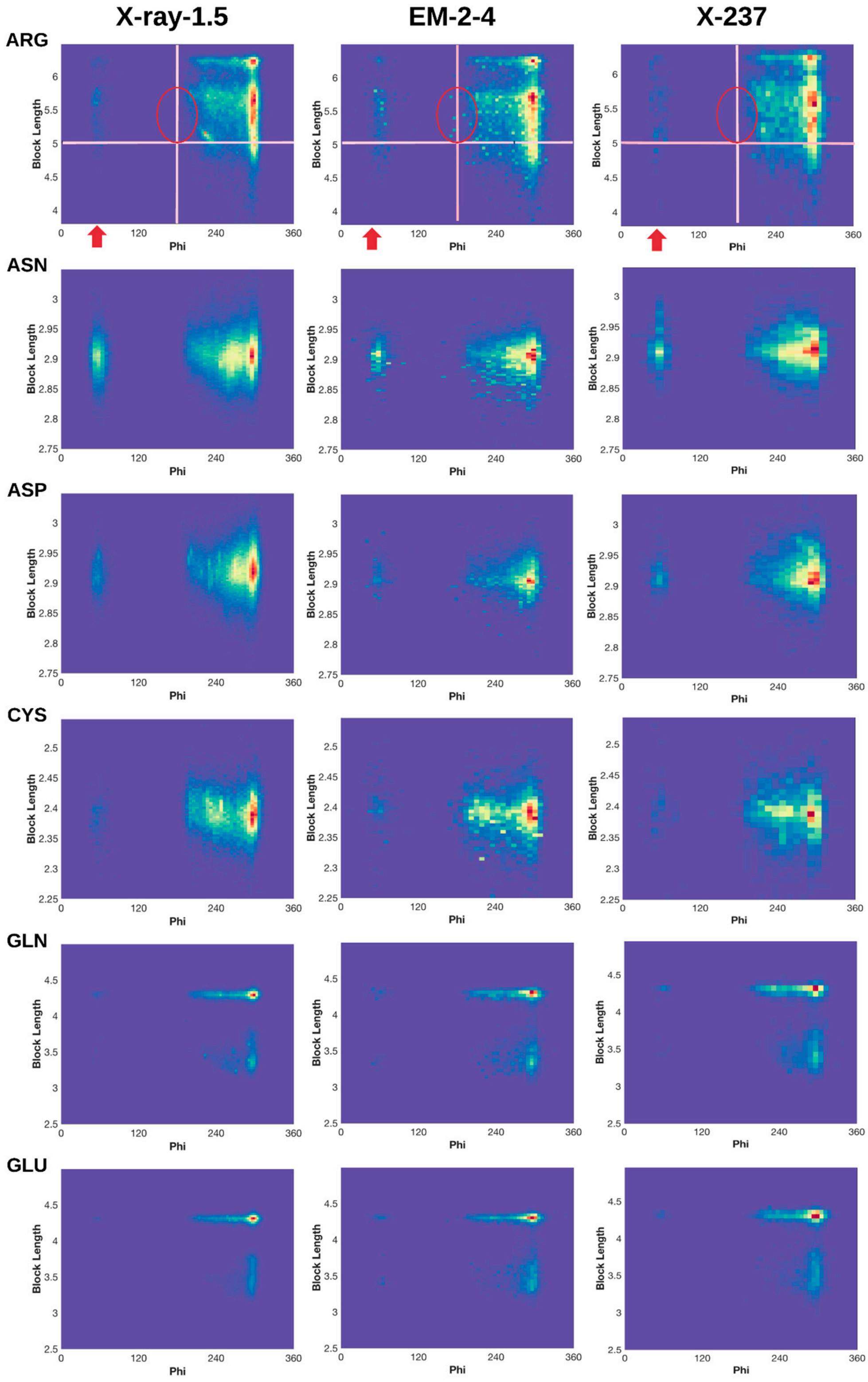

Figure 3. *Cont.*

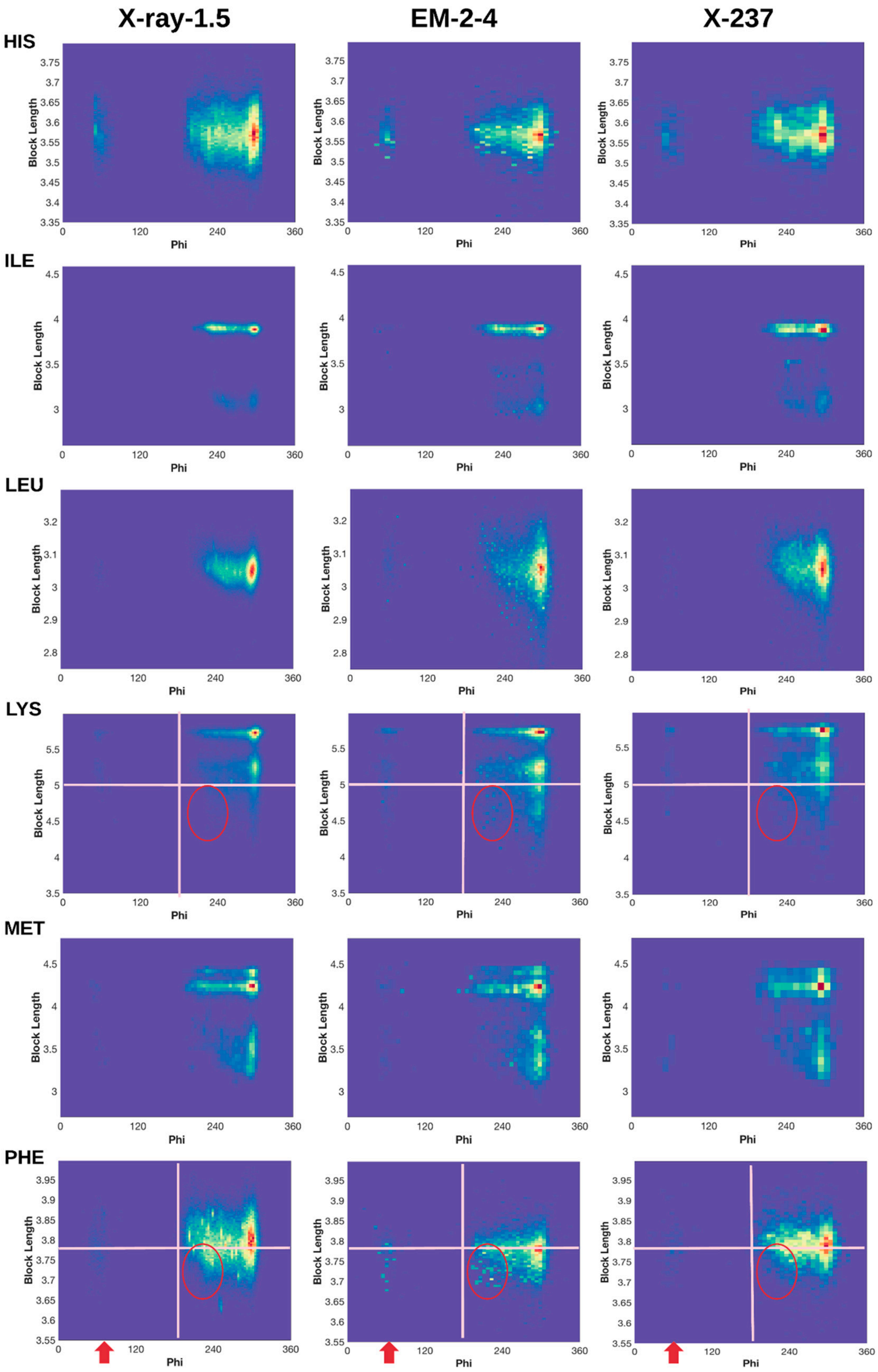

Figure 3. *Cont*.

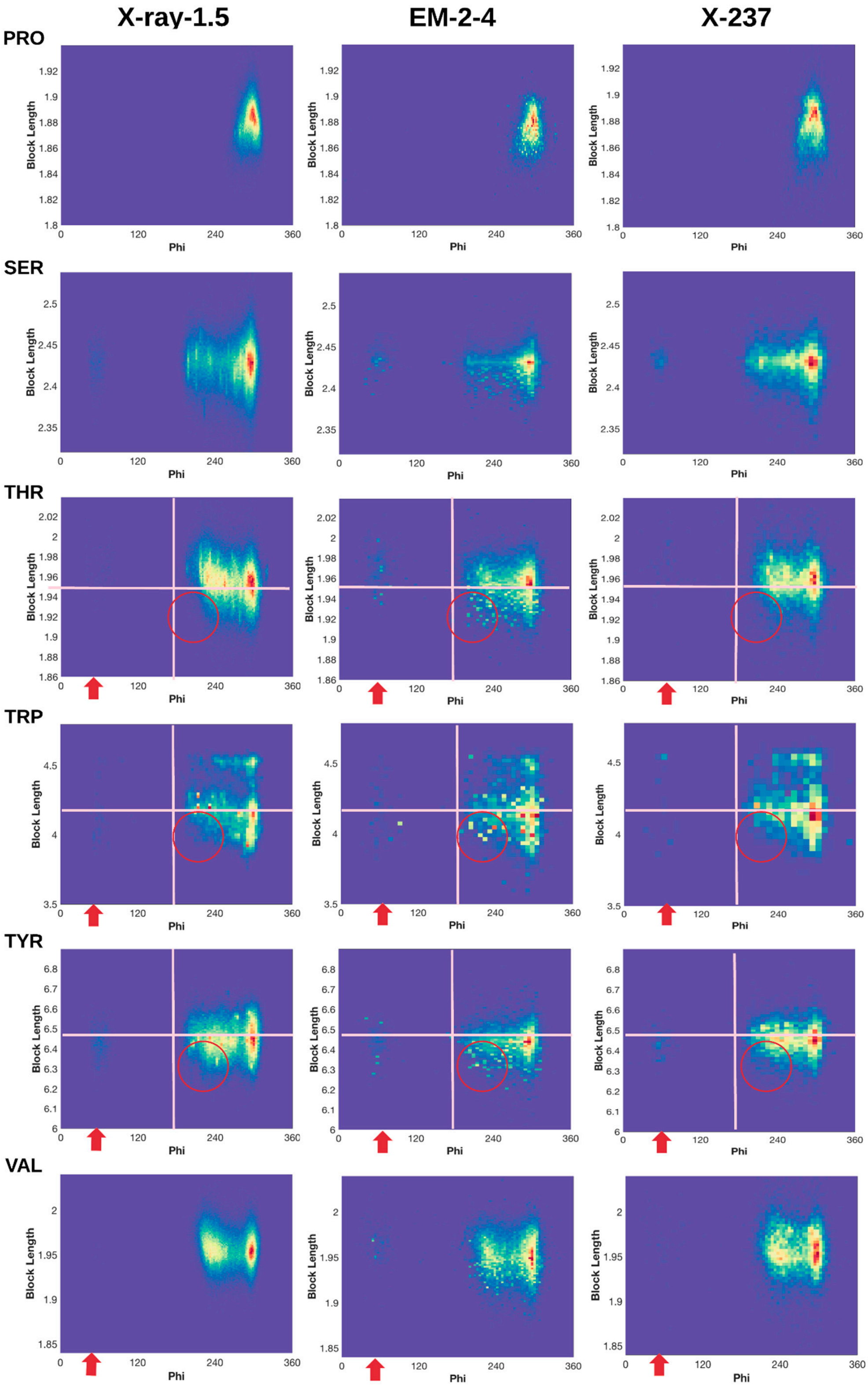

Figure 3. Normalized 2D histogram of (ϕ, $\mathbf{d_{Block}}$) for 18 residue types (Alanine (ALA) and Glycine (GLY) not included). Normalized 2D histograms of X-ray-1.5 data set (left column), EM-2-4 data set (middle column), and X-ray-237 data set (right column) were plotted using the ndhist function of MATLAB such that the most popular (ϕ, $\mathbf{d_{Block}}$) is colored red and the least popular is in dark blue [41]. Visually observed difference regions are labeled with red arrows and circles. Cross lines are drawn to assist comparison.

2.3. 2D Histograms of Combined Features Using the Block Length and Backbone Torsion ϕ

We characterized (ϕ, d_{Block}) profile for each of the 18 residues (Figure 3), since the 2D histogram reflects the dependency between the backbone torsion angle ϕ and the side-chain conformation. The two residues that are not characterized for block length are the two small residues GLY and ALA. A normalized 2D histogram was obtained by normalizing using the value of the highest peak (shown in red). Therefore, the color at each point of the 2D histogram represents a population ratio with respect to the population of the most popular conformation. The 2D histogram of (ϕ, d_{Block}) demonstrates a characteristic pattern for each residue. For example, the pattern of ARG is quite different from that of ASN (Figure 3 row 1 and row 2), suggesting that the 2D histogram captures the distribution of conformations despite of its simplicity.

2.3.1. Conformations of Side-Chains When ϕ is Near 60°

Although the backbone torsion angle ϕ is mostly negative in the classical view of Ramachandran plot in which ϕ is between −180° to 180°, positive ϕ exist and they mostly cluster around 60°. The eighteen 2D histograms derived from 9131 ultra-high-resolution structures show that the population of ϕ near 60° exists but it is extremely low, as seen from the dark blue or blue color for all eighteen residues except ASN, ASP, and HIS (Figure 3 left column). In the case of ARG, those side-chain conformations are highly restricted in specific clusters when ϕ is about 60° (Figure 1A). The low population of such conformations may present a challenge in assignment of such conformations during structural determination. Higher population ratio of ϕ was observed in the EM-2-4 data set for ARG, PHE, THR, TRP, and TYR when ϕ is about 60°. As an example for ARG, there is a more obvious "vertical belt" at about 60° in the histogram of EM-2-4 (middle panel of row 1) than that in the histogram of X-ray-1.5 (left panel of row 1) (Figure 3). In fact, a few spots in the "vertical belt" appear yellow indicating higher population ratio than those points in blue. The visual difference at this region is less likely due to the difference in total number of proteins included in the two data sets, since the "vertical belt" is not as obvious in the 2D histogram of X-ray-237 set that has the same number of proteins as in the EM-2-4 set. The "vertical belt" observed in the 2D histogram aligns with the observation in the 3D scatter plot of ARG (Figure 1B) in which more points appear to scatter at the region of ϕ near 60°.

Normalized 2D histogram of PHE shows that the population ratio at ϕ ~60° is higher in the EM-2-4 set than in the X-ray-1.5 or the X-ray-237 set. We observe a "belt" at the region where ϕ is between 40° and 70° in the EM-2-4 set (Figure 3). In fact, some of the points in the belt are light yellow. The most popular point (ϕ, d_{Block}) in the "belt region" of EM-2-4 set is (58.04°, 3.79Å), and the most popular point of PHE is (301.8°, 3.78 Å). The population ratio between the two points is 18%, meaning the most popular configuration in the belt is 18% of the population of the most popular configuration of PHE. For X-ray-1.5 set, the most popular point in the belt is at (69.28°, 3.86 Å), and the most popular point (red peak in the left column of PHE row in Figure 3) is at (303.5°, 3.81 Å). The population ratio is 4.1%, more than 1/4 times lower than that of the EM-2-4 set.

2.3.2. Shorter Side-Chains When ϕ is Near 210°

Each residue has a range of block lengths representing the overall lengths of different conformations. As an example, LYS has a block length ranging from about 4 Å to 5.8 Å (row LYS of Figure 3). When comparing the normalized 2D histograms of three data sets (the X-ray-1.5, EM-2-4, and X-237) four residues LYS, PHE, TRP, and TYR, show a higher population ratio at the region of ϕ near 210° for the shorter block lengths. For example, the population ratio of LYS is very low for those conformations with block length between 4 and 5 Å and ϕ between 200° and 230°, according to the X-ray-1.5 data set (the left column of row LYS in Figure 3). However, the population ratio is higher for about eight spots (in the circle) when the 2D histogram of EM-2-4 is compared with that of the X-ray-1.5 and X-ray-237. Since the ratio is calculated between the population at a point and the population at the peak, a higher ratio suggests either there is higher population in this region or the population is low at the peak. In this case, the block length at the peak is 5.8 Å that corresponds to an extended conformation of LYS.

Similar observations are with PHE, TRP, and TYR when the residue has shorter block lengths and when ϕ is near 210°. For example, TYR block lengths range from about 6.2 Å to 6.8 Å. The population ratio at about (210°, 6.3 Å) is quite small according to X-ray-1.5 data set (left column of TYR row in Figure 3). However, the population ratio is higher at this region for the EM-2-4 data set. Our investigation suggests that one needs to be cautious when assigning a conformation with a short block length of the residue with backbone ϕ near 210°, since most of the residues have low probability of adopting such conformations.

3. Materials and Methods

The statistical analysis was implemented using protein structures downloaded from RCSB (www.rcsb.org). Four features, backbone torsion angle Phi ϕ and Psi ψ, side-chain torsion angle χ_1, and block length d_{Block} were used in characterization of backbone-dependent side-chain conformations (Figure 4). Note that the range of ϕ in this paper is 0°–360°, instead of typical range of −180°–180° in the Ramachandran plot. In fact, $\phi = \phi_R$ when $\phi \in [0, 180°]$, and $\phi = \phi_R + 360°$ when $\phi \in (180°, 360°)$, where ϕ and ϕ_R are the torsion angle Phi in this paper and in the Ramachandran plot respectively.

A side-chain is divided into blocks and the position of the distal block was used to represent the conformation of a side-chain for 18 of the 20 residues [40]. GLY and ALA are not included due to the small size of their side-chains. The position and orientation of the distal block have been used previously to characterize conformations of side-chains in an energy function [40]. A side-chain block length is the distance between Cα and the mass center of the distal block of a residue (Figure 4).

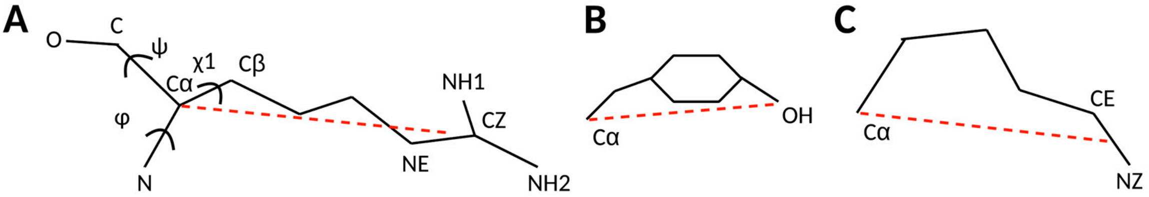

Figure 4. The four features—ϕ, ψ, χ_1, and block length. The block length (dashed line) was calculated using the center of the distal block of the side chain, illustrated for ARG (**A**), TYR (**B**) and LYS (**C**).

Four data sets (X-ray-1.5, EM-2-4, EM-4-6 and X-ray-237) were created (see Results). For those structures solved using X-ray crystallography, the first chain of each protein was used and the first conformation of each side-chain was used if there are alternative conformations. Since the X-ray-1.5 set contains 9131 protein structures, the number of residues should be sufficient for statistical analysis. EM-2-4 set contains 237 protein structures. In order to include as many conformations as possible, we used all chains in each PDB file in the study. Although some chains are related by non-crystallographic symmetry and are expected to have overall similar structures, structural difference has been observed between different chains that are related by non-crystallographic-symmetry (NCS) [42]. However, further characterization with more EM structures is needed as they are becoming available.

For each of the four features, a probability density function (pdf) was generated for each of the 18 residues using X-ray-1.5 and EM-2-4 data sets respectively. A pdf was generated using a bin size of 5° for ϕ, ψ and χ_1 and 0.05 Å for d_{Block}. Normalized probability density functions (npdf) were derived using two probability density functions, one from the X-ray-1.5 set, and the other from the EM-2-4 set. Normalized pdf was calculated by dividing each pdf with the value of the highest peak of the X-ray-1.5 pdf. A total of 72 npdfs were generated for 18 types of residues and 4 features for each type. The python scripts and MATLAB scripts in the study have been deposited to Github at https://github.com/lin-chen-VA/structures. The plots and histograms will be maintained at http://www.cs.odu.edu/~jhe. The 2D histograms were plotted for each pair of features (ϕ, χ_1), (ψ, χ_1), (ϕ, ψ), (ϕ, d_{Block}) and (ψ, d_{Block}) using MATLAB ndhist function [41].

4. Conclusions

Precise characterization of the conformation of a residue involves the use of both backbone and side-chain torsion angles. Although there are sufficient (9131) X-ray structures in the X-ray-1.5 data set for statistical characterization, there are only 237 protein structures in the EM-2-4 data set. In order to sample conformations with sufficient statistics, we used the length of the distal block to represent the overall side-chain conformation. Although not as precise as using all torsion angles of the side-chain, it is a simple parameter to distinguish between a folded and an extended conformation of the side-chain. We characterized side-chain conformations using the length of the distal block for 18 of the 20 amino acids, in addition to torsion angle ϕ, ψ, and χ_1 for the side-chain and backbone respectively.

The distribution of block lengths shows that the npdf of block lengths is similar between the X-ray-1.5 set and EM-2-4 set for most of the 18 residues. However, differences were observed for seven residue types – ARG, GLU, ILE, LYS, PHE, TRP, and TYR. The most popular block length is 5.7 Å and 3.88 Å for LYS an ILE respectively. Our results suggest that it is more likely to observe those conformations with block lengths between 4 Å to 5.2 Å for LYS and between 2.9 Å to 3.5 Å for ILE in the EM-2-4 set than in the X-ray-1.5 set. The most popular block length for PHE and TYR is 3.78 Å and 6.43 Å respectively. Our results show that it is less likely to observe those PHE and TYR conformations with block lengths longer than the most popular lengths in EM-2-4 set than in X-ray-1.5 set. Although it is generally challenging to assign side-chains accurately from a density map beyond 2 Å in resolution for both X-ray and cryo-EM density maps, our analysis suggests that those seven residues show statistical bias towards certain shorter side-chains. It is not clear if the bias toward shorter chain is related to the fitting process in which the density closer to the backbone may have dominated the assignment. Further investigation is needed for more accurate assignment of side-chains.

Due to limited data available, the analysis of EM-2-4 set is based on 237 protein structures with all chains available in the PDB files. Ideally, those chains that are NCS-related should not be included, although they are often not identical. The first alternative is included if multiple alternatives are available. Though side-chain polymorphism [43] is common in models from high-resolution X-ray maps, alternative side-chain conformations are not prevalent in models from EM maps. Wlodawer et al. [44] reported differences between selected high-resolution density maps produced from X-ray crystallography and cryo-EM. This paper aims to characterize individual residues from a statistical perspective.

The number of atomic structures solved using cryo-EM technique has grown rapidly. Although structural validation guidelines are recommended for the newly solved cryo-EM structures, it is important to understand any statistical difference between the cryo-EM structures and most accurately determined X-ray structures. The ultra-high-resolution data of X-ray was used as a reference in this study. We report the systematic bias towards shorter side-chains in details for seven residues. The finding in this paper should be further verified when more cryo-EM structures are available.

Supplementary Materials: The following are available online.

Acknowledgments: The work in this paper is supported by NSF DBI-1356621, NIH R01-GM062968, NIH 4T34GM100831-05.

Author Contributions: Lin Chen and Jing He participated in the method design and writing, Lin Chen wrote the scripts, Salim Sazzed helped data collection and diagram plotting, Rayshawn Walker participated in the study as an undergraduate research trainee.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. PDB RCSB. RCSB Statistics Release. 2018. Available online: https://www.rcsb.org/stats/growth/em. (accessed on 26 January 2018).
2. Europe PDBi. EMDB Statistics Release. 2018. Available online: http://www.ebi.ac.uk/pdbe/emdb/statistics_releases.html/ (accessed on 26 January 2018).
3. Quinlan, R.A.; Zhang, Y.; Lansbury, A.; Williamson, I.; Pohl, E.; Sun, F. Changes in the quaternary structure and function of MjHSP16.5 attributable to deletion of the IXI motif and introduction of the substitution, R107G, in the α-crystallin domain. *Philos. Trans. R. Soc. B Biol. Sci.* **2013**, *368*, 20120327. [CrossRef] [PubMed]
4. Harauz, G.; van Heel, M. Exact Filters for General Geometry Three Dimensional Reconstruction. *Optik* **1986**, *73*, 146–156.
5. Rawson, S.; Bisson, C.; Hurdiss, D.L.; Fazal, A.; McPhillie, M.J.; Sedelnikova, S.E.; Baker, P.J.; Rice, D.W.; Muench, S.P. Elucidating the structural basis for differing enzyme inhibitor potency by cryo-EM. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 1795–1800. [CrossRef] [PubMed]
6. Autzen, H.E.; Myasnikov, A.G.; Campbell, M.G.; Asarnow, D.; Julius, D.; Cheng, Y. Structure of the human TRPM4 ion channel in a lipid nanodisc. *Science* **2018**, *359*, 228–232. [CrossRef] [PubMed]
7. Shalev-Benami, M.; Zhang, Y.; Rozenberg, H.; Nobe, Y.; Taoka, M.; Matzov, D.; Zimmerman, E.; Bashan, A.; Isobe, T.; Jaffe, C.L.; et al. Atomic resolution snapshot of Leishmania ribosome inhibition by the aminoglycoside paromomycin. *Nat. Commun.* **2017**, *8*, 1589. [CrossRef] [PubMed]
8. Kater, L.; Thoms, M.; Barrio-Garcia, C.; Cheng, J.; Ismail, S.; Ahmed, Y.L.; Bange, G.; Kressler, D.; Berninghausen, O.; Sinning, I.; et al. Visualizing the Assembly Pathway of Nucleolar Pre-60S Ribosomes. *Cell* **2017**, *171*, 1599-1610.e14. [CrossRef] [PubMed]
9. Guo, J.; She, J.; Zeng, W.; Chen, Q.; Bai, X.C.; Jiang, Y. Structures of the calcium-activated, non-selective cation channel TRPM4. *Nature* **2017**, *552*, 205–209. [CrossRef] [PubMed]
10. Bränd'en, C.-I.; Alwyn Jones, T. Between objectivity and subjectivity. *Nature* **1990**, *343*, 687–689. [CrossRef]
11. Kleywegt, G.J.; Jones, T.A. Homo Crystallographicus—Quo Vadis? *Structure* **2002**, *10*, 465–472. [CrossRef]
12. Jones, T.A.; Zou, J.Y.; Cowan, S.W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **1991**, *47 Pt 2*, 110–119. [CrossRef] [PubMed]
13. Kleywegt, G.J.; Jones, T.A. Phi/Psi-chology: Ramachandran revisited. *Structure* **1996**, *4*, 1395–1400. [CrossRef]
14. Kleywegt, G.J.; Jones, T.A. Where freedom is given, liberties are taken. *Structure* **1995**, *3*, 535–540. [CrossRef]
15. Laskowski, R.A.; MacArthur, M.W.; Moss, D.S.; Thornton, J.M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291. [CrossRef]
16. Hooft, R.W.; Vriend, G.; Sander, C.; Abola, E.E. Errors in protein structures. *Nature* **1996**, *381*, 272. [CrossRef] [PubMed]
17. Lovell, S.C.; Davis, I.W.; Arendall, W.B., 3rd; de Bakker, P.I.; Word, J.M.; Prisant, M.G.; Richardson, J.S.; Richardson, D.C. Structure validation by Calpha geometry: Phi, psi and Cbeta deviation. *Proteins* **2003**, *50*, 437–450. [CrossRef] [PubMed]
18. Bruno, I.J.; Cole, J.C.; Kessler, M.; Luo, J.; Motherwell, W.D.; Purkis, L.H.; Smith, B.R.; Taylor, R.; Cooper, R.I.; Harris, S.E.; et al. Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2133–2144. [CrossRef] [PubMed]
19. Davis, I.W.; Murray, L.W.; Richardson, J.S.; Richardson, D.C. MOLPROBITY: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **2004**, *32*, W615–W619. [CrossRef] [PubMed]
20. Kleywegt, G.J.; Harris, M.R.; Zou, J.Y.; Taylor, T.C.; Wahlby, A.; Jones, T.A. The Uppsala Electron-Density Server. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2240–2249. [CrossRef] [PubMed]
21. Zwart, P.H.; Grosse-Kunstleve, R.W.; Adams, P.D. Xtriage and Fest: Automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newslett.* **2005**, *43*. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.378.2281&rep=rep1&type=pdf (accessed on 26 January 2018).

22. Chen, V.B.; Arendall, W.B., III; Headd, J.J.; Keedy, D.A.; Immormino, R.M.; Kapral, G.J.; Murray, L.W.; Richardson, J.S.; Richardson, D.C. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66 Pt 1*, 12–21. [CrossRef] [PubMed]
23. Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.J.; Chen, V.B.; et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **2018**, *27*, 293–315. [CrossRef] [PubMed]
24. Hintze, B.J.; Lewis, S.M.; Richardson, J.S.; Richardson, D.C. Molprobity's ultimate rotamer-library distributions for model validation. *Proteins* **2016**, *84*, 1177–1189. [CrossRef] [PubMed]
25. Weichenberger, C.X.; Pozharski, E.; Rupp, B. Twilight reloaded: The peptide experience. *Acta Crystallogr. D Struct. Biol.* **2017**, *73 Pt 3*, 211–222. [CrossRef] [PubMed]
26. Read, R.J.; Adams, P.D.; Arendall, W.B., 3rd; Brunger, A.T.; Emsley, P.; Joosten, R.P.; Kleywegt, G.J.; Krissinel, E.B.; Lütteke, T.; Otwinowski, Z.; et al. A new generation of crystallographic validation tools for the protein data bank. *Structure* **2011**, *19*, 1395–1412. [CrossRef] [PubMed]
27. Henderson, R.; Sali, A.; Baker, M.L.; Carragher, B.; Devkota, B.; Downing, K.H.; Egelman, E.H.; Feng, Z.; Frank, J.; Grigorieff, N.; et al. Outcome of the first electron microscopy validation task force meeting. *Structure* **2012**, *20*, 205–214. [CrossRef] [PubMed]
28. Montelione, G.T.; Nilges, M.; Bax, A.; Güntert, P.; Herrmann, T.; Richardson, J.S.; Schwieters, C.D.; Vranken, W.F.; Vuister, G.W.; Wishart, D.S.; et al. Recommendations of the wwPDB NMR Validation Task Force. *Structure* **2013**, *21*, 1563–1570. [CrossRef] [PubMed]
29. Gore, S.; Velankar, S.; Kleywegt, G.J. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68 Pt 4*, 478–483. [CrossRef] [PubMed]
30. Stewart, A.; Grigorieff, N. Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy* **2004**, *102*, 67–84. [CrossRef] [PubMed]
31. Scheres, S.H. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.* **2012**, *415*, 406–418. [CrossRef] [PubMed]
32. Leschziner, A.E.; Nogales, E. The orthogonal tilt reconstruction method: An approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.* **2006**, *153*, 284–299. [CrossRef] [PubMed]
33. Radermacher, M. Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *J. Electron Microsc. Tech.* **1988**, *9*, 359–394. [CrossRef] [PubMed]
34. Tang, G.; Peng, L.; Baldwin, P.R.; Mann, D.S.; Jiang, W.; Rees, I.; Ludtke, S.J. EMAN2: An extensible image processing suite for electron microscopy. *J. Struct. Biol.* **2007**, *157*, 38–46. [CrossRef] [PubMed]
35. Penczek, P.A.; Frank, J.; Spahn, C.M. A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *J. Struct. Biol.* **2006**, *154*, 184–194. [CrossRef] [PubMed]
36. Brünger, A.T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **1992**, *355*, 472–475. [CrossRef] [PubMed]
37. Rossmann, M.G.; Bernal, R.; Pletnev, S.V. Combining electron microscopic with x-ray crystallographic structures. *J. Struct. Biol.* **2001**, *136*, 190–200. [CrossRef] [PubMed]
38. Ramachandran, G.N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95–99. [CrossRef]
39. Shapovalov, M.V.; Dunbrack, R.L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19*, 844–858. [CrossRef] [PubMed]
40. Chen, L.; He, J. A distance- and orientation-dependent energy function of amino acid key blocks. *Biopolymers* **2014**, *101*, 681–692. [CrossRef] [PubMed]
41. MathWorks. MATLAB Central. Available online: https://www.mathworks.com/matlabcentral/fileexchange/45325-efficient-2d-histogram--no-toolboxes-needed (accessed on 26 January 2018).
42. He, J.; Schmid, M.F.; Zhou, Z.H.; Rixon, F.; Chiu, W. Finding and using local symmetry in identifying lower domain movements in hexon subunits of the herpes simplex virus type 1 B capsid. *J. Mol. Biol.* **2001**, *309*, 903–914. [CrossRef] [PubMed]

43. Miao, Z.; Cao, Y. Quantifying side-chain conformational variations in protein structure. *Sci. Rep.* **2016**, *6*, 37024. [CrossRef] [PubMed]
44. Wlodawer, A.; Li, M.; Dauter, Z. High-Resolution Cryo-EM Maps and Models: A Crystallographer's Perspective. *Structure* **2017**, *25*, 1589–1597.e1. [CrossRef] [PubMed]