Protein Engineering, Design & Selection, 2016, 1–9
doi: 10.1093/protein/gzw027
Original Article

OXFORD

Original Article

Steric interactions determine side-chain conformations in protein cores

D. Caballero[1,2], A. Virrueta[2,3], C.S. O'Hern[1,2,3,4], and L. Regan[2,5,6,7,*]

[1]Department of Physics, Yale University, New Haven, CT 06520, USA, [2]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA, [3]Department of Mechanical Engineering and Materials Science, Yale University, New Haven, CT 06520, USA, [4]Graduate Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA, [5]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [6]Department of Chemistry, Yale University, New Haven, CT 06520, USA, and [7]Raymond and Beverly Sackler Institute for Biological, Physical, and Engineering Sciences, Yale University, New Haven, CT 06520, USA

*To whom correspondence should be addressed. E-mail: lynne.regan@yale.edu
Edited by Kent Kirshenbaum

Received 13 April 2016; Revised 11 June 2016; Accepted 12 June 2016

Abstract

We investigate the role of steric interactions in defining side-chain conformations in protein cores. Previously, we explored the strengths and limitations of hard-sphere dipeptide models in defining sterically allowed side-chain conformations and recapitulating key features of the side-chain dihedral angle distributions observed in high-resolution protein structures. Here, we show that modeling residues in the context of a particular protein environment, with both intra- and inter-residue steric interactions, is sufficient to specify which of the allowed side-chain conformations is adopted. This model predicts 97% of the side-chain conformations of Leu, Ile, Val, Phe, Tyr, Trp and Thr core residues to within 20°. Although the hard-sphere dipeptide model predicts the observed side-chain dihedral angle distributions for both Thr and Ser, the model including the protein environment predicts side-chain conformations to within 20° for only 60% of core Ser residues. Thus, this approach can identify the amino acids for which hard-sphere interactions alone are sufficient and those for which additional interactions are necessary to accurately predict side-chain conformations in protein cores. We also show that our approach can predict alternate side-chain conformations of core residues, which are supported by the observed electron density.

Key words: electron density, protein crystal structures, protein design, side-chain conformations, steric interactions

Introduction

One of the most incisive insights into the physical basis of protein structure was the work of Ramachandran and colleagues in the 1960s. They showed that steric interactions alone (i.e. the repulsive part of the Lennard–Jones interatomic potential) in an alanine dipeptide determine the allowed backbone dihedral angle (ϕ and ψ) combinations (Ramachandran *et al.*, 1963). Subsequently, these predicted backbone dihedral angle combinations were confirmed by protein crystal structures (Ramakrishnan and Ramachandran, 1965). Even today, agreement between the observed ϕ–ψ backbone dihedral angles and the predictions of the Ramachandran plot is a key metric of the quality of protein structures (Laskowski *et al.*, 1993; Chen *et al.*, 2010).

Although the Ramachandran hard-sphere dipeptide approach defines the sterically allowed backbone ϕ–ψ combinations, it does not specify which of the allowed backbone conformations will be adopted by a particular amino acid in a given protein. The repulsive part of the Lennard–Jones interatomic potential is just one contribution in a more complete potential energy function that would include, for example, hydrogen-bonding, hydrophobic, electrostatic and other interactions

with particular relative weights between them (Beauchamp *et al.*, 2012). However, given the importance of steric interactions that was so convincingly demonstrated in the work of Ramachandran and colleagues, we will employ a similar approach to specify the side-chain conformations of amino acids in protein cores.

In this manuscript, we delineate for which residues the hard-sphere model is able to recapitulate the observed side-chain conformations and, equally importantly, for which residues is it necessary to include additional interactions. We believe that this approach provides new insights into the dominant forces that determine the structure of protein cores. Thus, this approach both enhances our fundamental understanding of protein structure and provides new computational methods for protein design applications (Eriksson *et al.*, 1992; Peterson *et al.*, 2014).

In a previous work, we demonstrated that the hard-sphere dipeptide model was sufficient to recapitulate the observed (in a database of high-resolution protein crystal structures) side-chain dihedral angle distributions of all of the non-polar, aromatic and polar amino acids (Zhou *et al.*, 2011, 2012, 2013). This work gave several additional insights: (i) The hard-sphere dipeptide model is sufficient to recapitulate the observed side-chain dihedral angle distribution $P(\chi_1)$ for the polar amino acids Ser and Thr, without including hydrogen-bonding interactions (Zhou *et al.*, 2014); (ii) The hard-sphere model in the context of a regular α-helix (rather than a dipeptide mimetic) improved the quantitative agreement between the predicted and observed side-chain dihedral angle distributions for several amino acids, such as Ile and Phe (Zhou *et al.*, 2014); (iii) The hard-sphere dipeptide model identifies mechanisms for transitions between different allowed main chain and side-chain dihedral angle conformations (Caballero *et al.*, 2014, 2015); (iv) Although the hard-sphere dipeptide model correctly predicts the observed side-chain dihedral angle distributions $P(\chi_1)$ and $P(\chi_2)$ for Met, weak attractive interactions between hydrogens must be included to recapitulate the observed χ_3 distributions (Virrueta *et al.*, 2016). These studies provide the scientific underpinning for the work we describe here, where we explore the strengths and limitations of the hard-sphere model with both intra- and inter-residue steric interactions to predict the side-chain dihedral angle conformations adopted by particular amino acids in the context of protein cores.

Below, we show that the hard-sphere model with both intra- and inter-residue steric interactions predicts to within 20° the side-chain dihedral angle conformation of 97% of Leu, Ile, Val, Phe, Tyr, Trp and Thr core residues. We also gained several other interesting insights. Although the hard-sphere dipeptide model can correctly predict the allowed side-chain dihedral angle distributions $P(\chi_1)$ for both Thr and Ser, the hard-sphere model with both intra- and inter-residue steric interactions does not accurately predict the side-chain conformations for Ser in protein cores, even though it does for Thr. From this observation, we conclude that the positioning of Thr side chains is dominated by steric interactions (which we rationalize as being due to the presence of a bulky methyl group on C_β). In contrast, the positioning of the smaller Ser side chain is more significantly influenced by other forces, for example, hydrogen-bonding.

In addition to predicting high-occupancy side-chain conformations that match those in reported crystal structures, in some cases we predict additional allowed conformations. We further investigated this observation by analyzing the electron density distributions around those side chains. We found that for a few structures in the database of high-resolution protein crystal structures we studied, models with multiple side-chain conformations had been deposited. In other examples, although the deposited model included only a single conformation for that residue, when we calculated electron density from the deposited structure factor, we sometimes observed electron density corresponding to an alternate conformation. In all cases, where the electron density supports the existence of alternate conformations, our method predicts them.

In this manuscript, we will demonstrate that steric repulsion dominates the energetics and effectively specifies the side-chain conformations of amino acids in protein cores (Ponder and Richards, 1987; Lim and Sauer, 1989; Joh *et al.*, 2009). Our studies also reveal in which amino acids this effect is apparently significantly offset by other forces, and the power of this approach in revealing unappreciated alternative side-chain conformations.

Materials and methods

Datasets of protein crystal structures

In this study, we employed two ultra-high-resolution databases of protein crystal structures: 'Dunbrack 1.0 Å' (Shapovalov and Dunbrack, 2011) and 'HiQ54' (Leaver-Fay *et al.*, 2013). For both databases, the crystal structures possess few bond length, bond angle and backbone dihedral angle outliers. The HiQ54 database is composed of 54 non-redundant, single-chain, monomeric proteins that possess between 60 and 200 residues and do not include tightly bound or large ligands. All of the proteins have both a resolution and MolProbity score ≤1.4 Å (Chen *et al.*, 2010). The Dunbrack 1.0 Å database includes 220 proteins from the protein data bank (PDB) with a resolution of ≤1.0 Å, R-factors ≤0.2, side chain B-factors ≤30 Å^2 and sequence identity ≤50%. We tested the predicted side-chain dihedral angle distributions from the hard-sphere model against the distributions observed in the HiQ54 database. The Dunbrack 1.0 Å database was used to construct distributions of bond lengths and bond angles, which were used to construct the side chains for each core residue in the HiQ54 database.

Identification of core residues

Our analyses in this manuscript focus on residues in protein cores. Our definitions of core atoms and residues are similar to the ones we employed in a recent study of packing in protein cores (Gaines *et al.*, 2016). For an atom to be classified as a core atom, it cannot have empty space around it where a probe about the size of a water molecule (a sphere of radius $R = 1.4$ Å) can fit. We identify all points that are not located inside atoms and are a distance >1.4 Å from the surface of all atoms in each protein using Monte Carlo sampling. The closest atom to each of these points is then designated as a non-core atom. (See Supplementary Fig. S3 for a schematic that illustrates the method for identifying core atoms.) For a residue to be considered a core residue, it must only contain core atoms (including the hydrogens). Using this classification method, we find that as expected Ile, Leu, Val and Phe have the largest percentages of residues that are classified as core. The fractions of core residues for the eight amino acids (Leu, Ile, Val, Tyr, Phe, Trp, Thr and Ser) that we study are given in Table I. Cys and Met also occur in protein cores, but we did not include studies of Cys because of its ability to form disulfide bonds and we did not include studies of Met because weak attractive interactions are necessary to accurately predict the side-chain dihedral angle distribution $P(\chi_3)$ (Virrueta *et al.*, 2016).

Hard-sphere model

We obtain predictions for the side-chain conformations of residues in protein cores using two models: (1) the hard-sphere dipeptide mimetic model that includes only intra-residue steric interactions (Fig. 1a) and

Table I. Frequently occurring residues in protein cores

Amino acid	Total	Core	Core percentage (%)
Ile	360	72	20
Leu	565	103	18
Val	472	75	16
Phe	286	47	16
Trp	123	8	6.5
Tyr	281	13	5
Thr	436	23	5
Ser	439	22	5

Total number, and the number and percentage of residues designated as core for the neutral, non-polar and aromatic residues Leu, Ile, Val, Tyr, Phe, Trp, Thr and Ser in the HiQ54 database (Leaver-Fay *et al.*, 2013).

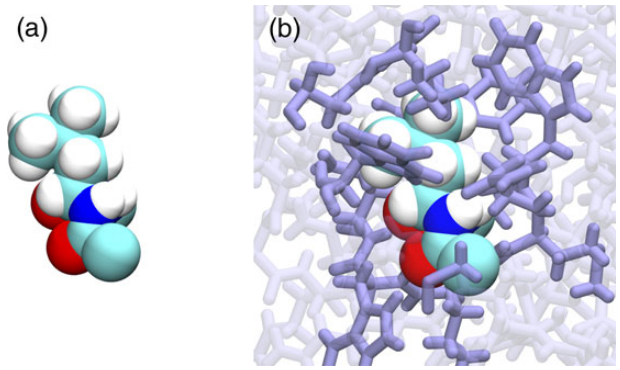

Fig. 1 Comparison of (**a**) the hard-sphere dipeptide model for Leu 31 in PDB 2NWD and (**b**) the hard-sphere model that includes both inter- and intra-residue steric interactions for Leu 31 in PDB 2NWD. The atoms in Leu 31 are shaded teal (carbon), blue (nitrogen), red (oxygen) and white (hydrogen). In (b), residues within 2.8 Å of Leu 31 are displayed in purple, whereas the remaining residues have been faded.

(2) the hard-sphere model for a residue in the context of its protein environment that includes both intra- and inter-residue steric interactions (Fig. 1b). For (1), we model each core residue in HiQ54 as a dipeptide mimetic (Caballero *et al.*, 2015). A dipeptide mimetic is a single amino acid (labeled i) plus the C_α, C and O atoms of the preceding amino acid ($i-1$) and the N, H and C_α atoms of the proceeding amino acid ($i+1$). Each atom is represented as a sphere with radius $\sigma_i/2$. We included seven atom types, N: 1.3 Å, O: 1.4 Å, hydroxyl O_H: 1.45 Å, C^3_{sp}: 1.5 Å, C^2_{sp}: 1.3 Å, H: 1.1 Å and amide hydrogen H_N: 1.0 Å. The atomic radii were obtained by minimizing the difference between the side-chain dihedral angle distributions predicted by the hard-sphere dipeptide model and those observed in protein crystal structures for a small set of amino acids. The atomic radii are similar to values of van der Waals radii reported in earlier studies (Caballero *et al.*, 2015; Gaines *et al.*, 2016). All atom representations of Leu, Ile, Val, Tyr, Phe, Trp, Thr and Ser residues that we study are shown in Supplementary Figs. S1 and S2.

To sample bond length and bond angle fluctuations, we performed Langevin dynamics (LD) simulations of hard-sphere models of dipeptide mimetics at temperature T for each core residue in the HiQ54 database (Caballero *et al.*, 2015). Atoms in the dipeptides interact via four potentials: U^{bl}, U^{ba}, U^{da} and U^{rlj}. The harmonic potential $U^{bl} = (1/2)\sum_{ij} K^{bl}_{ij}(r_{ij} - l_{ij})^2$ constrains the average separation r_{ij} between bonded atoms i and j to be equal to the bond length l_{ij} with standard deviation $\Delta l_{ij} = \sqrt{(k_BT)/K^{bl}_{ij}}$. The harmonic potential $U^{ba} = (1/2)\sum_{ijk} K^{ba}_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2$ constrains the average bond angle θ_{ijk} between bonded atoms i, j and k to be equal to θ^0_{ijk} with standard deviation $\Delta\theta_{ijk} = \sqrt{(k_BT)/(K^{ba}_{ijk})}$. The ω-backbone dihedral angle potential $U^{da} = \sum_{ijkl} K^{da}_{ijkl}(\cos\omega_{ijkl} - \cos\omega^0_{ijkl})^2$ constrains $\omega^0_{ijkl} \approx 180^\circ$ with standard deviation $\Delta\omega_{ijkl} = \sqrt{(k_BT)/(K^{da}_{ijkl})}$. The values of l_{ij}, θ^0_{ijk}, ω^0_{ijkl}, Δl_{ij}, $\Delta\theta_{ijk}$ and $\Delta\omega_{ijkl}$ for each pair, triple and quadruple of bonded atoms in the dipeptide were obtained from the Dunbrack 1.0 Å database. The interaction potentials allow all of the bond lengths and angles, as well as the backbone dihedral angle ω_{ijkl} of a given residue to fluctuate simultaneously around average values obtained from high-resolution protein crystal structures in the Dunbrack 1.0 Å database. In addition, during the LD simulations the backbone dihedral angles ϕ and ψ of a given residue fluctuate, but remain within ±10° of the crystal structure values, even though there are no explicit backbone dihedral angle interaction potentials for ϕ and ψ. We also include repulsive Lennard–Jones interactions (Weeks *et al.*, 1971) between all non-bonded atom pairs k and l in the dipeptide, $U^{rlj} = \varepsilon_R \sum_{k>l}(1 - ((\sigma_{kl})/(r_{kl}))^6)^2\Theta(1 - (r_{kl})/(\sigma_{kl}))$, where ε_R is the energy scale of the repulsive interaction, $\sigma_{kl} = (\sigma_k + \sigma_l)/2$, and Θ is the Heaviside step function that ensures that non-bonded atoms do not interact when they are not in contact. The hydrogen atoms were added using the REDUCE software program (Word *et al.*, 1999), which sets the bond lengths for C–H, N–H and S–H to 1.1, 1.0 and 1.3 Å, respectively, and the bond angles to 120° and 109.5° for bond angles involving C^2_{sp} and C^3_{sp} atoms. Additional dihedral angle degrees of freedom involving hydrogen atoms were chosen to minimize steric clashes. For model (1), for which we consider only intra-residue interactions, we fixed the terminal C_α atoms at locations $i-1$ and $i+1$, while all other atoms were allowed to fluctuate. The LD simulations were performed using a velocity Verlet integration scheme (Allen and Tildesley, 1987) for each dipeptide for 10^5 time steps with $\Delta t = 10^{-4}\, t_0$, where $t_0 = \sigma_H\sqrt{(m_H)/(\varepsilon_R)}$, m_H is the mass of hydrogen, and $N_s = 10^3$ 'snapshots' at equal time intervals were saved. These snapshots provide an ensemble of configurations of the given residue with different bond length and bond angle combinations and backbone dihedral angle values near those for the crystal structure. The temperature scale $k_BT/\varepsilon_R = 10^{-2}$ was chosen to be low enough such that the predicted side-chain dihedral angle distributions were independent of T (Caballero *et al.*, 2014).

To efficiently sample the space of side-chain dihedral angle conformations, we take each of the snapshots, fix the bond lengths, bond angles and backbone dihedral angles, rotate the side chain to sample each side conformation $(\chi_1, \ldots, \chi_n)$ in 5° intervals for each χ, and evaluate the total repulsive Lennard–Jones energy $U^{rlj}(\chi_1, \ldots, \chi_n)$. For each snapshot s and residue α, we calculate the Boltzmann weight, $P_{s,\alpha}(\chi_1,\ldots,\chi_n) \propto e^{(-U^{rlj}_{s,\alpha}(\chi_1,\ldots,\chi_n))/(k_BT)}$, average over snapshots $P_\alpha(\chi_1,\ldots,\chi_n) = 1/(N_s)\sum_s P_{s,\alpha}(\chi_1,\ldots,\chi_n)$, and normalize such that $\int P_\alpha(\chi_1,\ldots,\chi_n)\, d\chi_1,\ldots,d\chi_n = 1$ to determine the side-chain dihedral angle distribution for each residue. We can also calculate an average over residues to obtain the side-chain dihedral angle distribution $P(\chi_1,\ldots,\chi_n)$ for each residue type.

For the hard-sphere model (2), for which we consider each residue in the context of its protein environment, the methodology is the same as for model (1), except the total repulsive Lennard–Jones potential U^{rlj} includes all non-bonded atom pairs k and l, where k and l can be located on the same residue or different residues. As a side chain is rotated about its dihedral angles, the rest of the protein structure is held fixed. A comparison of $P_\alpha(\chi_1,\ldots,\chi_n)$ obtained from models (1) and (2) allows us to assess the importance of the protein environment in determining side-chain conformations.

Comparison of observed and predicted side-chain conformations

For model (2), for each residue α, we determined the side-chain conformation $(\chi_1^{HS},\ldots,\chi_n^{HS})$ with the largest value of $P_\alpha(\chi_1,\ldots,\chi_n)$. As a measure of the accuracy of the prediction, we calculated the difference in the side-chain conformations $\Delta\chi = \sqrt{(\chi_1^{xtal}-\chi_1^{HS})^2+\ldots+(\chi_n^{xtal}-\chi_n^{HS})^2}$, where $(\chi_1^{xtal},\ldots,\chi_n^{xtal})$ is the side-chain conformation of the residue in the protein crystal structure. For each residue type, we also calculated the cumulative probability distribution $C(\Delta\chi)=\int_0^{\Delta\chi}P(\Delta\chi')d(\Delta\chi')$ of side-chain conformation differences between the predicted and observed values less than $\Delta\chi$, where $P(\Delta\chi)$ is the probability distribution of side-chain conformation differences $\Delta\chi$.

Analysis of electron density maps

For each protein in the HiQ54 dataset, we used the software package PHENIX (Adams *et al.*, 2010) to extract the observed electron density F_o. Using the Computational Crystallography Toolbox (CCTBX) library (Grosse-Kunstleve *et al.*, 2002), we analyzed F_o for each core residue. We identify the grid points at which the electron density was above 1.5 standard deviations and set the density at grid points that do not satisfy this condition to zero. For each core residue, we overlay side-chain conformations (with bond lengths and angles given by the crystal structure) onto the grid. We then used a tri-linear interpolation to estimate the electron density at each heavy atom location in the side chain. We rotate the side chain to sample all conformations $(\chi_1,\ldots,\chi_n)$ in 5° intervals for each χ and calculate the geometric mean $F(\chi_1,\ldots,\chi_n)=\sqrt[N]{\prod_k F_o(C_k)}$, where $F_o(C_k)$ is the electron density evaluated at the kth carbon, and the product over $k=1,\ldots,N$ includes all carbons in the side chain after C_β. The geometric mean F was computed to eliminate signal redundancy arising from structural symmetries in the side-chain conformations. $F(\chi_1,\ldots,\chi_n)$ was then filtered by setting any values that were below half of the global maximum in F to zero. We then normalized the integral of $F(\chi_1,\ldots,\chi_n)$ over side-chain dihedral angles to unity, $\int d\chi_1,\ldots,d\chi_n F(\chi_1,\ldots,\chi_n)=1$.

Results

As discussed in the introduction, the hard-sphere dipeptide model with intra-residue, but not inter-residue steric interactions (Fig. 1), is able to predict the multiple *possible* side-chain conformations of uncharged residues observed in proteins. However, to predict the *specific* side-chain conformation of a particular residue in the context of the protein core, one must also include inter-residue steric interactions. As an example, in Fig. 2 we compare the side-chain dihedral angle distribution for Ile residues (a) observed in the cores of protein structures in the Dunbrack 1.0 Å database to the (b) predicted distribution for these same residues using the hard-sphere dipeptide model. In the observed distribution for Ile, the three most highly probable rotamer boxes are 6, 3 and 4. For the predicted distributions, the same boxes are most probable, but the rotamer probabilities differ quantitatively; for box 6, the difference is roughly 20%.

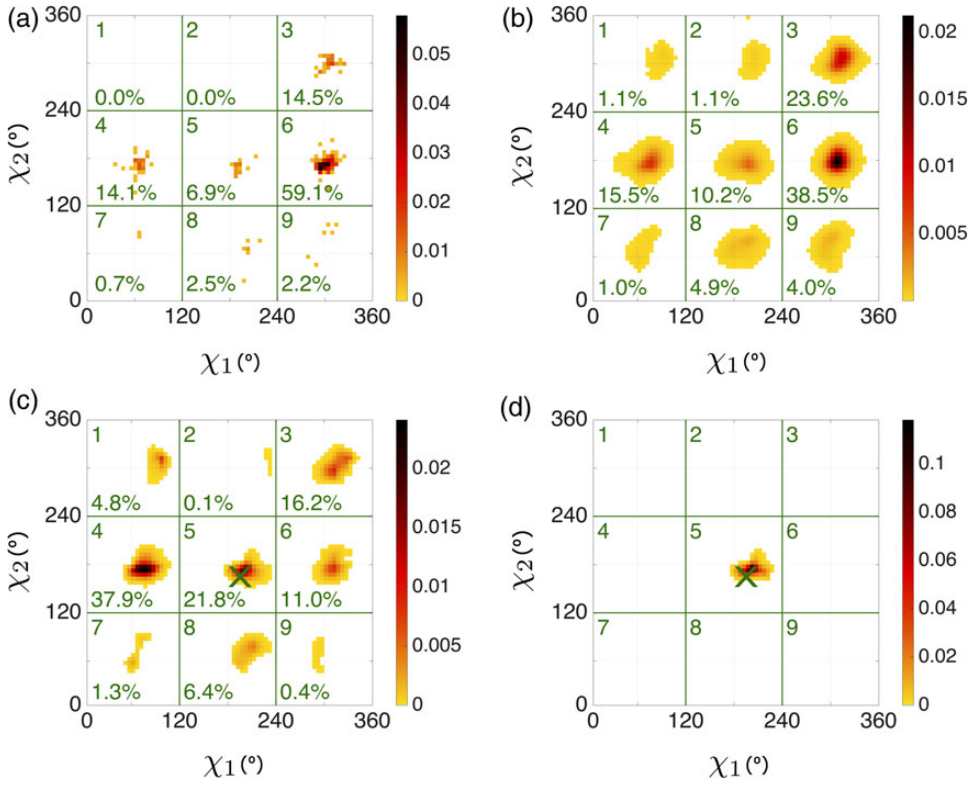

Fig. 2 Comparison of the side-chain dihedral angle distributions $P(\chi_1,\chi_2)$ for the 276 core Ile residues observed in the Dunbrack 1.0 Å database (Shapovalov and Dunbrack, 2011) (**a**) and the distribution predicted using the hard-sphere dipeptide model for the same residues (**b**). Comparison of the side-chain dihedral distributions $P(\chi_1,\chi_2)$ for Ile 56 in PDB 2NWD predicted using (**c**) the hard-sphere dipeptide model and (**d**) the hard-sphere model that includes both intra- and inter-residue steric interactions. In (c) and (d), the green cross indicates the side-chain conformation of Ile 56 in 2NWD. In all panels, the side-chain dihedral angle distribution is normalized such that $\int d\chi_1 d\chi_2 P(\chi_1,\chi_2)=1$, the probabilities increase from light to dark, and the percentages give the fractional probabilities in each of the nine square bins.

We show that, for residues in protein cores, the hard-sphere model with both intra- and inter-residue steric interactions can predict with high accuracy their specific side-chain dihedral angle conformations. As an example, in Fig. 2d we show that the predicted side-chain dihedral angle distribution for the hard-sphere model for Ile 56 in PDB 2NWD (with backbone dihedral angles $\phi = -65°$ and $\psi = -29°$) is strongly peaked near $\chi_1 = 195°$ and $\chi_2 = 165°$, which is essentially identical to the crystal structure values (indicated by the green cross). In contrast, when using the hard-sphere dipeptide model of this residue (with crystal structure values of ϕ and ψ), side-chain conformations in box 4 are predicted to be the most probable (with finite probabilities also predicted in boxes 3, 5, 6 and 8). In Fig. 3, we compare the results from the hard-sphere dipeptide model and the hard-sphere model that includes both intra- and inter-residue steric interactions for a residue with an aromatic ring, Phe 94 from PDB 1SAU (with $\phi = -126°$ and $\psi = 87°$), and an amino acid with a smaller side-chain and only one side-chain dihedral angle, Val 57 from PDB 1X6X (with $\phi = -109°$ and $\psi = 143°$). The results are similar to those in Fig. 2c and d for Ile: the hard-sphere dipeptide model predicts multiple highly probable side-chain conformations, whereas the hard-sphere model with both intra- and inter-residue steric interactions predicts a single strongly peaked side-chain dihedral angle distribution located near the crystal structure value. These examples illustrate that dense packing of residues in protein cores selects the particular side-chain conformations that occur for each residue (Gaines *et al.*, 2016).

In Fig. 4a and b, we compare the side-chain dihedral angle distribution obtained for the 72 core Leu residues (Table I) in the HiQ54 database to the distribution predicted using the hard-sphere model with both intra- and inter-residue steric interactions. The observed and predicted distributions are very similar. In fact, the observed and predicted probabilities in each rotamer box differ by <1%. The observed side-chain dihedral angle distributions for the core Ile residues in the Dunbrack 1.0 Å (Fig. 2a) and HiQ54 (Fig. 4a) databases differ quantitatively, but not qualitatively. For both databases, the most probable rotamer boxes in order of decreasing probability are 6, 4, 3 and 5. However, the values of the rotamer box probabilities differ quantitatively because the backbone dihedral angle distributions are different for the two databases (Dunbrack and Karplus, 1994; Dunbrack and Cohen, 1997). We find that the hard-sphere model with both intra- and inter-residue steric interactions is also able to predict the rotamer box probabilities observed in the Dunbrack 1.0 Å database to within 1%.

In addition to comparing the predicted and observed side-chain dihedral angle distributions, we calculated the difference between the predicted and observed side-chain conformations $\Delta\chi = \sqrt{(\chi_1^{\rm xtal} - \chi_1^{\rm HS})^2 + \cdots + (\chi_n^{\rm xtal} - \chi_n^{\rm HS})^2}$ for each individual core residue in the HiQ54 database, where $(\chi_1^{\rm xtal}, \ldots, \chi_n^{\rm xtal})$ is the side-chain conformation of the residue in the protein crystal structure and $(\chi_1^{\rm HS}, \ldots, \chi_n^{\rm HS})$ is the most probable side-chain conformation predicted by the hard-sphere model with both intra- and inter-residue steric interactions. In Fig. 5, we show the cumulative probability distributions $C(\Delta\chi)$ (as defined in the Materials and Methods section) for all instances of the amino acids Val, Leu, Ile, Phe, Tyr and Trp that occur in protein cores in the HiQ54 database. The data in Fig. 5 show excellent agreement between the observed and predicted side-chain conformations for nearly all instances of these six residues. Specifically, for 97% of all core residues studied, the predicted and observed side-chain conformations differ by <20°. For Val and Trp, all instances of these residues are correctly predicted to within ≈10°. This level of accuracy for the hard-sphere model is significantly higher than that reported using other scoring functions for rotamer recovery applications (Peterson *et al.*, 2014).

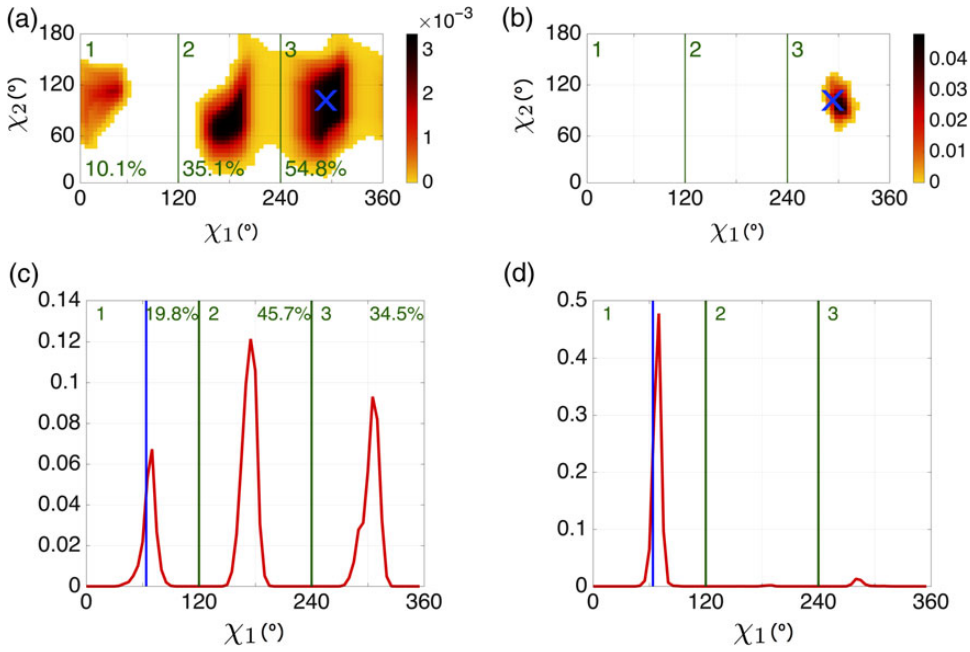

Fig. 3 Comparison of the predicted side-chain distributions $P(\chi_1,\chi_2)$ for Phe 94 in PDB 1SAU using (**a**) the hard-sphere dipeptide model and (**b**) the hard-sphere model including both intra- and inter-residue steric interactions. We also compare the predicted $P(\chi_1)$ for Val 57 in PDB 1X6X obtained from (**c**) the hard-sphere dipeptide model and (**d**) the hard-sphere model including both intra- and inter-residue steric interactions. In (a) and (b), the blue crosses indicate the side-chain conformation of Phe 94 in 1SAU, the side-chain dihedral angle distribution is normalized such that $\int d\chi_1 d\chi_2 P(\chi_1,\chi_2) = 1$, and the probabilities increase form light to dark. In (c) and (d), the solid blue vertical lines indicate the side-chain conformation of Val 57 in 1X6X and the side-chain dihedral angle distribution is normalized such that $\int d\chi_1 P(\chi_1) = 1$. In (a) and (c), the percentages give the fractional probabilities in each of the nine or three rotamer bins, respectively.

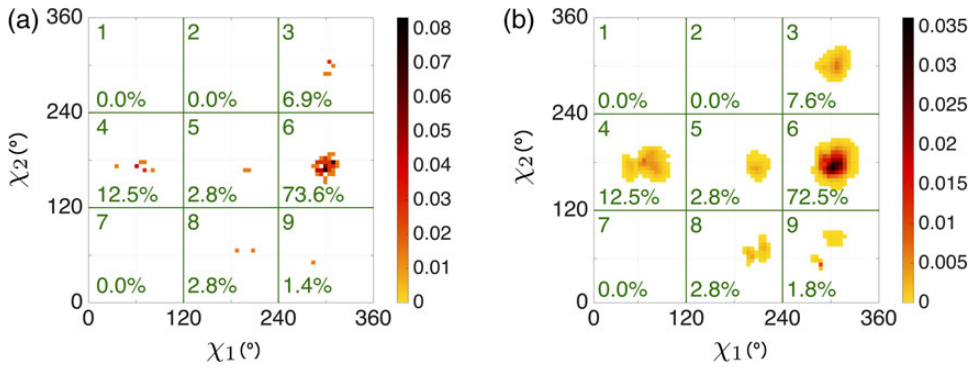

Fig. 4 Comparison of the side-chain dihedral angle distributions $P(\chi_1,\chi_2)$ for the 72 core Ile residues (**a**) observed in the HiQ54 database (Leaver-Fay *et al.*, 2013) and (**b**) predicted using the hard-sphere model with both intra- and inter-residue steric interactions. (See the description for calculating the observed and predicted $P(\chi_1,\chi_2)$ in the Materials and Methods section.) In (a) and (b), the side-chain dihedral angle distributions are normalized such that $\int d\chi_1 d\chi_2 P(\chi_1,\chi_2) = 1$ and the probabilities increase from light to dark. The percentages give the fractional probabilities found in each of the nine rotamer boxes, which are nearly identical for the observed and predicted side-chain dihedral angle distributions.

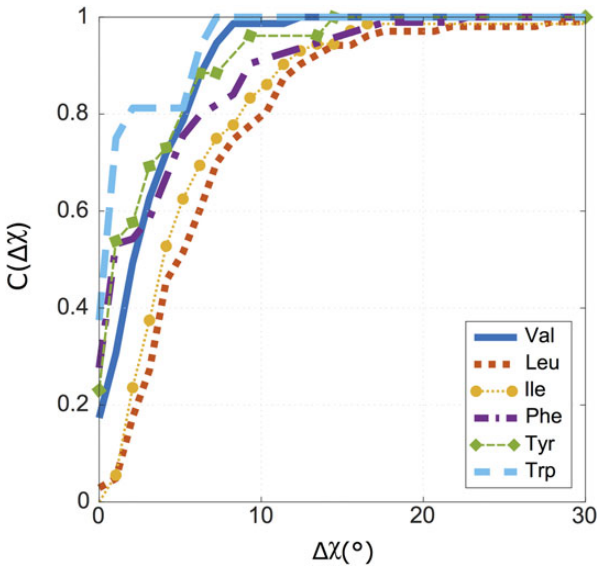

Fig. 5 Cumulative distribution $C(\Delta\chi)$ of the difference $\Delta\chi$ between the side-chain conformation for a given core residue (Val, Leu, Ile, Phe, Tyr or Trp) in each crystal structure in the HiQ54 database and that predicted using the hard-sphere model that includes both intra- and inter-residue steric interactions. For these six residues, we can predict the side-chain conformations within ~20° for more than 97% of the core residues in the HiQ54 database.

In addition to the six amino acids (Val, Leu, Ile, Phe, Tyr and Trp) studied in Fig. 5, we performed the same analysis for the two polar amino acids with hydroxyl groups in their side-chains, Thr and Ser. For both Thr and Ser, we find that the hard-sphere dipeptide model is able to recapitulate the main features of the observed side-chain dihedral angle distribution $P(\chi_1)$. In Fig. 6b, we show that the hard-sphere model with both intra- and inter-residue steric interactions predicts the probabilities that are observed in the three rotamer bins to <1% for Thr. The results for Ser are not as accurate, but the hard-sphere model predicts the probabilities observed in the three rotamer bins to within 5% (Fig. 6c). These results are qualitatively similar to those presented in Zhou *et al.* (2014) for the hard-sphere dipeptide model for Thr and Ser.

However, we find a significant difference between the accuracy of the hard-sphere predictions between Thr and Ser when we consider the cumulative distribution $C(\Delta\chi)$. In Fig. 6a, we show that the hard-sphere model with both intra- and inter-residue steric interactions predicts the observed side-chain dihedral angle χ_1 to within 15° for all instances of core Thr residues in the HiQ54 database. This accuracy is similar to that found for the six previously studied amino acids: Val, Leu, Ile, Phe, Tyr and Trp. In contrast, the hard-sphere model is only able to correctly predict the observed side-chain dihedral angle χ_1 to within 20° for <60% of the core Ser residues in the HiQ54 database (Fig. 6a). While the hard-sphere model can recapitulate the main features of the observed side-chain dihedral angle distribution for Ser on average, the hard-sphere model cannot accurately specify the side-chain conformation for each individual core Ser residue.

In contrast to those for Val, Leu, Ile, Phe and Trp, the side chains for Tyr, Ser and Thr include a hydroxyl group, which can form hydrogen bonds with other residues (Baker and Hubbard, 1984; McGregor *et al.*, 1987; Pace *et al.*, 2001). Hydrogen bonds can significantly decrease the oxygen–hydrogen separation below the sum of the oxygen and hydrogen radii $(\sigma_O + \sigma_H)/2$ used in the hard-sphere model. Thus, in the current hard-sphere model, hydrogen bonds are strongly disfavored. We speculate that because the Tyr side-chain includes a bulky aromatic ring, the number of sterically allowed side-chain conformations for Tyr is small, and because this effect dominates (Pace *et al.*, 2001), the hard-sphere model can accurately predict the side-chain conformations for Tyr in protein cores. Similarly, the side chain for Thr includes C_β, a hydroxyl group and a methyl group, while the side chain for Ser only includes C_β and a hydroxyl group. We speculate that the methyl group on the Thr side-chain significantly limits the number of sterically allowed side-chain formations for Thr residues in protein cores and thus the hard-sphere model can accurately predict its side-chain conformations. In contrast, Ser side chains are smaller than those for Tyr and Thr and are not similarly constrained by steric interactions. As a result, $C(\Delta\chi)$ for Ser is much below that for Thr and Tyr. This observation is important, because it indicates that the inclusion of hydrogen bonding will not be equally important for all amino acids. In future work, we will determine the effect of the addition to the hard-sphere model of favorable energy contributions from

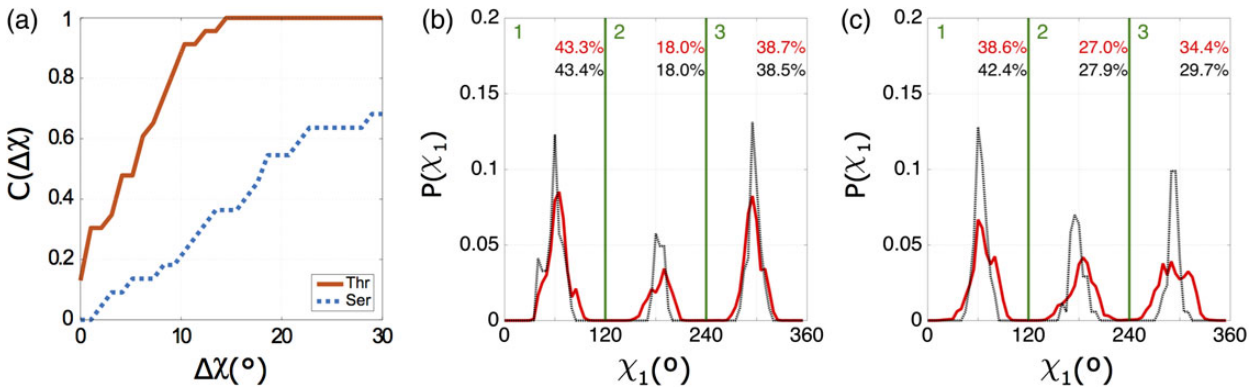

Fig. 6 (**a**) Cumulative distribution $C(\Delta\chi)$ of the difference $\Delta\chi = |\chi_1^{xtal} - \chi_1^{HS}|$ between the side-chain conformation for a given core Thr or Ser residue observed in each crystal structure in the HiQ54 database and that predicted using the hard-sphere model that includes both intra- and inter-residue steric interactions. In (**b**) and (**c**), we show the side-chain dihedral angle distribution $P(\chi_1)$ for core Thr and Ser residues, respectively, observed in the Dunbrack 1.0 Å database (red lines) and predicted (dotted black lines) using the hard-sphere model with both intra- and inter-residue steric interactions. The percentages on top (bottom) give the fractional probabilities for the observed (predictions) distributions in each of the three rotamer bins.

hydrogen-bonding to the predictions of the side-chain dihedral angle conformations for Ser, Thr and Tyr.

For most of the core residues in the HiQ54 database, the side-chain distribution predicted by the hard-sphere model possesses a single, strong peak located near the side-chain dihedral angle conformation observed in the crystal structure (e.g. Leu 25 in PDB 1JBE shown in Fig. 7a). Out of the 341 core residues we studied (excluding Ser), the side-chain dihedral angle distributions predicted by the hard-sphere model for 31 of those residues (9%) possess multiple peaks: 25 Leu, 5 Ile and 1 Val. An example of a core residue for which the hard-sphere model predicts multiple possible side-chain conformations is Leu 74 in PDB 2OSS as shown in Fig. 7b.

Since the hard-sphere model occasionally predicts multiple side-chain conformations for core residues in HiQ54, we also analyzed the deposited electron density maps to determine whether there is experimental evidence that these residues do indeed sample multiple conformations. For most of the residues where the hard-sphere model predicts a single, strong peak (e.g. Leu 25 in PDB 1JBE shown in Fig. 7a), we find that the electron density displays only one side-chain conformation that agrees with the predicted value. In the left column of Fig. 7b and c, we show two cases (Leu 74 in 2OSS and Leu 68 in 2V1M) where the hard-sphere model predicts two side-chain conformations. For Leu 74 in 2OSS, two models for the side-chain conformation were deposited that fit the electron density, and these match the two side-chain conformations predicted by the hard-sphere model. For 2V1M, a model with only a single side-chain conformation for Leu 68 has been deposited ($\chi_1 = 292°$, $\chi_2 = 67°$), yet our analysis shows that there is electron density corresponding to both conformations ($\chi_1 = 292°$, $\chi_2 = 67°$ and $\chi_1 = 313°$, $\chi_2 = 190°$) predicted by the hard-sphere model (center and right columns of Fig. 7c). Out of the 31 residues for which the hard-sphere model predicts multiple side-chain conformations, our analysis of the electron density suggests that six of these residues sample multiple conformations. However, multiple side-chain conformations have been deposited in the PDB for only one of these six residues.

There are a total of seven non-Met core hydrophobic residues in the HiQ54 database for which multiple side-chain conformations have been deposited in the PDB. The hard-sphere model predicts the same multiple side-chain conformations that have been deposited for three of the seven residues. The hard-sphere model shows that one of these residues (Ile 37 in PDB 1Z2U) should not be modeled with two conformations using an electron density threshold of at least 1.5 standard deviations (see Supplementary Fig. S4). Multiple side-chain conformations for the remaining four residues are not predicted by the hard-sphere model. A more detailed comparison of the predictions of the hard-sphere model and the identification of multiple conformations in the electron density maps, for example as a function of temperature, will be carried out in a future work.

Discussion

Our work represents an important step in a systematic approach for quantifying and understanding the dominant forces that specify protein structure. In prior work on dipeptide mimetics (Zhou *et al.*, 2014; Caballero *et al.*, 2015), we demonstrated that steric interactions specify the allowed side-chain conformations in non-polar, aromatic and polar residues. Our predictions based on dipeptides mostly correspond with the side-chain dihedral angle distributions observed in protein crystal structures. By investigating in detail where the predicted and observed distributions differ, we gain new insights into the energetics of protein structure.

In this manuscript, we employ the hard-sphere model in the context of the environment of protein cores. Whereas in prior work, we focused on identifying all sterically allowed conformations as determined in a dipeptide mimetic, here we predict the particular side-chain conformations that individual Leu, Ile, Val, Phe, Tyr, Trp, Thr and Ser residues adopt in protein cores.

Our studies have revealed four fundamental insights: (i) Steric interactions are the dominant force in specifying side-chain conformations of residues in protein cores. (ii) For seven of the amino acids frequently (≈81%) found in protein cores (Leu, Ile, Val, Phe, Tyr, Trp and Thr) steric considerations alone allow us to predict 97% of their side-chain conformations to within 20°. (iii) The hard-sphere model with both intra- and inter-residue steric interactions provides accurate predictions for Thr, but not for Ser. We speculate that steric interactions dominate for Thr residues because of the methyl group on C_β and thus hydrogen-bonding plays a minor role in specifying Thr side-chain conformations. Conversely, we find that steric interactions

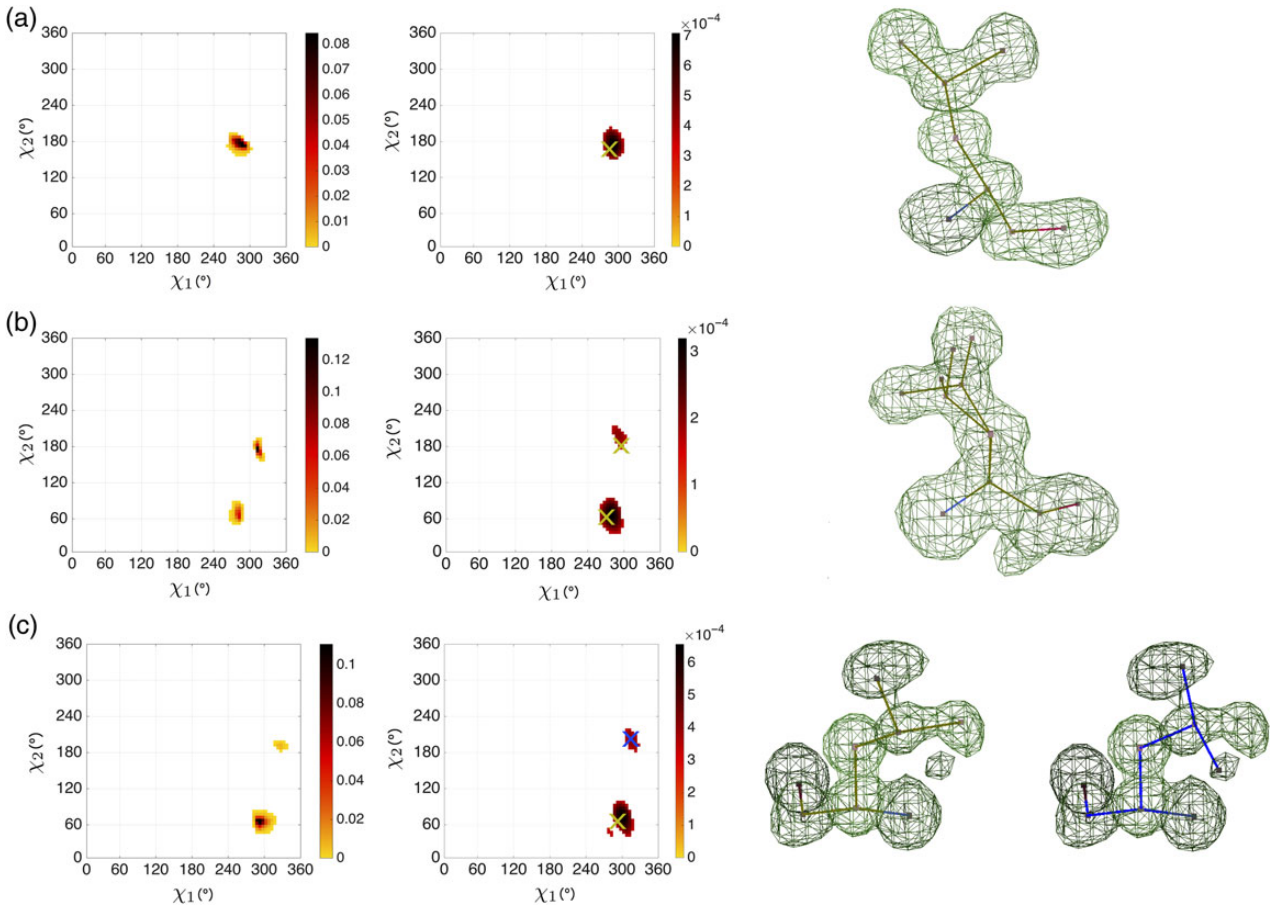

Fig. 7 The side-chain dihedral angle distribution $P(\chi_1,\chi_2)$ (left column) predicted using the hard-sphere model with both intra- and inter-residue interactions for Leu 25 in 1JBE (top row), Leu 74 in 2OSS (middle row) and Leu 68 in 2V1M (bottom row). In the middle column, we show the probability that the observed electron density is above a threshold of 1.5 standard deviations as a function of χ_1 and χ_2 for each of the three Leu residues (from top to bottom). In the left and middle columns, the distributions are normalized such that the integral over both side-chain dihedral angles is unity. The brown crosses in the middle column indicate the side-chain conformations of the models that have been deposited in the PDB. For Leu 25 (top row) and Leu 68 (bottom row), only one conformation has been deposited, whereas two model conformations have been deposited for Leu 74 (middle row). In the right column, we display the observed electron density F_o (green mesh) for the three Leu residues (with thresholds at 3, 2 and 3.5 standard deviations for Leu 25, 74 and 68, respectively), as well as brown connections between side-chain atom centers that indicate the side-chain conformations for each deposited model. For Leu 68 (bottom row), the hard-sphere model predicts an alternate conformation at $\chi_1 = 313°$, $\chi_2 = 190°$, indicated by the blue cross. In this case, the observed electron density shows an alternate side-chain conformation (indicated by blue connections between side-chain atom centers) even though only one model has been deposited in the PDB.

are insufficient to specify the side-conformations of Ser residues in protein cores. Steric interactions are not as important for Ser residues because the side chain is smaller and hydrogen-bonding interactions can play a more significant role. (iv) Our analysis not only predicts high-occupancy side-chain conformations, but also reveals alternate conformations (Lang *et al.*, 2010). In some cases, models have been deposited with multiple side-chain conformations for a particular residue. For these, the hard-sphere model also predicts the multiple conformations and an analysis of the electron density is consistent with the predictions. In other cases, we have identified residues for which the hard-sphere model predicts multiple conformations and the electron density is consistent with multiple conformations, yet the deposited crystal structure model includes only one conformation. In some cases, there is no electron density corresponding to alternate conformations predicted by the hard-sphere model. For these, we speculate that transitioning between the two conformations is prohibitive at temperatures for which the structure was crystallized. It will therefore be important to investigate the temperature dependence of multiple occupancy side-chain conformations in protein crystal structures and compare such experimental results with our predictions (Rasmussen *et al.*, 1992; Tilton *et al.*, 1992; van den Bedem *et al.*, 2013; Keedy *et al.*, 2015).

These results have a number of important implications for protein structure prediction and design, which are the focus of our current work. First, we are now able to perform direct comparisons of side-chain predictions (i.e. $C(\Delta\chi)$) in protein cores of the HiQ54 dataset from the hard-sphere model against those of other modeling strategies, for example, the widely used Rosetta modeling software. Such studies will include two modes for sampling the side-chain dihedral angle conformation space: (i) single rotations as performed in the present study, where we rotate the side chain of a single residue (with the rest of the protein held fixed) and (ii) collective rotations, where we rotate the side chains of multiple residues simultaneously. In future studies, we will also test our hard-sphere model predictions of side-chain conformations against the results from mutation studies in protein cores and at protein–protein interfaces.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgements

The authors thank R. L. Dunbrack, Jr., and J. S. and D. C. Richardson for providing the Dunbrack 1.0 Å and HiQ54 databases of protein crystal structures, respectively, and for their interest in this work and helpful discussions. The authors also thank J. C. Gaines for providing the algorithm for determining the core residues of protein crystal structures.

Funding

The authors acknowledge support from the National Science Foundation (Grant NSF-PHY-1522467 to L.R., D.C. and C.S.O.; Grant NSF-DMR-1307712 to L.R.); the Ford Foundation Pre-Doctoral Fellowship program (to A.V.); the National Science Foundation Graduate Research Fellowships program (to A.V.); and the Raymond and Beverly Sackler Institute for Biological, Physical and Engineering Sciences (to D.C., C.S.O., L.R. and A.V.). The authors benefited from the Facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center and acknowledge the National Science Foundation (Grant No. CNS 08-21132) that in part funded acquisition of the computational facilities.

References

Adams,P.D., Afonine,P.V., Bunkoczi,G., *et al*. (2010) *Acta Crystallogr.*, **D66**, 213–221.

Allen,M.P. and Tildesley,D.J. (1987) *Computer Simulations of Liquids*. Oxford University Press, New York.

Baker,E.N. and Hubbard,R.E. (1984) *Prog. Biophys. Mol. Biol.*, **44**, 97–179.

Beauchamp,K.A., Lin,Y.-S., Das,R. and Pande,V.S. (2012) *J. Chem. Theory Comput.*, **8**, 1409–1414.

Caballero,D., Määttä,J., Zhou,A.Q., Sammalkorpi,M., O'Hern,C.S. and Regan,L. (2014) *Prot. Sci.*, **23**, 970–980.

Caballero,D., Smith,W.W., O'Hern,C.S. and Regan,L. (2015) *PROTEINS*, **83**, 1488–1499.

Chen,V.B., Arendall,W.B., Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral, G.J., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2010) *Acta Cryst. D*, **66**, 12–21.

Dunbrack,R.L. and Cohen,F.E. (1997) *Prot. Sci.*, **6**, 1661–1681.

Dunbrack,R.L. and Karplus,M. (1994) *Nat. Struct. Mol. Biol.*, **1**, 334–340.

Eriksson,A.E., Baase,W.A., Zhang,X.J., Heinz,D.W., Blaber,M., Baldwin,E.P. and Matthews,B.W. (1992) *Science*, **255**, 178–184.

Gaines,J.C., Smith,W.W., Regan,L. and O'Hern,C.S. (2016) *Phys. Rev. E*, **93**, 032415.

Grosse-Kunstleve,R.W., Sauter,N.K., Moriarty,N.W. and Adams,P.D. (2002) *J. Appl. Cryst.*, **35**, 126–136.

Joh,N.H., Oberai,A., Yang,D., Whitelegge,J.P. and Bowie,J.U. (2009) *J. Am. Chem. Soc.*, **131**, 10846–10847.

Keedy,D.A., Kenner,L.R., Warkentin,M., *et al*. (2015) *eLife*, **4**, e07574.

Lang,P.T., Ng,H.L., Fraser,J.S., Corn,J.E., Echols,N., Sales,M., Holton,J.M. and Alber,T. (2010) *Prot. Sci.*, **19**, 1420–1431.

Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) *J. Appl. Crystalogr.*, **26**, 283–291.

Leaver-Fay,A., O'Meara,M.J., Tyka,M., *et al*. (2013) *Methods Enzymol.*, **523**, 109–143.

Lim,W. and Sauer,R.T. (1989) *Nature*, **339**, 31–36.

McGregor,M.J., Islam,S.A. and Sternberg,M.J. (1987) *J. Mol. Biol.*, **198**, 295–310.

Pace,C.N., Horn,G., Hebert,E.J., Bechert,J., Shaw,K., Urbanikova,L., Scholtz,J.M. and Sevcik,J. (2001) *J. Mol. Biol.*, **312**, 393–404.

Peterson,L.X., Kang,X. and Kihara,D. (2014) *PROTEINS*, **82**, 1971–1984.

Ponder,J.W. and Richards,F.M. (1987) *J. Mol. Biol.*, **193**, 775–791.

Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) *J. Mol. Biol.*, **7**, 95–99.

Ramakrishnan,C. and Ramachandran,G.N. (1965) *Biophys. J.*, **5**, 909–933.

Rasmussen,B.F., Stock,A.M., Ringe,D. and Petsko,G.A. (1992) *Nature*, **357**, 423–424.

Shapovalov,M.S. and Dunbrack,R.L. (2011) *Structure*, **19**, 844–858.

Tilton,R.F., Jr., Dewan,J.C. and Petsko,G.A. (1992) *Biochemistry*, **31**, 2469–2481.

van den Bedem,H., Bhabha,G., Yang,K., Wright,P.E. and Fraser,J.S. (2013) *Nat. Med.*, **10**, 896–902.

Virrueta,A., O'Hern,C.S. and Regan,L. (2016) *PROTEINS*, **84**, 900–911.

Weeks,J.D., Chandler,D. and Andersen,H.C. (1971) *J. Chem. Phys.*, **54**, 5237–5247.

Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) *J. Mol. Biol.*, **285**, 1735–1747.

Zhou,A.Q., O'Hern,C.S. and Regan,L. (2011) *Prot. Sci.*, **20**, 1166–1171.

Zhou,A.Q., O'Hern,C.S. and Regan,L. (2012) *Biophys. J.*, **102**, 2345–2352.

Zhou,A.Q., Caballero,D., O'Hern,C.S. and Regan,L. (2013) *Biophys. J.*, **105**, 2403–2411.

Zhou,A.Q., O'Hern,C.S. and Regan,L. (2014) *PROTEINS*, **82**, 2574–2584.