The oxygen-oxygen distance of water in crystallographic data sets

Luigi Leonardo Palese

University of Bari “Aldo Moro”, Department of Basic Medical Sciences, Neurosciences and Sense Organs (SMBNOS), Bari, 70124, Italy

Article info

Article history:
Received 29 November 2019
Received in revised form 20 December 2019
Accepted 23 December 2019
Available online 7 January 2020

Keywords:
Water
Proton
X-ray crystallography
Water cluster
Confinement
Protein surface

Abstract

Water is a key component of cellular biochemistry and numerous water molecules are visible in crystallographic structures. Here we report a series of data sets of crystallographic water: a high resolution data set, a cytochrome c oxidase (subunit I) data set and a carbonic anhydrase data set. These data support the evidence that short distance water molecule pairs are present both at the surface and inside the cavities of proteins. These data are related to an article entitled “Oxygen-oxygen distances in protein-bound crystallographic water suggest the presence of protonated clusters” (Palese, 2020) [1].

© 2020 The Author. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Data

Fig. 1 reports the atomic radial pair distribution function (RDF) in the region between 2.15 and 2.85 Å of the high resolution (HR) data set and of the HR subset not refined by SHELX (see Ref. [1]).

Fig. 2 reports the RDF for the water oxygen atoms in the sodium free HR data set [1].

Fig. 3 shows the RDF of the water oxygen atoms in the sodium free and not refined by SHELX subset described in Ref. [1] (see also Table 4 below).

DOI of original article: https://doi.org/10.1016/j.bbagen.2019.129480.
E-mail address: luigileonardo.palese@uniba.it.
Fig. 4 shows the scatter plot of the RDFs relative to the no SHELX subset and the sodium free, no SHELX subset [1]. The Figure reports also the line of best fit, whose equation is $y = 1.0191x - 0.0001 (R^2 = 0.9966)$.

Fig. 5 reports the number of water molecules in cytochrome c oxidase (CcO) subunit I vs the structure resolution. Data have been obtained from the PDB entries 5B1A, 5B1B, 5B3S, 5XDQ, 5ZCP and 5ZCQ (diffraction temperature 50 K; resolution 1.5 Å, 1.6 Å, 1.68 Å, 1.77 Å, 1.65 Å, 1.65 Å, respectively).

---

**Value of the Data**

- X-ray crystallography has shown that proteins contain numerous water molecules
- The role of these protein-bound water molecules is still not fully understood
- Water molecules in large data sets of high resolution structures are reported
- Positions of candidate protonated water clusters in some model proteins are reported
and 1V54, 2DYR, 3AG2 and 3AG3 (diffraction temperature 100 K; all these structures have a resolution of 1.8 Å). Both subunits of the same type have been considered, labeled as A and N in the original pdb file.

Figs. 6 and 7 report the Euclidean distance distribution of oxygen-oxygen (O–O) pairs for the data set of CcO structures obtained at 50 K and 100 K, respectively. Two major peaks are present in both distributions, centered at 2.73 and 2.88 Å.

Table 1 reports the pdb codes for the entries of the HR data set [1], containing 469 elements.
Table 2 reports the pdb codes of the subset of the HR data set containing structures not refined by SHELX (the no-SHELX HR data set discussed in Ref. [1]).

Table 3 reports the pdb codes of the subset of the HR data set containing structures in which no sodium is declared in the crystallization methods (the sodium free HR data set discussed in Ref. [1]).

Table 4 reports the subset of the HR data set containing structures not refined by SHELX and in which the use of sodium in the crystallization conditions is not reported.

Table 5 reports the statistics of water pairs analyzed in human carbonic anhydrase II (hCA II), as detailed in Ref. [1].

Fig. 4. Scatter plot of the RDFs relative to the no SHELX subset (horizontal axis) and the sodium free, no SHELX subset (vertical axis) described in Ref. [1]. Solid black line is the best fit.

Fig. 5. Number of crystallographic water molecules in CcO subunit I structures as a function of the resolution of the relative PDB entry. Structures diffracted at 100 K and 50 K have been considered for this analysis.
Table 6 reports the statistics of water pairs analyzed in subunit I of bovine CcO, as detailed in Ref. [1]. Table 7 contains the pseudo-code for the RDF calculations (see the Experimental Design, Materials, and Methods section).

In the water_pairs.csv file in Supplementary data all the short distance water molecules (SDWMs) in the model proteins discussed in Ref. [1] are listed. These are pairs of water molecules whose O–O distance is in the range 2.29 Å - 2.50 Å, confirmed by inspection of the electron density maps at 1.0 and 4.0 sigma as described in Ref. [1]. Columns in this csv file are: the PDB id of the molecule, the residue number in the original pdb file of the two water molecules in the pair (two columns), and the O–O distance in this water pair obtained from the pdb coordinates. Water molecules in Figs. 2 and 4 in Ref. [1] are listed in this file.

Fig. 6. Distribution of the Euclidean oxygen-oxygen (O–O) distances of water molecules in the CcO subunit I data set. Data are obtained from the 50 K structures. Only water pairs with O–O distances <5 Å are reported.

Fig. 7. Distribution of the Euclidean oxygen-oxygen (O–O) distances of water molecules in the CcO subunit I data set. Data are obtained from the 100 K structures. Only water pairs with O–O distances <5 Å are reported.
In the Supplementary file raw_data.zip there are the raw data used in this work. The files contained in this zipped archive are:

- raw_data.txt, raw_RDF.csv, raw_RDF_CcO.csv, water_resolution.csv, 50K_distance.txt, 100K_distance.txt, which are described in detail below.

- The file raw_data.txt lists the URLs for all the *.pdb, *.cif and *.ccp4 files considered in this work.
- The file raw_RDF.csv contains all the raw RDF data used for calculating the plots shown in Figs. 1–4; these data were also used for the calculation of Fig. 1 in Ref. [1]. Each row corresponds to a crystallographic structure (PDB codes are in the first column).
- The file raw_RDF_CcO.csv contains all the raw RDF data used for calculating the plot reported as Fig. 3 in Ref. [1]. Each row corresponds to a subunit I structure (A and N subunits in the PDB entries reported in the first column).
- The file water_resolution.csv contains the number of water molecules in the CcO data set; these are the raw data for Fig. 5.
- The file 50K_distance.txt reports all the Euclidean O–O distances of crystallographic water molecules in subunits I of CcO structures obtained by X-ray diffraction at 50 K (A and N subunits in the PDB entries 5B1A, 5B1B, 5B3S, 5DXQ, 5ZCQ and 5ZGZ). Distances <5 Å, but above the reported experimental resolution in the relative PDB record, were considered. These are the raw data used to calculate the distribution shown in Fig. 6.
- The file 100K_distance.txt reports all the Euclidean O–O distances of crystallographic water molecules in subunits I of CcO structures obtained by X-ray diffraction at 100 K (A and N subunits in the PDB entries reported in Table 1).
entries 1V54, 2DYR, 3AG2 and 3AG3). Distances < 5 Å, but above the reported experimental resolution in the relative PDB record, were considered. These are the raw data used to calculate the distribution shown in Fig. 7.

2. Experimental Design, materials, and methods

The X-ray structures were obtained from the Protein Data Bank (PDB) [2] (Tables 1 e 4; direct URLs to these data are in the raw_data.txt file in the Supplementary data). The HR data set was obtained by searching in the PDB for structures corresponding to the following constraints: resolution/C20 1 Å; X-ray only; protein only; monomer only. After this, we considered only structures whose diffraction pattern was obtained at 100 K. Some entries have been discarded at the radial distribution function (RDF) calculation stage (typically small structures with few water molecules that give anomalous RDF). After these steps, the HR data set contained 469 entries. Two subset have been obtained from the HR data set (Table 2).

Table 2
The no-SHELX HR data set.

| Entry | Code |
|-------|------|
| 1EB6  | 1OA1 |
| 2AT8  | 2AYW |
| 2VFR  | 2PDK |
| 20Q7  | 2ZQA |
| 32YQ  | 3IO3 |
| 3QL9  | 3QMS |
| 3VIG  | 3VNC |
| 4AQO  | 4B70 |
| 4HVU  | 4HVE |
| 4MTY  | 4M2C |
| 4R5R  | 4REK |
| 4XQJ  | 4XZH |
| 5CMT  | 5DGD |
| 5HB7  | 5IG6 |
| 5LPW  | 5LXW |
| 5N13  | 5Q5A |
| 5TIF  | 5U3A |
| 5ZGW  | 5ZGY |
| 6FI0  | 6FMC |
| 6QZY  | 6Q49 |
| 7A3H  | 8A3H |

Table 3
The sodium free HR data set.

| Entry | Code |
|-------|------|
| 180Y  | 1DY5 |
| 1MTQ  | 1MCJ |
| 1TGO  | 1T8G |
| 1ZK2  | 1AZG |
| 2FYV  | 2GEW |
| 2FPH  | 2PVE |
| 3C78  | 3D1P |
| 3134  | 3JUD |
| 3QL9  | 3QPA |
| 3VII  | 3VQJ |
| 4ACJ  | 4AQ0 |
| 4GA2  | 4CA6 |
| 4MG7  | 4M6U |
| 4TKH  | 4TKJ |
| 5A8C  | 5B28 |
| 5DKC  | 5JIG |
| 52ZG  | 5GZG |
| 6NFR  | 6Q49 |

2.3. Results

The X-ray structures were obtained from the Protein Data Bank (PDB) [2] (Tables 1 e 4; direct URLs to these data are in the raw_data.txt file in the Supplementary data). The HR data set was obtained by searching in the PDB for structures corresponding to the following constraints: resolution ≤ 1 Å; X-ray only; protein only; monomer only. After this, we considered only structures whose diffraction pattern was obtained at 100 K. Some entries have been discarded at the radial distribution function (RDF) calculation stage (typically small structures with few water molecules that give anomalous RDF). After these steps, the HR data set contained 469 entries. Two subset have been obtained from the HR data set.
by adding additional constraints: (a) the absence of sodium in all buffer and solution declared in the deposited methods (the sodium free HR data set), or (b) the absence of the SHELX program in the software reported in the deposited refinement methods (the no SHELX HR data set). The entries in this last data set reported as hCA II are those considered as models for this enzyme (see Ref. [1]). From the HR data set, a further subset has been obtained, containing only entries obtained in absence of sodium in the crystallization protocol and not refined by SHELX.

### Table 4
The sodium free, no-SHELX HR data set.

| Entry | #H2O | #OeO | #OHO |
|-------|------|------|------|
| 1RZM  | 398  | 439  | 3     |
| 1SY2  | 214  | 333  | 1     |
| 1SY3  | 264  | 96141| 1     |
| 1UG6  | 439  | 373  | 9     |
| 1V0L  | 96141| 441  | 429   |
| 1W0N  | 333  | 441  | 429   |
| 2AT3  | 267  | 97020| 429   |
| 2AT8  | 12   | 91806| 429   |
| 2AYW  | 1    | 83436| 429   |
| 2CNQ  | 3    |       | 429   |
| 2GG2  | 10   |       | 429   |
| 2GGC  | 2    |       | 429   |
| 2H3L  | 1    |       | 429   |

### Table 5
The number of water molecules in each considered hCA II is reported (#H2O). The table reports also the number of calculated oxygen-oxygen Euclidean distances (# OeO) and the number of oxygen-oxygen pairs considered as putative, charged (protonated or deprotonated), water clusters as defined in Ref. [1] (# OHO).

| Structure | #H2O | #OeO | #OHO |
|-----------|------|------|------|
| 4MTY      | 398  | 439  | 3     |
| 4Q78      | 214  | 333  | 1     |
| 4YXI      | 264  | 96141| 1     |
| 5JLT      | 439  | 373  | 9     |
| 5Ogo      | 96141| 441  | 429   |
| 5Y2R      | 333  | 441  | 429   |
| 5Y2S      | 373  | 97020| 429   |
| 6B00      | 441  | 91806| 429   |

### Table 6
The number of water molecules in each considered CcO subunit I is reported (#H2O). The table reports also the number of calculated oxygen-oxygen Euclidean distances (# OeO) and the number of oxygen-oxygen pairs considered as putative, charged (protonated or deprotonated), water clusters as defined in Ref. [1] (# OHO).

| Structure | #H2O | #OeO | #OHO |
|-----------|------|------|------|
| 5B1A_A    | 297  | 286  | 2     |
| 5B1A_N    | 290  | 273  | 2     |
| 5B1B_A    | 289  | 257  | 2     |
| 5B1B_N    | 286  | 267  | 2     |
| 5B3S_A    | 273  | 235  | 2     |
| 5B3S_N    | 286  | 203  | 2     |
| 5XDO_A    | 273  | 228  | 2     |
| 5XDO_N    | 257  | 215  | 2     |
| 5ZCP_A    | 267  | 228  | 2     |
| 5ZCP_N    | 235  | 215  | 2     |
| 5ZCQ_A    | 203  | 235  | 2     |
| 5ZCQ_N    | 228  | 215  | 2     |

### Table 7
The RDF pseudo-code.

```plaintext
set num_mol [molinfo num]
for {set i 0} {$i <$num_mol} {incr i} {
    set name_prot [molinfo $i get name]
    set sel [atomselect $i “water”]
    set gr [measure gofr $sel $sel]
    set outfile [open gofr_$name_prot.dat w]
    set r [lindex $gr 0]
    set gr1 [lindex $gr 1]
    set igr [lindex $gr 2]
    foreach j $r k $gr1 l $igr {
        puts $outfile "$j $k $l"
    }
    close $outfile
}
```

by adding additional constraints: (a) the absence of sodium in all buffer and solution declared in the deposited methods (the sodium free HR data set), or (b) the absence of the SHELX program in the software reported in the deposited refinement methods (the no SHELX HR data set). The entries in this last data set reported as hCA II are those considered as models for this enzyme (see Ref. [1]). From the HR data set, a further subset has been obtained, containing only entries obtained in absence of sodium in the crystallization protocol and not refined by SHELX.
For the bovine CcO data set, only structures with a resolution of at least 1.8 Å were considered for analysis. The high resolution data set of CcO contained the PDB entries (see Ref. [1]): 5B1A (fully oxidized state, pH 6.8), 5B1B (fully reduced state, pH 6.8), 5B3S (carbon monoxide-bound mixed-valence, pH 6.8), 5XDQ (fully oxidized state, pH 7.3), 5ZCP and 5ZCQ (azide-bound states obtained by long time exposure to 20 mM or 10 mM azide solutions, respectively, pH 6.8). All these structures, obtained at 50 K, are characterized by a resolution between 1.65 and 1.50 Å (5XDQ has been considered here, even if its resolution is 1.77 Å, because it is the structure characterized by the higher resolution at alkaline pH). We have also considered a data set of structures obtained at 100 K, with a resolution of 1.8 Å: 1V54, 2DYR, 3AG2 and 3AG3, which are respectively fully oxidized, fully reduced, carbon monoxide-bound fully reduced and nitric oxide-bound fully reduced species, all at pH 6.8 [3–5].

The data sets containing a single type of protein were analyzed in detail as specified below. From the pdb files, the atomic coordinates of all atoms belonging to the subunit of interest were used to make a new pdb file (consider that bovine CcO is in dimeric form in crystals, and the type I subunits are labeled as A and N in the files retrieved from the PDB). Sequence and structural analyses have been performed as described previously [1,6–8]. The mutual Euclidean distance between all the oxygen atoms of the water molecules contained into the protein, at the protein surface or near lipids [9] was calculated by means of a Tcl program in a VMD environment [10]. The RDF for each pdb file has been calculated in VMD using a code similar to that reported in Table 7, and further analyzed in a Jupyter environment (see below). Electron density maps at 1.0 and 4.0 sigma of the model proteins discussed in Ref. [1] were obtained using the *.cif and *.ccp4 files in Jmol (http://www.jmol.org/). The URLs for these files are reported in the raw_data.txt file in the Supplementary data where * is the PDB code of the protein of interest (4mty, 4q78, 4yxi, 5ljt, 5ogo, 5y2r, 5y2s, 6b00, 5b1a, 5b1b, 5b3s, 5xdq, 5zcp, 5zcq).

The mean displacement of atoms U was calculated considering that the B-factor is given by $B = 8\pi^2 U^2$.

Numerical calculations were implemented in Python (www.python.org) in an IPython/Jupyter environment, using the NumPy numerical software library, the Scipy and the Matplotlib packages [11–14]. Final editing of images was performed by means of the GNU Image Manipulation Program (The GIMP team, GIMP 2.8.10, www.gimp.org) or the ImageMagik (imagemagick.org) software packages.

Acknowledgments

This work has been supported by the fund for research at the University of Bari.

Conflict of Interest

The author declares that he has no known financial interests or competing personal relationships that could have appeared influence the work reported in this document.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.105076.

References

[1] L.L. Palese, Oxygen-oxygen distances in protein-bound crystallographic water suggest the presence of protonated clusters, Biochim. Biophys. Acta 1864 (2020) 129480, https://doi.org/10.1016/j.bbadis.2019.129480.
[2] H.M. Berman, J. Westbrook, Z. Feng, et al., The protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242, https://doi.org/10.1093/nar/28.1.235.
[3] T. Tsukihara, K. Shimokata, Y. Katayama, et al., The low-spin heme of cytochrome c oxidase as the driving element of the proton-pumping process, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 15304–15309, https://doi.org/10.1073/pnas.2635097100.
[4] K. Shinzawa-Itoh, H. Aoyama, K. Muramoto, et al., Structures and physiological roles of 13 integral lipids of bovine heart cytochrome c oxidase, EMBO J. 26 (2007) 1713–1725, https://doi.org/10.1038/sj.emboj.7601618.
[5] K. Muramoto, K. Ohta, K. Shinzawa-Itoh, et al., Bovine cytochrome c oxidase structures enable O2 reduction with minimization of reactive oxygens and provide a proton-pumping gate, Proc. Natl. Acad. Sci. U.S.A. 107 (2010) 7740–7745, https://doi.org/10.1073/pnas.091040107.
[6] F. Bossis, A. De Grassi, L.L. Palese, C.L. Pierri, Prediction of high- and low-affinity quinol-analogue-binding sites in the aa3 and bo3 terminal oxidases from Bacillus subtilis and Escherichia coli, Biochem. J. 461 (2014) 305–314, https://doi.org/10.1042/B20140082.

[7] L.L. Palese, Protein states as symmetry transitions in the correlation matrices, J. Phys. Chem. B 120 (2016) 11428–11435, https://doi.org/10.1021/acs.jpcb.6b09216.

[8] L.L. Palese, Conformations of the HIV-1 protease: a crystal structure data set analysis, Biochim. Biophys. Acta 1865 (2017) 1416–1422, https://doi.org/10.1016/j.bbapap.2017.08.009.

[9] S. Lobasso, L.L. Palese, R. Angelini, A. Corcelli, Relationship between cardiolipin metabolism and oxygen availability in Bacillus subtilis, FEBS Open Bio 3 (2013) 151–155, https://doi.org/10.1016/j.fob.2013.02.002.

[10] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38, https://doi.org/10.1016/0263-7855(96)00018-5.

[11] F. Pérez, B.E. Granger, IPython: a system for interactive scientific computing, Comput. Sci. Eng. 9 (2007) 21–29, https://doi.org/10.1109/MCSE.2007.53.

[12] T.E. Oliphant, Python for scientific computing, Comput. Sci. Eng. 9 (2007) 10–20, https://doi.org/10.1109/MCSE.2007.58.

[13] S. Van Der Walt, S.C. Colbert, G. Varoquaux, The NumPy array: a structure for efficient numerical computation, Comput. Sci. Eng. 13 (2011) 22–30, https://doi.org/10.1109/MCSE.2011.37.

[14] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.