Journal of
Bioinformatics and Intelligent Control
Vol. 3, 309–314, 2014

SARAMAint: The Complementarity Plot for Protein–Protein Interface

Sankar Basu[1], Dhananjay Bhattacharyya[2], and Björn Wallner[1,*]

[1] *Bioinformatics Division, Department of Physics, Chemistry and Biology, University of Linkoping, Linkoping 58183, Sweden*
[2] *Computational Science Division, Saha Insititute of Nuclear Physics, Kolkata 700064, India*

Following our previous report of a graphical structural validation tool for single domain globular proteins namely the Complementarity Plot (SARAMA, available at http://www.saha.ac.in/biop/www/sarama.html), here we report the further development of the software (SARAMAint) for protein–protein interfaces available in the same webpage as a separate download link (http://www.saha.ac.in/biop/www/db/local/sarama/SARAMAint.tar.gz) to be effectively used not only to estimate the overall quality of a protein–protein complex but also to assess the individual quality of packing and electrostatics of residues embedded at the interface, in many cases arising due to coordinate errors, especially in low resolution structures. The plot could also be useful for the detailed residue-wise investigation of interfaces in realistic atomic models built for protein complexes in the absence of actual experimental data.

Keywords: Complementarity, Packing and Electrostatics, Protein–Protein Interfaces.

COMMUNICATION

Structure validation is a crucial component not only for experimentally solved protein structures but also for atomic models computationally built by homology or *ab-initio* modeling. There are quite a few standard validation tools[1–3] available in the public domain with a wide range of parameters that can be computed and analyzed from the three dimensional atomic coordinates of a given protein, e.g., the Ramachandran plot,[4] distribution of side-chain rotamers,[5] deviation in bond lengths and angles from their corresponding ideal values,[6] steric clashes,[2] packing defects, and unfulfilled hydrogen bonds[3] are the most common. The Complementarity Plot (CP)[7–9] has previously been reported to be an important inclusion in this already available repertoire. CP analyzes the quality of packing and electrostatic balance of interior residues of single domain globular proteins with respect to their local and non-local atomic neighborhood. The software is freely available as a standalone suite of programs (SARAMA)[8,9] from http://www.saha.ac.in/biop/www/sarama.html. Here, we report the further development and adaptation of the plot for protein–protein interfaces (SARAMAint). The basic methodology of the construction of the plot and the design of the associated scores is similar to the previous report[7–9] whereas the current version of the software analyzes packing and electrostatics of interfacial residues alone from a given protein–protein complex. In conjugation with SARAMA, SARAMAint could thus be effectively used as a structure validation tool for protein complexes.

The quality of residues at the interface are characterized by the packing and electrostatic balance measured by the two functions shape (S_m^{SC}) and electrostatic complementarities (E_m^{SC}), the detailed computational protocol are described in previous publications.[7–9] The Complementary Plot visualize these two values, coming from short- (S_m^{SC}) and long-range (E_m^{SC}) forces sustaining the native fold, against each other in a two-dimensional scatter plot. The software provides estimates of the quality of packing and electrostatics for individual residues and also a global quality estimate based on the distribution of points in the plot for the entire structure. As a validation technique, CP is probabilistic in nature and works best when applied over the full chain. It has been shown to be effective in the detection of erroneous side-chain torsion angles, low-intensity errors in main-chain geometrical parameters diffused over the entire polypeptide chain, packing anomalies and unbalanced partial charges in the protein interior.[9] As was demonstrated previously, the method could also find successful large-scale applications in homology modeling and protein design.[9]

Thus, CP is a sensitive indicator of the harmony or disharmony of interior residues of a globular protein with regard to the short- and long-range forces sustaining the native fold. It was also proposed earlier that

*Author to whom correspondence should be addressed.

complementarity could serve as a common conceptual platform between binding and folding.[7]

Protein interiors and protein–protein interfaces vary significantly in their physicochemical characteristics and also in their local environments. With the exception of dimers; interfaces share more similarity to protein surfaces than to interiors, both in composition and in the spatial distribution of the residues.[10] Hydrophobic residues are generally found to form clusters within protein interiors, whereas nonpolar residues are found in isolation at protein–protein interfaces, surrounded by polar or charged amino acids. However, despite these differences, both interfacial[11,12] and interior atoms[7,13] have to satisfy fairly stringent constraints both with regard to shape and electrostatic complementarity with their local and non-local neighborhood.

As has been reported in previous studies, CP requires the shape (S_m^{SC}) and electrostatic (E_m^{SC}) complementarity to be computed for buried residues.[7–9] Since there is practically no packing constraints (i.e., no nearest neighborhood to pack against) for residues completely exposed to the solvent (Bur > 0.30) they are disregarded from the above mentioned complementarity calculations. Likewise, interfacial residues are also generally found to be buried upon complexation and thus the CP methodology could be applied to analyze them. It was previously demonstrated that regardless of their source and type (antigen-antibody interactions, protein-inhibitor complexes etc.) protein–protein interfaces as two interacting rigid bodies generally satisfy high shape correlation[11] and optimum anti-correlation in their surface electrostatic potential.[12]

In order to detect atoms at the protein–protein interface, the solvent accessible surface area (ASA) of all atoms from each partner molecule was calculated using NACCESS[14] in their free and bound (complexed) conformations. The atoms having a net non-zero change in ASA ($\Delta ASA \neq 0$) were considered as interfacial atoms and a residue having at least one interfacial heavy atom was treated as an interfacial residue. Both the interacting molecules were then considered as a single (pseudo) molecular unit and shape and electrostatic complementarities were computed for the 'interfacial' residues against this entire biomolecular unit as a neighborhood. It was a judicious choice to consider the entire biomolecular unit consisting of both the partner molecules as the neighborhood rather than only the interfacial atoms, because first, a non-negligible fraction (in terms of surface area) of the interfacial residues can be packed against the local neighborhood coming from the source molecule, and second, the long-range balance of surface electrostatic potential of these residues are maintained by electric fields originating from the whole molecular unit and not just the interface.

The detailed construction of the complementarity plot is discussed elsewhere.[7–9] Briefly, subsequent to identifying the interfacial residues (by the methodology described above), the extent of burial (Bur) of every amino acid residue with respect to the solvent was quantified by the ratio of the ASA of the residue (X) embedded in the polypeptide chain to that of an identical residue located in a Gly-X-Gly peptide fragment, in a fully extended conformation. Only those interfacial residues with the burial ratio (Bur) ≤ 0.30 were considered for the complementarity plot. The van der Waals surface was calculated[13] for the entire polypeptide chain, sampled at 10 dots/Å^2 and shape (S_m^{SC}) and electrostatic (E_m^{SC}) complementarities have been computed[7] for all completely ($0.00 \leq Bur \leq 0.05$) or partially buried ($0.05 < Bur \leq 0.30$) residues from a database **(DB2)** of 400 highly resolved (resolution better than 2 Å, R-factor $\leq 20\%$, homologues with sequence identity greater than 30% removed) protein crystal structures. The plot of S_m^{SC} on the *X*-axis and E_m^{SC} on the *Y*-axis (spanning -1 to 1 in both axes) constitutes the 'Complementarity Plot' (CP), which is actually divided into three plots based on the burial ranges: $0.00 \leq Bur \leq 0.05$ (CP1), $0.05 < Bur \leq 0.15$ (CP2) and $0.15 < Bur \leq 0.30$ (CP3). All the buried residues from **DB2** were plotted in the CPs according to their burial and each of the plots were then divided into square grids (of width 0.05×0.05). The center of every square grid was assigned an initial probability (P_{grid}) equal to the number of points in each grid point divided by the total number of points in the plot. The technique of bilinear interpolation was then implemented to estimate the final probability of a residue to occupy a specific position in the plot. In the original CP (for the interior) each of three plots was contoured based on the initial probability values ($P_{grid} \geq 0.005$ for the first contour level and $P_{grid} \geq 0.002$ for the second) thus dividing the plot into three distinct non-overlapping regions. The region within the first contour was termed 'probable,' between the first and the second contour, 'less probable' and outside the second contour, 'improbable'[7–9] (see Fig. 1). Visualized in this way residues with suboptimal S_m^{SC} and E_m^{SC} are easily identified. Furthermore, the plots do not only visually display the distribution of residues in terms of (S_m^{SC}, E_m^{SC}) but also individually list the status of each (buried or partially buried) residue with regard to their location in the corresponding plot. In addition, two associated scores (Complementarity Score: CS_l and Accessibility Score: *rGb*) were defined, as was detailed in an earlier report.[9]

CS_l was designed in order to quantify the distribution of a given set of points (residues) spanning all the three CPs. First, all points in each plot were partitioned into two sets, those with zero and non-zero probabilities. Occurrence of any point with zero probability (essentially in the improbable region) implies that the corresponding residue exhibits suboptimal packing and/or electrostatics with respect to the rest of the protein and therefore should be penalized. The score, CS_l thus consists of two terms, one of which is essentially the average of the non-zero log probabilities and the other being the fraction of residues with zero-probability multiplied by a penalty.[9]

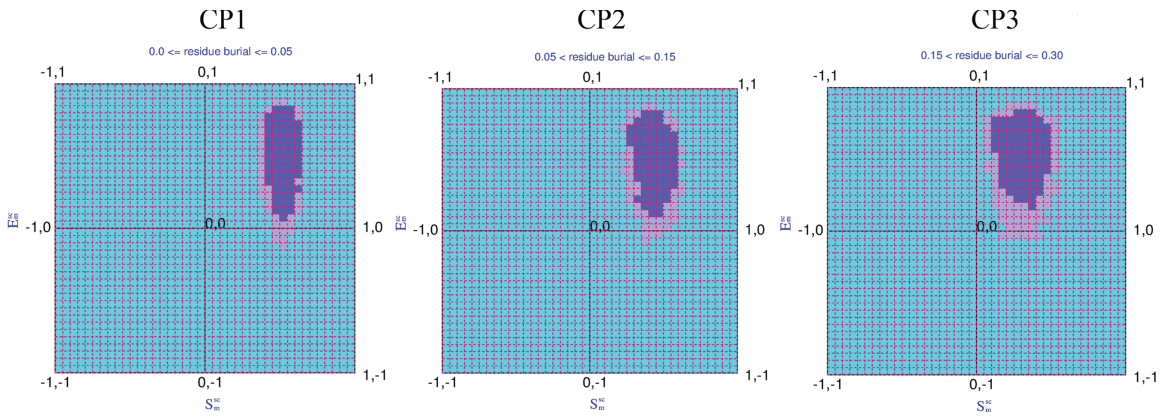

Fig. 1. The complementarity plots, CP1, CP2 and CP3 for burial bins 1, 2 and 3 respectively. 'Probable,' 'less probable' and improbable' regions of the plot are colored in purple, mauve and sky-blue respectively.

rGb was designed to check the expected distribution of amino acid residues with respect to their burial. Residues from a given polypeptide chain were first distributed in four burial bins (the three bins mentioned above and a fourth bin containing residues exposed to the solvent, Bur > 0.30) and the score is calculated as the logarithm of propensities of residues with respect to their burial averaged over the entire polypeptide chain. In contrast to CS_l, *rGb* was computed for the entire biomolecular unit (consisting of both the partner molecules).

In order to test the validity of the previously delineated contours (which segregate the plots into 'probable,' 'less probable' and 'improbable regions,' obtained from **DB2**), for the interface version of the CPs, we assembled another database **(DB3)** of 1,651 high resolution 'native' protein–protein complex crystal structures from the repository DOCKGROUND[15] (http://dockground.compbio.ku.edu/) with resolution better than 2 Å, and at least 10 residues at the interface. For complexes with more than two chains, the two largest interacting chains were considered for the calculation. The S_m^{SC}, E_m^{SC} values for the interfacial residues from this database, were plotted in each of the three plots (CP1, CP2, CP3) according to their particular burial. The P_{grid} values (as defined earlier) for each of the 1600 square grids (of width 0.05×0.05) in each of the three plots (CP1, CP2, CP3) were calculated and compared to the distribution derived for the interior. The interfacial plots contained slightly more points than interior, for CP1, 28,593 versus 23,850, for CP2, 18,521 versus 10624 and for CP3, 18,263 versus 13,255. The overlap between interior and interface was found to be 87.4%, 88.2% and 87.3% for CP1, CP2, and CP3, respectively. Based on the agreement between the interior and interface distributions (see Fig. 2), the original contours obtained from the interior CPs for the interfacial plots were retained. Subsequent to plotting the interfacial residues in the CPs according to their burial, CS_l was computed for these residues alone whereas *rGb*

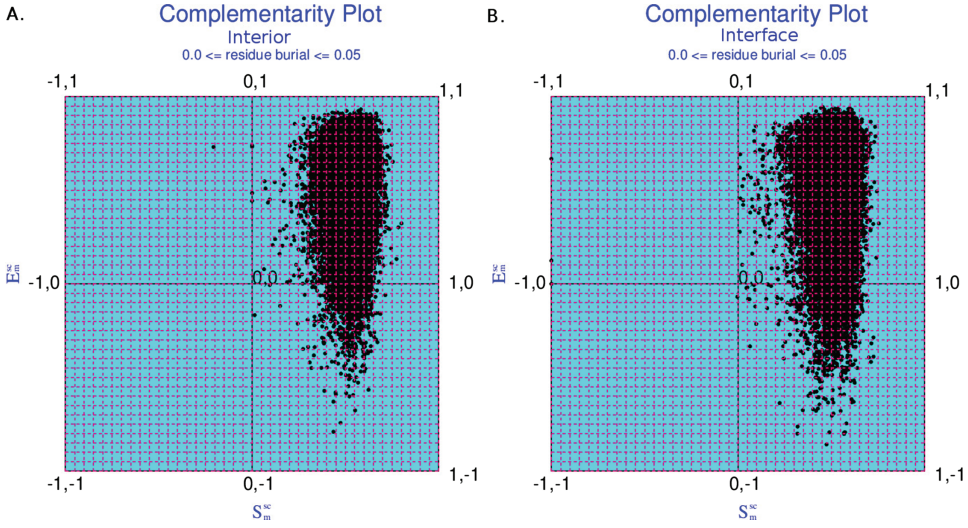

Fig. 2. Distribution of (A) interior and (B) interfacial completely buried residues ($0.00 \leq$ Bur ≤ 0.05) from databases **DB2** and **DB3** respectively in the Complementarity Plot (CP1). The overlap between the corresponding grid probabilities was found to be 87.3%.

COMMUNICATION

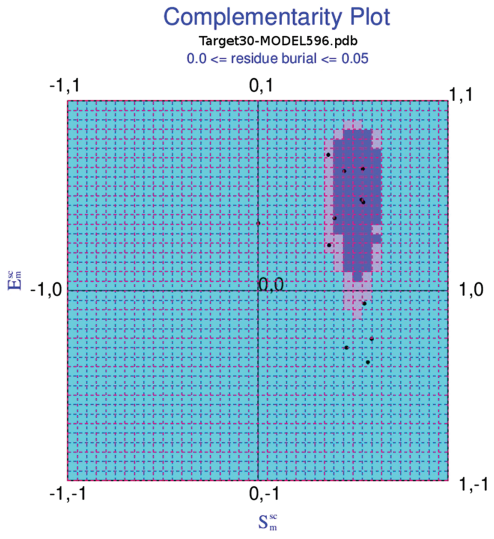

Fig. 3. Distribution of all completely buried interfacial residues in the Complementarity Plot (CP1) of the CAPRI model number 596 of target 30. The overall surface and electrostatic complementarities between the two interacting surfaces have been found to be Sc: 0.432, EC: −0.711. As could be seen from the distribution of the points in the plot, the residues clearly have suboptimal electrostatic complementarities (many of them falling in the negative E_m axis) in spite of retaining optimum shape complementarities.

was computed for the entire (pseudo) molecular unit. As a matter of convention, for structures with no interfacial residues found to be falling in either of the three plots (i.e., all of them effectively being exposed to the solvent; Bur > 0.30), only *rGb* was calculated and CS_l was set to zero.

The corresponding (complementarity and accessibility) scores for the interior and interface were very similar, CS_l: 2.24 (±0.48), *rGb*: 0.055 (±0.022) for the interior and CS_l: 2.29 (±0.71), *rGb*: 0.059 (±0.022) for the interface. This result also quantitatively supports the idea that together shape and electrostatic complementarity could indeed serve as a common conceptual platform to discuss binding and folding.

In order to investigate how the residue-level complementarity of individual interfacial amino acids contribute to the overall complementarity attained at the interface of two interacting proteins, the overall shape (S_c) and electrostatic complementarity (EC) of the complete interface considering the molecular pair as two interacting rigid bodies were calculated using the methodology described by Lawrence and Colman[11] and McCoy et al.[12] To test the performance of CP on realistic models, the plot was run on 16,111 CAPRI[16] models built for 15 targets downloaded from http://cb.iri.univ-lille1.fr/Users/lensink/Score_set/. It was noteworthy to encounter that 35% of these models (5,625 out of 16,111 models) actually had positive values for shape complementarity (Sc > 0) whereas negative values for electrostatic complementarity (EC < 0) which is consistent with the general notion that for oligomer formation, shape complementarity is a necessary condition[11] whereas electrostatic complementarity is sufficient.'[12, 17] For these cases, the suboptimal residues had fairly good shape complementary but a strong electrostatic imbalance falling into the fourth quadrant of the CP (see Fig. 3).

To investigate if the CPs could be useful to globally discriminate between high and low resolution protein–protein complexes solved by X-ray crystallography,

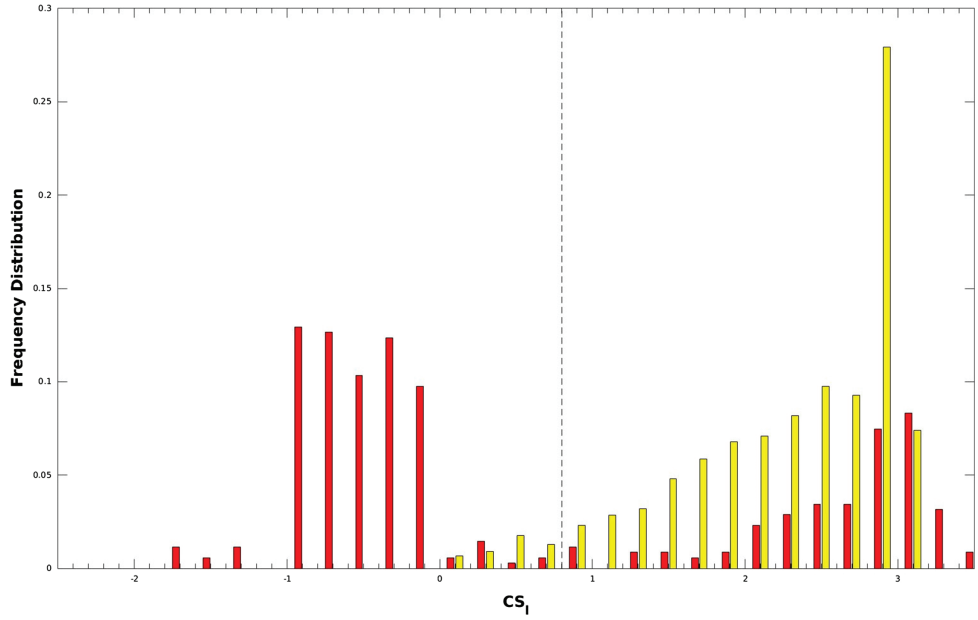

Fig. 4. Normalized frequency distribution of CS_l values for low (red bars) and high resolution (yellow bars) structures. The dashed line represents the CS_l cutoff (0.80) above which structures are considered to be validated successfully.

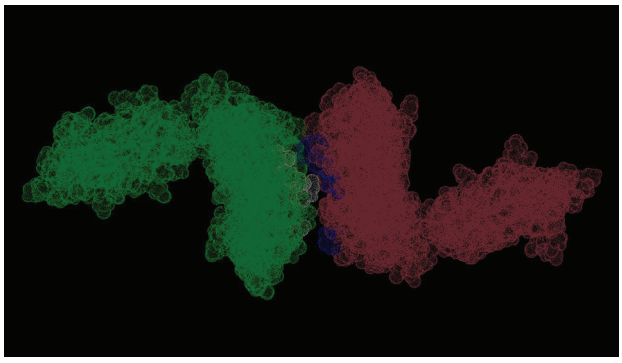

Fig. 5. Van der Waals surface (represented as dots) of a low resolution (3.2 Å) complex 1A9B having a 'small' interface. The interfacial surface (containing 8 residues, all of them having a burial > 0.30) is drawn in white and blue on the background of green and pink for the two partner molecules respectively.

a low-resolution (> 3 Å) protein complex set was culled from DOCKGROUND[15] with identical culling criteria (at least 10 residues at the interface). After removal of mutants, CA-only templates and DNA/RNA binding complexes a total of 357 structures was obtained and the Complementarity Plots were run for each of these. The frequency distribution (Fig. 4) of these low-resolution structures was found to be bimodal (possibly suggesting a mixed population of 'good and bad' interfaces) compared to a unimodal distribution obtained for the high-resolution structures. A careful investigation of the bimodal distribution (for the low resolution set) suggested that the two humps corresponded to 65% and 35% of structures respectively below and above the CS_l threshold (0.80) for successful validation.[9] The same fractions for the high resolution set were found to be 5% and 95% respectively. The average *rGb* score for the low resolution set was found to be 0.019 (±0.030), also significantly less than the set of high-resolution structures 0.059 (±0.022).

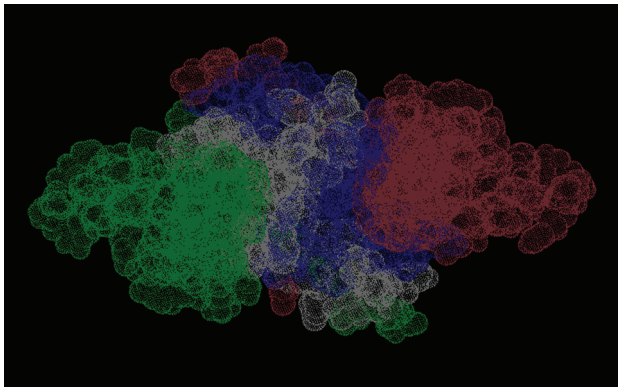

Fig. 6. Van der Waals surface (represented as dots) generated for a protein–protein complex (PDB ID: 4A5N) as displayed by a RasMol script generated by the software SARAMAint. The interfacial surface is drawn in white and blue on the background of green and pink for the two partner molecules respectively.

However, for complexes with really small interfaces (<10 residues) with all residues found to be exposed to (Bur > 0.30) the solvent and therefore not falling in either of the three plots, a CS_l score can not be computed and thus the validation remains limited to the *rGb* score alone. Such an example is given in Figure 5. However, a detail and systematic analyses of this kind lies out side the scope of the current study.

The results clearly suggests that the complementary plot for the interface (SARAMAint) in conjugation with the complementary plot for the interior (SARAMA) can be used to globally discriminate between high- and low-resolution structures and could be effectively used for structure validation of protein–protein complexes.

Another added feature to the current software is that it produces Rasmol scripts to view the different molecular surfaces (e.g., interface and others) colored differently (see Fig. 6). The utility of the 'complementarity' method in modeling, scoring and predicting protein–protein complexes are currently being investigated.

References and Notes

1. R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, Procheck: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283 **(1993)**.
2. R. W. Hooft, C. Sander, and G. Vriend, Positioning hydrogen atoms byoptimizing hydrogen-bond networks in protein structures. *Proteins* 26, 363 **(1996)**.
3. I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, III, J. Snoeyink, J. S. Richardson, and D. C. Richardson, MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucl. Acids. Res.* 35, W375 **(2007)**.
4. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7, 95 **(1963)**.
5. R. L. Dunbrack, Jr. and M. Karplus, A backbone dependent rotamer library for proteins: Application to sidechain prediction. *J. Mol. Biol.* 230, 543 **(1993)**.
6. R. A. Engh and R. Ber, International Tables for Crystallography. In International Tables for Crystallography, edited by M. G. Rossmann and E. Arnold, Kluwer Academic Publishers, Dordrecht, The Netherlands **(2001)**, pp. 382–392.
7. S. Basu, D. Bhattacharyya, and R. Banerjee, SARAMA: A standalone suite of programs for the complementarity plot—A graphical structure validation tool for proteins. *J. Bioinf. Intell. Control* 2, 321 **(2013)**.
8. S. Basu, D. Bhattacharyya, and R. Banerjee, Applications of complementarity plot in error detection and structure validation of proteins. *Indian Journal of Biochemistry and Biophysics* 51, 188 **(2014)**.
9. S. Basu, D. Bhattacharyya, and R. Banerjee, Self-complementarity within proteins: Bridging the gap between binding and folding. *Biophys. J.* 102, 2605 **(2012)**.
10. S. Jones and J. M. Thornton, Principles of protein–protein interactions. *Proc. Natl. Acad. Sci. USA* 93, 13 **(1996)**.
11. M. C. Lawrence and P. M. Colman, Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* 234, 946 **(1993)**.
12. A. J. McCoy, V. C. Epa, and P. M. Colman, Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.* 268, 570 **(1997)**.
13. R. Banerjee, M. Sen, P. Saha, et al. The jigsaw puzzle model: Search for conformational specificity in protein interiors. *J. Mol. Biol.* 333, 211 **(2003)**.

14. B. Lee and F. M. Richards, The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* 55, 379 **(1971)**.
15. I. Anischenko, P. J. Kundrotas, A. V. Tuzikov, and I. A. Vakser, Protein models: The Grand Challenge of protein docking. *Proteins* 82, 278 **(2014)**.
16. M.F. Lensink and S. J. Wodak, Score_set: A CAPRI benchmark for scoring protein complexes. *Proteins* 82, 3163 **(2014)**.
17. Y. Tsuchiya, K. Kinoshita, and H. Nakamura, Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity protein engineering. *Design and Selection* 19, 421 **(2006)**.

Received: 25 March 2014. Accepted: 15 July 2014.