Complementarity and Accessibility Scores

Following are the description and rationalization of the scores associated to the Complementarity Plots (CPs). Essentially there are five scores (***CS_l*, *rGb*, *P_{Sm}*, *P_{Em}*, *P_{count}***) designed for different purposes the first four of which are global and the last one is local in nature. ***CS_l*** and ***rGb*** should be used in conjunction and the thresholds for successful validation for each of them are given in details in the README page.

CS_l:

CS_l is a complementarity score designed to quantify the quality of the plots wherein all points in each plot were first 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 thus consists of two terms, the first essentially the average of the non-zero log probabilities and the second, the fraction of residues with zero-probability multiplied by a penalty (*Pen*). Thus the score would be expected to decrease with increase in the points in the improbable regions of the plot. For a particular plot (say CP1) the score can be defined as:

$$Sl = \left[\frac{1}{N}\sum_{i=1}^{N}\log_{10}(P_i)\right] - Pen\cdot\left(\frac{N_{zero}}{N_{tot}}\right)$$

$$= Sl_{non\text{-}zero} + Sl_{zero}$$

where N_{tot} is the total number of points in the plot which can be partitioned into points which fall in square grids of non-zero probability (N) with grid probabilities P_i and those located in grids of zero probability (N_{zero}). For the first term it was assumed that the probability assigned to one point (Pi) is independent of the others, leading to a multiplication of probabilities (P_1, P_2, …) and converted into a summation by taking log ($\sum_{i=1}^{N}\log_{10}(P_i)$). There is some measure of arbitrariness in assigning the value for *Pen* which was computationally optimized. Even for accurately determined structures from the training database, **DB2**, generally 10% of the residues (per chain) would be located in the improbable regions of the plots. It was thus decided that for correctly folded proteins (of the kind found in **DB2**), the ratio of the two terms ($R_{Sl} = Sl_{zero} / Sl_{non\text{-}zero}$) should optimally be in the range 0.30, greater than which it would unjustifiably begin to dominate the overall score whereas too low a value (say less than 0.10) would compromise the sensitivity of the score to structural errors. Several values of *Pen* were tested on DB2 where the two terms (Sl_{zero} & $Sl_{non\text{-}zero}$) were estimated for each polypeptide chain in the database; initially applying the same *Pen* for all the three plots (CP1, CP2, CP3; **Table R1**). For uniform penalties applied to all the three plots it was observed that RSl tended to increase from CP1 to CP3 as relaxation in packing constraints (with corresponding increase in solvent exposure) increased the relative fraction of points in the zero probability grids from CP1 to CP3 (N_{zero}/N_{tot} for CP1: 0.026 (± 0.029), CP2: 0.037 (± 0.048), CP3: 0.045 (± 0.043)). Thus, to introduce some measure of uniformity, *Pen* was modulated (CP1: 25; CP2: 20; CP3: 15) such that R_{Sl} was in the range 0.30 – 0.35 for all the three plots. Understandably, the ratios of the penalties (*Pen*) in the three plots (CP1/CP2: 25/20 = 1.25; CP1/CP3: 25/15 = 1.67) were correlated to the corresponding ratios of N_{zero}/N_{tot} (CP2/CP1: 0.037/0.026 = 1.42; CP3/CP1: 0.045/0.026 =1.73).

Finally,

$$CS_l = K + \sum_{j=1}^{3} wb_j \cdot Sl_j$$

As has been mentioned, scores for deviant structures are expected to decrease in value. So for convenience of interpretation, K was empirically set to 5.0 so as to obtain an overall positive score from 0 to 5 in case of a favorable distribution spanning the three plots. It follows that such a constant merely acts as a scale factor universally applied to all CS_l scores. wb_j is the number of points in the j^{th} plot divided by the total number of points in the three plots and the (weighted) summation is over CP1, CP2 and CP3.

The sensitivity of $\boldsymbol{CS_l}$ was also tested **(Table R1)** for different combinations of penalties by computing its mean and standard deviations for all chains in **DB2**. Standard deviations were especially high (1.17 to 2.33) for uniform penalties 100, 75, 50 whereas for different combinations of penalties in the range of 5 to 30, $\boldsymbol{CS_l}$ was found to be fairly stable with standard deviations falling in range of 0.14 to 0.60 **(Table R1)**, and $\boldsymbol{CS_l}$ was confirmed to be well behaved for the selected penalty values (*Pen* = 25, 20, 15 for CP1, CP2, CP3 respectively).

rGb:

In order to check the expected distribution of amino acid residues w.r.t. burial, the following score was defined.

$$rGb = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \log_{10}(\Pr_i)$$

where N_{res} is the total number of residues in a polypeptide chain and Pr_i is the propensity of a particular (i^{th}) amino acid (Val, Leu etc) in that chain to acquire a particular degree of solvent exposure (corresponding to buried residues in the three burial bins and a 4th bin composed of exposed residues

(Bur > 0.30)).

$$\mathrm{Pr}_j = \frac{P(\mathrm{Res}(j) \mid Bur(k))}{\left(\dfrac{N(\mathrm{Res}(j))}{N_{DB}}\right)}$$

where k = 1, 2, 3, 4 corresponds to four burial bins with fraction of exposed solvent accessible areas **[1]** 0.0 – 0.05, 0.05 – 0.15, 0.15 – 0.30 , > 0.30 respectively and j = 1,2,3,4….19 corresponds to the 19 amino acids excluding glycine. For any residue the fraction of exposed solvent accessible area (SAA) was estimated as the ratio of SAA of the residue (X) in the polypeptide chain to its SAA in a fully extended Gly – X – Gly conformation. *P(Res(j)|Bur(k))* is the conditional probability of *Res(j)* (say Val) to acquire a given burial, *Bur(k)* and *N(Res(j))* is the number of residues of identity *Res(j)* found in the **DB2** consisting of a total of N_{DB} residues.

P_{Sm}, P_{Em}:

To quantify the individual contributions of S_m^{sc} and E_m^{sc} **[2]**, two additional (global) scores P_{Sm} and P_{Em} were further defined. The normalized frequency distribution separately for each burial bin was used to assign discrete probabilities (P (x < S_m^{sc} < (x+0.05))) to S_m^{sc} divided into intervals of 0.05. Three such probability distributions were computed one for each burial bin and a similar procedure was adopted for E_m^{sc}. Then, for each polypeptide chain, the individual probabilities were averaged over all buried or partially buried residues, giving rise to the two following measures:

$$P_{Sm} = \frac{\sum_{i=1}^{N_b} \log_{10}(P_i(S_m^{sc}))}{N_b}; \qquad P_i(S_m^{sc}) \neq 0$$

$$P_{Em} = \frac{\sum_{i=1}^{N_b} \log_{10}(P_i(E_m^{sc}))}{N_b}; \qquad P_i(E_m^{sc}) \neq 0$$

where N_b is the total number of buried or partially buried residues in a given polypeptide chain.

P_{count}:

A local score (P_{count}) was also defined simply as the number of points in the improbable regions divided by the total number of points spanning the three plots.

Table R1. Sensitivity of CS_l to different values of penalty (*Pen*). The quantum of penalty *(Pen)* applied to CP1, CP2, CP3 is indicated in the first column of the table. $R_{Sl} = Sl_{zero} / Sl_{non\text{-}zero}$

Pen			**R_{Sl}**			**CS_l**
CP1	**CP2**	**CP3**	**CP1**	**CP2**	**CP3**	
100	100	100	1.31 (±1.44)	1.75 (±2.22)	2.02 (±1.93)	-0.54 (± 2.33)
75	75	75	0.98 (± 1.08)	1.31 (± 1.66)	1.52 (± 1.45)	0.33 (± 1.75)
50	50	50	0.66 (± 0.72)	0.88 (± 1.11)	1.01 (± 0.96)	1.19 (± 1.17)
30	30	30	0.39 (± 0.43)	0.53 (± 0.66)	0.61 (± 0.58)	1.89 (± 0.71)
25	25	25	0.33 (± 0.36)	0.44 (± 0.55)	0.51 (± 0.48)	2.06 (± 0.59)
20	20	20	0.26 (± 0.29)	0.35 (± 0.44)	0.41 (± 0.39)	2.23 (± 0.48)
15	15	15	0.20 (± 0.22)	0.26 (± 0.33)	0.31 (± 0.29)	2.40 (± 0.36)
10	10	10	0.13 (± 0.14)	0.18 (± 0.22)	0.20 (± 0.19)	2.58 (± 0.25)
5	5	5	0.07 (± 0.07)	0.09 (± 0.11)	0.10 (± 0.10)	2.75 (± 0.14)
30	25	20	0.39 (± 0.43)	0.44 (± 0.55)	0.41 (± 0.39)	2.06 (± 0.60)
25	20	15	0.33 (± 0.36)	0.35 (± 0.44)	0.31 (± 0.29)	2.24 (± 0.48)
20	15	10	0.26 (± 0.29)	0.26 (± 0.33)	0.20 (± 0.19)	2.41 (± 0.37)

1. Lee, B., and Richards, F.M., **The interpretation of protein structures: Estimation of static accessibility.** ***J. Mol. Biol.*** 1971 (55) 379-400.

2. Basu, S., Bhattacharyya, D., Banerjee, R., **Self-Complementarity within Proteins: Bridging the Gap between Binding and Folding.** ***Biophys. J.*** 2012, (102) 2605-2614.