PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

Title:

Distance-depending hydrophobic-hydrophobic contacts in protein folding simulations

Authors:

Angelo Onofrio[1§], Giovanni Parisi[1§], Giuseppe Punzi[1§], Simona Todisco[1], Maria Antonietta Di Noia[2], Fabrizio Bossis[1], Antonio Turi[1], Anna De Grassi[1**#], Ciro Leonardo Pierri[1*#]

Author affiliations and contributions:

[1]Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Via Orabona 4, 70125, Bari, Italy.
[2]Department of Sciences, University of Basilicata, via N. Sauro 85, 85100 Potenza, Italy.

[§] Equal contribution

[#] Joint senior authors

Corresponding author information:

* Ciro Leonardo Pierri

Department of Biosciences, Biotechnologies and Biopharmaceutics

University of Bari

Via Orabona 4, 70125, Bari, Italy

tel: +39-0805442772

fax: +39-0805442770

email: ciroleopierri@gmail.com; ciro.pierri@uniba.it

Corresponding author information:

** Anna De Grassi

Department of Biosciences, Biotechnologies and Biopharmaceutics

University of Bari

Via Orabona 4, 70125, Bari, Italy

tel: +39-0805443614

fax: +39-0805442770

email: anna.degrassi@uniba.it

Keywords: secondary structure elements and small contact regions; no reverse self avoiding walk; alpha-helix, beta-sheet; hydrophobic driving force; very short-range interactions, short-range interactions, medium-range interactions, long-range interactions; lattices for protein folding

Abbreviations: bcc, body center cubic ; fcc, face center cubic; HH, hydrophobic-hydrophobic; PP, polar-polar; HP, hydrophobic-polar

Abstract

Successful prediction of protein folding from an amino acid sequence is a challenge in computational biology. In order to reveal the geometric constraints that drive protein folding, highlight those constraints kept or missed by distinct lattices and for establishing which class of intra- and inter-secondary structure element interactions is the most relevant for the correct folding of proteins, we have calculated inter alpha carbon distances in a set of 42 crystal structures consisting of mainly helix, mainly sheet or mixed conformations. The inter alpha carbon distances were also calculated in several lattice "hydrophobic-polar" models built from the same protein set. We found that helix structures are more prone to form "hydrophobic-hydrophobic" contacts than beta-sheet structures. At distance lower than or equal to 3.8 Å (very short-range interactions), "hydrophobic-hydrophobic" contacts are almost absent in the native structures, while they are frequent in all the analyzed lattice models. At distances in-between 3.8 and 9.5 Å (short-/medium-range interactions), the best performing lattices is the body-center-cubic lattice for reproducing mainly helix structures. If protein structures contain sheet portions, lattice performances get worse with few exceptions observed with double-tetrahedral and body-center-cubic lattices. Finally, we can observe that *ab initio* protein folding algorithms, i.e. those based on the employment of lattices and Monte Carlo simulated annealings, can be improved simply and effectively by preventing the generation of "hydrophobic-hydrophobic" contacts shorter than 3.8 Å, by monitoring the "hydrophobic-hydrophobic/polar-polar" contact ratio in short-/medium distance ranges and by using preferentially body-center-cubic lattice.

Introduction

Predicting a fold mirroring the native one is an exceptionally demanding task as protein folding is a complicated and multi-step process. Although the protein folding problem field is moving towards all-atom off-lattice work, it is still far from having a definitive solution for proteins longer than 100 amino acid residues [1-4]. The difficulty in making accurate calculations, which are sufficiently sophisticated and computationally manageable, has been the limiting factor in these approaches [4, 5]. Thus, lattices and simplified protein models appear to be still necessary to investigate mechanisms and forces that drive protein folding [6, 7]. In this study we aim to reveal the geometric constraints that drive protein folding and highlight those constraints kept or missed by distinct lattices. Furthermore, we show that protein folding is driven by specific intra- and inter-secondary structure element interactions that fall in specific distance ranges. The folding of a protein consists of three main stages: a) formation of ordered backbone structures (secondary structure elements) by short-range interactions between amino acids, b) formation of small contact regions by medium-range interactions, and c) association of the small contact regions to form the native structure by long-range interactions [8]. It is not clear to what extent, the formed secondary structure modules interact with each other during the folding process and which is the mechanism leading to the final stage of protein folding. All the classes of the inter-residue interactions (short-, medium- and long-range) have been shown to be important for the protein folding process [8-10]. It is therefore necessary to understand the exact role of each class of those interactions to develop algorithms (i.e. those employed in Monte Carlo simulated annealings [11-14]) for better predictions.

A powerful method to study this problem is based on using a simplified theoretical hydrophobic-polar (HP) protein model [15, 16] to be fitted on a 3D lattice. In this model a "protein macromolecule" is represented by a no reverse self-avoiding walk (NR-SAW, a specific SAW that cannot go back on itself, [5, 17, 18]) consisting of n units connected by $n-1$ bonds of the same length as the studied lattice's constant. These units consist of the alpha-carbons (Cαs) coordinates of each residue recovered from the crystal structure. The hydrophobic (H) or hydrophilic (P) character of each

residue can be conferred on the basis of the Kyte-Doolittle hydrophobicity scales [19] and the units can be placed on the vertices of lattices of different geometry [5, 16, 20-29]. The resulting protein model can adopt a variety of conformations depending on the lattice geometry and based on bond distances, valence and torsion angles between lattice vertices on which protein units are placed. We had previously extracted Cαs coordinates ("true" models) of 42 crystallized structures (14 mainly alpha-helical, 14 mainly beta-sheets, and 14 mixed conformations) and generated lattice models from the same set of protein structures [5] by placing the Cαs of each "true" model on the vertices of 8 distinct lattices, i.e. cubic [23], quadrilateral [5], cubic × 2unit [5], tetrahedral [30], double-tetrahedral [31], face center cubic (fcc, [32, 33]), body center cubic (bcc, [34]) and 210-lattice [35] . Herein, we have relied on those protein models to estimate the number of residues in reciprocal contacts at different ranges of distance. Only hydrophobic-hydrophobic (HH) contacts [16] were counted in light of the assumption that the hydrophobic force drives protein folding [10, 36-42]. An HH contact consists of each pair of hydrophobic Cαs that are not adjacent within the primary structure (or the amino acid sequence) and thus not connected by the protein backbone (Fig. 1) [16]. First, we have established which ranges of distances are more represented in "true" models and are consequently more important for protein folding. Second, we have estimated the fidelity of each lattice in reproducing the correct number of short-, medium-, and long-range interactions. We also observed that in *ab initio* protein folding simulations very short-range interactions should be discouraged in favor of longer (short- or medium-range) interactions. In summary, the present study furnishes novel and general rules for improving the prediction of protein folding by using simplified models and lattices.

Materials and methods

Protein and lattice data

The 42 crystallized proteins analyzed in the present study are extensively described in our previous study [5] and consist of 14 mainly alpha helical protein structures, 14 mainly beta sheet protein structures and 14 mixed protein structures. Our previous algorithm recovered all the protein Cαs

coordinates from the investigated crystal structures and generated the so-called "true" models consisting of the set of all the connected Cαs for each protein structure. ~~This~~ Each generated "true" model resulted in the best obtainable off-lattice theoretical model. The Cαs coordinates of each "true" model were then used to generate HP lattice models as previously described [5], i.e. by choosing the coordinates that are closest to the native set of Cαs coordinates in order to obtain the best model that is theoretically possible for each lattice. The eight lattices are formed by some of the Bravais Lattices [43] or combination of them: cubic, tetrahedral, double-tetrahedral, quadrilateral, cubic × 2unit, bcc, fcc and 210-lattices.

Calculation of the maximum number of contacts

All residues of the analyzed proteins were considered as hydrophobic or polar (see Table S1), based on the Kyte-Doolittle hydrophobicity scale, which is derived from the physico-chemical properties of the amino acid side chains [19]. We chose to use the Kyte-Doolittle hydrophobicity scale because our predictions are based purely on protein primary structures without information from 3D homologous templates [44]. Notably, after transforming a protein sequence into a HP binary sequence, the following formula was developed to calculate the maximum number of HH contacts (n_{HH}) and polar-polar (PP) contacts (n_{PP}) between all pairs of hydrophobic and polar residues, respectively, not connected by the protein backbone:

eq. (1) $n_{HH} = (n_{Htot}/2) \times (n_{Htot} - 1) - n_{Hb.res}$

eq. (2) $n_{PP} = (n_{Ptot}/2) \times (n_{Ptot} - 1) - n_{Pb.res}$

n_{Htot} and n_{Ptot} indicate the total number of hydrophobic and polar residues, respectively, in the HP binary sequence; $n_{Hb.res}$ and $n_{Pb.res}$ indicate the total number of peptide bonds between consecutive hydrophobic residue pairs and polar residue pairs, respectively, along the HP binary sequence. For example, in the binary sequence $H_1H_2H_3P_4P_5H_6P_7$ $n_{Htot} = 4$ (H_1, H_2, H_3, H_6) and $n_{Hb.res} = 2$ (H_1-H_2, H_2-H_3) resulting in $n_{HH} = 4$ (H_1-H_3, H_1-H_6, H_2-H_6, H_3-H_6). Analogously, $n_{Ptot} = 3$ (P_4, P_5, P_7) and $n_{Pb.res} = 1$ (P_4-P_5) resulting in $n_{PP} = 2$ (P_4-P_7, P_5-P_7). The non redundant 2D obtainable NR-SAWs

corresponding to the above cited binary sequence are reported in Figure 1.

Calculation of distances between Cαs

The inter-Cα distances of close hydrophobic residues in the crystallized structures were classified according with their location within each helix (intra-helical interactions), within two close helices (inter-helical interactions), within a sheet-pair (intra-sheet interactions), between two close sheet-pairs (inter-sheets interactions) and between two close mixed secondary structure elements (for example a helix and a sheet). Cα-Cα distance (3.8 Å) was chosen as our protein folding metric because Cα represents with good approximation the centre of mass of each natural alpha amino acid. Cα atoms represent a fold invariant feature of each residue in protein folding [5, 45, 46]. Notably, bcc, fcc and 210-lattices show edges of the same length (long 3.8 Å for our simulation) but also edges of different lengths (multiple or submultiple of 3.8 Å for our simulations, see the "Appendix" of Pierri et al. [5] for a list of the direction vectors of the investigated lattices), whereas all the other cited lattices have a single length constant (3.8 Å for our simulations, see the "Appendix" of Pierri et al. [5]). In each HP model, inter-Cα interactions were evaluated as the distances between any possible pair of hydrophobic residues forming HH contacts (see Supp. Info.).

Results

Definition of interactions and interaction ranges within and between secondary structure elements

The upper-bound of short range interactions (i.e. those participating in the formation of a secondary structure element [8]) was set to 9.5 Å, as previously reported [47]. Interactions occurring at distances greater than 9.5 Å can be considered long-range interactions [47]. In order to discriminate the different classes of interactions falling in the 0-9.5 Å range, we measured all the inter-Cαs distances in a turn of helix or between two heads of a sheet pair and in a pair of helices or in a couple of sheet pairs (Figure 2). In a turn of helix or between two heads of a sheet pair (intra- secondary-structure

interactions), two consecutive Cαs in the backbone are 3.8 Å distant, as expected, whereas the other Cαs are 3.8 – 6.5 Å distant from each other (Figure 2). Interactions between Cαs being equal or below 3.8 Å in length were therefore defined as very-short range interactions, while those between 3.8 and 6.5 Å were defined as short-range interactions. The remaining interactions, in-between 6.5 and 9.5 Å, were named medium-range interactions (inter-secondary-structure interactions) and participate in the packing of a pair of secondary structure elements (for example a pair of helices, a pair of two stranded parallel beta-sheet fragments, a pair of mixed secondary structure elements) (Figure 2).

Examples of structural location of very short- and short-range hydrophobic interactions

The main forces driving the folding of a protein are due to the formation of hydrophobic, van der Waals, H-bond and electrostatic interactions. By the way, at coarse-grained level we can assume that residues of a generic protein can be classified as hydrophobic and polar and the hydrophobic force represents the main force that drives protein folding in a such simplified model. In order to investigate the role played by hydrophobic forces in protein folding, we quantified the number of HH contacts within and between secondary structure elements of the analyzed crystal structures. In particular we highlighted HH contacts formed in the distance ranges 0-3.8 Å and 3.8-6.5 Å. HH contacts were detectable in the distance range between 5.1 and 6.4 Å i.e., within the ribbon representation of the helix extending from residue L127 to residue N115 of the ADP/ATP carrier (PDB_ID: 1okc, Figure 3, panel a), or from A53 to N68 of the hemoglobin (PDB_ID: 1a00, Figure 3, panel b). Similarly, in the beta sheet pairs extending from residue N19-I39 of the cytotoxin CTI (PDB_ID: 1zad, Figure 3, panel c), and from V57-V83 of the outer membrane protein OmpA (PDB_ID: 2jmm, Figure 3, panel d), HH contacts were detected in the distance range between 5.0 and 6.2 Å (with one HH contact observed at 4.1 Å in cytotoxin CTI (Figure 3, panel c)).

On the basis of those measurements, we found that both the analyzed alpha-helical and beta-sheet secondary structure elements do not host "interacting" hydrophobic residues (or HH contacts)

within the 0-3.8 Å distance range (very short-range interactions), whereas several short-range interactions are detectable between 5 and 6.5 Å (Figure 3). Thus we consider interactions falling in the 3.8-6.5 Å distance range as the main responsible interactions of the formation of the secondary structure elements (see Figure S1).

Examples of structural location of medium-range and long-range interactions

By searching for HH contacts between secondary structure elements of the analyzed crystal structures, in the distance range of 6.5-9.5 Å within the helix pair extending from residue T83 to residue L127 of the ADP/ATP carrier (Figure 4, panel a) or from P95 to T137 of the hemoglobin (Figure 4, panel b) five HH contacts were detected within the 5.9 - 9.4 Å distance range for each protein fragment. Similarly, for the sheet pairs extending along the T18-S50 and the T63-A96 portions of the cytotoxin CTI (Figure 4, panel c) and along the K258-T272 and the T232-N248 fragments of the outer surface protein OspA (PDB_ID: 2af5, Figure 4, panel d) we detected respectively seven and five (respectively) HH contacts within the 5.8-9.3 Å distance range. Similarly, for the mixed conformation structures extending along the P12-K38 of the agitoxin 2 (PDB_ID: 1agt, Figure 4, panel e) and along the V151-Y186 fragments of the Bcr-abl tyrosine kinase regulatory domain (PDB_ID: 2abl, Figure 4, panel f) at least eight and six HH contacts were detected between 4.3 and 9.4 Å in each protein fragment, respectively. Thus we consider interactions formed in the 6.5-9.5 Å range as the main responsible interactions of the formation of the small contact protein regions (i.e. groups of interacting secondary structure elements).

Maximum number and distribution of contacts.

The maximum number of HH contacts depends neither on the interatomic distance distribution along the entire 3D structure of a protein nor on the lattice geometry, but only on the number of hydrophobic residues not connected by the backbone. A formula was developed to calculate the maximum number of HH or PP contacts in any protein that is preliminarly converted into a HP

binary sequence (see Material and Methods). Such approach can be useful to predict the propensity of a protein to form HH contacts independently from any other parameter than the primary sequence. For example, the all helix p73 alpha domain (PDB_ID: 1dxs) is 57-residue long, out of which 30 residues are hydrophobic and 27 residues are polar. $n_{Hb.res}$ and $n_{Pb.res}$ are 11 and 14 respecitvely and, according to our equations (eq. (1) and eq. (2)), "1dxs" hosts 340 HH contacts and 421 PP contacts (HH/PP ratio = 0.81). The all sheet epidermal growth factor (PDB_ID: 1egf) is 53-residue long, out of which 15 residues are hydrophobic and 38 residues are polar. $n_{Hb.res}$ and $n_{Pb.res}$ are 9 and 14 respectively, thus, according to eq. (1) and eq. (2), "1egf" hosts 201 HH contacts and 477 PP contacts (HH/PP ratio = 0.42). By comparing the two HH/PP ratios, the first protein is expected to have a 2-fold higher propensity to form HH contacts than the second protein. In our dataset, the ratio between the maximum number of HH contacts (61318 for helix structures, 9251 for beta sheet structures) and PP contacts (94210 for helix structures, 26848 for beta sheet structures) is significantly higher in the analyzed alpha-helix (mean ratio = 0.83) in comparison with beta-sheet conformations (mean ratio = 0.56, two-tailed Wilcoxon test p-value = 0.029). Setting a threshold for the HH/PP ratio at 0.60, 72% of the analyzed mainly helical structures show a ratio above the threshold and 72% of the analyzed mainly sheet structures show a ratio below the threshold (see Table S2). The results of eq. (1) and eq. (2) were independently validated by detecting and enumerating all the contacts using our algorithm for the calculation of distances within small 2D/3D HP protein models (see Figure 1 for a 2D illustrative validation scheme).

The distributions of HH contact distances were generated to verify which range of distances are more frequent in each model. In the case of "true" models, all-helix and mixed conformations behave similarly. Two frequent distances were detected in the short-/medium- range (3.8-9.5 Å), i.e., (1.4-1.8) × 3.8 Å for all-helix conformations (or (1.6-1.8) × 3.8 Å for mixed conformations) and (2.2-2.6) × 3.8 Å for all-helix conformations (or (2.2-2.4) × 3.8 Å for mixed conformations (see Figure S4 and Figure S5). The "true" model of beta-sheet structures differs from the the other two types of conformations not only for the lower HH/PP contact ratio, as previously observed, but also

for a flatter and wider distribution of the HH contacts (see panel 1, Figure S6). Nevertheless, two frequent range of distances are still detectable also in beta-sheet conformations, i.e. (1.4-1.6) × 3.8 and (2.4-2.6) × 3.8 (see panel 1 of Figure S6). Considering the full distributions of distances, each lattice was significantly different from the "true" model (pval < 0.001, kolmogorov-smirnov test, see Supp. Info.).

Specific lattices dislocation

In order to understand how our lattice-models deviate from the native protein-folding pathway (i.e. from our "true" models), we classified the HH contacts in six ranges of distances (r1-r6) for each analyzed lattice. Those ranges were chosen in order to monitor very short- (r1: $0 < x \leq 3.8$ Å, where x is the measured inter-Cαs distance), short- (r2: $3.8 < x \leq 5.37$ Å; r3: $5.37 < x \leq 6.5$ Å), and medium-range (r4: $6.5 \leq x \leq 7.6$ Å; r5: $7.6 < x \leq 8.5$ Å; r6: $8.5 < x \leq 9.5$ Å) interactions.

Two connected residues can occupy two vertices of a side or a diagonal of the described lattices. Given that edges of lattices are 3.8 Å long, face diagonals are $2^{1/2} \times 3.8$ Å and depth diagonals are $3^{1/2} \times 3.8$ Å (Figure S7), the three corresponding distances are in ranges r1, r2 and r3, respectively. Notably, 210, bcc and fcc lattices have also edges of different length (see the "Appendix" of Pierri et al. [5]). Vertices on those edges could ideally be occupied by connected Cα, if they would exist in native proteins. It is worth noting that it is not possible to find 2 backbone connected residues at a distance lower than 3.8 Å. By the way, during the projection of HP "true" model-residues on lattices, some HP residues (although not connected by the backbone) could locate at a distance lower than 3.8 Å. In order to evaluate how medium-range interactions ($6.5 < x < 9.5$ Å) influence the dislocation of residues on the analyzed lattices we also classified HH contacts that are distant 2×3.8 Å, $5^{1/2} \times 3.8$ Å, $6^{1/2} \times 3.8$ Å(Figure S7) within lattices that correspond roughly to distance-ranges r4, r5 and r6.

Number of contacts in each distance range and model

We observed that the trend of protein Cαs to occupy specific positions in the space and within the specific distance ranges appears to be very similar for the analyzed all-helix structures and mixed conformation structures. In particular we observed that in those kind of structures, Cαs not connected by the protein backbone do not occupy positions at a distance lower than 3.8 Å (Figure 5 panels "a" and "b"). Furthermore, by screening "true" models we observed that Cαs have a high preference in occupying vertices in the range of distances r3 ($5.37 < x \leq 6.5$ Å) and r6 ($8.5 < x \leq 9.5$ Å) and much less frequently within the distance ranges r2, r4 and r5 (Figure 5 panels "a" and "b" and Figure 6). Unlike "true" models, lattice models tend to host a high number of Cαs in the ranges r1, r2 and r5, in addition to r6. Furthermore, our models appear to allocate a low number of Cαs in the ranges r3 and r4. Only the bcc and the fcc lattices appear to allocate a considerable number of Cαs, as allocated by "true" models, in the ranges r2 (only bcc), r3 (only bcc), r4 (only fcc), r5 (only fcc) and r6 (both lattices).

Differently from what was reported for mainly helix and mixed conformation structures, mainly beta-sheet Cαs show a high preference for occupying vertices mainly in the range of distances r3 and only at a lower extent in the range of distances r2, r4, r5 and r6 (Figure 5, panel "c"). Notably, in the range r5 the 210-lattice structures host a number of HH contacts comparable to that contained in the "true" lattice structures providing a better result than the ones obtained with fcc lattice (Figure 5, panel "c").

Reproducibility of the "true" model distances in each lattice

A similar number of contacts per distance range between "true" models and lattices does not necessarily imply that the contacts are the same, i.e. that they involve the same interacting residues, between models. Tests of correlation were therefore performed between the distances of the "true" model contacts and the distances of the same contacts in each lattice model. Despite the high overall correlation between the "true" model and each lattice for long-range distances, the correlation drops

down in r2-r6 distance ranges (Figure 6). No correlation analysis can be performed for range r1, for which almost no contacts were detected in the "true" models. Furthermore, correlations are much different among the three classes of protein conformations investigated. In mainly helix structures r3 is the best approximated distance range by most lattices, followed by r4 and r6, and bcc is the only lattice that shows a statistically significant correlation with the "true" model in four distance ranges, i.e. r3-r6 (Figure 6, panel "a"). In mixed conformation structures the best performing lattice is the double-tetrahedral, for which the correlation is statistically significant in three distance ranges, i.e. r2, r3 and r6, while bcc well preforms in r4 and r6 (see "d_tet" in Figure 6, panel "b"). Neverthless, mainly beta-sheets are the worst approximated structures, being the correlation statistically significant for double-tetrahedral in two distance ranges, i.e. r3 and r4, and for bcc only in r6 (Figure 6, panel "c").

Discussion

The study here presented aims to understand which constraints are kept or missed by distinct lattices and to reveal the geometric constraints that would allow a more correct *ab initio* protein folding prediction. Lattice HP models were built using eight classic lattices that we already screened for their overall precision in approximating a crystallized structure [5]. In this study we determined which are the most frequent intra- and inter-secondary structure element interactions as observed from "true" models built from a sample of 42 crystal structures. We also established which HP lattice models can better reproduce them. We observed that both alpha-helical and beta-sheet native secondary structure elements do not contain HH contacts within the 0-3.8 Å range. We can speculate that HH interactions in the 0-3.8 Å range (the typical distance among two Cα belonging to connected residues) are disfavored in order to free the space for the torsion and the bending of the backbone-connected Cα residues. Notably, conformational changes and the formation of a secondary structure element can be triggered by the constitution of specific/allowed ψ and ϕ angles that can be formed by atoms of close residues as described by Ramachandran et al.

[48]. We retain that very short-range interactions can appear during conformational changes, whereas the short-range interactions strongly participate in (and/or initiate) the folding of a secondary structure element (or part of it, for example a turn). The short-range interactions are also frequently observed among residues of distinctive structural elements (i.e. two close alpha-helices or two close beta-sheets) and we retain that they contribute to the stabilization of these contact regions. By the way, medium-range interactions appear to play a more important role in the formation of small contact regions between secondary structure elements, given their high number within the 9.5 Å distance range. Furthermore, we propose that the residues in the range r2-r6 can drive the protein folding process, particularly due to the strength of interactions for distance ranges r2-r3 and to the number of interactions (although weaker) for distance ranges r4-r6. Finally, concerning long-range interactions, which do not represent the main focus of this study, we can observe that HH contacts increases exponentially after 9.5 Å (beyond the distance range r6) till a maximum detectable around 40 Å (Figure S4-S6)**.** It is clear that the energy contribution of two hydrophobic residues 40 Å far from each other is very close to zero. The problem is that the number of this very low energy interactions is very high and we cannot quantify the contribution of those very low energy interactions to protein folding.

In the present analysis, we argued that maximizing the number of very-short range interactions on the different lattice geometries [5, 16, 24, 49] returns incorrect interactions. Very short-range interactions should be instead discouraged in favor of longer (short- or medium-range) interactions. For medium-range ($5.4 \leq x \leq 7.6$) interactions the bcc lattice ($5.4 \leq x \leq 6.5$) and fcc lattice ($6.5 \leq x \leq 7.6$) appear to be able to better approximate the number of HH contacts detected within real proteins. Notably, the distance ranges r3 (short-range interactions, 4.1-6.5 Å) and r6 (medium-range interactions, 6.5-9.5 Å) appear to be more populated than other distance ranges (see Figure 5 and the number of HH contacts calculated within "true" models and reported within brackets below the X axis in Figure 6). Thus we retain that the formation of HH contacts in r3 and r6 distance ranges should be encouraged in the algorithms (i.e. those employed in Monte Carlo simulated annealings)

that aim to reproduce a native folding pathway. The number of short- and medium- range interactions appears to be better maintained when a protein is fitted on the bcc lattice or on the fcc lattice, whereas all the other analyzed lattices fail in a more severe way. In comparison to the contact number analysis, the analysis about the reproducibility of the "true" model distances in each lattice strengths a good performance for bcc, but not for fcc, in reproducing the native protein folding, at least for all-helix structures. On the other side, the detection of double-tetrahedral as a good performing lattice only in the distance reproducibility analysis may suggest that this lattice mainly alters the real contact distances by moving them from a certain distance range to a close one. It is worth noting that the damage to the secondary structure formation by an incorrect dislocation of residues on a lattice can be quite severe (Figure S2 and Figure S3 and Table S3).

In general, the number of HH contacts and the reproducibility of the "true" model distances for each distance range should be taken in great consideration when we think about an algorithm for protein folding simulations to be trained/validated on crystal structures. Even if the total number of HH contacts depends only on the primary structure (i.e. residues order in the backbone and number of hydrophobic and polar residues), the number of HH contacts within each distance range can vary in the folding simulation and in the formation of small protein contact regions. In particular short- and medium-range interactions appear to play a key role in the protein folding process and the number of HH contacts in the analyzed r1-r6 ranges needs to be constantly controlled within a sphere of approximately 9.5 Å (see Figure S1) during protein folding simulations. In order to gain new insights about how the primary structure contains the information for the entire protein folding [1, 9, 42, 50-52], new algorithms should be tested by folding proteins of known crystal structures, both on lattices and off-lattice [49, 53-55], taking into account the formation of HH contacts at each distance range in comparison with what observed in the "true" models.

After developing a formula to calculate the maximum number of HH contacts and PP contacts, we found that the ratio between these two values varies between mainly helix conformations and mainly sheet conformations and can be therefore useful for the prediction of the protein/domain

conformation from the primary sequence. Furthermore, each protein is defined by a specific HH/PP contact ratio and we assume that this value should be preserved in each range of distances depending on the number and type of residues potentially involved in the formation of a secondary structure element (short-range interactions) and/or a small contact region (medium-range interactions), by Monte Carlo algorithms for protein folding simulation [11-14]. In this way it will be possible to understand how a protein can spontaneously fold in the same native/functional structure also after induced denaturation/renaturation events [42].

For example, it would be possible to understand how the structure/shape of the same protein domain involved in two different protein complexes can change due to the different close interactors (chaperons and other protein subunit). Notably, it is known that some proteins sharing more than 90% of identical residues can undergo different folding pathways [56, 57]. We retain that it would be possible to define an HH contact number and an HH/PP contact ratio also for a protein subunit within different protein complexes resulting in a different folding pathway and final conformation. It is worth noting that HH/PP ratio cannot be the sole predictor of the right protein folding/activity, given that "conservative" small mutations can affect severely protein function without altering the HH/PP ratio. Neverthless, "not conservative" missense mutations and nonsense mutations are known to be frequently more deleterious for protein function and this observation can be related to the alteration in local structure depending on a different newly established HH/PP ratio. The ability of bcc and double-tetrahedral lattices in reproducing the "true" model distances and the best HH/PP ratio (within short-/medium-ranges of distances) observed in native proteins would allow obtaining the most accurate protein models already at a coarse grained level (i.e. HP models), as a result of an algorithm based on *in vivo* driving forces and physical constraints [11-14]. This model could be converted through a rescaling procedure [5, 12], into an accurate all aton model that could be more easily relaxed through MD simulations in order to handle last structural problems coming from lattice anisotropies.

REFERENCES

1. A. Ben-Naim, *Open J Biophys*, 2012, **2**, 23-32.
2. A. V. Finkelstein, N. S. Bogatyreva and S. O. Garbuzynskiy, *FEBS Lett*, 2013, **587**, 1884-1890.
3. A. V. Finkelstein and S. O. Garbuzynskiy, *J Biomol Struct Dyn*, 2013, **31**, 1013-1015.
4. S. O. Garbuzynskiy, D. N. Ivankov, N. S. Bogatyreva and A. V. Finkelstein, *Proc Natl Acad Sci U S A*, 2013, **110**, 147-150.
5. C. L. Pierri, A. De Grassi and A. Turi, *Proteins*, 2008, **73**, 351-361.
6. J. J. Tsay and S. C. Su, *Proteome Sci*, 2013, **11**, S1-S14.
7. D. Shaw, A. Shohidull Islam, M. Sohel Rahman and M. Hasan, *BMC Bioinformatics*, 2014, **15**, S2-S7.
8. S. Tanaka and H. A. Scheraga, *Proc Natl Acad Sci U S A*, 1975, **72**, 3802-3806.
9. C. Levinthal, *J Chim. phys*, 1968, **65**, 44-45.
10. C. B. Anfinsen, *Biochem J*, 1972, **128**, 737-749.
11. A. Kolinski and J. Skolnick, *Proteins*, 1994, **18**, 338-352.
12. V. Villani, C.L. Pierri and A. Cascone, *Recent Research Development Macromolecules*, 2005, **8**, 47-72.
13. S. Cheon and F. Liang, *Biosystems*, 2011, **105**, 243-249.
14. U. H. Hansmann and Y. Okamoto, *Curr Opin Struct Biol*, 1999, **9**, 177-183.
15. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan, *Protein Sci*, 1995, **4**, 561-602.
16. B. Hayes, *Am Sci*, 1998, **86**, 314-319.
17. B. Hayes, *Am Sci*, 1998, **86**, 216-221.
18. G. Slade, *Math Intell*, 1994, **16**, 29-35.
19. J. Kyte and R. F. Doolittle, *J Mol Biol*, 1982, **157**, 105-132.
20. A. Sali, E. Shakhnovich and M. Karplus, *Nature*, 1994, **369**, 248-251.
21. A. Sali, E. Shakhnovich and M. Karplus, *J Mol Biol*, 1994, **235**, 1614-1636.
22. M. Karplus and A. Sali, *Curr Opin Struct Biol*, 1995, **5**, 58-73.
23. H. S. Chan and K. A. Dill, *J Chem Phys*, 1990, **92**, 3118-3135.
24. A. Kolinski, Lattice Polymers and Protein Models, *Multiscale Approaches to Protein Modeling*, 2011, 1-20.
25. T. Hoppe and J. Yuan, *J Phys Chem B*, 2011, **115**, 2006-2013.
26. A. Banerji and I. Ghosh, *Cell Mol Life Sci*, 2011, **68**, 2711-2737.
27. I. Shrivastava, S. Vishveshwara, M. Cieplak, A. Maritan and J. R. Banavar, *Proc Natl Acad Sci U S A*, 1995, **92**, 9206-9209.
28. A. Godzik, J. Skolnick and A. Kolinski, *Protein Eng*, 1993, **6**, 801-810.
29. A. Godzik, A. Kolinski and J. Skolnick, *J Comp Chem*, 1993, **14**, 1194-1202.
30. Y. Xia, E. S. Huang, M. Levitt and R. Samudrala, *J Mol Biol*, 2000, **300**, 171-185.
31. L. Toma and S. Toma, *Protein science : a publication of the Protein Society*, 1999, **8**, 196-202.
32. M. Peto, T. Sen, R. Jernigan and A. Kloczkowski, *J Chem Phys*, 2007, **127**, 044101.
33. P. Pokarowski, A. Kolinski and J. Skolnick, *Biophys J*, 2003, **84**, 1518-1526.
34. B. H. Park and M. Levitt, *J Mol Biol*, 1995, **249**, 493-507.
35. A. Kolinski, M. Milik and J. Skolnick, *J Chem Phys*, 1991, **94,** 3978-3985.
36. K. A. Dill, *Biochemistry*, 1990, **29**, 7133-7155.
37. C. N. Pace, H. Fu, K. L. Fryar, J. Landua, S. R. Trevino, B. A. Shirley, M. M. Hendricks, S. Iimura, K. Gajiwala, J. M. Scholtz and G. R. Grimsley, *J Mol Biol*, 2011, **408**, 514-528.
38. C. N. Pace, B. A. Shirley, M. McNutt and K. Gajiwala, *FASEB J*, 1996, **10**, 75-83.
39. C. Tanford, *Protein Sci*, 1997, **6**, 1358-1366.
40. L. Pauling, R. B. Corey and H. R. Branson, *Proc Natl Acad Sci U S A*, 1951, **37**, 205-211.

41. D. Chandler, *Nature*, 2005, **437**, 640-647.
42. C. B. Anfinsen, *Science*, 1973, **181**, 223-230.
43. B. Z. Lu, B. H. Wang, W. Z. Chen and C. X. Wang, *Protein Eng*, 2003, **16**, 659-663.
44. J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky and C. DeLisi, *J Mol Biol*, 1987, **195**, 659-685.
45. A. Mittal and B. Jayaram, *J Biomol Struct Dyn*, 2011, **28**, 443-454.
46. H. Hasegawa and L. Holm, *Curr Opin Struct Biol*, 2009, **19**, 341-348.
47. K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana and D. E. Shaw, *J Am Chem Soc*, 2012, **134**, 3787-3791.
48. C. Ramakrishnan and G. N. Ramachandran, *Biophys J*, 1965, **5**, 909-933.
49. K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich and K. A. Dill, *Proc Natl Acad Sci U S A*, 1995, **92**, 325-329.
50. J. R. Banavar, T. X. Hoang, F. Seno, A. Trovato and A. Maritan, *Journal of Statistical Physics*, 2012, **148**, 636-645.
51. M. Karplus, *Nat Chem Biol*, 2011, **7**, 401-404.
52. K. A. Dill and H. S. Chan, *Nat Struct Biol*, 1997, **4**, 10-19.
53. S. Teso, C. Di Risio, A. Passerini and R. Battiti, in *Pattern Recognition in Bioinformatics - Lecture Notes in Computer Science*, 2010, **6282**, 368-379.
54. M. Mann, R. Saunders, C. Smith, R. Backofen and C. M. Deane, *Adv Bioinformatics*, 2014, **2014**, 1-10.
55. M. Garza-Fabre, E. Rodriguez-Tello and G. Toscano-Pulido, *J Comp Sci and Tech*, 2013, **28**, 868-889.
56. A. Simpson, O. Bateman, H. Driessen, P. Lindley, D. Moss, S. Mylvaganam, E. Narebor and C. Slingsby, *Nat Struct Biol*, 1994, **1**, 724-734.
57. G. Wistow, J. Mulders and W. de Jong, *Nature*, **326**, 622-624.

Acknowledgments

The authors are grateful to dr. Aleksandra Włodek for critical reading and revising the manuscript. 3D lattice models will be available upon request**.** The computational work has been executed on the IT resources made available by ReCaS, a project financed by the MIUR (Italian Ministry for Education, University and Re-search) in the "PON Ricerca e Competitività 2007-2013 - Azione I - Interventi di rafforzamento strutturale" PONa3_00052, Avviso 254/Ric. Authors have no conflict of interests to declare. A.D.G., A.T. AND C.L.P. designed the research; A.O., G.P., G.P., S.T., M.A.D.N., F.B., A.T., A.D.G., and C.L.P. performed the analyses; A.O., A.D.G. and C.L.P. analyzed the data; and A.D.G. and C.L.P. wrote the paper.

Figure Legends

Figure 1. Scheme representation of 2D HP models. Putative NR-SAW representing the folding of the sequence 1-HHHPHPP-7 are reported on the 2D square lattice. The protein backbone is indicated by the black line. H and P residues are labeled and indicated by coloured circles. Putative HH contacts are highlighted by dashed coloured lines.

Figure 2. Inter-Cαs distances measured within and between different secondary structure elements in protein fragments from known crystal structures. Panel a) residues 7-11 extracted from "1dxs" are reported; panel b) residues 95-100 and 115-119 from "1okc"; panel c) residues 18-19 and 56-57 from "1btg"; panel d) residues 20-22, 46-48, 65-67, 92-94 from "1pdg". Protein fragments are reported by grey cartoon representation. In panel a) and b) inter-Cαs distances in the 3.8-6.5 Å are reported by means of black dashed lines. In panel c) and d) inter-Cαs distances in the range 6.5-9.5 Å are reported by blue dashed lines. Other pink dashed lines indicate inter- Cαs distances beyond the 6.5-9.5 Å distance range.

Figure 3. Examples of intra-secondary structure HH contacts. Panel a) the crystal structure of the all helix "1okc" is reported in grey cartoon and orange-blue ribbon representations. In Panel b), c) and d) the crystal structures of the all helix "1a00", the all sheet "1zad" and the all sheet "2jmm" are reported, respectively, with the same coloring scheme described for "1okc". Green and black dashed lines highlight HH interactions below 5.4 Å and those in the 5.5 - 6.4 Å range, respectively. The reported "1okc" helix includes residues N115-L127 (115-NLASGGAAGATSL-127); the "1a00" helix includes residues A53-N68 (53-AQVKGHGKKVADALTN-68); the "1zad" sheet pair includes residues N19-I39 (19-NLCYKMFMMSDLTIPVKRGCI-39); the "2jmm" sheet pair includes residues V57-Y65 and D74-V83 (57-VQLTAKLGY-65 / 74-DIYTRLGGMV-83).

Figure 4. Examples of inter-secondary structure HH contacts. Panel a) the crystal structure of

two helices of "1okc" is reported in grey cartoon and orange-blue ribbon representations. Cartoon and orange-blue ribbon representations are also reported for two helices from "1a00" (panel b); two sheet pairs from "1pdg" (panel c) and "2af5" (pandel d) and for mixed secondary structure elements from "1agt" (panel e) and "2abl" (panel f). Green dashed lines highlight HH interactions below 5.4 Å; black dashed lines highlight HH interactions in the 5.5 - 6.5 Å range; orange dashed lines highlight HH interactions in the 6.6 - 7.6 Å; cyan dashed lines highlight HH interactions in the 7.7-8.4 Å range; bordeaux dashed lines highlight HH interactions in the 8.5 - 9.4 range. The "1okc" helix pair reported includes residues N115-L127 (115-NLASGGAAGATSL-127) and T83-G100 (83-TQALNFAFKDKYKQIFLG-100). The "1a00" helix pair includes residues P95-A111 (95-PVNFKLLSHCLLVTLAA-111) and V121-T137 (121-VHASLDKFLASVSTVLT-137). The "1pdg" pair of sheet pairs includes residues T18-E24 (18-TRTEVFE-24); P42-S50 (42-PCVEVQRCS-50); T63-K74 (63-TQVQLRPVQVRK-74); K85-C97 (85-KKATVTLEDHLA-96). The "2af5" pair of sheet pairs includes residues T232-I237 (232-TLSKNI-237), V243-N248 (243-VSVELN-248), K258-W262 (258-KTAAW-262) and T268-T272 (268-TLTIT-272). The "1agt" mixed structure includes residues P12-K38 (12-PQCIKPCKDAGMRFGKCMNRKCHCTPK-38). The "2abl" mixed structure includes residues V151-S151 (151-VSRNAAEYLLS-161); F168-E172 (168-FLVRE-172); R180-Y186 (180-RSISLRY-186).

Figure 5. Number of HH contacts at increasing range of distances in the investigated structures. The number of HH contacts is reported for six distance ranges corresponding to very short- (r1), short- (r2, r3), and medium- (r4-r6) range interactions in the "true" models and in all the analyzed lattice HP models generated from mainly helix structures (panel "a"), mixed conformation structures (panel "b") and mainly sheet structures (panel "c").

Figure 6. Reproducibility of "true" model HH contact distances by lattices. The Pearson correlation coefficients between the HH contact distances in "true" models and the distances of the

same contacts in each lattice model are indicated for very short- (r2, r3), medium- (r4-r6) and long range interactions. Number in brackets indicate the number of HH contacts (present in "true" models) used to calculate the correlation coefficients. Stars indicate the p-value resulting from the test of correlation, i.e. * for p-value < 10^{-3} and ** for p-value < 10^{-6}.

FIGURES

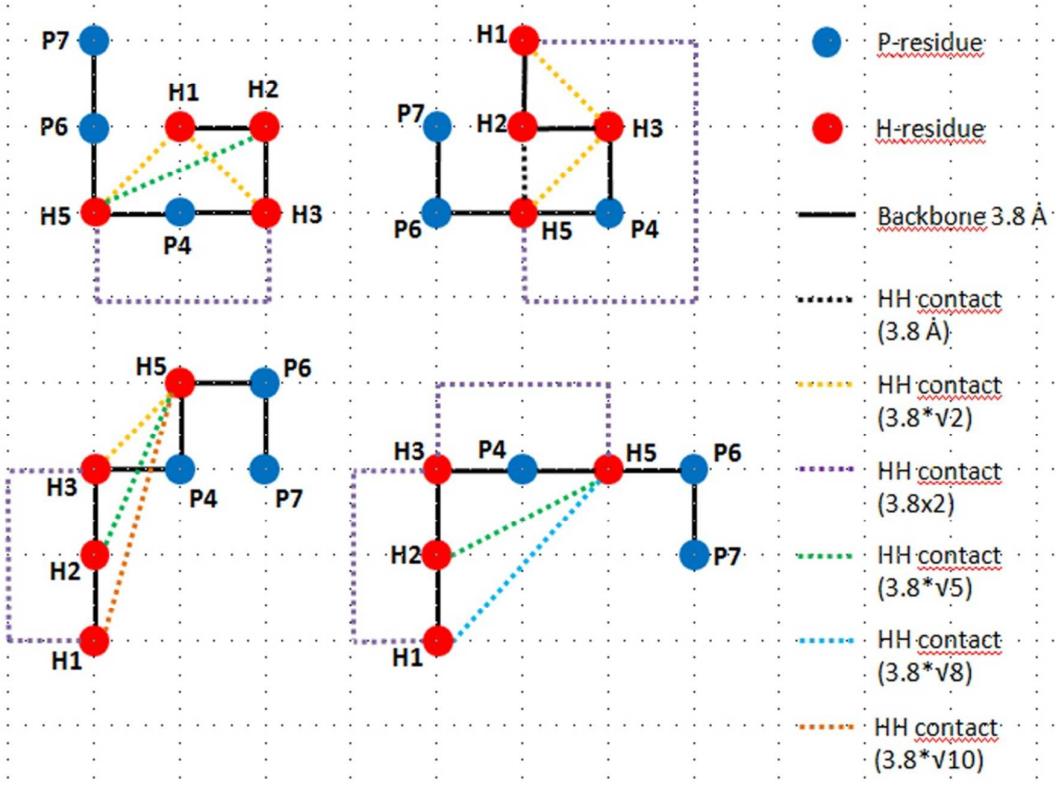

Figure 1

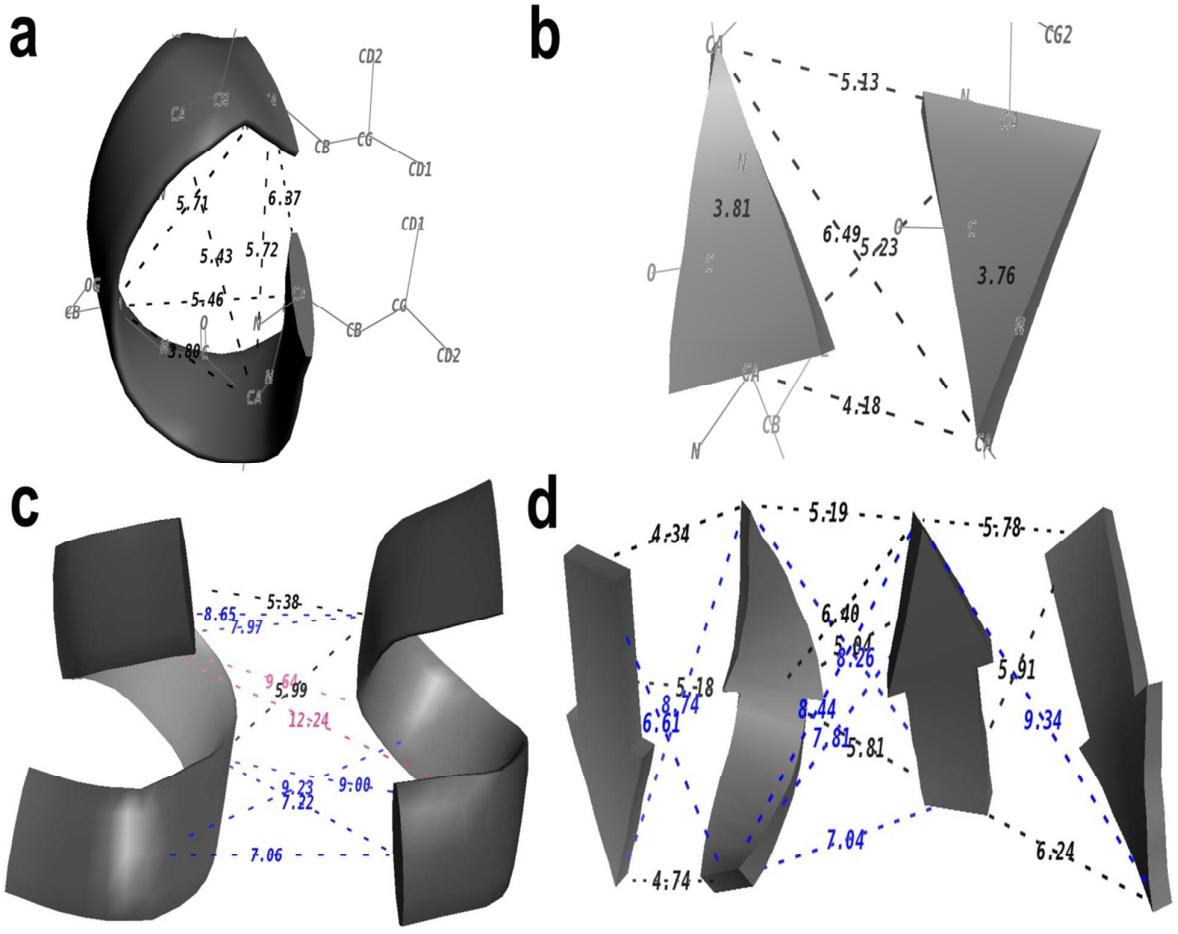

Figure 2

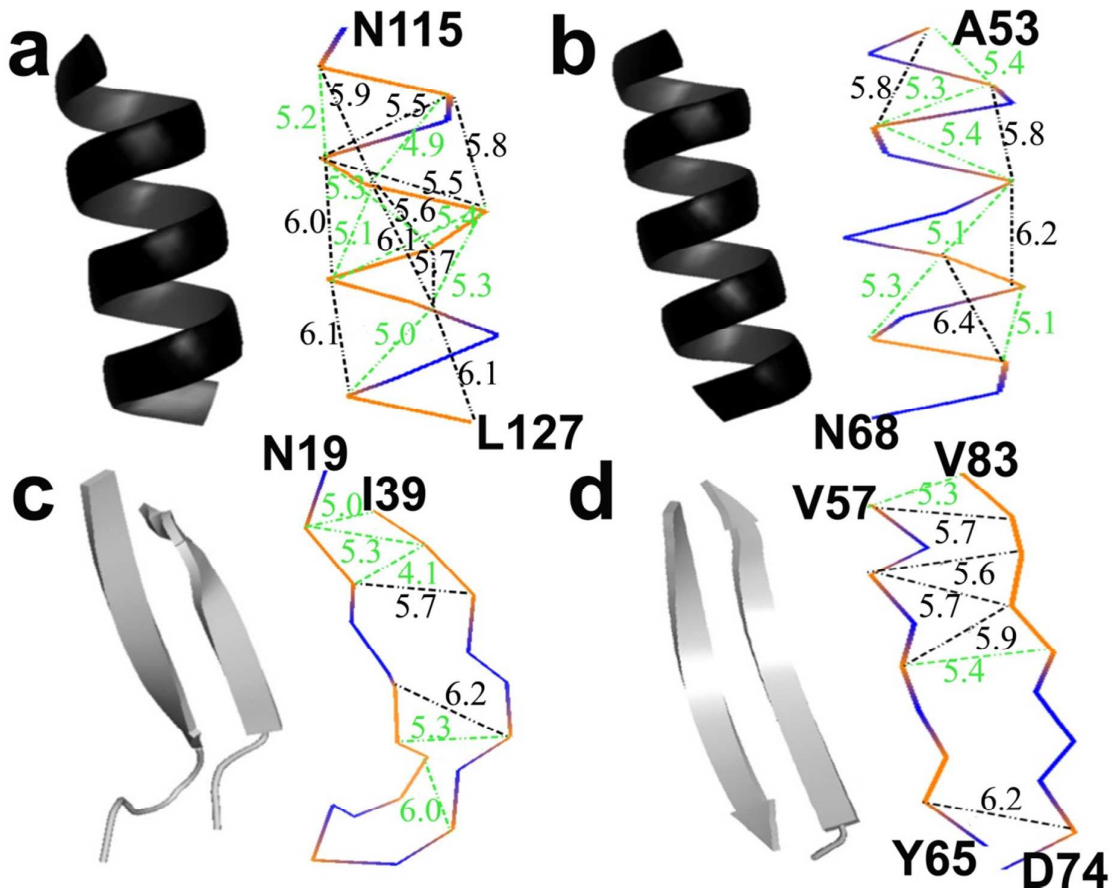

Figure 3

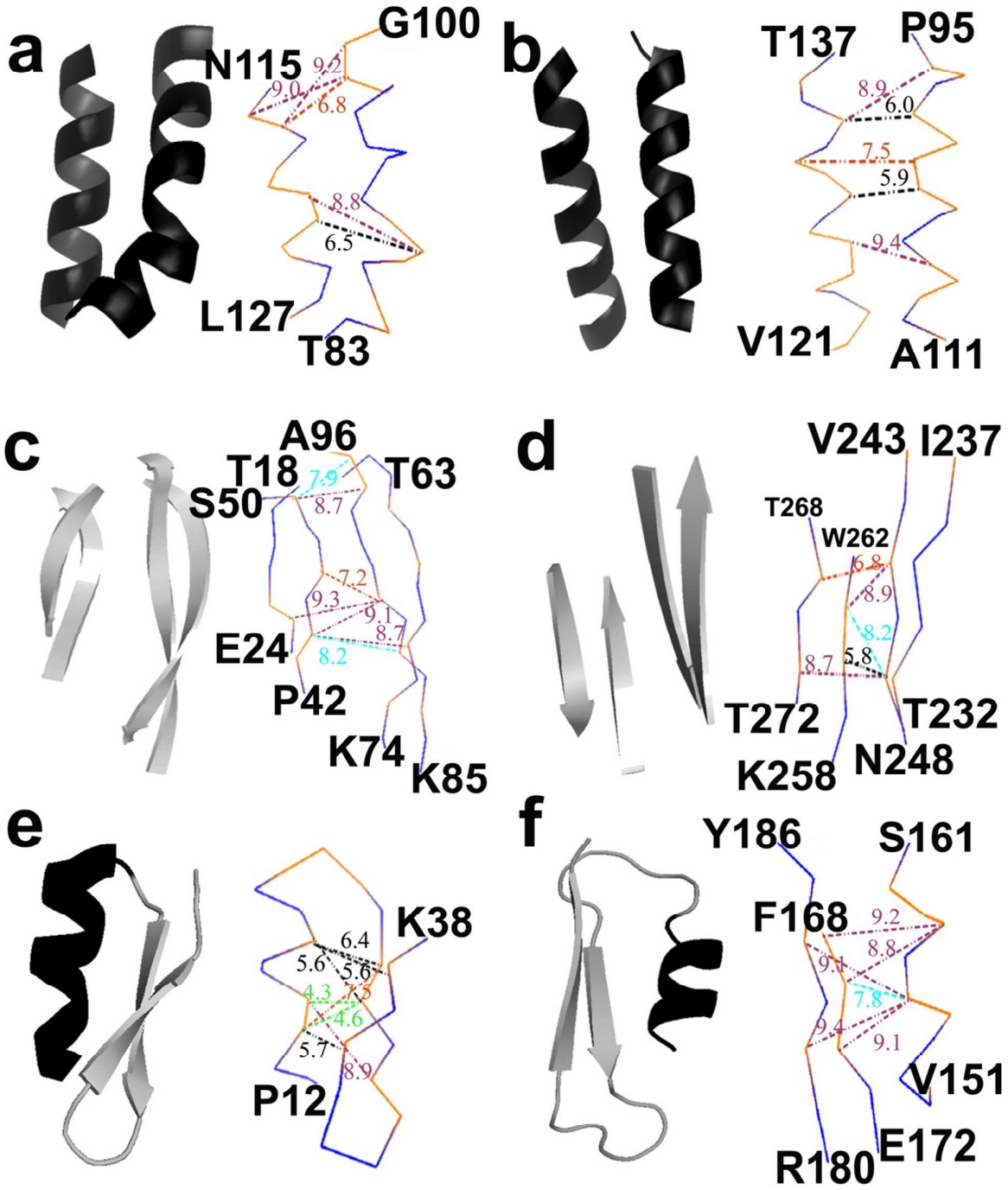

Figure 4

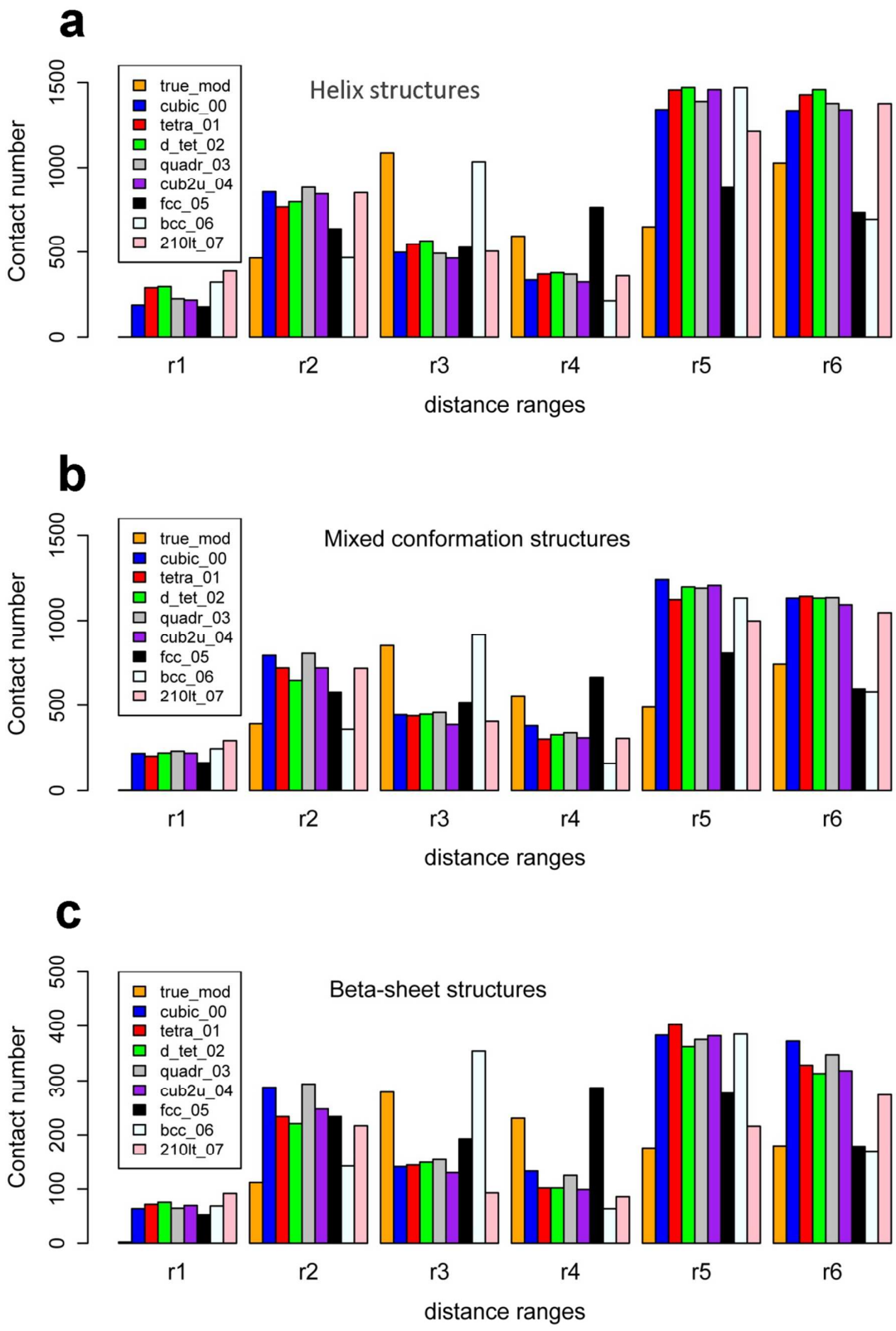

Figure 5

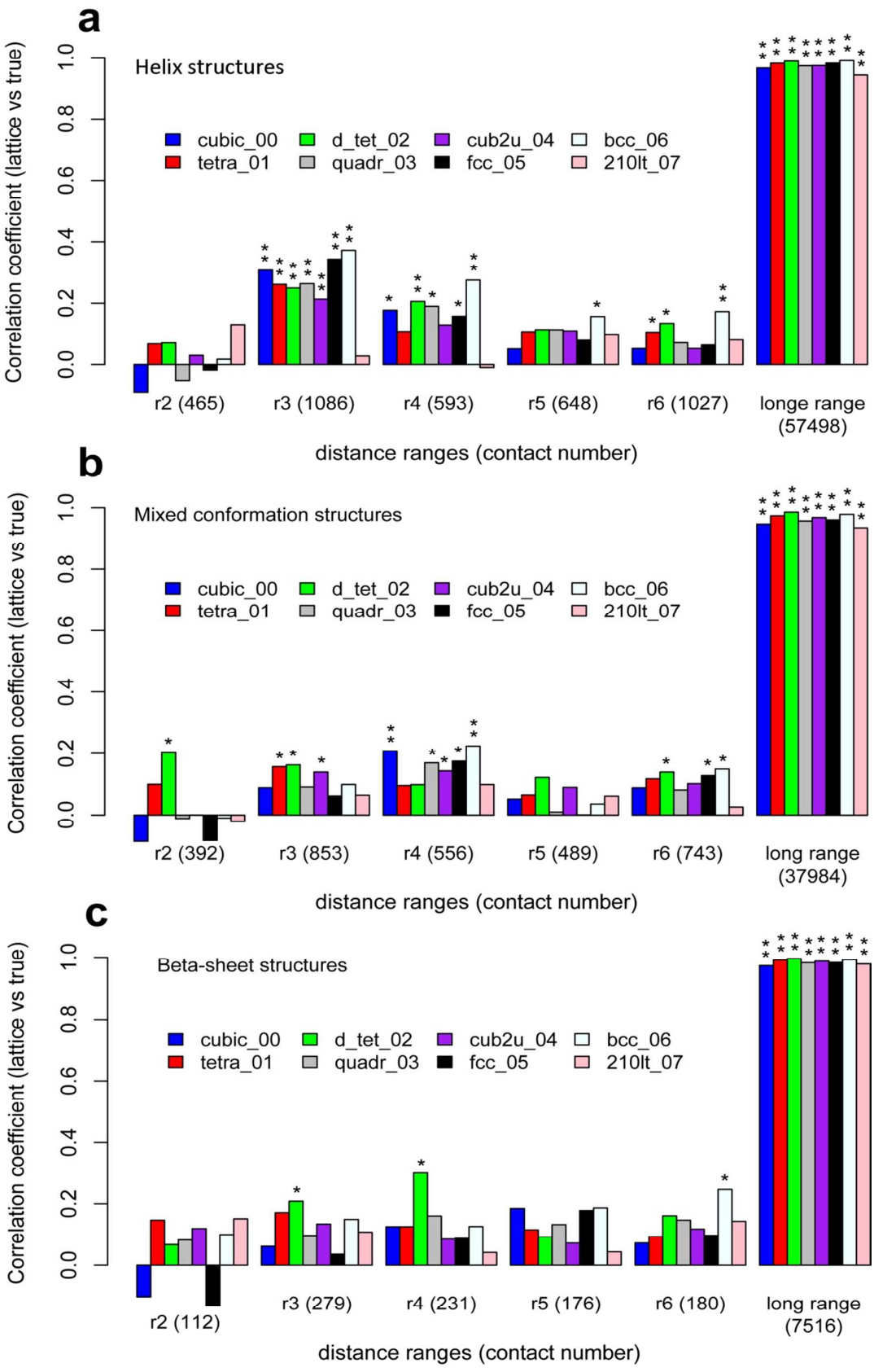

Figure 6