JOURNAL OF APPLIED CRYSTALLOGRAPHY

Volume 50 (2017)

Supporting information for article:

How many waters are detected in X-ray protein crystal structures?

Marco Gnesi and Oliviero Carugo

Table S1 List of the variables that were extracted or computed for each crystal structure. The symbols of hose which were retained for the Poisson multiple regression analysis are indicated in parentheses.

Variabile	Description
y	*(H_2O/aa)* Ratio between the number of water molecules and the number of amino acid residues in the asymmetric unit
x_1	Space group (each space group is coded by a number, from 0 to 73)[(a)]
X_2	Deposition date (year)
X_3	*(Res*) crystallographic resolution (Å)
X_4	*(R*) working R-factor
X_5	Free R-factor
X_6	*(T*) temperature of the data collection (K)
X_7	Number of amino acids in the asymmetric unit[(b)]
X_8	Number of missing amino acids in the asymmetric unit[(c)]
X_9	Volume of the unit cell (Å^3)[(d)]
x_{10}	*(Solv%*) Percentage of solvent in the crystal[(d)]
x_{11}	*(Bave*) Average B-factor of the protein atoms(Å^2)
x_{12}	Average B-factor of the water atoms (Å^2)
x_{13}	Average B-factor of the hetero atoms different from water (Å^2)
x_{14}	Oligomerization state (integer that indicates the number of protomers)[(e)]
x_{15}	Number of polypeptide chains in the asymmetric unit
x_{16}	Percentage of amino acid residues in helices[(f)]
x_{17}	Percentage of amino acid residues in strands[(f)]
x_{18}	*(aaLoops%*) Percentage of amino acid residues in loops[(f)]
x_{19}	*(Sasaaa*) Average solvent accessible surface areas of the amino acid residues (Å^2)[(g)]
X_{20}	*(GRAVY*) grand average of hydropathy of the proteins present in the asymmetric unit[(h)]
x_{21}	*(Tectp*) Total electric charge of the proteins in the asymmetric unit
x_{22}	*(Het/aa)* Ratio between the number of heteroatoms that are not water molecules and the number of residue in the asymmetric unit.
x_{23}	*(Software)* Type of software used to refine the structure[(i)]

[(a)] Space groups were labelled with a tag, which must not be intended as an integer variable, as it follows: 0 (P 21 21 21); 1 (P 43 2 2); 2 (C 2 2 21); 3 (P 1 21 1); 4 (C 1 2 1); 5 (P 21 21 2); 6 (I 4); 7 (P 65); 8 (P 32); 9 (I 4 2 2); 10 (P 43 21 2); 11 (I 2 2 2); 12 (P 1); 13 (P 61 2 2); 14 (P 4 3 2); 15 (H 3 2); 16 (P 41 21 2); 17 (P 4 2 2); 18 (P 31 2 1); 19 (P 32 2 1); 20 (I 21 3); 21 (P 43); 22 (P 3); 23 (P 63); 24 (P 62); 25 (F 2 2 2); 26 (P 41); 27 (P 61); 28 (P 65 2 2); 29 (I 41 2 2); 30 (I 41); 31 (I 2 3); 32 (P 4 21 2); 33 (P 42 2 2); 34 (I 4 3 2); 35 (P 42 3 2); 36 (P 3 2 1); 37 (H 3); 38 (P 6 2 2); 39 (P 6); 40 (P 63 2 2); 41 (P 64); 42 (P 42 21 2); 43 (P 32 1 2); 44 (P 41 2 2); 45 (P 31); 46 (C 2 2 2); 47 (P 31 1 2); 48 (P 21 3); 49 (P 62 2 2); 50 (P 64 2 2); 51 (P 1 2 1); 52 (P 43 3 2); 53 (P 2 3); 54 (I 41 3 2); 55 (I 21 21 21); 56 (F 4 3 2); 57 (P 2 2 21); 58 (P 42); 59 (P 41 3 2); 60 (F 2 3); 61 (P 4); 62 (P 3 1 2); 63 (F 41 3 2); 64 (P 2 2 2); 65 (P -1); 66 (I 1 2 1); 67 (P 2 21 21); 68 (P 1 1 21); 69 (I 41/a); 70 (P 21 2 21); 71 (P 1 21/c 1); 72 (P 21 2 2); 73 (I -4 2 d).

[b] Only the first conformation was considered in case of conformational disorder; hydrogen atoms were disregarded.

[c] The absence of a residues was deduced from the "REMARK 465" lines of the PDB file.

[d] Cell volume and percentage of solvent in the crystal were computed with the routine RWCONTENTS of the CCP4 software suite [Ref_1]

[e] When in a PDB file there are two or more proteins with different oligomeric states (for example a dimer and a trimer) the highest oligomeric state was arbitrarily retained (the trimeric in this example).

[f] Secondary structure assignments were done with Stride [Ref_2]; all helical types (α,π and 3_{10}) were considered into a single category; all extended structures (E, B or b) were considered in a single category; all the non-helical and non-extended secondary structures were considered to be a single category.

[g] Solvent accessible surface areas were computed with Stride [Ref_2], according with the algorithm published by Eisenhaber and Argos [Ref_3].

[h] Hydrophobicity values were taken from reference [Ref_4].

[i] This information was extracted automatically from the PDB files. Six types of software packages were observed (REFMAC, PHENIX, CNS, BUSTER, X-PLOR, SHELXL). When more than a single software is reported in the PDB file, only the first was considered.

[Ref_1] Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel E.B., Leslie, A.G., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A., Wilson, K.S. (2011) Acta Crystallogr. D, 67, 235-242.

[Ref_2] Frishman, D., Argos, P. (1995) Knowledge-based protein secondary structure assignment. Proteins 23, 566-579.

[Ref_3] Eisenhaber, F., Argos, P. (1993). Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. J. Comput. Chem. 14, 1272-1280.

[Ref_4] Carugo, O. (2003) Prediction of Polypeptide Fragments Exposed to the Solvent. In Silico Biology 3, 0035.

Table S2 Spearman correlation coefficients between all pairs of independent variables.

Variable	*Res*	*R*	*Solv%*	*Bave*	*aaLoops%*	*Sasaaa*	*GRAVY*
Res	/	/	/	/	/	/	/
R	0.602 p < .0001	/	/	/	/	/	/
Solv%	0.507 p < .0001	0.310 p < .0001	/	/	/	/	/
Bave	0.7723 p < .0001	0.580 p < .0001	0.498 p < .0001	/	/	/	/
aaLoops%	- 0.059 p < .0001	- 0.145 p < .0001	- 0.020 p = 0.0442	- 0.113 p < .0001	/	/	/
Sasaaa	0.0462 p < .0001	0.236 p < .0001	0.050 p < .0001	0.140 p < .0001	- 0.201 p < .0001	/	/
GRAVY	0.0027 p = 0.7886	- 0.029 p = 0.0037	0.0367 p < .0001	- 0.011 p = 0.2595	0.037 p = .0002	- 0.304 p < .0001	/
Het/aa	- 0.1013 p < .0001	- 0.240 p < .0001	- 0.058 p < .0001	- 0.109 p < .0001	0.094 p < .0001	0.026 p = 0.0111	- 0.063 p < .0001
Tectp	0.0361 p = 0.0003	0.0365 p = 0.0003	0.010 p = 0.3069	0.0385 p = .0001	0.042 p < .0001	- 0.037 p = 0.0003	- 0.005 p = 0.6305

Table S3 Distribution-free Tolerance intervals for in-sample predicted H_2O/aa ratio (Murphy, 1948). A tolerance interval indicates the range in which one is expecting to find a certain percentage of data (the coverage); here, we adopted coverages of 50%, 75% and 90%. This interval refers to the reference population (in our case, all the structures) but it is inferred from sample data by taking into account sampling errors (in our case, at a confidence level of 95%).

Coverage	**Tolerance Interval***
50%	0.3641 – 0.8803
75%	0.2403 – 1.1266
90%	0.1583 – 1.4304

* Confidence level: 95%

Figure S1 Graphical representation of bivariate relationships between observed *H_2O/aa* ratio and other variables.

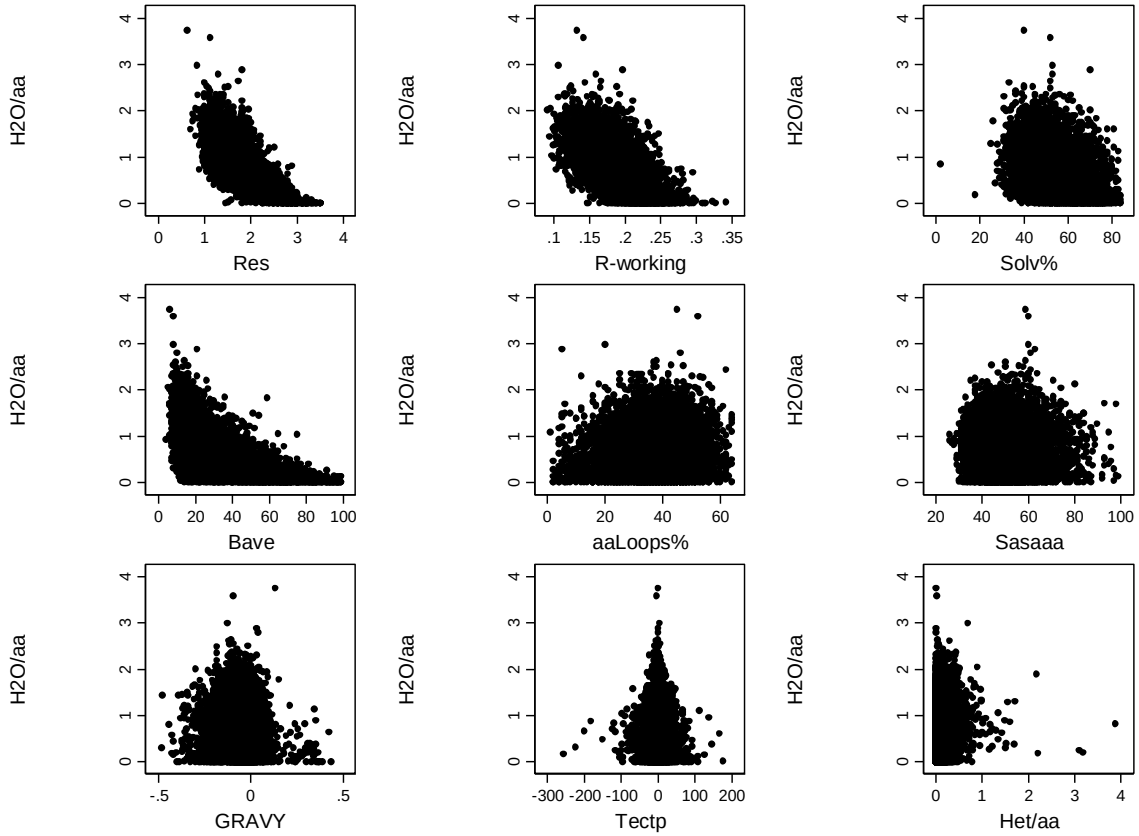

Figure S2 Predicted values of H_2O/aa ratios at the observed GRAVY values within the "training dataset".

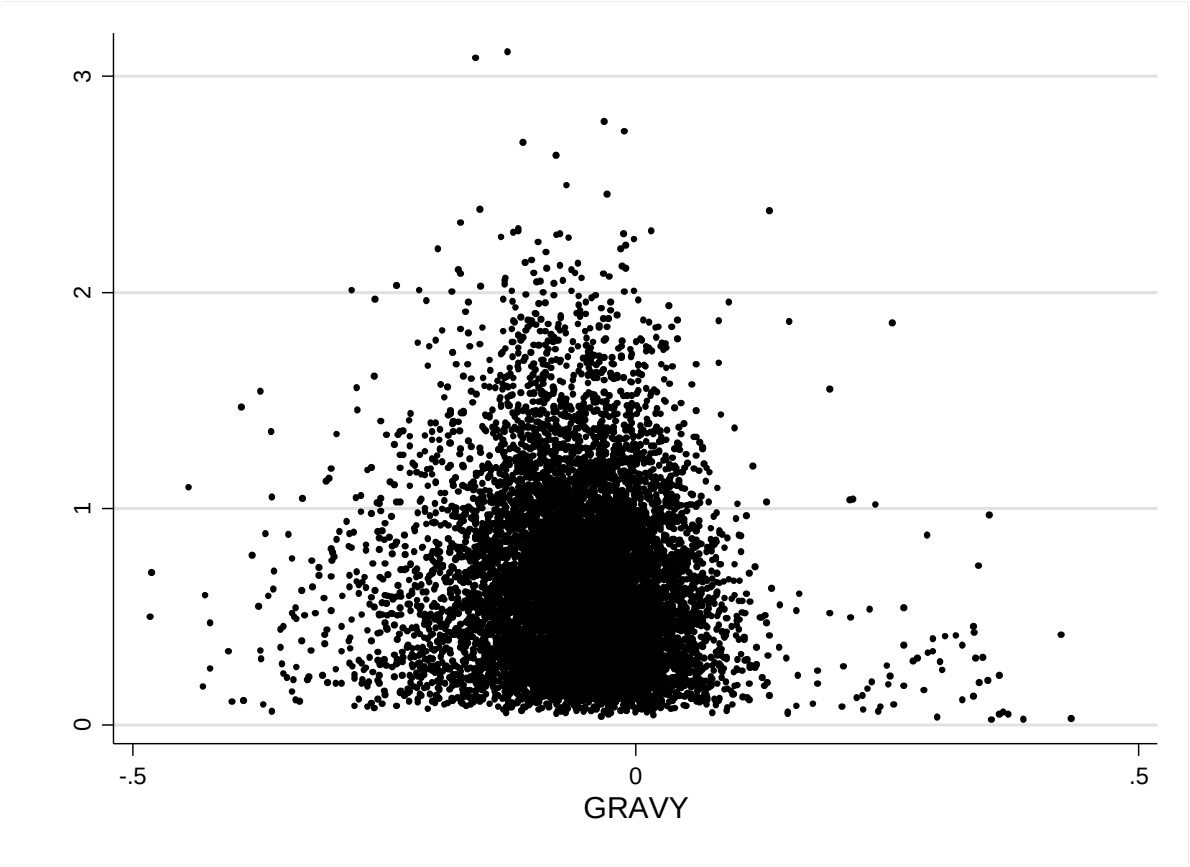

Figure S3 Histograms showing the distributions of the eight independent variables that allows the estimation of the relative number of water molecules observed in protein crystal structures: resolution (*Res*), R-factor (*R*), percentage of solvent in the crystal (*Solv%*), average B-factor of the protein atoms (*Bave*), percentage of amino acid residues in loops (*aaLoops%*), average solvent accessible surface area of the amino acid residues (*Sasaaa*), grand average of hydropathy of the protein(s) in the asymmetric unit), and normalized number of heteroatoms that are not water molecules (*Het/aa*). These distributions must be considered to estimate the expected reliability of the predictions, which is higher in the middle zone of the distributions and lower outside the distributions range.

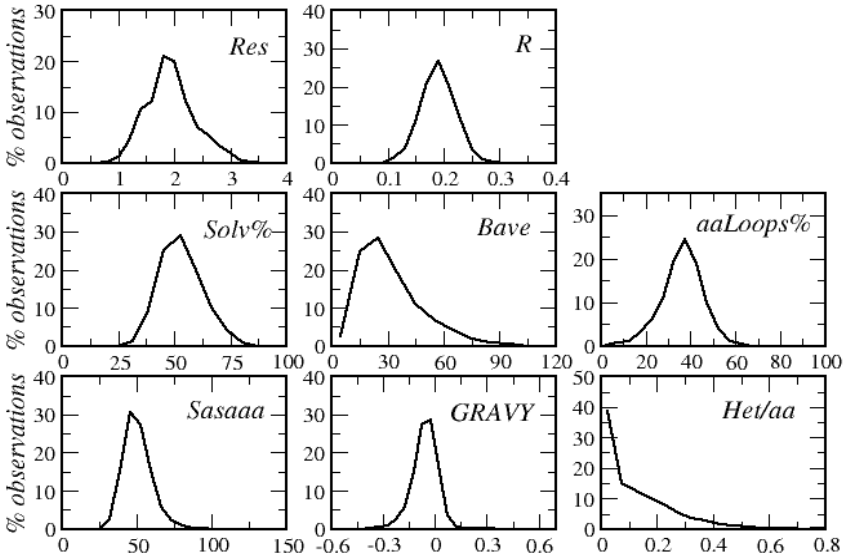

Figure S4 Predicted vs. observed values of the H2O/aa ratio within the "test dataset" (external validation)