THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

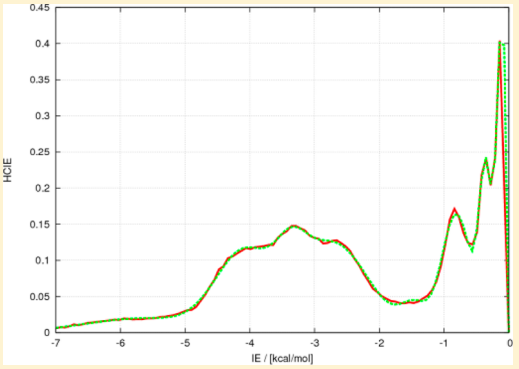
Optimal Definition of Inter-Residual Contact in Globular Proteins Based on Pairwise Interaction Energy Calculations, Its Robustness, and Applications

Boris Fačkovec and Jiří Vondrášek*

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo nam. 2, 166 10 Prague 6, Czech Republic

Supporting Information

ABSTRACT: Although a contact is an essential measurement for the topology as well as strength of non-covalent interactions in biomolecules and their complexes, there is no general agreement in the definition of this feature. Most of the definitions work with simple geometric criteria which do not fully reflect the energy content or ability of the biomolecular building blocks to arrange their environment. We offer a reasonable solution to this problem by distinguishing between "productive" and "non-productive" contacts based on their interaction energy strength and properties. We have proposed a method which converts the protein topology into a contact map that represents interactions with statistically significant high interaction energies. We do not prove that these contacts are exclusively stabilizing, but they represent a gateway to thermodynamically important rather than geometry-based contacts. The process is based on protein fragmentation and calculation of interaction energies using the OPLS force field and relies on pairwise additivity of amino acid interactions. Our approach integrates the treatment of different types of interactions, avoiding the problems resulting from different contributions to the overall stability and the different effect of the environment. The first applications on a set of homologous proteins have shown the usefulness of this classification for a sound estimate of protein stability.



INTRODUCTION

In a simplified way, the structure and stability of proteins in a water environment are determined by the flexibility of their backbones, the non-covalent interactions between the side chains of the composing amino acids, their interactions with solvent, and the hydrophobic effect, also driving the process of protein folding. Among the various non-covalent interaction motifs stabilizing biomolecules or their complexes, the hydrogen bonds, salt bridges, and vdW interactions play the most important role, but their origin is fundamentally different and their proportions are not easy to set. The question is how to assess the importance of these different contributions for overall protein stability and how to properly take into account non-homogeneous and non-uniform environments of the interacting amino acids.

A large number of studies have analyzed the available 3D structural data in the Protein Data Bank (PDB) and have shown that side chains have preferred interaction geometries; their packing is not entirely random.[1−3] The potential energy functions of proteins are most often approximated as a sum of the electrostatic charge−charge and Lennard-Jones contributions including the exchange-repulsion and dispersion terms. Molecular mechanics energy functions and the distributions of amino-acid pairs and their geometries observed in protein structures suggest that the intrinsic pairwise interaction energies indeed contribute to the packing of side chains in proteins rather than being overwhelmed by the numerous interactions with other atoms within the protein and with the solvent. As a protein folds into a stable 3D structure, residues, regardless of their distance in the sequence, mutually interact and come into "contact". Although "contact" is a fundamental concept of protein structure analysis, there is no general agreement as to how it should be defined.

A contact is a Boolean quantity determined usually by two steps for each pair of residues from the 3D structure of the protein. The first step is the quantification of their interaction—a function which takes two sets of atomic coordinate vectors as an input and produces a real number as an output. The full set of the atomic coordinates is often reduced to merely a single vector of three Cartesian coordinates, usually the geometry of an α carbon, beta carbon, or side-chain center of mass. The interaction between the two residues is then calculated only between such points. Alternative methods use mutual-surface-area calculations or the minimal distance of any pair of atoms and some other variants of this attempt.[4−6]

Received: March 31, 2012
Revised: September 17, 2012
Published: September 18, 2012

The second step in the definition of a contact is the selection of the threshold value for the calculated interaction quantity to be considered as a contact. Gromiha and Selvaraj presented in a review[7] an interesting survey of how many distance thresholds it is possible to use. Most researchers use arbitrary thresholds accepted in the field and justified by reasonable but heterogeneous assumptions.[9−10] Other ways are to perform analyses using different definitions and discuss their effect on the results. There have been several attempts[9,10] to establish a standard threshold distance value for a contact.

It is usually accepted that proximity in a 3D structure can be considered as a sign of a thermodynamically important interaction having an impact on protein stability. It seems plausible to assume that the contacts in protein chains could be useful for the search for hydrophobic clusters[11] or the development of statistical potentials.[12,13] Simple geometry definitions of a contact are satisfactory for studies which use contact maps as alternative structure representations of proteins.[14−16] Other applications would also significantly benefit from a sophisticated definition of a contact based rather on energy than on simple geometry criteria. The contact by means of the energy content depends on the nature of the interacting atoms and their environment. In order to identify the key contacts and key residues in protein structures by computational chemistry methods, the interaction energy matrix (IEM) concept was introduced[17] and further developed[18,19] to bring a new context into protein structure analysis. Still, further justification is needed, specifically sorting the contacts into categories of "productive" and "non-productive". Such a new methodology also needs a reasonable computational method capable of describing the interacting amino acids properly. The originally used quantum mechanics calculations demand an artificial fragmentation strategy[17] and are computationally too expensive. Fortunately, it has recently been found that in some cases including the intramolecular interactions of biomolecular building blocks the available empirical potentials are in very good agreement with the benchmark interaction energy calculations determined at the highest *ab initio* level.[20]

In this study, we use the empirical potential energy function to quantify the interaction between any two residues of a protein. We suggest treating the backbone and side-chain separately so that the "contact" is expressed by a value of the non-covalent interaction energy between both the backbones and side-chains. As the solvation energy of the ions, dipoles, and quadrupoles of the residues in question is different, we assume that using only one uniform dielectric constant for scaling all types of interactions would not reliably model the effect of the environment. On the other hand, we cannot simply neglect the effect of the environment when evaluating the interactions between heterogeneous groups of amino acids in the gas phase. Therefore, we decided to classify the inter-residual non-covalent interactions based on their physical-chemical characteristics and sort them into corresponding groups reflecting their interaction properties. Besides the classification, it would be very useful to separate the interactions based on their contribution to the overall stability of a protein. We have therefore defined the productive and non-productive interactions as a measure of their importance to stabilize significantly or merely buffer other factors contributing to protein stability.

Additivity is a very helpful property of molecular mechanics force field interaction energies. As we construct an independent optimal contact definition separately for each type of interactions, we implicitly assume that the whole stabilizing energy can be easily decomposed. An objection might be raised that, as the interactions are not independent, their free energies are not additive. Nevertheless, the potential energy contributions are additive in a single microstate. We model the native state ensemble for a protein with just one well-resolved experimental geometry structure. We further assume that there is an interaction compensation in the unfolded state ensemble and an entropic compensation for each type of interactions which determines the properties of the native-state interactions. The contact definitions presented in this work are the statistical property of the native state of a protein only. The definitions enable us to merge all of the inter-residual non-covalent interactions into one desired quantity—a contact.

To follow the construction process, we first introduced representations of the non-covalent interaction-energy distributions—the cumulative distribution and its derivative function (the histogram of the contributions of the interaction energies—HCIE). We subsequently present a tenable classification of the amino-acid side-chains in globular proteins based on the similarity of their HCIE functions. Next, we discussed the number of the productive contacts of the amino acids in each class in order to find reasonable limits for an optimum contact definition. Finally, we present contact definitions for the derived classes of inter-residual interactions as the statistically significant values on the HCIE curves and discussed their properties.

METHODS

The method of the structure set construction, protein fragmentation, and calculation of the interaction energies is the same as in our characterization of the residue interaction energy (RIE) distributions in globular proteins.[21] The X-ray structures with a resolution below 2.0 Å of the single-chain proteins with no ligands were obtained from the PDB[22] (Jan 31, 2011). Structures with a 70% sequence identity and higher were eliminated. The database filter yielded 1531 structures. This number was slightly reduced by inconveniences with file processing to 1358.

After optimization of hydrogen atom positions in the whole structure, the pairwise non-covalent interaction energies for 2*N* fragments (*N* side-chains and *N* backbone fragments) using an OPLS[23−25] force field were calculated, excluding those between the backbones of subsequent amino acids and the side-chain and backbone of the same AA, which were set to zero. All the calculations were repeated using the CHARMM27[26] force field for comparison. Utilization of the OPLS or CHARMM force fields guarantees that the backbone (including C_α atoms) and side-chain fragments are neutral. The interactions were calculated as the sum of the interatomic Lennard-Jones and Coulombic contributions in the gas phase ($\epsilon_r = 1$, see eqs 1 and 2). Only the interactions of an absolute value exceeding 0.05 kcal/mol were considered throughout the work in order to prevent sampling zeros. Backbone atoms of terminal residues were not taken into consideration, as their backbones do not fit any group.

$$U_{\text{coulomb}} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \quad (1)$$

dx.doi.org/10.1021/jp303088n | J. Phys. Chem. B 2012, 116, 12651−12660

$$U_{\mathrm{vdW}} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} 4\epsilon_{i,j}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad (2)$$

Construction of the HCIE curves, which are interaction energy histograms multiplied by interaction energy (IE), was done as follows. First, all of the interaction energies of the selected type of absolute value higher than 0.05 kcal/mol were sampled from all the proteins and sorted. For each interaction energy value, all of the lower values were summed to obtain the cumulative HCIE, which was then differentiated. Differentiation was done by binning the cumulative HCIE and least-squares fitting the lines to data points in the bins. This method of HCIE curve smoothing is not biased toward Gaussians, which ensures that the observed nature of HCIE is genuine.

Distance matrices were calculated for the four most commonly used definitions of inter-residual distance. Only heavy atoms were considered in all of the contact-matrix (CM) calculations. "CA" distances were defined as geometrical distances between Cα atoms of residues. "CB" distances were defined as geometrical distances between Cβ atoms except for glycines, for which Cα atom positions were used instead of Cβ. The "center" distances were defined as the geometrical distances between the centers of geometry for residues which were calculated from positions of all the heavy backbone and side chain atoms. Finally, the "minimum" distances were defined as the distances between the two closest heavy atoms, one from each residue. When comparing the contact matrices, we set the contact definition values for each distance definition so that the number of all the contacts was equal to the number of all contacts in our structure set determined by OPLS inter-residual interaction energy. The similarity of contact matrices i and j was defined as

$$s_{ij} = \frac{|\mathrm{A} \cap \mathrm{B}|}{\sqrt{|\mathrm{A}||\mathrm{B}|}} \quad (3)$$

where A and B are sets of contacts in contact matrices i and j, $\cap$ denotes the set intersection, and |A| and |B| are the numbers of elements of sets A and B.

To demonstrate the applicability of the contact definition, we decided to analyze the thermal stability on a set of homologous proteins—hyperthermophiles and their mesophilic counterparts. The application of the strategy described above was straightforward. The structures of 22 pairs of known homologous proteins from thermophilic and mesophilic organisms were downloaded from the PDB database (see Table 3) according to the work of Kannan et al.[27] For the NMR structures, the first model was considered; if any other atom occupancy was present in the PDB file, the first occupied position was always considered. The gas-phase optimization of the hydrogen atoms in proteins was performed in GROMACS using the OPLS force field. The interaction energy matrices were calculated, and the contact matrices were constructed using our contact definitions. The numbers of residue−residue contacts of all the types were summed and divided by the number of residues.

To explore more deeply the utilization of the energy defined contacts for thermodynamic stability of mutational variants of a protein, we focus on the lysozyme—one of the most populated proteins in the PDB. There are five papers[28−32] dealing with change of thermodynamic stability or melting temperature of the protein upon series of mutations. Only full atom X-ray structures of mutants were taken into account for this purpose, and the same routine was used for optimization of the hydrogens as well as for construction of the contact matrices. In this case, we could only analyze data paper by paper, since the experimental conditions and methods differed so the comparison of the whole set would not be consistent.

RESULTS AND DISCUSSION

The substantial difference between the interaction energy (IE) and previously defined residue interaction energy (RIE)[21] distributions results from the fact that the number of pairwise interactions grows quadratically with the number of amino-acid residues in a protein, whereas the number of productive interactions grows approximately linearly with the chain length. The limit of an IE histogram in principle diverges at IE → 0. Its finiteness, which is observed in reality, arises from the finite diameter of protein molecules. Therefore, the identification of the optimum definition of residue−residue contact from IE histograms is not straightforward.

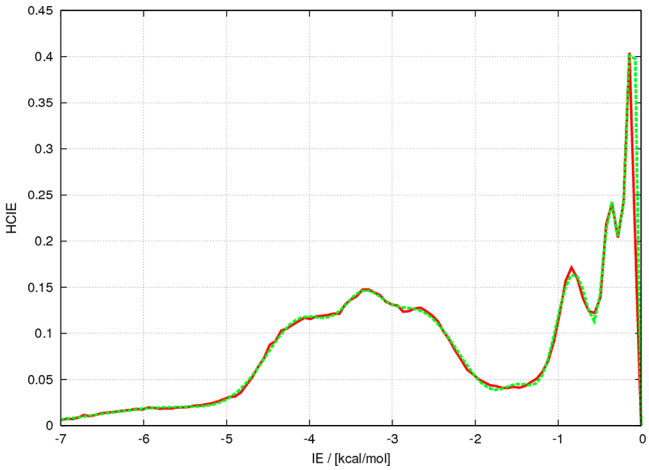


Figure 1. The HCIE curve for backbone interactions constructed from the calculated data (red) and the function (eq 4) fitted to these data (green). m = 8 Gaussians were used, 7 of which for productive interactions. The calculated data are very well described excluding HCIE at IE > −0.2 kcal/mol, where the Gaussian is a wrong approximation of the function diverging to ∞.

A useful alternative approach is the multiplication of an IE histogram by an IE value, i.e., construction of a HCIE curve. The HCIE function represents the contribution of IEs in an IE interval to the sum of all the IEs and is characterized by the following properties

$$\lim_{\mathrm{IE}\to-\infty} \mathrm{HCIE} = 0$$

$$\lim_{\mathrm{IE}\to+\infty} \mathrm{HCIE} = 0$$

$$\mathrm{HCIE}(0) = 0$$

$$\mathrm{HCIE} < 0 \;\forall\; \mathrm{IE} > 0$$

$$\lim_{\mathrm{IE}\to 0^-} \mathrm{HCIE} \propto |\mathrm{IE}|^x, \quad 0 < x < 1 \quad (4)$$

Integral $\int_{-\infty}^{x} \mathrm{HCIE}\, \mathrm{d}\,\mathrm{IE}$ equals the contribution of all interactions lower than x to the stabilization enthalpy. The multiplication by IE ensures convergence in the case of short-ranged interactions like dispersion and multipole−multipole, whose density of state goes to IE^x, $x \in \{-1, 0\}$, at IE → 0.

dx.doi.org/10.1021/jp303088n | J. Phys. Chem. B 2012, 116, 12651−12660

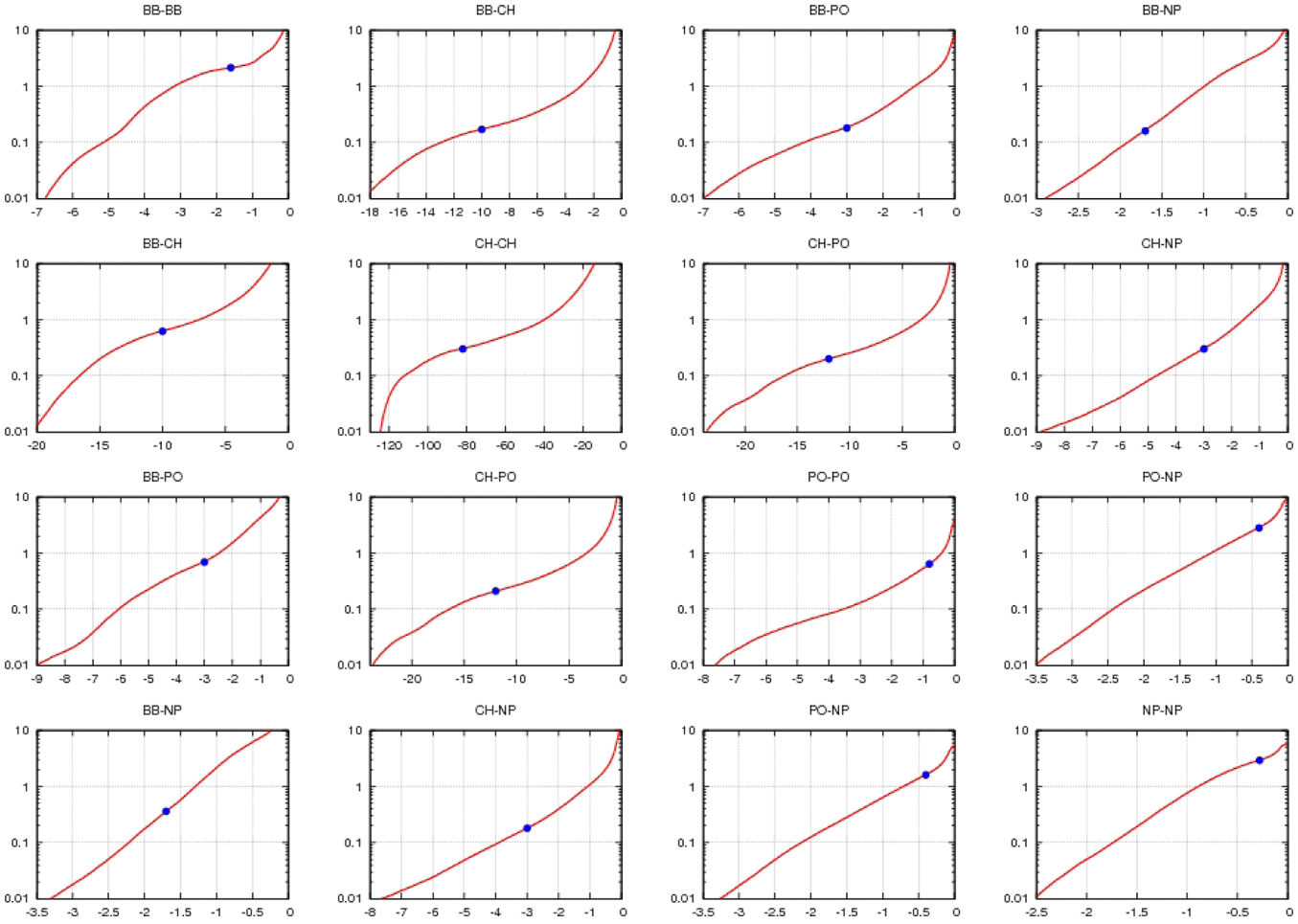


Figure 2. The number of contacts that one residue of a particular type (types in rows—first row BB, second row CH, third row PO, fourth row NP) participates on average from a particular type of interaction as a function of the interaction energy contact definition. Results for the OPLS force field are in red. The determined optimum contact definitions (marked blue) are very close to the inflection points with the lowest first derivatives.

Long-ranged interactions compensate by a mechanism similar to that in ionic crystals and because of the finite diameters of proteins.

The IE value X for which

$$\int_{-\infty}^{X} \mathrm{HCIE}\ \mathrm{dIE} = \int_{-\infty}^{+\infty} \mathrm{HCIE}\ \mathrm{dIE} \tag{5}$$

seems to be a natural energetic definition of a residue−residue contact, because X defines the point where the weaker attractive interactions are compensated by all of the repulsive ones. However, since we understand productive interactions as exceptionally strong and not only attractive, the sum of the bulk interactions should be non-positive but not necessarily zero. Therefore, it provides a useful upper boundary for productive contact definition.

HCIE has an interesting shape with the local minima and maxima corresponding to specific interaction patterns. The interactions between residues in pairs rarely reach their energy minima, because the optimum positions are rarely met.[19] As some types of interactions are required by global protein topology, the local density of states is deformed. An example of this effect are the interactions between non-polar residues inside a protein, which are strongly affected by their tendency to cluster owing to the minimization of the exposed hydrophobic surface area. Therefore, the IE of each interaction pattern can be approximated by a random variable with normal distribution. The HCIE can be reliably approximated by a sum of Gaussians and a function (diverging to ∞ at IE → 0) corresponding to the bulk interactions, all multiplied by IE. We approximated the function corresponding to bulk interactions by a sum of Gaussians (n Gaussians for bulk interactions and $m - n$ for the productive ones in eq 4).

$$\mathrm{HCIE} = \mathrm{IE}\left(\sum_{i=1}^{n} a_i \mathrm{e}^{-(\mathrm{IE}/\sigma_i)^2} + \sum_{j=n+1}^{m} a_j \mathrm{e}^{-((\mathrm{IE}-\mathrm{IE}_j)/\sigma_j)^2}\right) \tag{6}$$

The HCIE is very well described by the proposed function (see the fit in Figure 1) in the IE region of productive contacts but quite poorly in the bulk IE region. In our studies, we have found that fitting the proposed function on the obtained data leads to vast errors in the determined parameters, especially in the case of long-ranged interactions because of the large contribution of the bulk interactions. Therefore, we attempted to identify local minima or points with significant change of slope—the intersection of two Gaussians corresponding to different interaction patterns.

Classification of Amino Acids. The major difficulty of the pairwise interaction energy concept results from the huge compensation of the electrostatic interactions. Therefore, only interaction energies undergoing the same compensation by solvation and with the same distance scaling can be directly compared. We therefore propose the classification of amino-

dx.doi.org/10.1021/jp303088n | J. Phys. Chem. B 2012, 116, 12651−12660

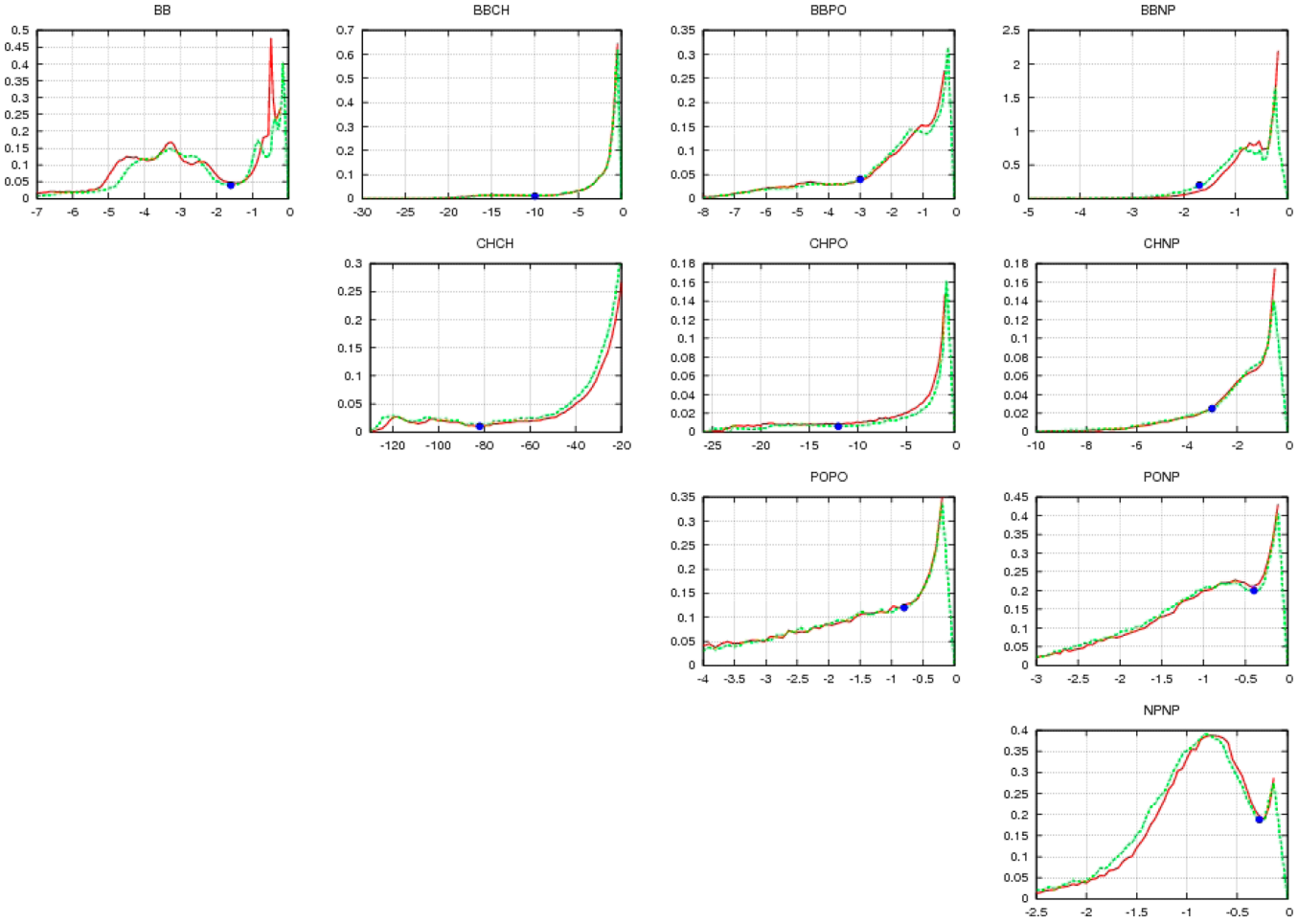


Figure 3. The HCIE curves for all 10 types of interactions calculated using OPLS (green) and CHARMM (red) force fields. The contact definitions are marked in blue.

Table 1. The Calculated Contact Definitions and Their Properties[a]

#IE	CD OPLS	BIE	compens	compCD	CD charmm
BBBB	−1.6	−0.2	1.14	0.72	−1.5
BBCH	−10	−3.5	4.04	0.39	−10
BBPO	−3	−0.4	1.26	0.28	−3.5
BBNP	−1.7	−0.1	1.05	0.08	−1.5
CHCH	−82	−69	11.70	0.81	−82
CHPO	−12	−4	3.03	0.47	N/A
CHNP	−3	−1.4	2.50	0.44	−3
POPO	−0.8	−0.5	1.29	0.90	−0.7
PONP	−0.4	−0.1	1.06	0.82	−0.4
NPNP	−0.3	−0.2	1.09	0.96	−0.3

[a]The BIE (abbreviation for boundary interaction energy) is an IE value, for which all the weaker interactions compensate with positive interactions (see eq 5). The compensation of the interactions with positive and negative IE values ("compens", fourth column) is always expressed as a ratio of the sum of all the negative interactions to the sum of all the interactions. compCD is the ratio of the energy content of the productive and energy content of all the interactions. All the energy values are in kcal/mol.

Table 2. Similarity of Contact Matrices[a]

	opls	charmm	CA	CB	center	mindist
opls	1	0.95	0.49	0.44	0.52	0.51
charmm	0.95	1	0.49	0.43	0.52	0.50
CA	0.50	0.49	1	0.71	0.77	0.79
CB	0.44	0.43	0.71	1	0.77	0.65
center	0.52	0.52	0.77	0.77	1	0.72
mindist	0.51	0.50	0.79	0.65	0.72	1

[a]"opls" denotes contact matrices calculated from interaction energy matrices using an OPLSAA force field and our contact definitions. "charmm" denotes the same using the CHARMM27 force field. "CA" contact matrices are based on Calpha atom distances, and "CB", on C beta atom distances. "center" contact matrices are based on the geometry centers of residues calculated from positions of heavy atoms. "mindist" contact matrices are based on the minimum heavy atom distance. See the Methods section.

acid fragments based on their multipolar characters—charged (CH), polar (PO), and non-polar (NP) side chains. In addition to these, the backbone fragments (BB) are so numerous and their interactions are so specific in proteins that they need to be treated as a separate class.

Our key hypothesis is that a residue−residue non-covalent interaction of a certain class with a lower IE value is stronger and more stabilizing than the one with a higher (less negative) value. In other words, we suppose that each IE distribution corresponds to the free-energy distribution which has a similar shape with significantly strong interactions which can be considered as contacts. Although all types of interactions have

dx.doi.org/10.1021/jp303088n | J. Phys. Chem. B 2012, 116, 12651−12660

Table 3. The Application of Energy and Distance Defined Contacts (Minimum Distance Criteria) on Thermostable and Mesostable Protein Homologues[a]

PDB id		OPLSAA		mindist		PDB id		OPLSAA		mindist		OPLSAA	mindist
thermophile	N	CN	CN/N	CN	CN/N	mesophile	N	CN	CN/N	CN	CN/N	delta CN/N	delta CN/N
1THL	316	784	2.48	1370	4.34	1NPC	317	785	2.48	1345	4.24	0.005	0.093
1LDN	316	816	2.58	1341	4.24	1LDM	329	835	2.54	1326	4.03	0.044	0.213
3PFK	319	838	2.63	1322	4.14	2PFK	301	823	2.73	1255	4.17	−0.107	−0.025
1RIL	146	315	2.16	576	3.95	2RN2	155	387	2.50	603	3.89	−0.339	0.055
1BMD	327	916	2.80	1352	4.13	4MDH	333	844	2.53	1360	4.08	0.267	0.050
2PRD	174	417	2.40	652	3.75	1INO	175	359	2.05	666	3.81	0.345	−0.059
1PHP	394	985	2.50	1637	4.15	3PGK	415	670	1.61	1517	3.66	0.886	0.499
1THM	279	739	2.65	1217	4.36	1ST3	269	659	2.45	1152	4.28	0.199	0.079
1BTM	251	621	2.47	1066	4.25	1TIM	247	518	2.10	937	3.79	0.377	0.453
1YNA	193	478	2.48	731	3.79	1XYN	178	405	2.28	693	3.89	0.201	−0.106
1XYZ	320	1000	3.13	1406	4.39	2EXO	312	895	2.87	1359	4.36	0.256	0.038
1CAA	53	108	2.04	196	3.70	6RXN	45	114	2.53	157	3.49	−0.496	0.209
1BRF	53	106	2.00	190	3.58	1RB9	51	110	2.16	184	3.61	−0.157	−0.023
1GD1	332	947	2.85	1420	4.28	1GPD	333	610	1.83	1195	3.59	1.021	0.689
1TIB	269	716	2.66	1134	4.22	1LGY	265	795	3.00	1093	4.12	−0.338	0.091
1ZIP	217	558	2.57	910	4.19	1AK2	220	582	2.65	881	4.00	−0.074	0.189
1FFH	287	830	2.89	1230	4.29	1FTS	295	791	2.68	1269	4.30	0.211	−0.016
1PCZ	183	477	2.61	738	4.03	1VOK	192	513	2.67	741	3.86	−0.065	0.173
1OBR	323	945	2.93	1422	4.40	2CTC	307	882	2.87	1353	4.41	0.053	−0.005
1PHN	162	447	2.76	679	4.19	1CPC	162	437	2.70	668	4.12	0.062	0.068
1TMY	118	320	2.71	462	3.92	3CHY	128	369	2.88	504	3.94	−0.171	−0.022
1GTM	417	1177	2.82	1807	4.33	1HRD	449	1249	2.78	1920	4.28	0.041	0.057
averages			2.59		4.10				2.48		3.99	0.105	0.115

[a]The PDB code of thermophiles can be found in the first column, the number of their residues in the second column, the numbers of contacts and contacts per residue ("N and Cont/N") for OPLS AA and minimum distance criteria are in columns 3−4 and 5−6, respectively. Columns 7−12 contain the same data for mesophilic proteins. The last two columns are calculated differences between the numbers of contacts evaluated by both methods per residue.

different IE scales, their free-energy distributions should have scales comparably similar, because the experimentally observed effects of these interactions on the protein stability are very similar. Additionally, the forces forming the IE distributions of productive Gaussian interactions are of similar character and therefore similar in magnitude. It is plausible to suggest that the contact definition values can be understood as values scaling the IE distributions to sort out the free energy distributions and separating interactions with significant interaction free energies from the negligible ones. We require additivity of the contacts, so the number of contacts for a particular amino acid quantifies the stabilization of a protein by residue−residue interactions of this amino acid.

The proper classification of fragments was derived from similarity of the HCIE curves for all the amino-acid pairs. Pairs with similar HCIE curves are supposed to belong to the same fragment class. All 210 HCIE curves can be found in the Supporting Information. We propose the following classification. Each residue is cut into two fragments—side-chain and backbone (BB). The side-chains are classified as charged (CH: Asp, Glu, Lys, Arg, His), polar (PO: Asn, Gln, Thr, Ser, Tyr, Trp), or non-polar (NP: Ala, Leu, Ile, Val, Pro, Cys, Met, Phe), yielding 4 types of fragments and therefore defining 10 types of mutual pairwise interactions. The only exception is Gly with no side chain. The His was always treated as double protonated and charged. This simplification of the His protonation state should not be critical, and in the case where His is not charged, it could be treated as polar. Trp and Tyr residues are ambivalent: on the one hand, they can form hydrogen bonds and have a relatively strong dipole moment and therefore strongly interact with charged residues, but on the other hand, they are very often located in the hydrophobic core of proteins and interact with non-polar residues via short-ranged and relatively strong van der Waals interactions. We still face the problem of the proper description of some Cys residues which seem usually to subdue non-covalent interactions to covalent Cys−Cys bonds.

Average Number of Contacts. Having classified the side-chains into groups based on their HCIE, we can characterize each inter-residual interaction energy type by plotting the number of contacts possessed by one amino acid against the contact definition in log scale. The average number of contacts per one amino acid of a particular type is a sum of its four average numbers of contacts (contributed by the interactions with each type of fragment). We assumed that the reasonable sum of four average numbers of contacts should lie between 0.01 and 1.

Productive Contact Definition. Identification of the local minimum or point with significant change of slope separating productive and bulk interactions is shown in Figure 3. For the justification of the stationary points, the average number of contacts was also taken into consideration (Figure 2) for each minimum or inflection point used for the contact definition. We have excluded all of the inflection points and minima with an unreasonable average number of contacts. For example, −0.5 kcal/mol in the case of BBNP would suggest that one backbone fragment has more than three productive BBNP interactions.

Table 4. The Application of the Energy Defined Contacts on Stability of Sets of Lysozyme Mutants[a]

PDB id	CN/N	ΔG (kcal/mol)	Tm (°C)
1UIG	3.14	10.1	
1UIC	3.11	9.9	
1UID	3.09	9.7	
1UIE	3.05	9.3	
1UIF	3.23	8.8	
1LSN	3.14		73.6
1LSM	3.22		77.3
1IOR	3.19		69.5
1IOQ	3.21		70.4
1IOS	3.19		66.8
1IOT	3.10		64.2
1FLQ	3.18		61.0
1FLU	3.11		62.7
1FLW	3.13		64.4
1FLY	3.20		65.6
1HEM	3.21		77.5
1HEN	3.21		74.5
1HEO	3.13		71.2
1HEP	3.17		73.4
1HEQ	3.23		75.5
1HER	3.19		73.0

[a]The first column is the PDB id, the second is the CN/N ratio, and the third and fourth columns are corresponding ΔG of unfolding or Tm of a mutant melting temperature.

In the case of charged−charged interactions, the level of compensation productive/non-productive IEs is higher. This is because the long-ranged non-productive interactions are more important for electrostatic than for the vdW interactions, whose strength decreases much faster with distance. The fact that the contribution of the bulk interactions is much higher in the case of long-ranged interactions (Figure 3) can be attributed to the higher compensation of the positive and weak negative interactions (see the BIE values in Table 1).

In the case of backbone fragments and their interactions, we can see peaks representing particular structural motifs reflecting most probably their distance in sequence. The peak with separation in sequence = 4 corresponds to helices with IE ∼ −4 kcal/mol, separation in sequence >7 for beta-sheet interactions with IE ∼ −3.2 kcal/mol, separation in sequence = 3 and separation in sequence = 2 for interactions in loops with IE ∼ −2.5 kcal/mol and IE ∼ −1 kcal/mol, respectively.

As shown in Table 1, the consideration of Tyr and Trp as polar residues fits into our classification schema very well. It is demonstrated by the fact that the PONP and NPNP contact definition values are very similar.

As one contact is shared by two residues, the numbers of contacts per residue from Figure 2 must be divided by 2. The summation of all four contributions to each overall average number of contacts yields 1.34 for BB and 0.74, 2.25, and 2.56 for the CH, PO, and NP side-chains, respectively.

Comparison of Contacts Defined by Geometry and Energy. To assess the robustness and convertibility of the contact matrices defined by energy and by geometry criteria, we compared contact matrices constructed using our contact definition with matrices based on a different definition of the geometry criteria. It is clear (see Table 2) that geometry contact matrices are very sensitive to the way of their definition. Second, all of the geometry-based contact matrices are different from the contact matrices based on interaction energies, which can be attributed to the missing effect of mutual orientation and to different average distances between the residues for particular interaction types. A comparison of the contact matrices based on our contact definitions using OPLS and CHARMM force fields for interaction-energy calculations indicates that contact definition based on energy is robust to the utilization of a different force field. The sensitivity of energy-based contacts to a force field is even much lower than the sensitivity of geometry contacts to the way of definition used.

Application of Contact Definition to Thermal Stability Prediction. Test Case of Thermophilic Proteins and Lysozyme Mutants. To show a practical utilization of the suggested energy contact definition, we first decided to test a correlation of protein thermal stability with the average number of contacts in globular proteins from thermophilic organisms and from their mesophilic counterparts. There are many works dealing with the issue of thermostability, and we have to highlight here the review of Kumar and Nussinov published in 2001.[33] It is plausible to hypothesize that the highly stable protein homologues should have a higher number of contacts per amino acid. There are three main rebuttals to this hypothesis. First, the stabilization mechanism is probably not as simple as the number of intramolecular stabilization interactions or their strength. Second, the contacts in thermostable protein might just be stronger instead of more common. Third, the contacts might be enhanced or formed just at some location important for protein stability. On the other hand, it would be a great help for biochemists to acquire a quick qualitative orientation in protein stability.

There is also one problem we addressed which is connected with sparse data regarding thermostability of proteins. According to Kumar et al.,[34] there is a strong correlation (r^2 = 0.83) between folding free energy and melting temperature. Unfortunately, the level of confidence is low. Experimental data for folding free energies are sparse in general. We decided to correlate the number of contacts per residue with melting temperature mainly because it is more common and therefore probably useful but also because of the scarcity of ΔG data. We are aware of the error introduced by this effect, and it is our goal to develop the method toward the ΔG rather than to use the Tm.

The results for a set of 22 thermophilic proteins and their less stable mesophilic homologues are in Table 3. We must emphasize here that we just took our contact definition and applied it directly to a set of proteins taken from the work of Kannan et al.[35] In this special case, both type of proteins (thermophile and mesophile) share the same architecture and fold and both variants are approximately of the same length. In most of the cases, the number of contacts per residue rises as the stability of a protein (measured by Tm) increases. It would be naive to expect that the melting temperature Tm and number of contacts per residue are linearly proportional for all proteins in the studied set. Not only because a different stabilization mechanism could be characteristic for certain protein's fold but also because no quantitative measurement for Tm differences was applied. On the other hand, the presented correlation between Tm, the number of contacts per residue,

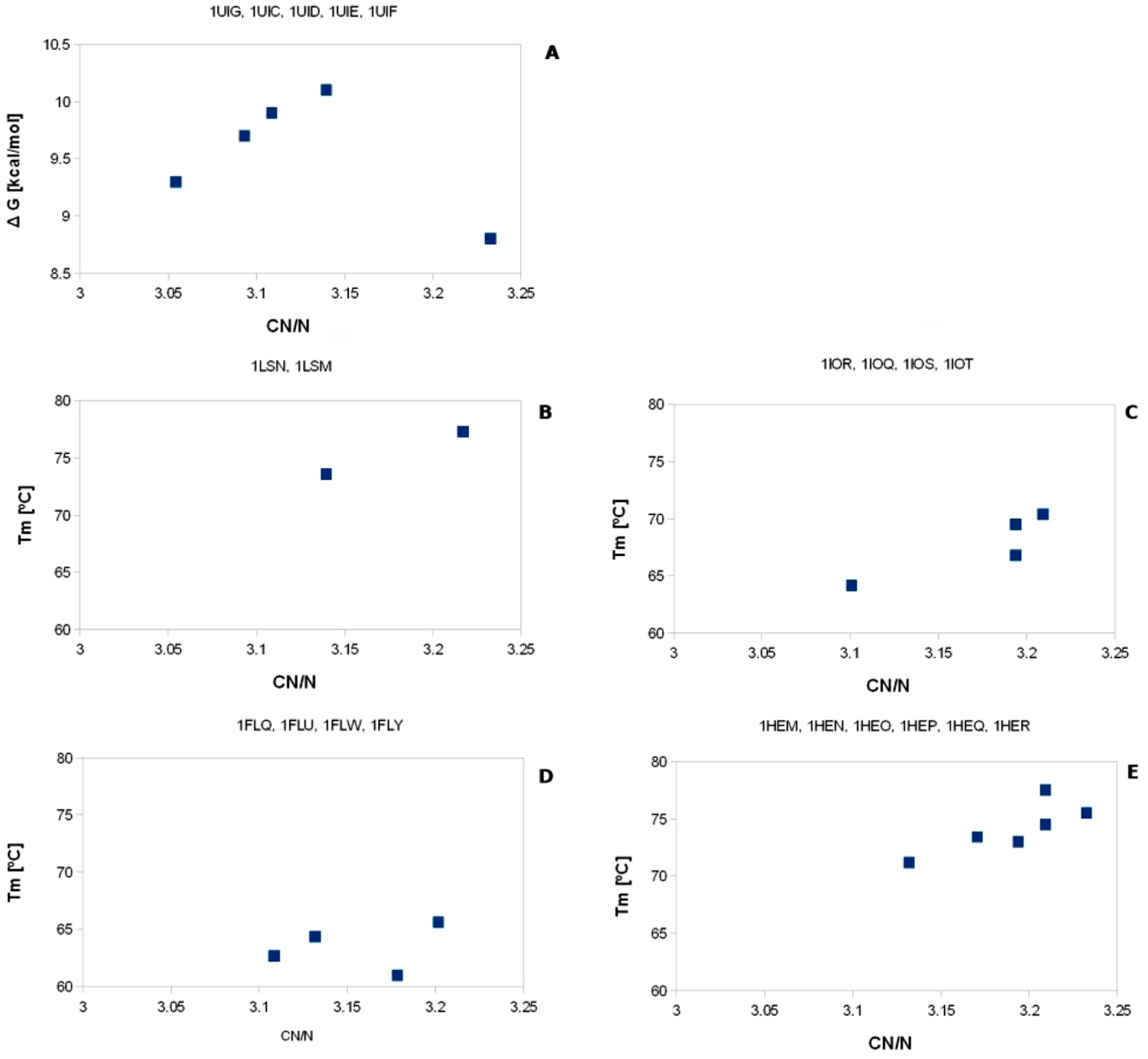


Figure 4. ΔG and Tm as a function of the CN/N for various mutants of lysozyme. (A) ΔG dependency on CN/N for 1UIG, 1UIC, 1UID, 1UIE, 1UIF; Tm dependency on CN/N for 1LSN, 1LSM (B); 1IOR, 1IOQ, 1IOS, 1IOT (C); 1FLQ, 1FLU, 1FLW, 1FLY (D) and 1HEM, 1HEN, 1HEO, 1HEP, 1HEQ, and 1HER.

and the average interaction energy per residue proves the usefulness of our contact-definition concept and its wise application.

The results in Table 3 also provide a comparison between the energy defined contacts and the minimal distance definition of contacts between residues. This geometry definition was chosen out of three commonly used geometry based criteria—minimum distance, C_α distance, and center of the residue (see jp303088n_si_002.pdf, Supporting Information). It is interesting that we did not get a consistent picture for both of the applied criteria, and we can only speculate if the reason for this inconsistency lies rather in threshold values for geometry defined contact or in the accuracy of the utilized force field. Interestingly, the minimal distance criterion provided much more contacts per residuum so the level of statistical significance was tested. As follows from such analysis, the thermophiles have on average more than 4% more energy based contacts and about 2% more geometry contacts. The hypothesis "thermophiles have at least 3% more contacts than mesophiles" can be accepted at 67% confidence level for OPLS contacts and at 66% confidence level for CHARMM contacts.

To test the applicability of the energy defined contacts on a different type of protein stability problem, we decided to perform our calculations for a set of lysozyme mutants. Lysozyme is one of the easiest crystallizing proteins, and the PDB contains more than 400 structures of the protein and its engineered variants with corresponding references. Some of these papers addressed the problem of thermal stability upon mutation of certain amino acids so the complete set of data were available—the PDB structure of the wild type molecule, crystal structure of mutants, and thermal data—either the ΔG of the denaturation process or Tm value usually determined by

dx.doi.org/10.1021/jp303088n | J. Phys. Chem. B 2012, 116, 12651−12660

the calorimetric measurements. We have found five heterogeneous sets for which we performed our calculations of stability based on energy defined contacts. Again, no other processing of the data except the optimization of hydrogens was performed. The results of these calculations are shown in Table 4 and corresponding Figure 4.

It is immediately visible that the expected trend of increasing stability upon a rising number of amino acid contacts is presented for all studied sets. The biggest discrepancy was found for one particular case in the first set for which ΔG of the denaturation process in guanidinium hydrochloride was measured (panel A in Figure 4). All the other graphs (panels B−E) for the sets where correlation between Tm and CN/N was plotted look surprisingly better. There is one important fact which has to be stressed here. The range of Tm was not large, we typically calculated CN/N for cases where Tm of mutants lies in the interval 0−10 °C, so the method is quite sensitive. The improvement of the method sensitivity might provide quite a powerful tool for mutagenesis studies in globular proteins.

CONCLUSION

We have proposed a method of a protein native geometry conversion into a map of contacts based on statistically significant interaction energies. The process is based on fragmentation and calculation using an empirical (OPLS or CHARMM) force field and relies on its pairwise additivity. Our approach unifies the treatment of different types of interactions, avoiding the problems arising from the different contributions to the overall stability and the different effect of the environment. On the one hand, we can better understand the values in interaction-energy matrices. On the other hand, we can say that interactions stronger than appropriate contact definition are the productive or the most stabilizing ones. The matrices of productive contacts can be used whenever the energy content of the contact instead of the geometric proximity is required. Utilization of our method with NMR analysis of proteins is also possible. We have shown that our contact definition is sufficiently robust and different from a geometry-based definition to replace them in such applications.

In this work, we have applied our contact definition to a naïve model of globular protein thermostability and shown that the number of energy-defined residue−residue contacts per residue is higher in thermophilic proteins than in their mesophilic counterparts. It is important to mention that we did not optimize our method toward the protein thermostability issue. There is still a space for improvement and applicability of this method for such a purpose. We can optimize contact definitions for some types of residue−residue interactions, since no significant threshold was localized. There is always a force field issue and its accuracy, and we currently work with other force fields (AMBER) to find if our contact definitions are robust and general.

The stability issue in proteins has quite a large application potential. It is possible that rational design of more stable variants of existing proteins could help in preparation of biotechnologically relevant materials. The method would also be beneficial for wet lab biochemists or structural biologist engineering proteins toward certain functions of structural aspects. This is very well documented on the test case of lysozyme and its mutants in this paper. A straightforward utilization of this application is currently in a phase of preparation of a publicly available web tool.

Probably the most perspective utilization of our method is in prediction of 3D structures of globular proteins. One of the possible scenarios would be an algorithm which discriminates a great number of examined folds in tertiary structure prediction processes coupled with secondary structure prediction algorithms. However, one of the most outstanding challenges in this field is the identification of a structure nearest to the native from a large ensemble of structures with very similar folds. There are already structure ensembles suitable for testing our methods (as for example Decoys 'R' Us), and it is our immediate aim to show that our method is able to distinguish between the native structure and the decoy.

ASSOCIATED CONTENT

Supporting Information

The full matrix of histograms of the contributions of the interaction energies (HCIE) for all 20 natural amino acid pairwise contacts and tables for CHARMM force field and geometry criteria defined contacs. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jiri.vondrasek@uochb.cas.cz.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by Grant No. P208/10/0725 from the Czech Science Foundation and by Grant No. LH11020 from the Ministry of Education, Youth and Sports (MSMT) of the Czech Republic. It was also a part of subvention for development of research organization RVO: 61388963.

REFERENCES

(1) Banerjee, R.; Sen, M.; Bhattacharya, D.; Saha, P. *J. Mol. Biol.* **2003**, *333*, 211−226.
(2) Bromberg, S.; Dill, K. A. *Protein Sci.* **1994**, *3*, 997−1009.
(3) Liang, J.; Dill, K. A. *Biophys. J.* **2001**, *81*, 751−766.
(4) Rodionov, M. A.; Galaktionov, S. G. *Mol. Biol.* **1992**, *26*, 773−776.
(5) Rodionov, M. A.; Galaktionov, S. G. *Mol. Biol.* **1992**, *2*, 777−783.
(6) Rodionov, M. A.; Gurevich, A. V.; Galaktionov, S. G. *Mol. Biol.* **1993**, *27*, 220−224.
(7) Gromiha, M. M.; Selvaraj, S. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235−277.
(8) Esque, J.; Oguey, C.; de Brevern, A. G. *J. Chem. Inf. Model.* **2011**, *51*, 493−507.
(9) Duarte, J. M.; Sathyapriya, R.; Stehr, H.; Filippis, I.; Lappe, M. *BMC Bioinf.* **2010**, *11*, 1−10.
(10) Faure, G.; Bornot, A.; de Brevern, A. G. *Biochimie* **2008**, *90*, 626−639.
(11) Kanna, N.; Vishveshwara, S. *J. Mol. Biol.* **1999**, *292*, 441−464.
(12) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623−644.
(13) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.
(14) Vendruscolo, M.; Subramanian, B.; Kanter, I.; Domany, E.; Lebowitz, J. *Phys. Rev. E* **1999**, *59*, 977−984.
(15) Vendruscolo, M.; Najmanovich, R.; Domany, E. *Phys. Rev. Lett.* **1999**, *82*, 656−659.
(16) Vendruscolo, M.; Domany, E. *Vitam. Horm.* **2000**, *58*, 171−212.
(17) Bendova-Biedermannova, L.; Hobza, P.; Vondrasek, J. *Proteins* **2008**, *72*, 402−413.
(18) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982−992.

(19) Berka, K.; Laskowski, R. A.; Hobza, P.; Vondrasek, J. *J. Chem. Theory Comput.* **2010**, *6*, 2191−2203.
(20) Kolar, M.; Berka, K.; Jurecka, P.; Hobza, P. *ChemPhysChem* **2010**, *11*, 2399−2408.
(21) Fackovec, B.; Vondrasek, J. In *Systems and Computational Biology - Molecular and Cellular Experimental Systems*; Yang, N.-S., Ed.; InTech publisher: Rijeka, Croatia, 2011; pp 69−82.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Jorgensen, W. L.; TiradoRives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(24) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(25) Jorgensen, W. L.; Tirado-Rives, J. *Abstr. Pap. Am. Chem. Soc.* **1998**, *216*, U696.
(26) Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E. *J. Chem. Theory Comput.* **2010**, *6*, 459−466.
(27) Kannan, N.; Vishveshwara, S. *Protein Eng.* **2000**, *13*, 753−761.
(28) Malcolm, B. A.; Wilson, K. P.; Matthews, B. W.; Kirsch, J. F.; Wilson, A. C. *Nature* **1990**, *345*, 86−89.
(29) Masumoto, K.; Ueda, T.; Motoshima, H.; Imoto, T. *Protein Eng.* **2000**, *13*, 691−695.
(30) Ohmura, T.; Ueda, T.; Motoshima, H.; Tamura, T.; Imoto, T. *J. Biochem.* **1997**, *122*, 512−517.
(31) Ohmura, T.; Ueda, T.; Ootsuka, K.; Saito, M.; Imoto, T. *Protein Sci.* **2001**, *10*, 313−320.
(32) Shih, P.; Holland, D. R.; Kirsch, J. F. *Protein Sci.* **1995**, *4*, 2050−2062.
(33) Kumar, S.; Nussinov, R. *Cell. Mol. Life Sci.* **2001**, *58*, 1216−1233.
(34) Kumar, S.; Tsai, C. J.; Nussinov, R. *Biochemistry* **2001**, *40*, 14152−14165.
(35) Kannan, N.; Vishveshwara, S. *Protein Eng.* **2000**, *13*, 753−761.