APPENDIX 1

4

Decomposition of Intramolecular Interactions Between Amino-Acids in Globular Proteins - A Consequence for Structural Classes of Proteins and Methods of Their Classification

Boris Fackovec[1,2] and Jiri Vondrasek[1]
[1]*Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague,*
[2]*Faculty of Natural Sciences, Charles University in Prague, Prague,*
Czech Republic

1. Introduction

An amino-acid in proteins shows two different, yet mutually dependent faces connected through the polymer character of a protein in the final product. They are the amino-acid side-chain and its corresponding backbone part. On the level of the side-chains, we often refer to specific structural arrangements such as hydrophobic cluster motifs, salt-bridge motifs or hydrogen-bond motifs characterizing various parts of a protein and usually assigned to a certain function. The backbone on the other hand offers limited, yet general structural motifs – α, β and random coil patterns. All of these mentioned amino-acid features contribute to the synergy demonstrated observably by protein stability and protein function.

Thermal stability is one of the most important features of the structure of a fully folded protein. It is defined as the difference in the Gibbs free energy between its native and denaturated states and as such is a function of temperature and implicitly a function of protein composition and the effect of the environment. Nevertheless, it is necessary to say that for this function we do not know yet the precise and general form which could be applicable for a large set of proteins. There have been many attempts to propose an intuitive, yet productive decomposition of Gibbs free stabilization energy (GFSE) into simple terms. One of the scenarios utilized for such purposes is that the total free energy is the sum of the free energies of various atomic groups and the hydrophobic effect. However, as the free energy is not additive and the fractionation of free energy to independent terms is difficult, this attempt has been quite unsuccessful.

The utilization of molecular modeling methodology and tools has opened a more systematic and perhaps more promising approach – the evaluation of the enthalpy term in the equation for Gibbs free energy with reasonable accuracy (Lazaridis, Archontis, & Karplus, 1995). The remaining entropy term could be obtained by fitting the corresponding analytical form to the experimental data. There are basically three different enthalpy contributions that we can separate. The first comes from the intramolecular interactions between the atoms of proteins, producing the largest stabilizing enthalpy contribution. The second comes from the interactions between the molecules of a solvent, and finally the third contribution is the result of the interactions between the atoms of the solute (protein) and the solvent.

It is commonly believed that the dominant force of protein folding and therefore the main stabilizing force of the native structure is the hydrophobic effect (Dill, 1990). However, it has been insightfully pointed out (Makhatadze et al. 1995) that a water environment destabilizes folded protein structures and the decomposition of enthalpy shows that the solvation models introduce significant errors. In these studies, it has been assumed that the denatured state of a protein can be identified with the fully unfolded state (Makhatadze et al. 1989), where residues do not interact with each other. Even in light of this hypothesis, the intramolecular interactions between amino-acids in a protein are expected to contribute significantly to its overall stability. However, the hypothesis has never been proved and the importance of the intramolecular interactions would be much higher if the unfolding were considered as "core melting" rather than "oil-droplet dissolution". Regardless of the denatured form, the intramolecular organization of a protein is the result of a subtle balance between the rigidity/flexibility of the protein backbone and the noncovalent interactions between protein's side-chains. This result in conformational unique and stable protein structures as well as the ratio between the importance of the backbone/side-chain contributions can vary for different proteins.

The main problem of the enthalpy (or the potential energy) approach is that we are unable to evaluate the enthalpy-entropy compensation; therefore, the theoretically determined enthalpy contribution should be adjusted in some other way. A realistic method is to correlate the calculated values with the experimental data obtained by microcalorimetry, where both the enthalpy and the entropy terms can be determined. On the level of particular amino-acids, we face the problem of their "denatured-state" definition for the reasonable decomposition of the free energy on individual amino-acids.

The dissection of the enthalpy contribution which the intra-molecular noncovalent interaction energy (part of the potential energy) is a component of seems to be a reasonable approach for the study of the role of the composing amino-acids in protein stabilization. We can decompose this energy into individual pairwise amino-acid contributions and determine their importance for protein stability. The evaluation of the interaction energy (of noncovalent origin) between biomolecules or between their parts is a traditional field of the symbiosis between experiment and theory, and the methodology is well described and highly developed (Müller-Dethlefs & P Hobza, 2000). The crucial condition for the success of the theoretical methodology is the accuracy of the methods utilized. Recently, it has been quite common to evaluate the potential energy of a protein at the suitable *ab initio* methodology level, but we are still severely limited by the size of the protein. Therefore, the Density Functional Theory methods (DFT) are the most utilized for such purposes (Riley, 2010). Unfortunately, the DFT methods fail to describe the noncovalent interactions reasonably mostly because of the missing electron correlation term. Even the new functionals recently introduced (Kolář, 2010) have failed to describe properly the noncovalent potential curve mostly in the repulsion and asymptotic regions. Such inaccuracies can be tolerated at the energy minima, but only a limited number of the interactions between amino-acids in proteins meet such a requirement. Therefore, only high-level *ab initio* methods can be utilized – at least for benchmark studies. As was shown on a set of representative interactions between amino-acid side-chains in proteins in 2009, empirical force fields (namely OPLS and AMBER) are suitable for the description of their interaction (Berka, 2009). Kolar (Kolář, 2010) tested the performance of the energy calculations using MM on a representative set, S22, and found quite satisfactory agreement between the empirical force fields and high-level *ab initio* methods. It was later shown that we can use the empirical force field with satisfactory accuracy also for the description of the intramolecular interaction-energy distributions for pairs of amino-acid side-chains (Berka, 2010). Still, one has to be aware

of the limitations of the force-field methods, namely for subtle cases of the interactions present in proteins. On the other hand, the utilization of empirical methods decreases the computational cost and provides an opportunity to investigate the trends presented in biomolecules if the highest accuracy is not the major issue.

The evaluation of the interaction energy between amino-acid residues resulted in the interaction energy matrix (IEM) concept being introduced in 2008 (Bendová-Biedermannová, 2008). The IEM approach was used to identify the key residues for protein stability in a model system – rubredoxin. The matrix carries information about the energy and the role of a residue in the protein structure, namely its interaction energy strength, which is more than the simple distance matrix concept. It also shows how much a certain residue is a hub within the context of the other interacting amino-acids. The IEM approach might also open new horizons for the investigations of proteins. The concept could be incorporated into the methods of protein-structure superpositions (similar to the DALI approach)(Holm & Sander, 1997) and can shed light on other protein-related issues – for example protein stability, folding kinetics, foldability and design.

The work presented in this study is based on the calculations of the amino-acid – amino-acid interaction energies (IEs) between all of the residues in approximately 1400 proteins to justify the roles of different amino-acids, their backbones and side-chains and their physical-chemical character for structural or stabilization preferences. We especially focused on the problem of how the interaction energy distributions are related to the secondary-structure content defined by the CATH (Orengo et al., 1997) and SCOP(Murzin, 1995) criteria.

2. Amino-acids in proteins and their distribution

2.1.1 Representative structure-set selection

All of the protein structures utilized in this study were obtained from the PDB database (download Jan 31, 2011). We selected only protein molecules with one chain, no ligands, resolved by the X-ray crystallography method at a minimum resolution of 2.0 A. We also omitted structures with a 70% sequence identity and higher. The database filter yielded 1531 structures. This number was slightly reduced by inconveniences with file processing to 1358. The characteristics of the set are illustrated in Figure 1 (size histogram, resolution histogram).

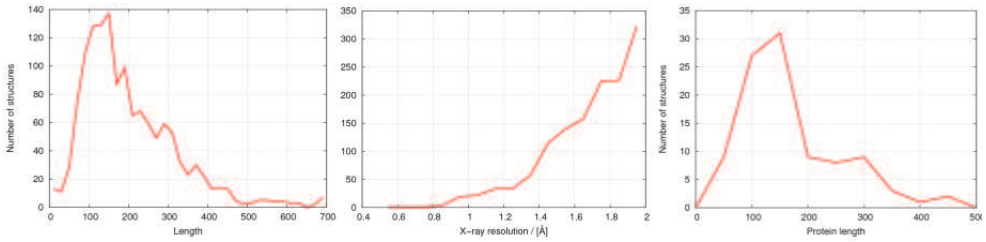

Fig. 1. a) Number of structures against protein length; binned by 20 AA; b) number of structures with a particular X-ray resolution; binned by 0.1 A; c) histogram of the sizes of the structures selected for secondary-structure studies.

Incomplete amino-acid side-chains (missing heavy atoms, disordered) were replaced by glycine in the cases where backbone atoms were available. Amino-acids with missing

backbone atoms would have discredited the whole set and were therefore omitted. The missing hydrogen atoms were added by the Xleap program from the AMBER (Case et al. 2010) simulation package for pH 7 and the parameters were assigned according to the OPLS FF (Jorgensen & Rives 1988). The ambiguity of protonation, mainly in the case of histidine, is discussed later. The structures were optimized using the GROMACS (Hess et al. 2008) molecular simulation package with the steepest descent algorithm being employed. The hydrogen atoms were optimized first and then the full optimization of the whole protein in the gas phase was performed.
To address the question of the residue selectivity for secondary structure motifs, the structures were classified according to the CATH and SCOP categories and four representative sets were selected. To prevent the interference of the size and secondary structure effect, we assured that the structure sets possess the same size distribution.
Hence, the structures pertaining to particular secondary-structure sets were binned according to their chain length (bin size 50, see Figure 2) and were randomly removed from the bins until the number of structures in the corresponding bins was the same for all the sets. This procedure resulted in four sets, each containing 99 structures.

2.1.2 The fragmentation of proteins

To differentiate between the particular types of interactions which every amino-acid can maintain, we assigned every atom of a residue to one of four attributes according to their occurrence in the backbone or to their occurrence in certain types of amino-acid side-chains. The attributes were as follows – BB – backbone atoms, CH – side-chain atoms of charged residues (asp, glu, lys, arg, his), PO – side-chain atoms of polar residues (asn, gln, thr, ser) and NP – side-chain atoms of nonpolar and aromatic residues (gly, ala, leu, ile, val, pro, cys, met, phe, tyr, trp). Such classification provides the lowest number of groups necessary to discern between interactions characterized by different distance dependencies and orders of magnitude (different physical characters). On the other hand, breaking residues into more parts is restrained by the resulting charges of the fragments which would introduce significant but artificial electrostatic energies. The OPLS force field guarantees that the backbone (which includes C_α) and side-chain fragments are neutral. The physical character of the interaction energies of the aromatic residues is close to those of nonpolar residues. Hence, taking into account digestibility of presented data, we decided not to increase the number of attributes.

2.1.3 The Interaction Energy Matrix (IEM) calculation

After all of the structural optimizations, the pairwise interaction energies for all of the residues at the OPLS level were calculated excluding those between backbones of adjacent amino-acid in primary structure which were set to zero. The interactions were calculated separately for the backbones and side-chains as the sum of the interatomic Lennard-Jones and Coulombic contributions in the gas phase (ε_r=1) using an in-house developed Python program utilizing the standard libraries. The classification of the amino-acid atoms in four groups resulted in ten types of mutual interactions – BB-BB, BB-CH, BB-PO, BB-NP, CH-CH, CH-PO, CH-NP, PO-PO, PO-NP, NP-NP – reflecting the attributes of the atoms involved. For example, CH-CH represents salt bridges and all of the interactions between the side-chains of charged residues regardless of their relative distance and charge sign.

Each type of interaction for one protein was represented by one interaction energy matrix, namely a NxN (where N denotes the number of residues) matrix containing the interaction energy between the atoms of residues i and j with particular attributes assigned. It is guaranteed that no interaction energy is counted twice, so the sum of all of the matrices provides the interaction energy between the corresponding residues.
In order to compare the residual energy content, we have introduced a residue interaction energy (RIE) characteristic for each residue. The RIE of a certain type is defined as the sum of all of the interactions the residue can maintain – the sum of all the numbers in a particular row (or column) in the IEM of that type. At the end, we have ten (NxN dimension, where N is the number of amino-acids) IEMs of different types in one protein. Most of the IEs are of course almost zero; some are set as zero by definition.

2.1.4 Representation of data – cumulative distribution functions and histograms

There are two main data representation schemes in this work. Those are as follows:
The distributions of RIEs of a certain type in one protein. For one specific type and one specific protein set (for example CH-CH in SCOP β), the following procedure was performed to acquire an average distribution representing the whole set. The non-zero RIEs calculated from appropriate IEM were sorted independently for each protein and the distributions were obtained as a plot of the RIE against the residue rank in the sorted list normalized to one. To enable the averaging of the distributions, we represented each one by 1001 equally distant (on the rank coordinate) points between 0 and 1 (instead of for example N in the case of RIE BB). The RIE for each point was obtained by linear interpolation using the nearest two points of the calculated distribution. The averaged distribution was obtained by averaging the RIEs of the corresponding points of the curves of all of the proteins pertaining to the set. The inverse of the averaged distribution is a quite smooth cumulative distribution function representing the average for the set.
The distributions of the RIEs of a certain type for a particular amino-acid were sampled from all of the 1358 proteins. The RIEs of a particular type and AA were sampled from all the proteins and binned to yield quite smooth histograms.

2.2 Secondary-structure dependence

The RIE distribution of a particular type in a protein describes the distribution of the energetic importance of the residues. An average distribution also characterizes the particular type of interaction in the ensemble – the fraction of the key residues, their importance, and the fraction of the residues with repulsive interactions. The magnitude interval of a distribution is a very important parameter. It contains information about the interaction strength in the native states of the proteins. Unfortunately, this information does not denote the contribution of particular interactions to stability as it lacks information on the denatured state.
The shape of the distribution determines the pressure exerted on a residue and might help estimate the actual contribution of the corresponding interactions to protein stability. It is not surprising that the BB RIEs correlate with the secondary structures as the classifications indirectly use the BB RIEs. However, the differences are smaller than one might expect. It is also clear that none of the interactions other than BB is affected by the secondary-structure content.
From Figure 2, it can be concluded that the difference between the CATH and SCOP classifications is more significant mainly in the case of α proteins. Figures 3 and 4 show all of the types of distributions for a nonpolar (ALA, Figure 3) and a polar (THR, Figure 4)

amino-acid. It is obvious that the BB RIE cumulative distributions are the only distributions to have their shape affected by the secondary-structure content and the particular AA RIE distributions show more than one peak. The distinctive peaks might be assigned to special structural features and their identification remains a task for future studies.

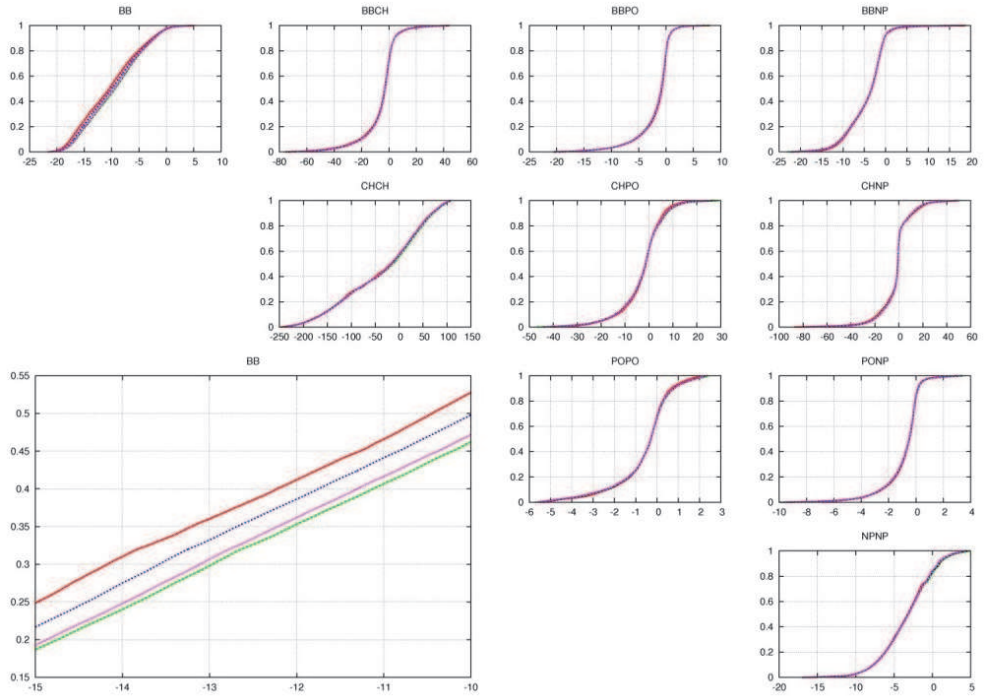

Fig. 2. The average RIE distributions of all ten types: a comparison of the secondary-structure classes. The colors of the lines correspond to the following structure sets: red – CATH α, blue – SCOP α, green – CATH β, magenta – SCOP β. The detail of the BB distribution in the bottom left corner is a zoom of the BB RIE distributions.

The fact that the CYS average NPNP RIE distribution is the only exception to the rule, because it has two peaks, can be explained by a different strength of the noncovalent interactions of the cystein SH group and cystine SS bridge.

The BB RIEs of particular AAs sampled through all of the structure sets are shown in Figure 5. There are remarkable differences between the shapes of the distributions corresponding to the α and β proteins as well as between the shapes of the distributions for particular AAs. Generally, the BB RIE distributions of the beta-structured proteins are shifted to a less attractive (less negative) noncovalent region.

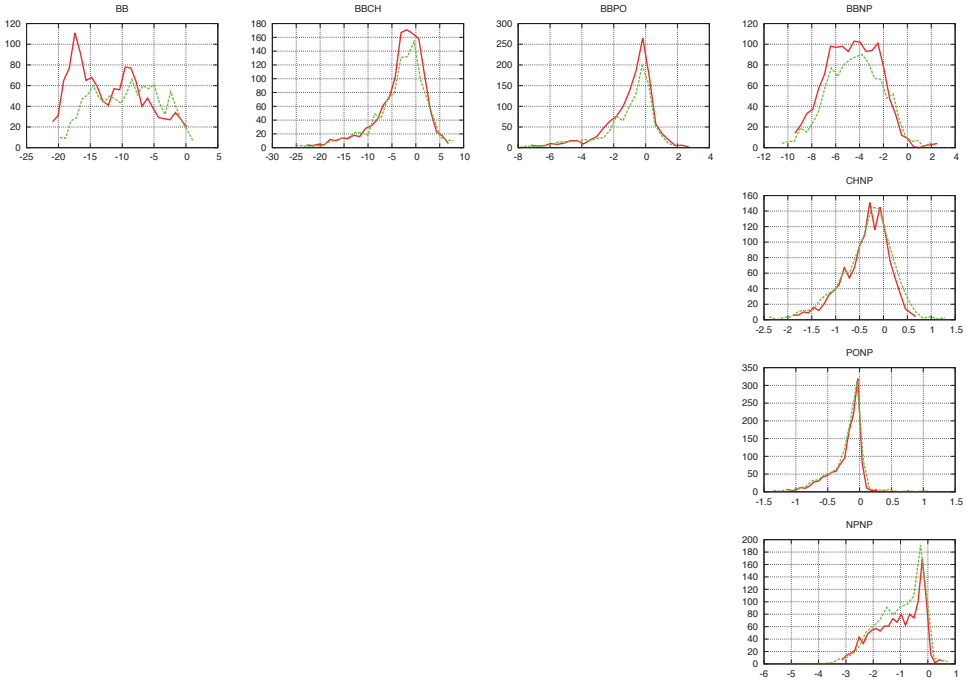

Fig. 3. All of the types of the RIE distributions of ALA. The red line corresponds to the CATH α set, the green line to the CATH β.

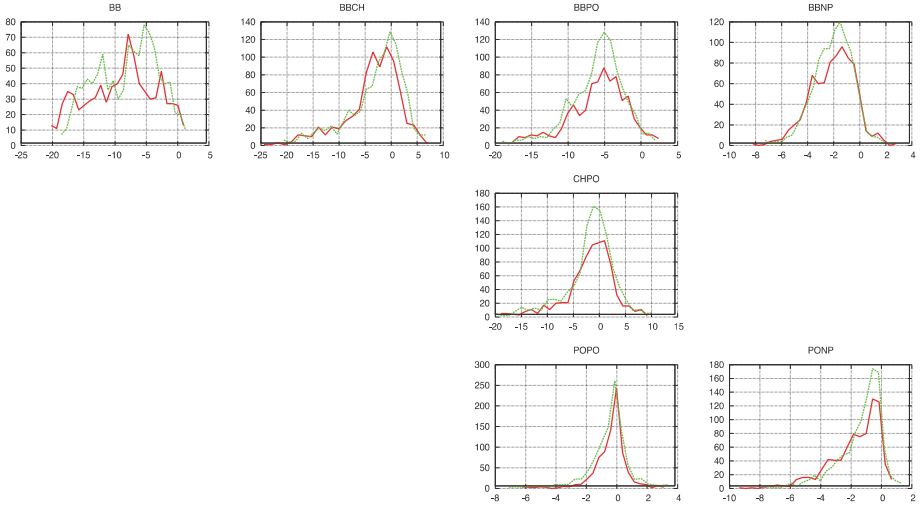

Fig. 4. All of the types of the RIE distributions of THR. The red line corresponds to the CATH α set, the green line to the CATH β.

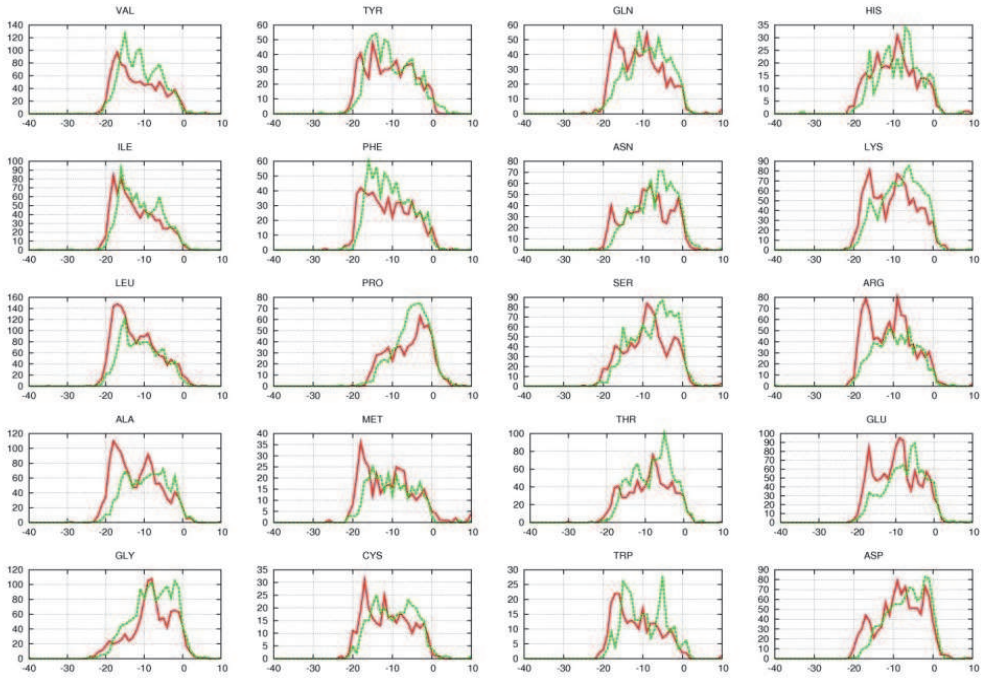

Fig. 5. The average BB RIE distributions for each AA. Sampled through proteins from the CATH α and CATH β sets. The red line corresponds to the CATH α set, the green line to the CATH β.

2.3 Size dependence

The proteins were selected based on their chain lengths up to fourteen groups regardless of their secondary-structure content. Their characteristics (chain-length range, average chain length, amino-acid type composition, number of proteins, number of residues of particular types, average surface area) are reviewed in Table 1.

min length	max length	average length	nonpolar resis	polar resis	charged resis	no. of structures	no. of residues
40	60	52.5	55.1	18.3	26.6	28	521
60	80	68.9	52.1	20.7	27.2	71	1342
80	100	90.7	53.0	18.5	28.5	105	1960
100	120	109.1	52.9	19.8	27.3	132	2492
120	140	129.1	53.9	18.8	27.3	125	2351
140	160	149.7	52.9	19.6	27.6	142	2655
160	180	169.6	53.8	19.5	26.6	85	1587
180	200	188.4	54.3	19.4	26.3	102	1909
200	220	209.9	53.1	19.9	26.9	65	1219
220	240	228.0	53.4	20.2	26.5	67	1253
240	260	249.3	54.3	18.9	26.8	57	1059
260	280	269.1	55.4	20.0	24.6	52	955
280	300	289.4	54.5	19.5	26.0	58	1061

Table 1. The characteristics of the structure sets used for the RIE-size dependence studies.

RIEs of a particular type were sampled from all of the proteins of a particular size group. The RIE averages were calculated separately for each interaction type of each size. The plots of the average RIEs against size are presented in four figures (Figures 6 to 9) in order to maintain the lucidity of the plots with lower magnitudes of average RIEs. The results reported in Figure 6 suggest that the RIE-size dependence varies significantly with the interaction type. On the one hand, the interaction of the polar residues with the backbone is almost independent of size. On the other hand, the interactions of the side-chains follow common rules, which are investigated later.

An interesting notion comes from a comparison of the magnitudes of the POPO and BBPO average RIEs. The lower RIE magnitudes in the case of POPO RIEs are probably caused by the lower probability of hydrogen-bond formation with polar side-chains in comparison with the backbone-polar side-chain because of the lower frequency of their occurrence.

A noticeable trend is the coupling of BBCH and CHPO interactions (see Figure 8). This binding may be ascribed to the same physical quality of these two types of interactions; they both represent charge–dipole interactions. The accuracy of the data can be estimated from the curve smoothness and is apparently lower in the case of charged residues. One possible reason for this trend is that the RIEs of charged residues are the products of a large compensation for the low amount of data.

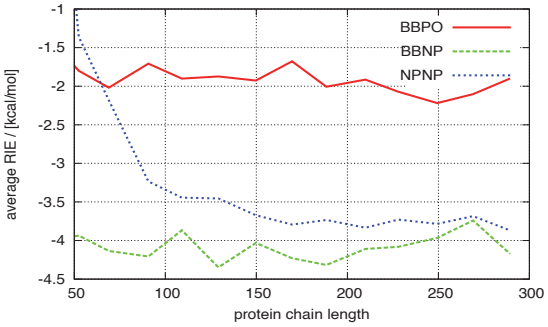

Fig. 6. The size dependence for BBPO, BBNP and NPNP interactions in the studied protein set. The NPNP differs significantly from the rest.

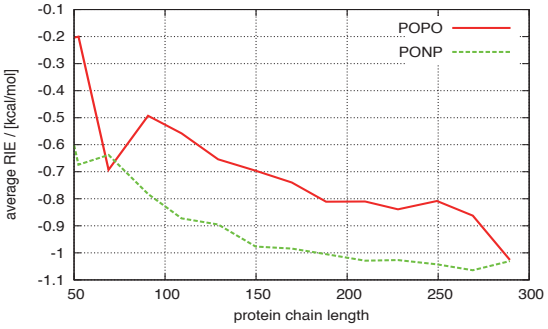

Fig. 7. The Size dependence for the POPO and PONP interactions in the studied protein set.

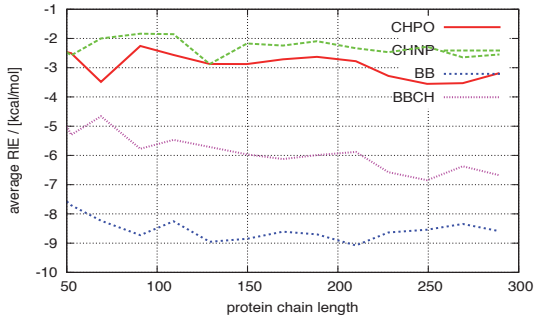

Fig. 8. The size dependence for the CHPO, CHNP, BB and BBCH interactions in the studied protein set.

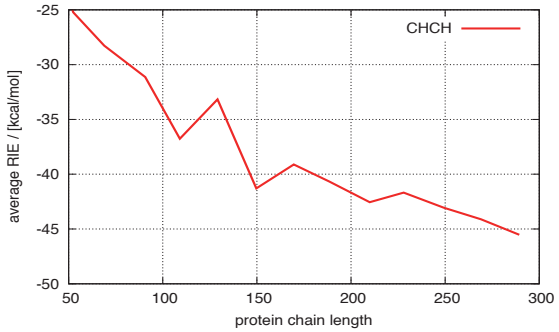

Fig. 9. The size dependence for the CHCH interactions in the studied protein set.

2.3.1 The model of size dependence for the interaction of nonpolar amino-acid side-chains

Two simple models were tested to explain the observed trends. In the first model, protein is assumed to be a sphere with nonpolar residues in its hydrophobic core and polar and charged residues forming its exterior shell. The size dependence of a NPNP RIE average is ascribed to the size dependence of the ratio between the core and surface residues. At infinite length, the NPNP RIE should reach its limit. The second model is more realistic in such a way that the core never behaves like a limitlessly increasing sphere and the volume occupied by the side-chains must reach its limit. This limits the NPNP RIE value which will not rise further with the increasing size of a protein and define a certain size of the most compact amino-acid arrangement.

2.3.1.1 The NPNP RIE Model 1

An average NP residue can be described by its characteristic length r, surface factor f_s (corresponding to its surface or interaction area $S_r = f_s r^2$), volume factor f_v (corresponding to its volume $V_r = f_v r^3$) and the NPNP RIE limit for the infinite bulk E_∞. As we are assuming that all of the nonpolar residues form the core which has a spherical shape, the core size is determined by the size of the protein and the ratio φ of nonpolar residues and all residues. A protein can

be described by its porosity ε (determining the ratio of the gap volume to the volume of the whole protein) and at least its length N. Assuming that all of these quantities except for N are constants, the volume of each protein can be expressed as $V_p = NV_r/(1-\varepsilon) = Nf_v r^3/(1-\varepsilon)$ and the core volume as $V_c = V_p\varphi = N\varphi f_v r^3/(1-\varepsilon)$. The interaction surface of the core residues can be considered as $S_i = N\varphi S_r$ and the core surface is $S_c = 4\pi r_c^2$. E can be calculated as

$$E = E_\infty \left(1 - kN^{-1/3}\right), \tag{1}$$

where

$$k = \frac{1}{f_S}\sqrt[3]{\frac{1024\pi f_v^2}{9\varphi(1-\varepsilon)^2}}.$$

The k and E_1 parameters were fitted to the calculated data using Equation (1). As can be seen in Figure 10, the fitted curve does not represent the data very well.

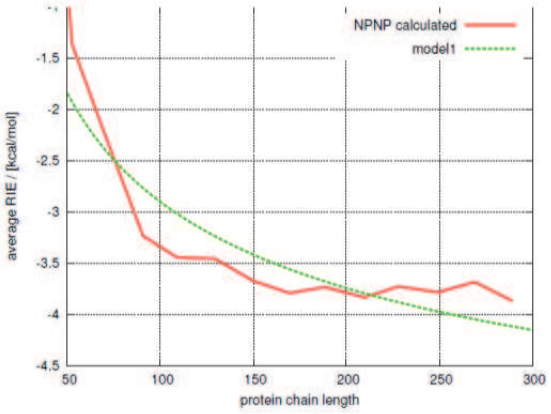

Fig. 10. The performance of Model 1

2.3.1.2 The NPNP RIE Model 2

The first model was extended by adding a new parameter, representing the domain size. The energy was represented by the following function:

$$E = \begin{cases} E = E_\infty\left(1 - kN^{-1/3}\right) & : N \le N_D \\ E_D & : N > N_D \end{cases}, \tag{2}$$

where N_D is the domain size and $E_D = E_\infty (1\text{-}kN_D^{-1/3})$ is NPNP RIE average at N_D. The parameters N_D, k and E_1 were fitted to the NPNP RIE averages. The agreement of the fitted curve with the data is satisfactory considering the simplicity of the model as one can see in Figure 11.

The coefficient k obtained by fitting the data is comparable to that obtained by a calculation using the estimated values of fv, fs, ε and the experimental value of φ. Other types of interactions seem to be unrelated to the domain size of a protein as there is no mechanism connected with size that we could follow.

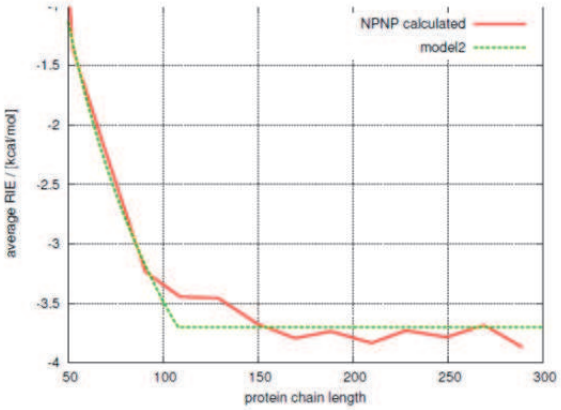

Fig. 11. The performance of Model 2

2.3.2 The reliability of the evaluated distributions

To adjust the reliability of our findings from computational point of view, we divided all of the proteins randomly into two groups. The distributions are indistinguishable, which proves that the distributions can be obtained by averaging even smaller sets of proteins. Additionally, we calculated the distributions using the OPLS force field in a Cα representation of the protein side-chains. Apparently (see Figure 12), the distributions for both FFs are the same. This not only proves that our results are robust against a FF parametrization error but also suggests that both FFs are within their limits equally good for RIE-distribution investigations.

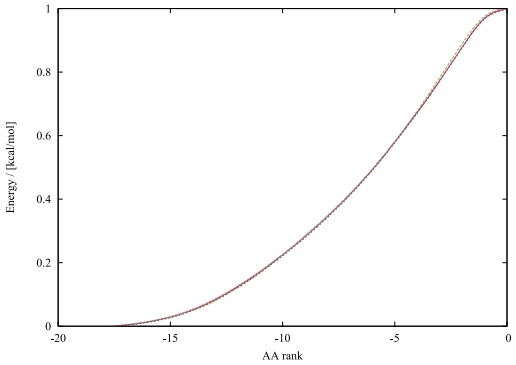

Fig. 12. A Comparison of the distributions obtained by averaging the distributions within the whole set using the OPLS Ca FF (dots) and Amber 03 C_a (full line) shows the robustness of the distributions against the FF used. The distributions obtained by averaging the distributions in two randomly chosen half-sets of structures calculated using the Amber C_a FF are indistinguishable, which proves that our set is sufficiently large.

3. Conclusion

RIE distributions in proteins, except for the BB RIE distributions, are not affected by secondary-structure content. The same applies for the distributions sampled for each amino-

acid separately. Hence, we can claim that the strength and selectivity of the SC-SC and SC-BB interaction do not correlate with the secondary-structure content.
The size dependence of the RIEs can be satisfactorily described by the second model proposed. Its three parameters can be fitted to the results obtained by FF calculations of a high number of protein structures. One of the parameters obtained by fitting to the NPNP RIE averages represents the optimum definition of the domain size in globular proteins. Although the models proposed apply for all types of NP and PO SC-SC interactions, the models fail in the description of the BB and CH interactions. Many interesting facts about the size dependence of the RIE averages were revealed. First, the BBCH and CHPO interactions seem to be bound by some as-yet unknown rule. Second, the PO interactions exhibit "strange" behavior at a protein chain length of approximately seventy residues. These findings need to be investigated more deeply.

4. Acknowledgment

This work was supported by Grant No. P208/10/0725 from the Czech Science Foundation and Grants LH11020 and LC512 from the Ministry of Education, Youth and Sports of the Czech Republic. It was also a part of research projects No. Z40550506 and No. SM6198959216.

5. References

Bendová-Biedermannová, L., Hobza, Pavel, & Vondrásek, J. (2008). Identifying stabilizing key residues in proteins using interresidue interaction energy matrix. *Proteins*, *72*(1), 402-13. doi: 10.1002/prot.21938.

Berka, K., Laskowski, R. a, Hobza, Pavel, & Vondrášek, J. (2010). Energy Matrix of Structurally Important Side-chain/Side-chain Interactions in Proteins. *Journal of Chemical Theory and Computation*, *6*(7), 2191-2203. doi: 10.1021/ct100007y.

Berka, K., Laskowski, R., Riley, K., & Hobza, Pavel. (2009). Representative amino-acid side-chain interactions in proteins. a comparison of highly accurate correlated ab initio quantum chemical and empirical potential. *Journal of Chemical*, *5*(4), 982-992. doi: 10.1021/ct800508v.

D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvai, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman (2010), *AMBER 11*, University of California, San Francisco.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, *29*(31), 7133–7155. ACS Publications.

Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* 2008;4(3):435-47

Holm, L., & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic acids research*, *25*(1), 231-4. Retrieved from

Jorgensen WL, TiradoRives J. The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *Journal of the American Chemical Society* 1988;110(6):1657-66

Kolář, M., Berka, K., Jurečka, K., Hobza, P. (2010). Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy

Calculations Using the Extended S22 Data Set. *Journal of Chemical Theory and Computation*, *6*, 2365-2376.

Kolář, M., Berka, K., Jurecka, P., & Hobza, Pavel. (2010). On the Reliability of the AMBER Force Field and its Empirical Dispersion Contribution for the Description of Noncovalent Complexes. *Chemphyschem : a European journal of chemical physics and physical chemistry*, *11*(11), 2399-408. doi: 10.1002/cphc.201000109.

Lazaridis, T., Archontis, G., & Karplus, M. (1995). Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. *Advances in Protein Chemistry*, *47*, 231-306.

Makhatadze, G. I., & Khechinashvili, N. N., with Venyaminov SYu & Griko YuV. (1989). Heat capacity and conformation of proteins in the denatured state. *Journal of molecular biology*, *205*(4), 737-50..

Makhatadze, G., & Privalov, P. (1995). Energetics of protein structure. *Advances in Protein Chemistry*, *47*, 307-425.

Müller-Dethlefs, K., & Hobza, P. (2000). Noncovalent interactions: a challenge for experiment and theory. *Chemical Reviews*, *100*(1), 143-167.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*, 536-540..

Orengo, C. a, Michie, a D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, *5*(8), 1093-108.

Riley, K. E., Pitoňák, M., Jurečka, P., & Hobza, Pavel. (2010). Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories. *Chemical Reviews*, *110*(9), 5023-63. doi: 10.1021/cr1000173.

APPENDIX 2

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

Journal:	*The Journal of Physical Chemistry*
Manuscript ID:	jp-2012-03088n
Manuscript Type:	Article
Date Submitted by the Author:	31-Mar-2012
Complete List of Authors:	Fackovec, Boris; Institute of Organic Chemistrty and Biochemistry, Czech Academy of Sciences, Bioinformatics Vondrasek, Jiri; Inst. Org. Chem. & Biochem., AS CR, Bioinformatics

Optimal definition of inter-residual contact in globular proteins based on pairwise interaction energy calculations, its robustness and applications.

Boris Fačkovec[1], Jiří Vondrášek[1,*]

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 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 using 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 emanating 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 estimation of protein stability.

Contact: jiri.vondrasek@uochb.cas.cz

INTRODUCTION

In a simplified way, the structure and stability of proteins in water environment are determined by the flexibility of their backbones, by the non-covalent interactions between the side chains of the composing amino acids, their interactions with solvent and by 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 landscapes 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 landscapes and the distributions of amino-acid pairs and their geometries observed in protein structures suggest that the intrinsic pair-wise 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 alpha carbon, beta carbon or the geometry of a side-chain centre of mass. The interaction between the two residues is then

calculated only between such points. More rigorous methods use mutual-surface-area calculations or the minimal distance of any pair of atoms and some other variants of this attempt[4-6]

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[8 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.

Simple geometry definitions of a contact are satisfactory for studies which use contact maps as alternative structure representations of proteins[11-13] . 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[14] or the development of statistical potentials[15,16] . 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[18-20] 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 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[21] .

In this study, we have used 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 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 article are the statistical property of the native state of a protein only. Such a definition enables 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 performed a 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 presented 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[20] . The X-ray structures with a resolution below 2.0 Å of the single-chain proteins with no ligands were obtained from PDB [22] (31st Jan 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 an energy optimization of the whole structure, the pairwise non-covalent interaction energies for 2N 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_a 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 (e_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. The terminal residues were not taken into consideration, as their backbones do not fit any group.

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

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

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-square fitting the lines to data points in the bins. This method of HCIE curve smoothing is not biased towards Gaussians, which ensures that the observed nature of HCIE is genuine.

Distance matrices were calculated for the four most commonly used definitions of residue-pair 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. The similarity of the two contact matrices was defined as

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

where A and B are sets of contacts in corresponding contact matrices, $\cap$ 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 twenty pairs of known homologous proteins from thermophilic and mesophilic organisms were downloaded from the pdb database (see Table 3). 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 with an 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.

RESULTS AND DISCUSSIONS

The substantial difference between the interaction energy (IE) and previously defined residue interaction energy (RIE)[20] 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.

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

$$\left.\begin{array}{l} \lim_{IE \to -\infty} HCIE = 0 \\ \lim_{IE \to +\infty} HCIE = 0 \\ HCIE(0) = 0 \\ HCIE < 0 \square IE > 0 \\ IE \to 0 \quad IE^{x}, \quad x \square \{-1, 0\} \end{array}\right\} \qquad (4)$$

Integral $\int_{-\infty}^{X} HCIE\, d\, IE$ 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 $IE \to 0$. 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

$$\int_{-\infty}^{X} HCIE \, \mathrm{d}\, IE = \int_{-\infty}^{+\infty} HCIE \, \mathrm{d}\, IE \qquad (5)$$

seems to be a natural energetic definition of a residue-residue contact, because the X defines the point where the weaker attractive interactions are compensated by all of the repulsive ones. However, as 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 the 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 $IE \to 0$) 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 Equation 4).

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

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 stationary points – the intersection of two Gaussians corresponding to different interaction patterns.

Figure 1

Groups of amino acids

The major difficulty of the pairwise interaction energy concept results from the huge compensation of the interactions of electrostatic origin. 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-acid fragments based on their multi-polar 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 positive) value. We also 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 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 interactions Gaussian 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 checked by comparing 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 supplementary material. 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 four types of fragments and therefore defining ten 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 a 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. 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.

Figure 2

Productive contact definition

Identification of the stationary point 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.

Figure 3

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 dist=4 corresponds to helices with IE ~ -4 kcal/mol, dist >7 for beta-sheet interactions with IE ~ -3.2 kcal/mol, dist=3 and dist=2 for interactions in loops with IE ~ -2.5 kcal/mol and IE ~ -1 kcal/mol, respectively.

As follows from 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.

Table 1.

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, 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.

Table 2

Application of contact definition to thermal stability prediction

To show a practical utilization of the suggested energy contact definition, we decided to test a correlation of protein thermal stability with the average number of contacts in globular proteins from thermophillic organisms and from their mesophilic counterparts. It is plausible to hypothesize that the highly stable protein homologs 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 quick qualitative orientation in protein stability. The results for a set of twenty highly stable proteins and their less stable homologs are in Table 3. We must emphasize that we just took our contact definition and applied it to a set of proteins taken from the work of[27] . In most of the cases, the number of contacts per residue rises as the stability of a protein increases. It is however impossible to correlate directly the melting temperature and number of contacts per residue

mainly because of the fact that a different stabilization mechanism is characteristic for a certain protein fold. Another limiting factor is that only protein homologs with the same fold and topology can be compared. On the other hand, the presented correlation between protein thermal stability, the number of contacts per residue and average interaction energy per residue proves the usefulness of our contact-definition concept and its wise application.

Table 3

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 understand the values in interaction-energy matrices. On the other hand, we hypothesize that these interactions 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. We have shown that our contact definition is sufficiently robust and different from a geometry-based definition to replace them in such applications. We have also 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 increased in thermostable homologs.

ACKNOWLEDGEMENT

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

Supporting Information Available: The full matrix of histograms of the contributions of the interaction energies (HCIE) for all 20 natural amino acids pair-wise contacts. This material is available free of charge via the Internet at http://pubs.acs.org.

Reference List

1. Banerjee, R.; Sen, M.; Bhattacharya, D.; Saha, P. The Jigsaw Puzzle Model: Search for Conformational Specificity in Protein Interiors? *J. Mol. Biol.* **2003,** *333*.

2. Bromberg, S.; Dill, K. A. Side-Chain Entropy and Packing in Proteins. *Protein Sci.* **1994,** *3* (7), 997-1009.

3. Liang, J.; Dill, K. A. Are proteins well-packed? *Biophys. J.* **2001,** *81* (2), 751-766.

4. Rodionov, M. A.; Galaktionov, S. G. Analysis of the 3-Dimensional Structure of Proteins in Terms of Residue Residue Contact Matrices .1. the Contact Criterion. *Mol. Biol.* **1992,** *26* (5), 773-776.

5. Rodionov, M. A.; Galaktionov, S. G. Analysis of the 3-Dimensional Structure of Proteins in Terms of Residue Residue Contact Matrices .2. Coordination Numbers. *Mol. Biol.* **1992,** *26* (5), 777-783.

6. Rodionov, M. A.; Gurevich, A. V.; Galaktionov, S. G. Analysis of the 3-Dimensional Structure of Proteins in Terms of Residue-Residue Contact Matrices .3. Residue Affinity. *Mol. Biol.* **1993,** *27* (2), 220-224.

7. Gromiha, M. M.; Selvaraj, S. Inter-residue interactions in protein folding and stability. *Progress in Biophysics & Molecular Biology* **2004,** *86* (2), 235-277.

8. Esque, J.; Oguey, C.; de Brevern, A. G. Comparative Analysis of Threshold and Tessellation Methods for Determining Protein Contacts. *Journal of Chemical Information and Modeling* **2011,** *51* (2), 493-507.

9. Duarte, J. M.; Sathyapriya, R.; Stehr, H.; Filippis, I.; Lappe, M. Optimal contact definition for reconstruction of Contact Maps. *BMC Bioinformatics* **2010,** *11*.

10. Faure, G.; Bornot, A.; de Brevern, A. G. Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie* **2008,** *90* (4), 626-639.

11. Vendruscolo, M.; Subramanian, B.; Kanter, I.; Domany, E.; Lebowitz, J. Statistical properties of contact maps. *Physical Review e* **1999,** *59* (1), 977-984.

12. Vendruscolo, M.; Najmanovich, R.; Domany, E. Protein folding in contact map space. *Phys. Rev. Lett.* **1999,** *82* (3), 656-659.

13. Vendruscolo, M.; Domany, E. Protein folding using contact maps. *Vitamins and Hormones - Advances in Research and Applications, Vol 58* **2000,** *58*, 171-212.

14. Kanna, N.; Vishveshwara, S. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **1999,** *292* (2), 441-464.

15. Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996,** *256*.

16. Miyazawa, S.; Jernigan, R. L. Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures - Quasi-Chemical Approximation. *Macromolecules* **1985,** *18* (3), 534-552.

17. Bendova-Biedermannova, L.; Hobza, P.; Vondrasek, J. Identifying stabilizing key residues in proteins using interresidue interaction energy matrix. *Proteins-Structure Function and Bioinformatics* **2008,** *72* (1), 402-413.

18. Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures. *Journal of Chemical Theory and Computation* **2009,** *5* (4), 982-992.

19. Berka, K.; Laskowski, R. A.; Hobza, P.; Vondrasek, J. Energy Matrix of Structurally Important Side-Chain/Side-Chain Interactions in Proteins. *Journal of Chemical Theory and Computation* **2010,** *6* (7), 2191-2203.

20. Fackovec, B.; Vondrasek, J. General Trends of Intramolecular Interactions between Amino Acid Side-Chains in Globular Proteins. *Sciences-New York*.

21. Kolar, M.; Berka, K.; Jurecka, P.; Hobza, P. On the Reliability of the AMBER Force Field and its Empirical Dispersion Contribution for the Description of Noncovalent Complexes. *Chemphyschem* **2010,** *11* (11), 2399-2408.

22. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235-242.

23. Jorgensen, W. L.; TiradoRives, J. The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *J. Am. Chem. Soc.* **1988,** *110* (6), 1657-1666.

24. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996,** *118* (45), 11225-11236.

25. Jorgensen, W. L.; Tirado-Rives, J. Development of the OPLS-AA force field for organic and biomolecular systems. *Abstracts of Papers of the American Chemical Society* **1998,** *216*, U696.

26. Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E. Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. *Journal of Chemical Theory and Computation* **2010,** *6* (2), 459-466.

27. Kannan, N.; Vishveshwara, S. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Engineering* **2000,** *13* (11), 753-761.

TABLES

Table 1: The calculated contact definitions and their properties. 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

#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	-1.8	-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

Table 2. Similarity of contact matrices. "opls" denotes contact matrices calculated from interaction energy matrices using an OPLSAA force field and our contact definitions. "charmm" denotes the same using CHARMM27 force field. "CA" contact matrices are based on Calpha atom distances, "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.

	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

Table 3. The application of contact matrix construction to thermostable and mesostable protein homologs. The PDB code of thermostable homolog can be found in the first column ("thermo"), the number of all contacts determined by our method in the second column ("Cont") and the length of its chain (number of residues) in the third column ("N"). The fourth column contains the number of contacts per residue ("Cont/N" = the value in the third column divided by the value in the second one). The next four columns contain the same characteristics for mesostable homologs. The last column contains the difference for the number of contacts per residue connected with the stability increase. In 16 out of 20 protein pairs, an increase of number of contacts per residue is found. The same result was obtained using the CHARMM force field.

thermo	Cont	N	Cont/N	meso	Cont	N	Cont/N	Delta
1THL	754	316	2.39	1NPC	729	317	2.30	0.09
1LDN	775	316	2.45	1LDM	729	329	2.22	0.24
1BMD	853	327	2.61	4MDH	777	333	2.33	0.28
2PRD	390	174	2.24	1INO	353	175	2.02	0.22
1PHP	973	394	2.47	3PGK	676	415	1.63	0.84
1THM	692	279	2.48	1ST3	631	269	2.35	0.13
1YNA	447	193	2.32	1XYN	401	178	2.25	0.06
1XYZ	937	320	2.93	2EXO	846	312	2.71	0.22
1CAA	107	53	2.02	6RXN	105	45	2.33	-0.31
1BRF	110	53	2.08	1RB9	100	52	1.92	0.15
1GD1	755	334	2.26	1GPD	569	333	1.71	0.55
1TIB	678	269	2.52	1LGY	719	265	2.71	-0.19
1ZIP	542	217	2.50	1AK2	528	220	2.40	0.10
1FFH	779	287	2.71	1FTS	786	295	2.66	0.05
1PCZ	440	183	2.40	1VOK	468	192	2.44	-0.03
1OBR	879	323	2.72	2CTC	819	307	2.67	0.05
1PHN	421	162	2.60	1CPC	414	162	2.56	0.04
1TMY	304	118	2.58	3CHY	346	128	2.70	-0.13
1GTM	1131	419	2.70	1HRD	1151	449	2.56	0.14
1HDG	800	332	2.41	1GD1	755	334	2.26	0.15

FIGURE CAPTIONS

Figure 1 The HCIE curve for backbone interactions constructed from the calculated data (red) and the function (Equation 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 ∞.

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 is in red. The determined optimum contact definitions (marked blue) are very close to the inflection points with the lowest first derivatives. The result calculated by CHARMM force field is in green.

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.

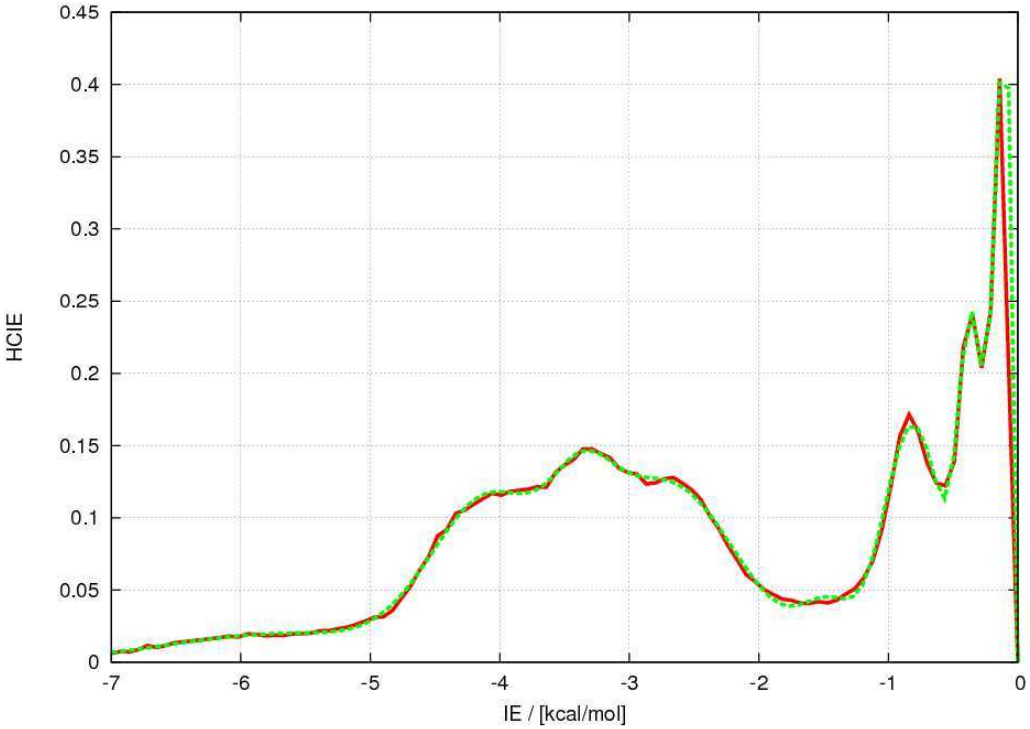

Figure 1

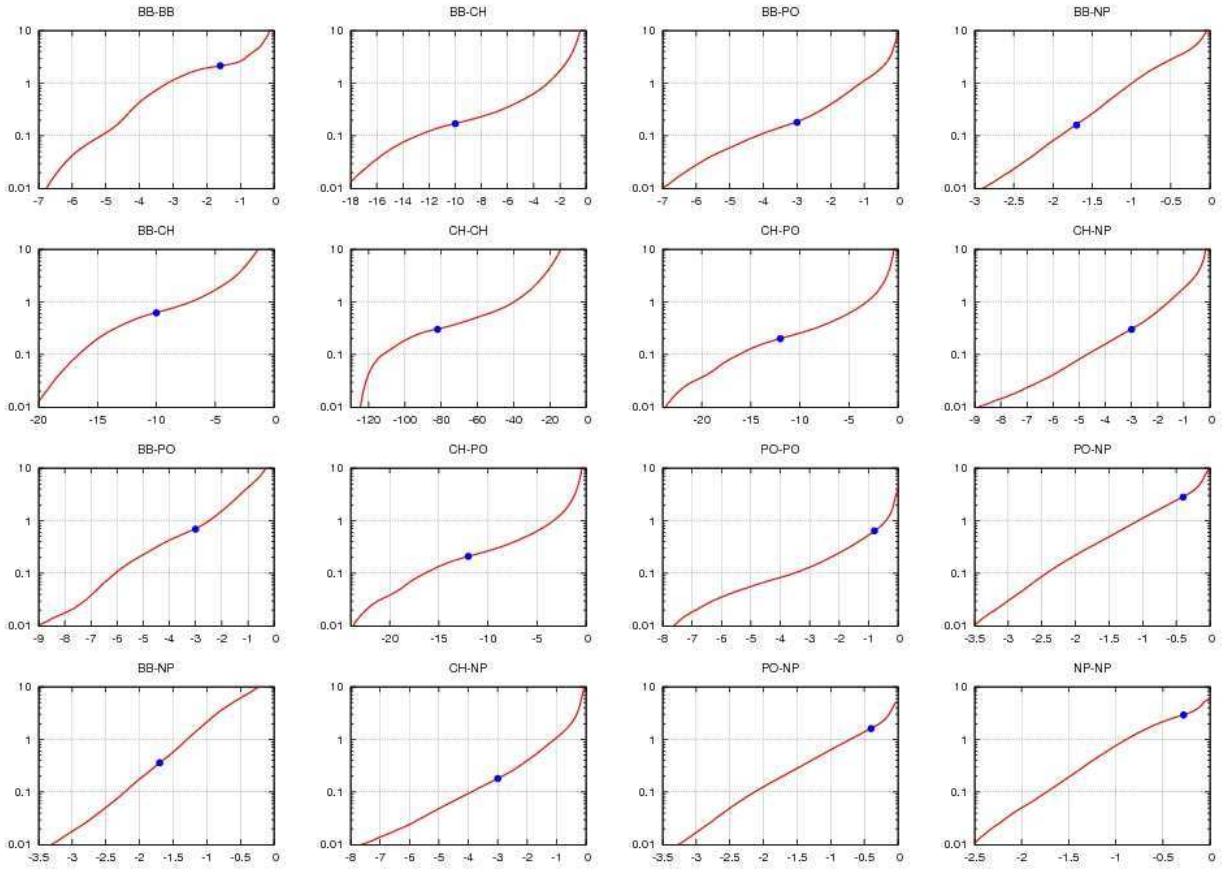
BB-BB
BB-CH
BB-PO
BB-NP
BB-CH
CH-CH
CH-PO
CH-NP
BB-PO
CH-PO
PO-PO
PO-NP
BB-NP
CH-NP
PO-NP
NP-NP

Figure 2

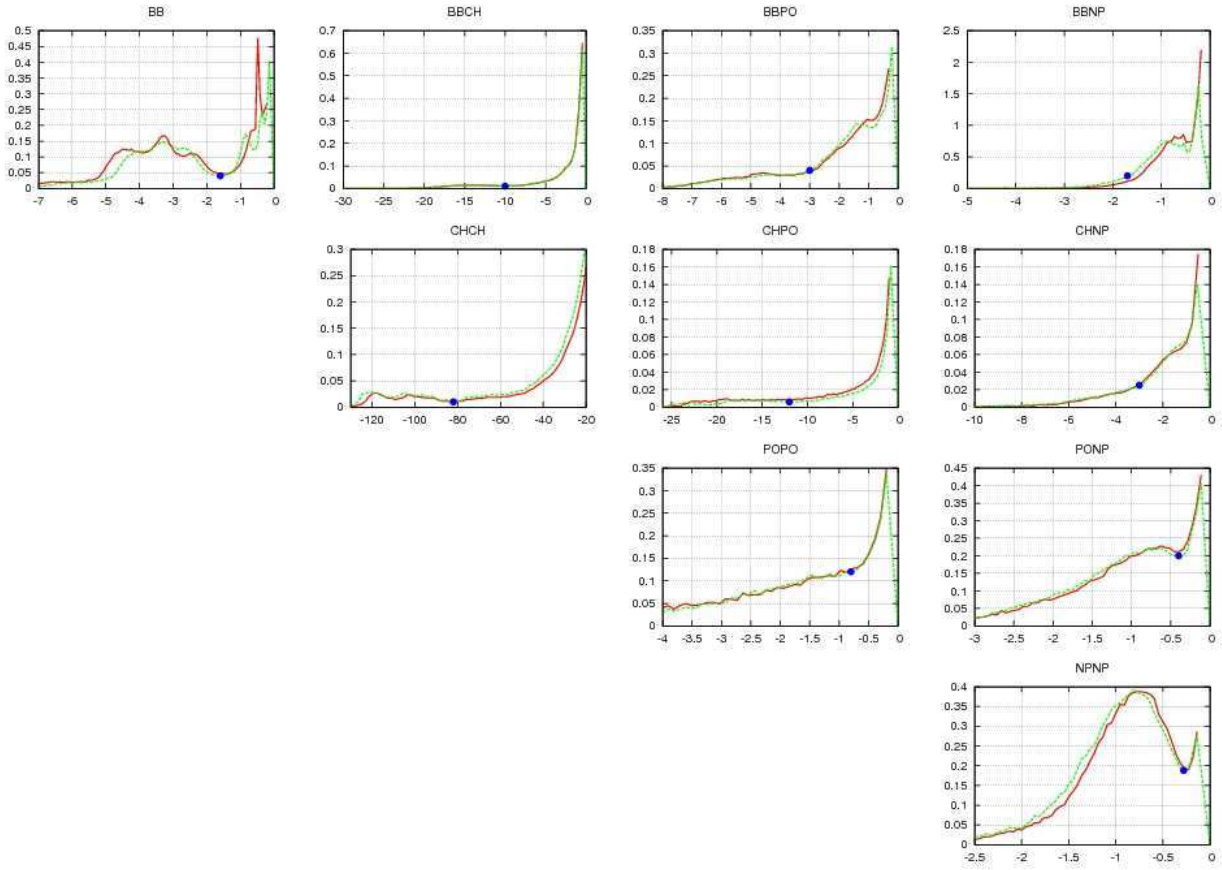

Figure 3.

TOC