Journal of
Bioinformatics and Intelligent Control
Vol. 2, 321–323, 2013

SARAMA: A Standalone Suite of Programs for the Complementarity Plot—A Graphical Structure Validation Tool for Proteins

Sankar Basu[1,*], Dhananjay Bhattacharyya[2], and Rahul Banerjee[1,*]

[1] Crystallography and Molecular Biology Division, Saha Institute of Nuclear Physics, Kolkata 700064, India
[2] Computational Science Division, Saha Institute of Nuclear Physics, Kolkata 700064, India

Structure validation is a crucial component not only in protein crystallography but also in model quality estimation in homology modeling, protein design and de-novo structure prediction. Two key attributes of a correctly determined atomic model are optimal packing between side-chains and absence of destabilizing unbalanced electric fields within the interior of a protein molecule. The complementarity plot (CP) combines them in a single unified measure. CP has now been compiled into a user friendly validation package and made available as a standalone suite of programs in the public domain (http://www.saha.ac.in/biop/www/sarama.html). The application of CP in the detection of wrong rotamer assignment has been surveyed.

Keywords: Complementarity, Packing and Electrostatics, Structure Validation.

COMMUNICATION

We report the free availability of a standalone suite of programs (Sarama) for the Complementarity Plot (Linux Platform) with detailed features and documentation available at the website: http://www.saha.ac.in/biop/www/sarama.html. The basic methodology has already been reported.[1] Briefly, the Complementarity Plot (CP) estimates the shape and electrostatic complementarity of interior residues of a globular protein and is a sensitive indicator of their harmony or disharmony with regard to the short and long range forces sustaining the native fold. A correctly determined natively folded protein structure should have optimal packing between its buried side-chains and absence of destabilizing unbalanced electric fields within the interior of the molecule. CP has already been demonstrated to be effective in detecting local regions of suboptimal packing or electrostatics which were found to be highly correlated to coordinate errors. CP has now been compiled into an user friendly validation package which should be an useful addition in the already existing repertoire of structure validation tools. A set of scores have now been included in the methodology which gives an estimate of the probabilities associated with the distribution of points in the plot and the propensities of specific residues to different degrees solvent exposure.

As has been reported previously[1] CP requires the surface (S_m^{sc}) and electrostatic (E_m^{sc}) complementarity to be computed for buried residues. In this regard, the extent of burial (Bur) of every amino acid residue with respect to the solvent was estimated by the ratio of the solvent accessible areas (probe radius: 1.4 Å)[2] of the residue (X) in the polypeptide chain to that of an identical residue in a Gly-X-Gly peptide fragment, in a fully extended conformation. Only those residues with the burial ratio (Bur) ≤ 0.30 were henceforth considered for the complementarity plot. The van der Waals surface was calculated for the entire polypeptide chain, sampled at 10 dots/Å^2 [3] and surface (S_m^{sc}) and electrostatic (E_m^{sc}) complementarities calculated for buried or partially buried side-chains.[1,3]

For surface complementarity (S_m^{sc}), only side-chain surface points of buried residues (target) were considered and their nearest neighboring surface points identified from the rest of the polypeptide chain (within a distance of 3.5 Å). Surface points essentially being area elements are characterized by their positions (x, y, z) and the direction cosines (dl, dm, dn) of their normals. Then, adapted from Lawrence and Colman,[4] the following expression was calculated:

$$S(a, b) = \mathbf{n}_a \cdot \mathbf{n}_b \cdot \exp(-w.d_{ab}^2) \quad (1)$$

where $\mathbf{n}_a$ and $\mathbf{n}_b$ are two unit normal vectors, corresponding to the dot surface point a (located on the side chain surface of the target residue) and b (the dot point nearest to a, within 3.5 Å) respectively, with d_{ab} the distance between them and w, a scaling constant set to 0.5. S_m^{sc} was defined as the median of the distribution $\{S(a, b)\}$

*Authors to whom correspondence should be addressed.

2326-7496/2013/2/321/003 doi:10.1166/jbic.2013.1059

calculated over all the dot surface points of the side-chain target residue.

For electrostatic complementarity (E_m^{sc}), the electrostatic potential of the molecular surface was estimated using the finite difference Poisson-Boltzmann method as implemented in DelPhi.[5] The potential on the side-chain surface points of a buried residue was then computed twice,[1] first, due to all atoms of the target residue and second as a function of all atoms from the rest of the polypeptide chain (excluding the target). Thus, each surface point was tagged with two values of electrostatic potential. Following McCoy et al.,[6] negative of the Pearson's correlation coefficient between these two sets of potential values over the side-chain dot surface points of the target residue was defined as E_m^{sc}

$$E_m^{sc} = -\left(\frac{\sum_{i=1}^{N}(\varphi(i)-\bar{\varphi})(\varphi'(i)-\bar{\varphi}')}{(\sum_{i=1}^{N}(\varphi(i)-\bar{\varphi})^2\sum_{i=1}^{N}(\varphi'(i)-\bar{\varphi}')^2)^{1/2}}\right) \quad (2)$$

where, for a given residue consisting of a total of N side-chain dot surface points, $\varphi(i)$ is the potential on its ith point realized due to its own atoms and $\varphi(i)$, due to the rest of the protein atoms, $\bar{\varphi}$ and $\bar{\varphi}$ are the mean potentials of $\varphi(i)$, $i = 1, \ldots, N$ and $\varphi'(i)$, $i = 1, \ldots, N$ respectively.

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). Initially, all the buried residues from a training database (DB2) consisting of 400 highly resolved protein crystal structures[1] were plotted in the CPs, which had been divided into square-grids (of width 0.05×0.05), and the center of every square grid was assigned an initial probability (P_{grid}) equal to the number of points in the grid divided by the total number of points in the plot. The probability of a residue to occupy a specific position in the plot was then estimated by bilinear interpolation from the probability values of its four nearest neighboring voxels. Each CP 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 regions. The cumulative probability of locating a point within the second (outer) contour for the three plots were 91%, 90%, 88% respectively whereas for the first (inner) contour, the probability gradually dropped with increasing solvent exposure (82%, 76%, 71%). Inspired by the Ramachandran Plot, the region within the first contour was termed 'probable,' between the first and second contour, 'less probable' and outside the second contour, 'improbable' (Fig. 1).

In such a plot residues with low S_m^{sc} and E_m^{sc} (<0.2 for both) are easily identified. The methodology has already been shown to detect errors in side-chain conformers in obsolete structures w.r.t. their upgraded counterparts.[1] Such side-chains were found to have suboptimal packing and/or electrostatics and thus predominantly lie in the improbable region of the plots. An example of such a calculation is given below.

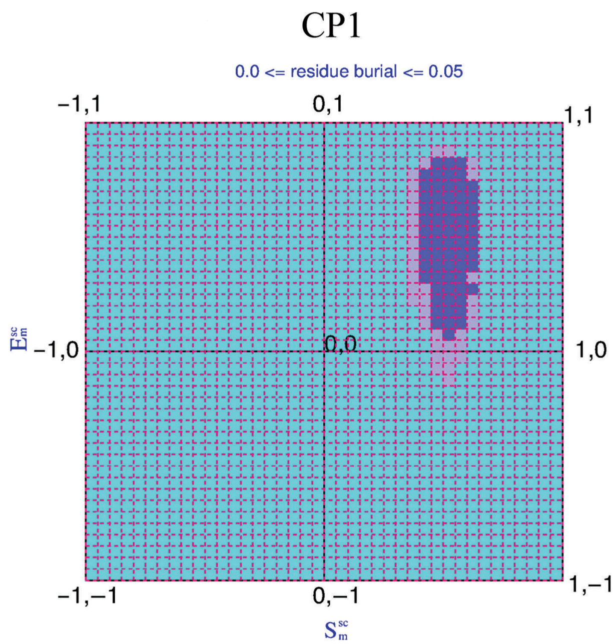

Fig. 1. CP1: The complementarity plot for the 1st burial bin. 'Probable,' 'less probable' and improbable' regions of the plot are colored in purple, mauve and sky-blue respectively.

110 pairs of obsolete and their corresponding upgraded counterparts were assembled from the PDB (ftp://ftp.wwpdb.org/pub/pdb/data/status/obsolete.dat). In order to ensure that the upgraded structure was genuinely better than its obsoleted counterpart, only those pairs were selected wherein the improvement in resolution and R-factor were better than 0.2 Å and 0.02 respectively. 222 deeply buried residues ($0.0 \leq Bur \leq 0.05$) from the upgraded structures were identified which were originally found to be located in the probable region of CP1, and whose counterparts in the corresponding obsolete structures differed by more than 40° (involving χ_1 and χ_2) though belonging to another valid rotamer combination.[7] They were then replaced by their corresponding counterparts from the obsolete structures. Subsequent to the replacement, 45% of the points were relocated in the improbable region of the plot, 16% were found in the less probable region whereas 39% were retained in the probable region (Fig. 2). Deviations from the expected distributions (DB2) were estimated by means of χ^2 (df = 3–1, probable, less probable, improbable; $\chi^2_{0.05} = 5.991$) subsequent to the replacement which was found to be 397.63. Thus, CP could have applications when dealing with low-resolution data where automated side-chain rebuilding methods generally do not work very efficiently.

The Complementarity Plot as a validation technique is probabilistic in nature and can be utilized either over the full chain, or on any distribution of points. Further, this is the only validation procedure which combines both packing

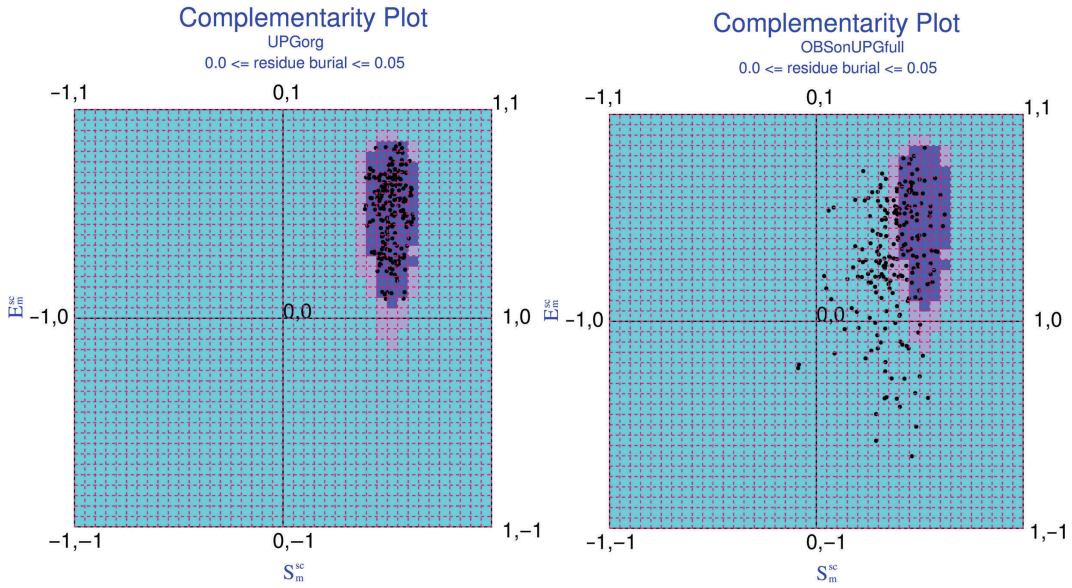

Fig. 2. Distributions (in CP1) for residues with native side-chain conformers from the upgraded structures and replaced by rotamers from corresponding obsolete counterparts. (A) Distribution of residues with native side-chains all falling into the probable regions of CP1 and (B) distribution subsequent to the replacement.

and electrostatics in a single unified measure and displays graphically (apart from actually listing) residues with faulty packing and/or electrostatics. Thus, CP should be a useful addition in the already existing repertoire of structure validation tools. The output of the program gives S_m^{sc} and E_m^{sc} of buried residues which can also be used for a wide range of other applications e.g., fold recognition, analysis of side-chain packing, detection of unbalanced partial charges within protein interiors, protein design and modeling.

The website contains detailed documentation of the different thresholds for successful validation for all the CP-scores for a given atomic model. The model might be experimentally or computationally derived but should definitely contain coordinates of (geometrically fixed) hydrogen atoms consistent with the format of REDUCE.[8] The suite has been successfully tested on Redhat Enterprise and open Suse linux platforms with PERL and Fortran compilers f90, f95, gfortran or ifort. DELPHI[5] must be preinstalled and running under the command: delphi_static.

References and Notes

1. S. Basu, D. Bhattacharyya, and R. Banerjee, Self-complementarity within proteins: Bridging the gap between binding and folding. *Biophys. J.* 102, 2605 **(2012)**.
2. B. Lee and F. M. Richards, The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* 55, 379 **(1971)**.
3. R. Banerjee, M. Sen, D. Bhattacharyya, and P. Saha, The Jigsaw puzzle model: Search for conformational specificity in protein interiors. *J. Mol. Biol.* 333, 211 **(2003)**.
4. M. C. Lawrence and P. M. Colman, Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* 234, 946 **(1993)**.
5. A. Nichollos and B. Honig, A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comput. Chem.* 12, 435 **(1991)**.
6. A. J. McCoy, V. C. Epa, and P. M. Colman, Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.* 268, 570 **(1997)**.
7. M. S. Shapovalov and R. L. Dunbrack, Jr, A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19, 844 **(2011)**.
8. J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285, 1735 **(1999)**.

Received: 10 August 2013. Accepted: 3 September 2013.