Supporting Information

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

Calculation of distances between Cαs

The number and value of the distance between the above cited Cαs were calculated by using custom perl scripts and were graphically represented by using R (http://www.r-project.org/) scripts. The distance between Cαs was manually measured for some small proteins and large protein fragments for validating our scripts by using the distance wizard implemented in the protein visualizer PyMol (http://www.pymol.org/). Notably, R was also used to perform Kolmogorov-Smirnov tests between the distributions of distances and to calculate and test the Pearson correlation coefficient (cor) between distances in "true" models and distances of the same contacts in each lattice.

Table S1 Code for the conversion of the analyzed structures in HP models. 0 indicates polar (P) residues, 1 indicates hydrophobic (H) residues based on the Kyte-Doolittle hydrophobicity scale.

ASP	0
GLU	0
ALA	1
ARG	0
ASN	0
CYS	1
PHE	1
GLY	1
GLN	0
ILE	1
HIS	0
LEU	1
LYS	0
MET	1
PRO	0
SER	0
TYR	0
THR	0
TRP	0
VAL	1

Table S2 Number of HH and PP contacts and HH/PP ratio for each analyzed protein sequence.

mainly alpha helix structures	**Number of residues**	**Resolution (Å)**	**HH contacts**	**PP contact**	**HH / PP ratio**	**mainly beta sheet structures**	**Number of residues**	**Resolution (Å)**	**HH contacts**	**PP contacts**	**HH / PP contacts**	**Mixed conf. str.**	**Number of residues**	**Resolution (Å)**	**HH contacts**	**PP contacts**	**HH / PP contacts**
9ica	327	3,00	8204	19383	0,423257	1btg	326	2,5	973	1918	0,5072	1ngf	382	2,17	14971	21415	0,699089
1w7b	327	1,52	7952	18223	0,436372	2af5	282	2,5	4539	17095	0,2655	1iep	274	2,10	6733	12153	0,55402
1avr	319	2,30	9668	15303	0,631772	3hb3-chainC	118	2,25	934	2660	0,3511	1i9b	205	2,70	2535	8695	0,291547
1okc	292	2,20	12313	8845	1,392086	1pdg	87	3	726	1104	0,6576	1no5	202	1,80	3451	6835	0,504901
1vin	252	2,20	6057	9792	0,618566	1hae	63	1,14	340	611	0,5564	2bh8	171	1,90	3961	3128	1,266304
1ad6	185	2,30	3449	4995	0,69049	1pi2	61	2,5	223	717	0,3110	2abl	163	2,50	1811	5091	0,355726
1pbw	184	2,00	3126	5298	0,590034	1zad	60	0,2485	393	450	0,8733	1m2e	135	0,78	2242	2175	1,030805
2blh	153	1,77	2388	3362	0,710291	1ju5-sh2	60	1,29	268	611	0,4386	1o8v	132	1,60	1515	2804	0,5403
1elk	153	1,50	2314	3439	0,67287	1egf	53	2,5	201	477	0,4213	1ju5-sh3	109	1,21	726	2370	0,306329
1a00	136	2,00	2321	2528	0,918117	1k36	46	1,51	246	245	1,0040	1o13	105	1,83	1457	1202	1,212146
1d9s	67	1,244	2665	1867	1,427424	1a3p	43	1,14	146	287	0,5087	1h7m	97	1,96	1153	1060	1,087736
1rg6	64	1,39	395	683	0,578331	1edm	39	1,5	132	221	0,5972	1koy	62	1,49	265	679	0,39028
1dxs	57	2,54	340	421	0,807601	2gwp	32	1,54	129	100	1,29	1dgz	37	1,07	101	219	0,461187
7ins	30	2,00	126	71	1,774648	1zr7	30	0.90	1	352	0,0028	1c6w	33	1,4	63	198	0,318182
total			61318	94210					9251	26848					40984	68024	
Tot HH/PP ratio					0,650865						0,35042						0,602493

Table S3a. RMSD calculated between Cαs from "true" and "lattice" models. Helices PDB accession numbers are reported for each analyzed crystal structure (from Pierri et al., Proteins 2008). For each protein HH contacts, PP contacts and HH/PP ratio are again reported.

mainly helices-ID	cubic_00	tetra_01	d_tet_02	quadr_03	cub2u_04	fcc_05	bcc_06	HH contacts	PP contacts	HH / PP ratio
9ica	4,28	2,74	2,18	4,01	3,32	3,26	2,36	8204	19383	0,423257
1w7b	3,5	2,78	2,46	3,12	3,44	2,74	2,24	7952	18223	0,436372
1avr	3,79	2,88	2,52	3,34	3,36	2,95	2,45	9668	15303	0,631772
1okc	3,16	2,48	2,1	2,9	2,93	2,3	1,91	12313	8845	1,392086
1vin	3,52	2,72	1,99	3,15	2,96	2,56	2,03	6057	9792	0,618566
1ad6	3,13	2,78	1,99	2,97	3,39	2,44	1,88	3449	4995	0,69049
1pbw	2,96	2,68	2,23	3	3,01	2,28	2	3126	5298	0,590034
2blh	3,58	2,76	2,12	3,1	3,15	2,34	1,87	2388	3362	0,710291
1elk	3,19	2,63	2,24	3,08	3	2,62	2,03	2314	3439	0,67287
1a00	3,29	2,42	1,95	2,7	3,17	2,36	1,78	2321	2528	0,918117
1d9s	3,8	2,8	2,3	2,97	3,03	2,97	2,31	2665	1867	1,427424
1rg6	2,99	2,46	1,92	2,97	3,04	2,47	2,09	395	683	0,578331
1dxs	2,9	2,34	2,04	2,88	2,93	2,4	1,93	340	421	0,807601
7ins	3,37	2,22	2,05	3,3	2,6	3,39	2,59	126	71	1,774648
MEDIAN	**3,33**	**2,7**	**2,11**	**3,04**	**3,035**	**2,515**	**2,03**			

Table S3b. RMSD calculated between Cαs from "true" and "lattice" models. Mixed conformations PDB accession numbers are reported for each analyzed crystal structure (from Pierri et al., Proteins 2008). For each protein HH contacts, PP contacts and HH/PP ratio are again reported.

mixed conformation	cubic_00	tetra_01	d_tet_02	quadr_03	cub2u_04	fcc_05	bcc_06	HH contacts	PP contacts	HH / PP ratio
1ngf	3,95	2,89	2,23	3,75	3,17	3,51	2,42	14971	21415	0,699089
1iep	3,52	2,83	2,15	3,31	3,18	2,92	2,3	6733	12153	0,55402
1i9b	4,92	3,15	2,64	4,12	3,69	4,76	3,25	2535	8695	0,291547
1no5	3,87	2,92	2,17	3,35	3,12	3,29	2,52	3451	6835	0,504901
2bh8	4,91	3,54	2,93	4,54	3,66	4,65	3,55	3961	3128	1,266304
2abl	3,77	2,59	1,99	3,06	2,93	2,97	2,15	1811	5091	0,355726
1m2e	4,02	2,45	2,04	3,6	3,14	3,14	2,25	2242	2175	1,030805
1o8v	4,6	2,96	2,24	3,89	3,29	3,71	2,99	1515	2804	0,5403
1ju5sh3	4,09	2,7	2,25	3,35	3,01	3,47	2,56	726	2370	0,306329
1o13	4,36	2,59	2,08	4,23	3,02	3	2,14	1457	1202	1,212146
1h7m	3,14	2,62	1,98	2,97	2,82	3,07	2,44	1153	1060	1,087736
1koy	3,79	2,35	3,09	2,03	2,93	2,09	1,78	265	679	0,39028
1dgz	3,19	2,68	1,95	2,94	2,96	2,81	2,11	101	219	0,461187
1c6w	3,18	2,63	2,21	3,24	3,28	2,88	2,18	63	198	0,318182
MEDIAN	**3,91**	**2,69**	**2,19**	**3,35**	**3,13**	**3,105**	**2,36**			

Table S3c. RMSD calculated between Cαs from "true" and "lattice" models. Beta-sheet PDB accession numbers are reported for each analyzed crystal structure (from Pierri et al., Proteins 2008). For each protein HH contacts, PP contacts and HH/PP ratio are again reported.

mainly sheets-ID	cubic_00	tetra_01	d_tet_02	quadr_03	cub2u_04	fcc_05	bcc_06	HH contacts	PP contacts	HH / PP ratio
1btg	7,71	3,94	3,96	4,99	4,45	4,96	4,3	973	1918	0,507299
2af5	4,98	3,43	2,31	4,48	3,61	4,31	2,92	4539	17095	0,265516
1pdg	6	3,52	3,24	4,57	4,21	5,8	4,29	726	1104	0,657609
1hae	3,66	2,48	2,29	3,04	3,12	2,77	2,76	340	611	0,556465
1ju5Sh2	2,8	1,99	1,83	6,09	3,12	2,01	2,25	393	450	0,873333
1zad	3,97	2,86	2,54	3,77	3,53	3,47	2,39	268	611	0,438625
1egf	4,4	2,59	2,52	4,04	3,27	3,23	2,54	201	477	0,421384
1epg	3,24	2,64	2,07	3,03	2,86	2,34	2,67	201	477	0,421384
3egf	3,29	2,79	2,49	3,36	3,03	3,23	2,36	201	477	0,421384
1k36	3,54	2,25	2,01	5,35	2,68	3,26	2,84	246	245	1,004082
1a3p	5,77	2,16	2,02	3,03	3,06	3,03	2,42	146	287	0,508711
1edm	3,58	3,24	2,83	3,49	3,5	3,29	3,23	132	221	0,597285
2gwp	5,5	2,55	2,59	5,47	3,56	3,44	2,74	129	100	1,29
1zr7	2,82	2,66	2,85	2,73	-	2,27	1,9	1	352	0,002841
MEDIAN	**3,815**	**2,65**	**2,505**	**3,905**	**3,27**	**3,245**	**2,705**			

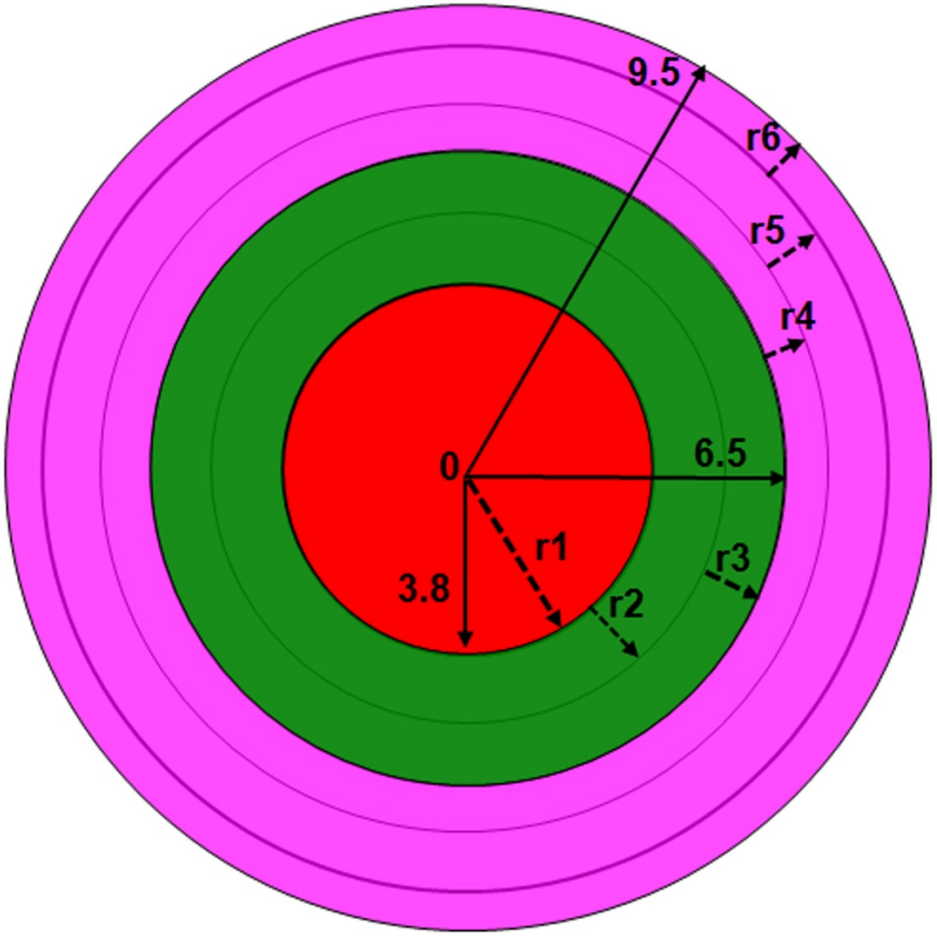

Figure S1. Schematic circular representation of the analyzed distance ranges. The three differently colored surfaces indicate the ranges of very short- (red), short- (green) and medium-range (purple) interactions, which are further classified into six sub-ranges (r1-r6, dashed arrows). Distances from the central Cα are indicated by solid arrows.

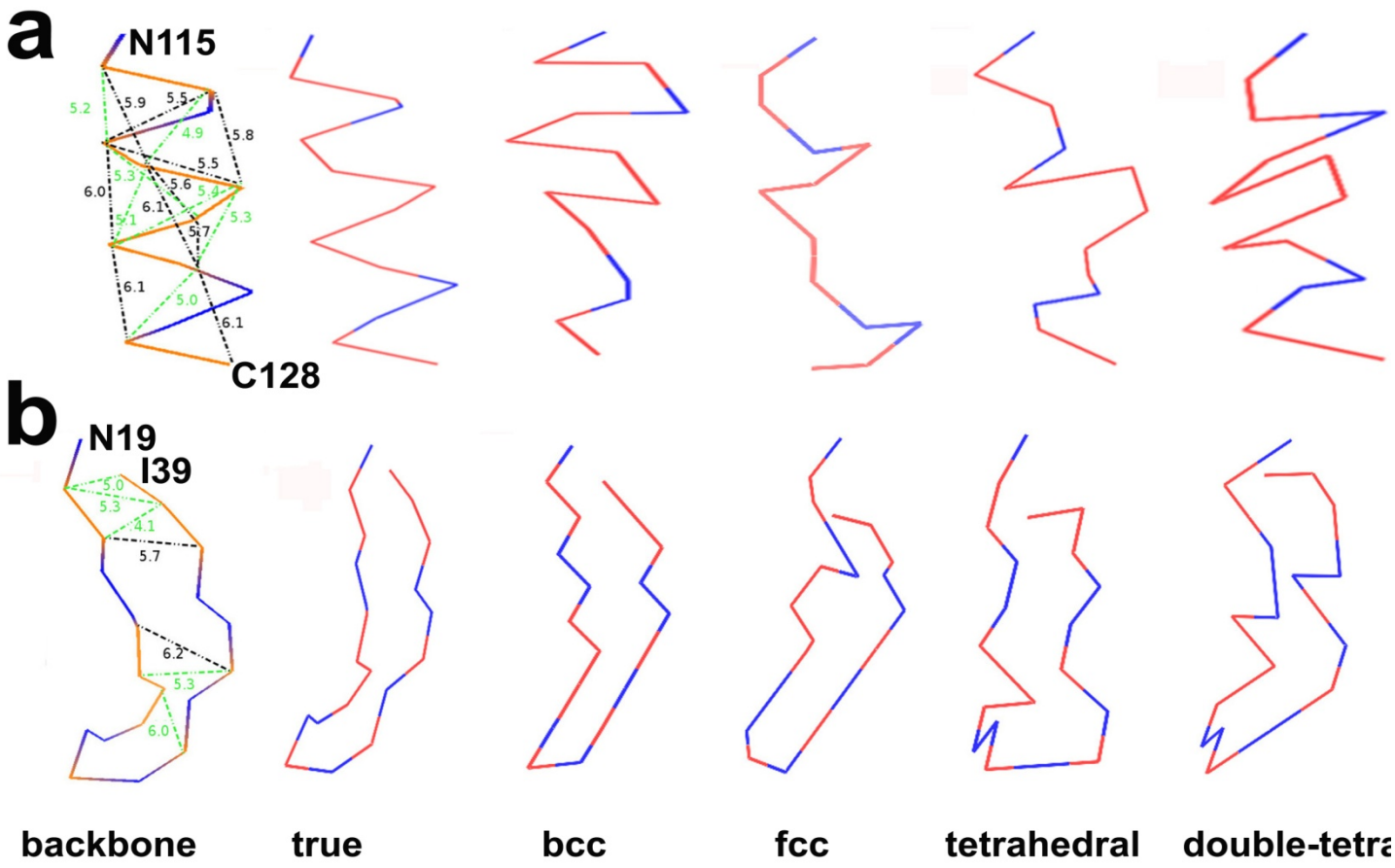

Figure S2. Fitting examples of intra-secondary structure interactions to lattices. a) All helix "1okc": ribbon representation for the protein fragment 115-NLASGGAAGATSL-127. b) All sheet "1zad" : ribbon representation for the protein fragment 19-NLCYKMFMMSDLTIPVKRGCI-39. Lattice models "true"; body center cubic, "bcc"; face center cubic "fcc", tetrahedral, double-tetrahedral, "double-tetra" are also reported as indicated on the bottom of the figures for both "1okc" and "1zad". All atom structures are reported in orange - blue ribbon representation (orange for hydrophobic and blue for polar residues) and side-chains are hidden. Lattice HP models are reported by ribbon representation, hydrophobic residues are reported in red, polar residues are reported in blue according to Kyte-Doolittle hydrophobicity scale. Green dashed lines highlight HH contacts in 5.4 Å distance range, black dashed lines highlight HH contacts in the 5.5 - 6.5 Å distance range.

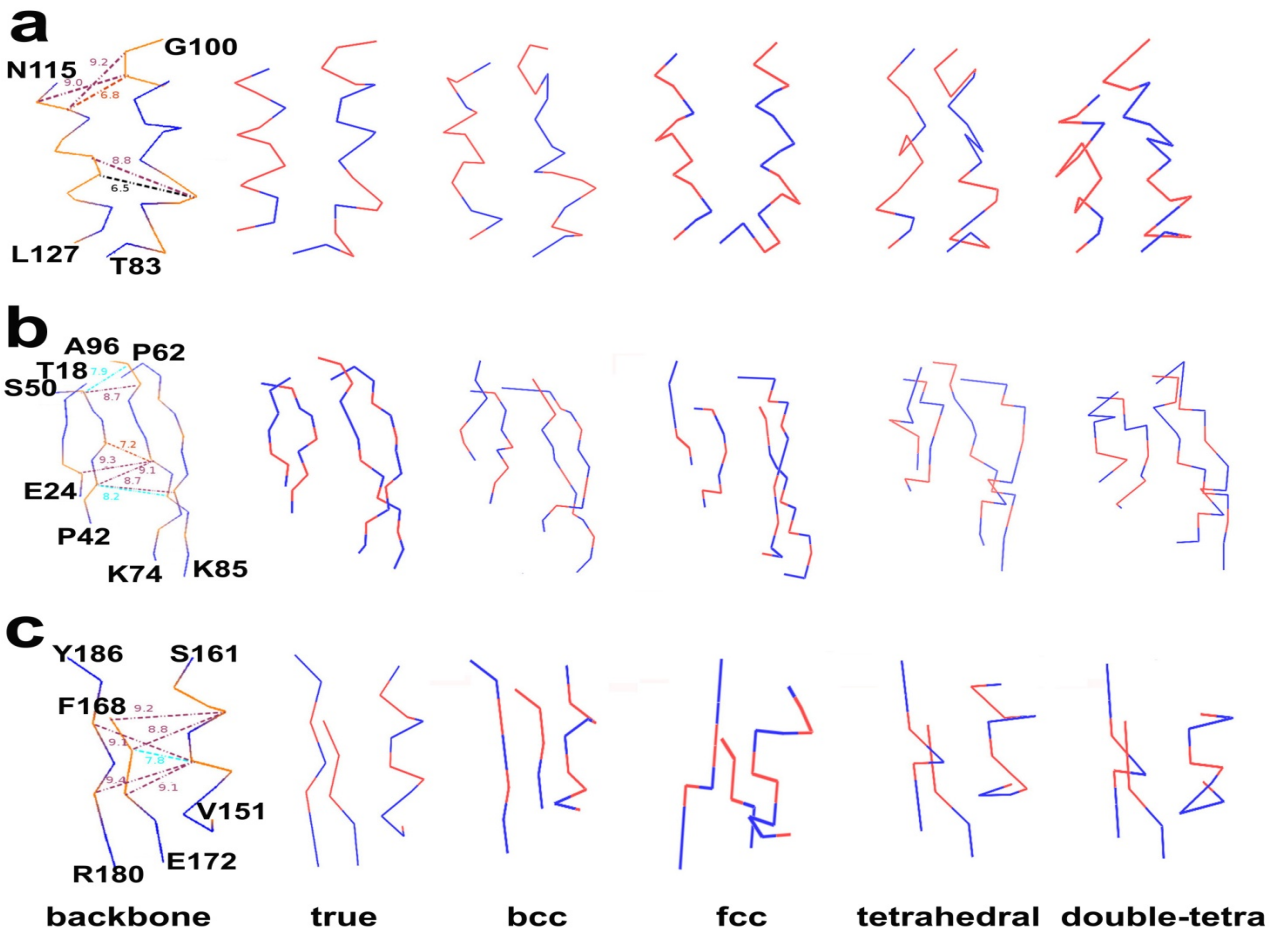

Figure S3. Fitting examples of inter-secondary structure interactions to lattices. a) All helix "1okc": ribbon representation for the protein fragments 83-TQALNFAFKDKYKQIFLG-100 and 115-NLASGGAAGATSL-127. b) All sheet "1pdg": ribbon representation for the protein fragments 18-TRTEVFE-24/42-PCVEVQRCS-50/62-PTQVQLRPVQVRK-74/85-KKATVTLEDHLA-96. c) Mixed conformation "2abl": 151-VSRNAAEYLLS-161/168-FLVRE-172/180-RSISLRY-186. Lattice models "true"; body center cubic, "bcc"; face center cubic "fcc", tetrahedral, double-tetrahedral, "double-tetra" are also reported as indicated on the bottom of the figures for both "1okc" and "1pdg" and "2abl".

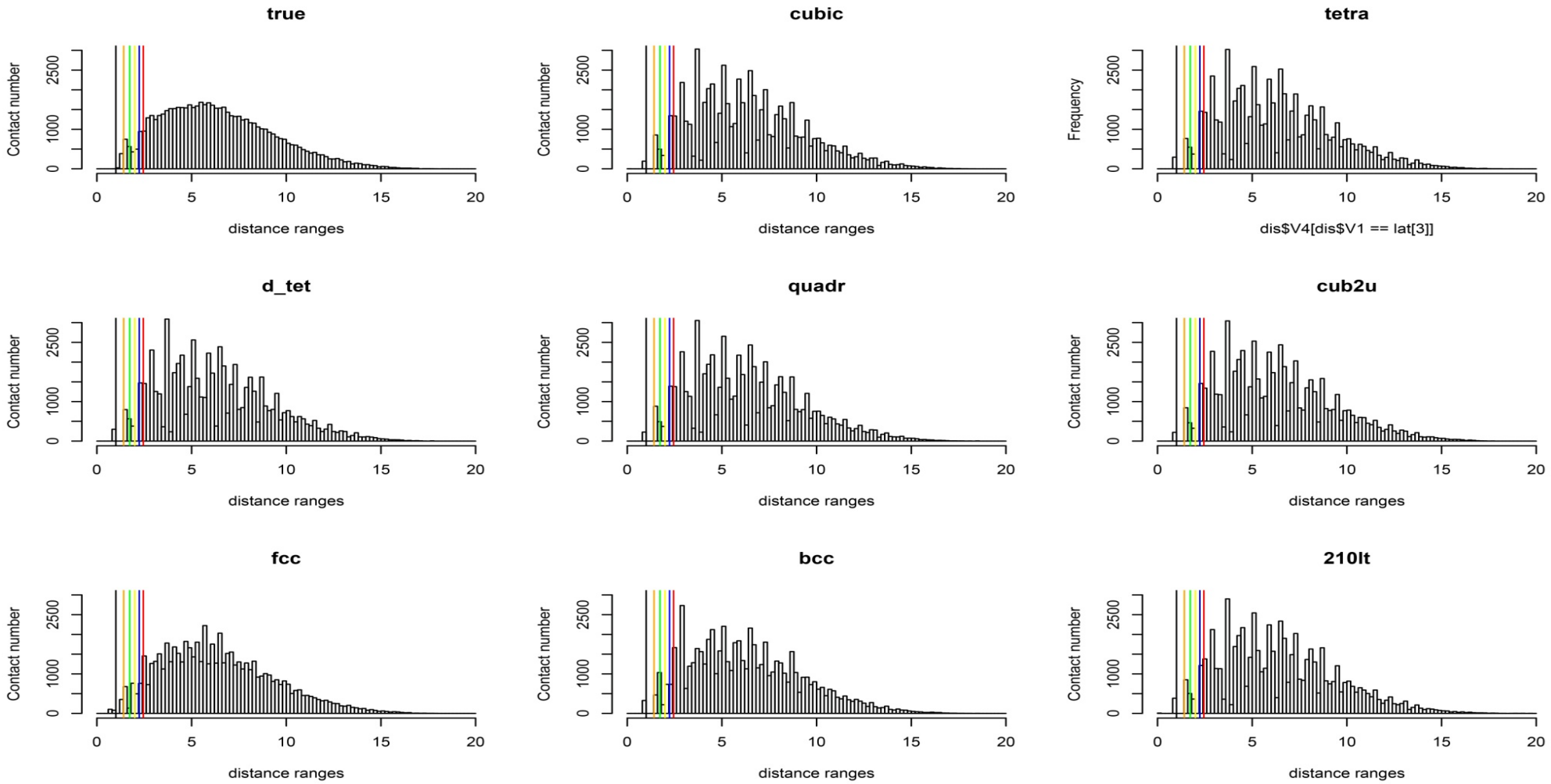

Figure S4. Distribution of all the HH contacts computable in the "true" models and in all the analyzed lattice HP models generated from mainly helix structures. The number of HH contacts counted through all true/lattice structures is reported on Y axis(contact number) for each of the represented bins covering the full range of measurable distances (reported on X axis as multiple of 3.8, i.e. from 0 x 3.8 to 20 x 3.8 Å) within the analyzed structures. Each bin covers a specific range of distances. Vertical colored lines indicate the extremities of distance range interactions containing very short- (r1, black line), short- (r2, r3, orange and green, respectively) and medium- (r4, r5, r6, yellow, blue and red, respectively) interactions.

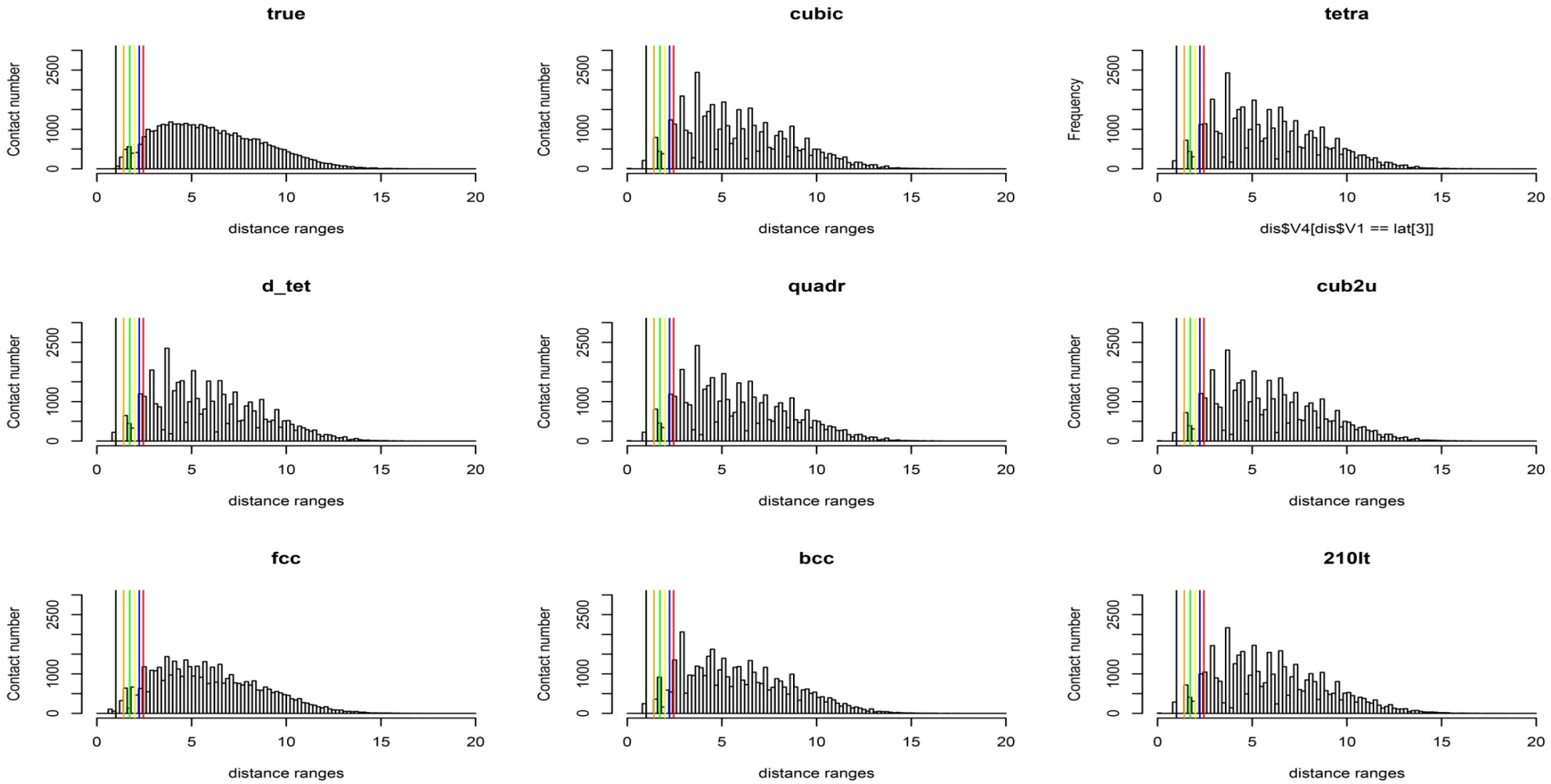

Figure S5. Distribution of all the HH contacts computable in the "true" models and in all the analyzed lattice HP models generated from mixed conformation structures. The number of HH contacts counted through all true/lattice structures is reported on Y axis (contact number) for each of the represented bins covering the full range of measurable distances (reported on X axis as multiple of 3.8, i.e. from 0 x 3.8 to 20 x 3.8 Å) within the analyzed structures. Each bin covers a specific range of distances. Vertical colored lines indicate the extremities of distance range interactions containing very short- (r1, black line), short- (r2, r3, orange and green, respectively) and medium- (r4, r5, r6, yellow, blue and red, respectively) interactions.

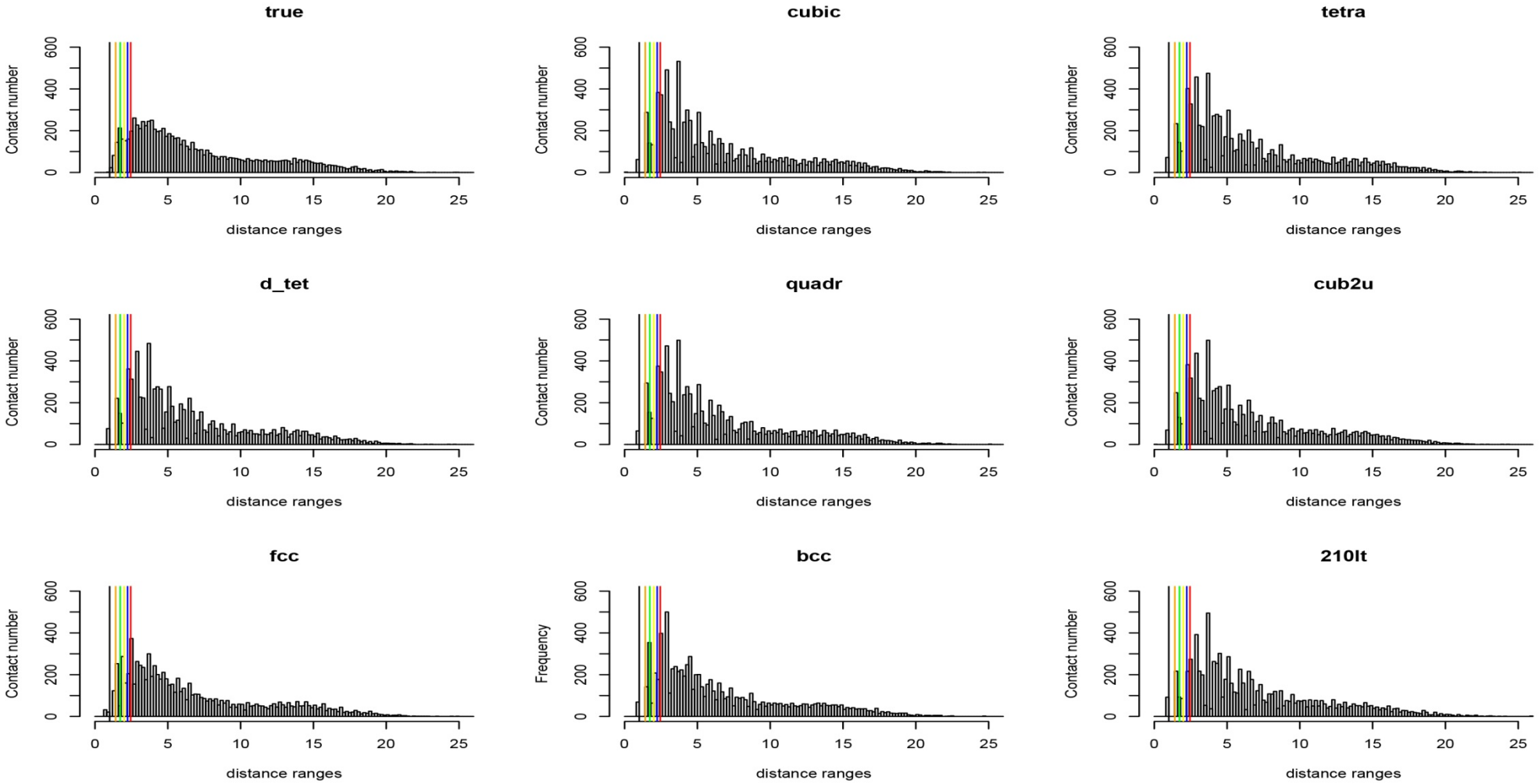

Figure S6. Distribution of all the HH contacts computable in the "true" models and in all the analyzed lattice HP models generated from mainly sheet structures. The number of HH contacts counted through all true/lattice structures is reported on Y axis (contact number) for each of the represented bins covering the full range of measurable distances (reported on X axis as multiple of 3.8, i.e. from 0 x 3.8 to 25 x 3.8 Å) within the analyzed structures. Each bin covers a specific range of distances. Vertical colored lines indicate the extremities of distance range interactions containing very short- (r1, black line), short- (r2, r3, orange and green, respectively) and medium- (r4, r5, r6, yellow, blue and red, respectively) interactions.

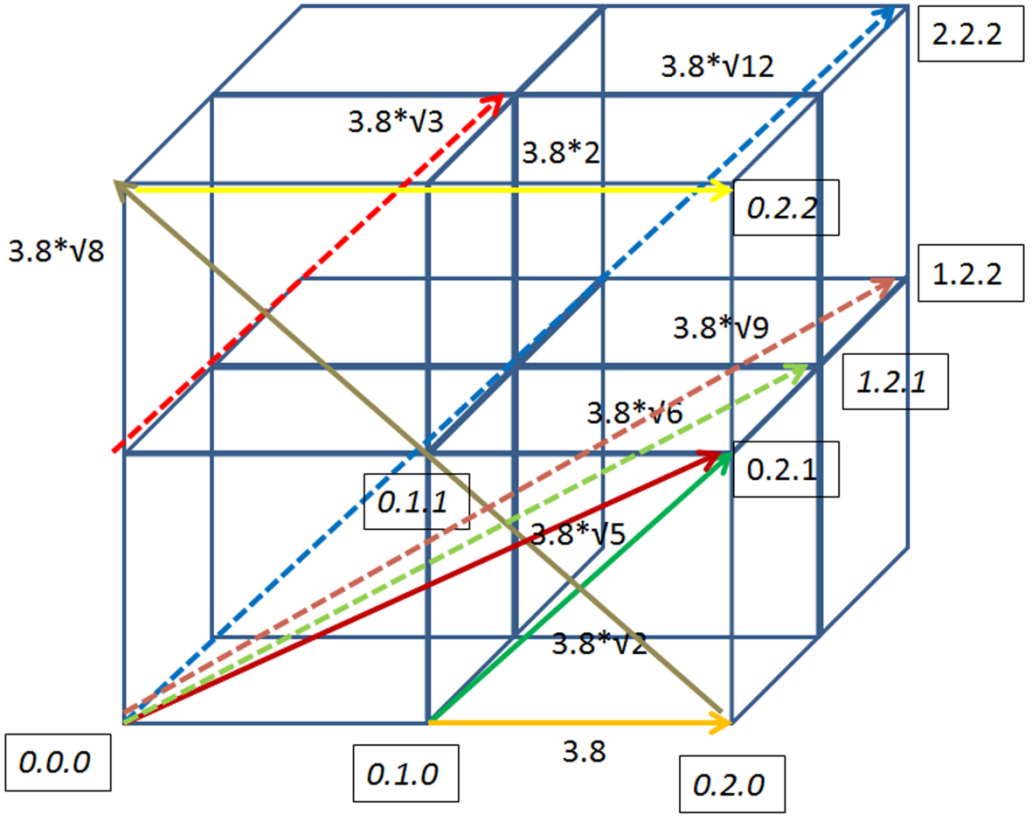

Figure S7. Examples of sides, face-diagonals (solid arrows) and depth-diagonals (dashed arrows) among vertices of a cubic lattice that could be occupied by the Cαs of a protein. Distances among vertices are indicated as multiple of 3.8 (or the distance between the Cαs of two residues connected by the backbone) and by labels. Cartesian coordinates (in units of 3.8 Å) of some vertices are indicated by labels within box.