Data in brief 28 (2020) 105076

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

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

A R T I C L E I N F O

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
Confined water
Protein surface

A B S T R A C T

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 article entitled "Oxygen-oxygen distances in protein-bound crystallographic water suggest the presence of protonated clusters" (Palese, 2020) [1].

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.

https://doi.org/10.1016/j.dib.2019.105076

Specifications Table

Subject	Biochemistry
Specific subject area	Biochemistry, biophysics, structural biology, bioenergetics.
Type of data	Table Figure Graph Text files
How data were acquired	Survey of the protein crystal structures obtained by X-ray diffraction deposited in the Protein Data Bank (PDB). Input data for analysis were obtained as pdb, cif or ccp4 files from public databases.
Data format	Raw: pdb files, ccp4 files, cif files and text files. Analyzed: table, csv files, text file, graph, figure.
Parameters for data collection	Raw pdb files were checked for quality (resolution). Atoms in cif files were compared with the electron density maps in ccp4 files.
Description of data collection	Raw data were analyzed by different computational protocols.
Data source location	Department of Basic Medical Sciences, Neurosciences and Sense Organs (SMBNOS), Bari 70124, Italy.
Data accessibility	All the data produced in this work are available within the article.
Related research article	Luigi Leonardo Palese, Oxygen-oxygen distances in protein-bound crystallographic water suggest the presence of protonated clusters, Biochemical et Biophysical Acta - General Subjects, https://doi.org/10.1016/j.bbagen.2019.129480.

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

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)

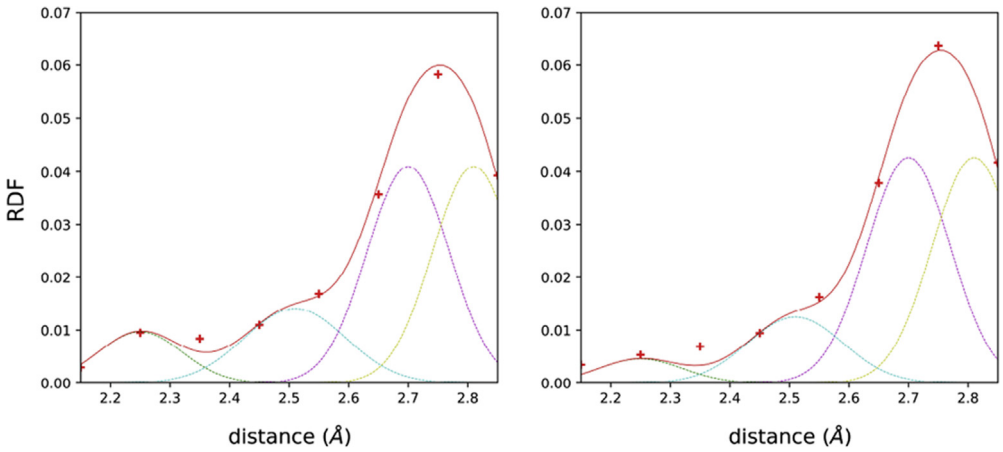

Fig. 1. The RDF of the protein-bound water oxygen atoms. Left panel reports the HR data set (see Table 1); right panel reports the HR data set not refined by SHELX (see Table 2). See Ref. [1] for details. Red crosses refer to the experimental RDF; colored curves indicated with dashed lines are the Gaussian curves used for the fitting, and their sum is reported as a thin red line.

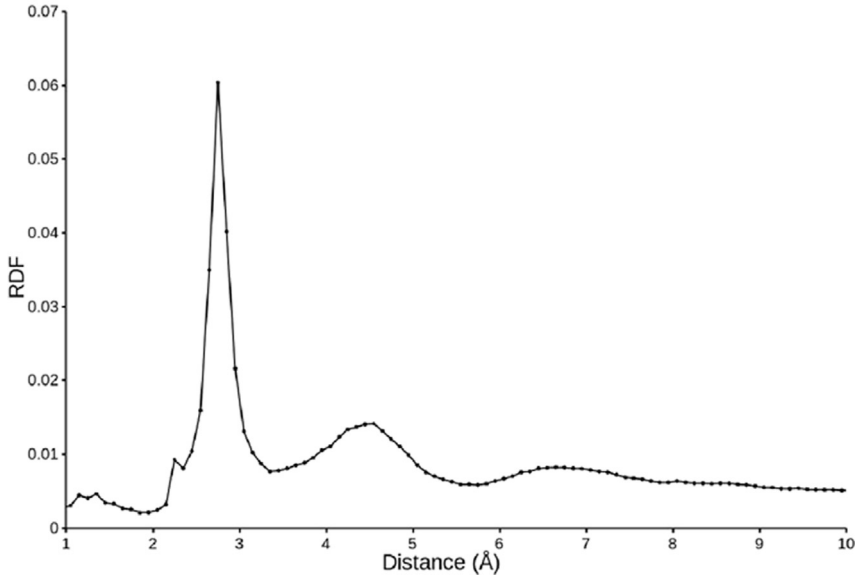

Fig. 2. The RDF of the water oxygen atoms in the sodium free HR data set (see Table 3) described in Ref. [1].

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.

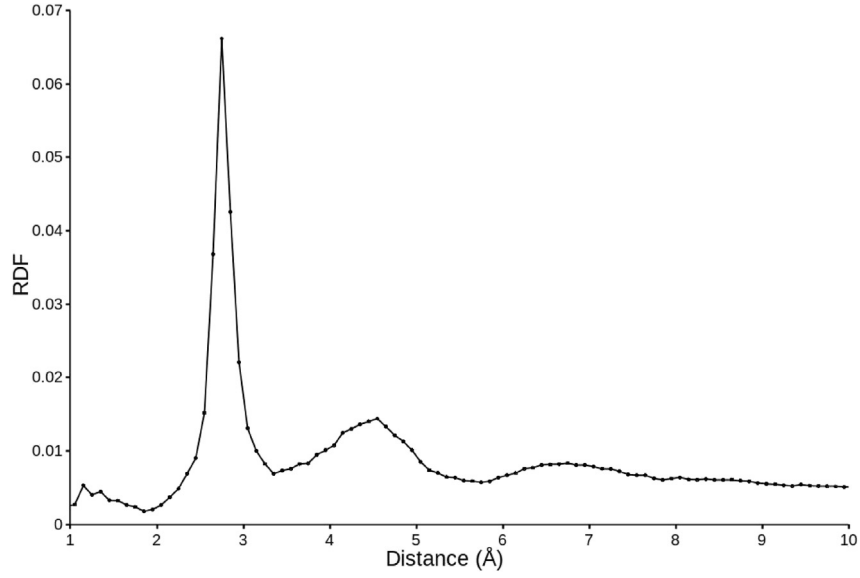

Fig. 3. The RDF of the water oxygen atoms in the sodium free, not refined by SHELX subset (see Table 4) described in Ref. [1].

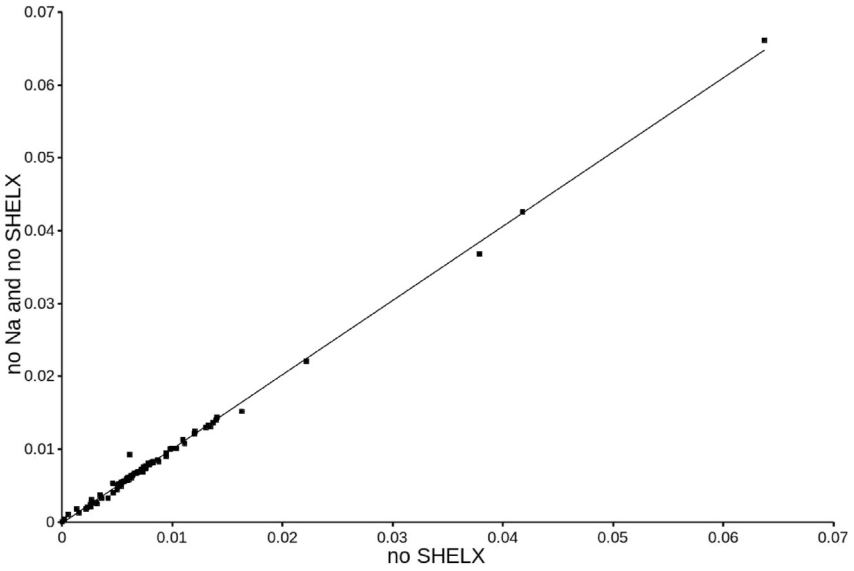

Fig. 4. Scatter plot of the RDFs relative to the no SHELX subset (horizontal axis) and the sodium free, no SHELX subset (vertical axis) described in Ref. [1]. Solid black line is the best fit.

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].

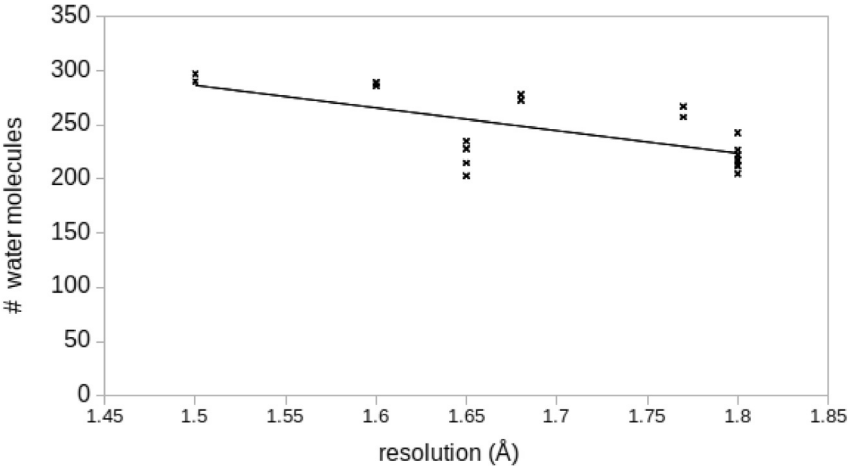

Fig. 5. Number of crystallographic water molecules in CcO subunit I structures as a function of the resolution of the relative PDB entry. Structures diffracted at 100 K and 50 K have been considered for this analysis.

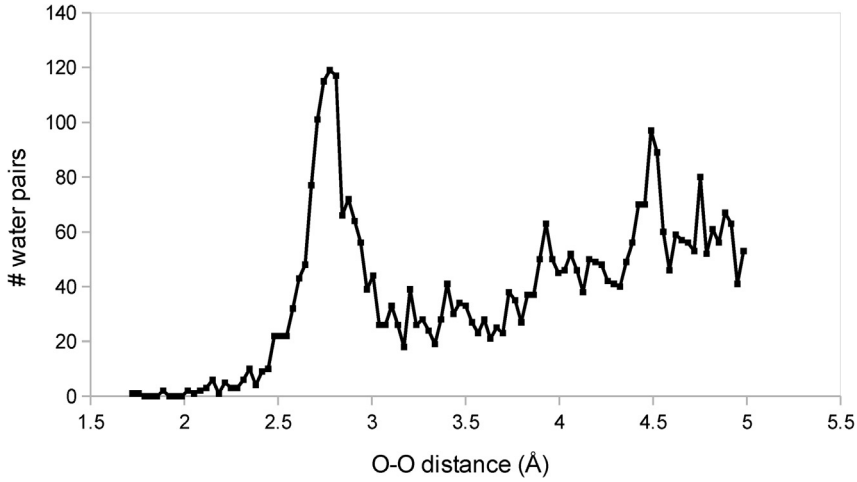

Fig. 6. Distribution of the Euclidean oxygen-oxygen (O—O) distances of water molecules in the CcO subunit I data set. Data are obtained from the 50 K structures. Only water pairs with O—O distances <5 Å are reported.

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.

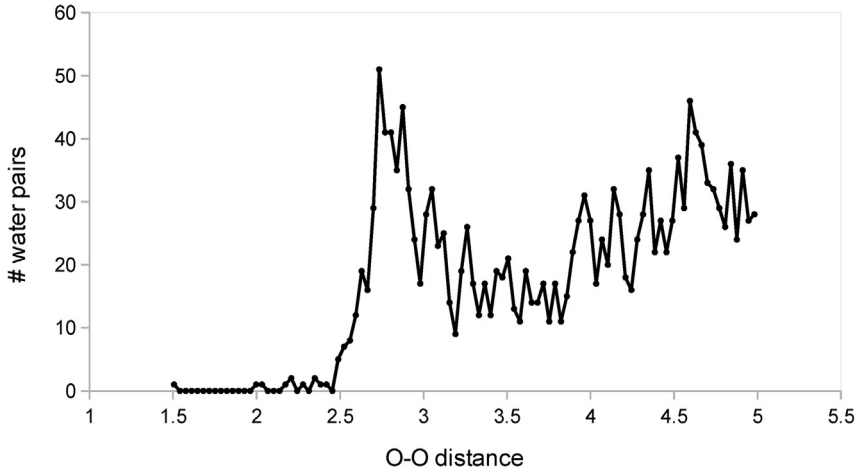

Fig. 7. Distribution of the Euclidean oxygen-oxygen (O—O) distances of water molecules in the CcO subunit I data set. Data are obtained from the 100 K structures. Only water pairs with O—O distances <5 Å are reported.

Table 1
The HR data set.

1A6M	1B0Y	1BXO	1C75	1DY5	1EA7	1EB6	1EXR	1F94	1FN8	1FY4	1FY5	1G4I
1G6X	1GA6	1GCI	1GDN	1GDQ	1GQV	1GVK	1HJ8	1I1W	1IQZ	1IR0	1IUA	1J0P
1JFB	1K2A	1K6U	1KWF	1L9L	1LNI	1LUG	1M1Q	1M40	1 MC2	1MJ5	1MN8	1MUW
1MXT	1N1P	1N4U	1N4V	1N4W	1N9B	1NQJ	1OAI	1OD3	1OEW	1OK0	1PQ5	1PQ7
1PQ8	1PWM	1R2M	1R6J	1RTQ	1SSX	1SY2	1SY3	1T2H	1TG0	1TQG	1TT8	1UG6
1UNQ	1US0	1V0L	1VB0	1VBW	1VL9	1VYR	1W0N	1X8P	1XMK	1XVO	1YK4	1YLJ
1YWA	1YWB	1YWC	1Z8A	1ZLB	1ZUU	1ZZK	2A6Z	2ABB	2AGT	2AT3	2AT8	2AYW
2B97	2BF6	2CE2	2CNQ	2CWS	2DDX	2E4T	2EWI	2EWK	2FDN	2FLA	2FMA	2FOU
2FVY	2GBA	2GEW	2GG2	2GGC	2GKG	2H3L	2H5C	2H5D	2I4A	2IDQ	2IDS	2IDT
2IDU	2IXT	2JFR	2JJJ	2NLS	2NRL	2NX0	2O7A	2O9S	2OFR	2OV0	2P5K	2P74
2PEV	2PF8	2PFH	2PND	2PNE	2PVB	2PVE	2PYA	2PZN	2QCP	2QDV	2QDW	2QSK
2RBK	2V1M	2V8B	2VB1	2VHK	2VHR	2VI3	2VK5	2VU6	2WFI	2WFJ	2WUR	2X46
2XFR	2XJP	2XOD	2XU3	2Z6W	2ZPM	2ZQ7	2ZQA	3A02	3A4R	3AJ4	3AKQ	3AKS
3AKT	3B3R	3BOI	3C78	3CNJ	3D1P	3D43	3DHA	3E4G	3E6Z	3EA6	3FSA	3FYM
3G9X	3GHR	3GHS	3GOE	3GYI	3GYJ	3H31	3HGP	3I2Y	3I30	3I34	3I37	3IP0
3JUD	3K34	3KFF	3KLR	3KS3	3LEP	3LL1	3LL2	3LZ5	3M4H	3MFJ	3 MI4	3NE0
3NED	3NIR	3O5P	3O5Q	3ODV	3P8J	3PUC	3QL9	3QM5	3QM6	3QM8	3QM9	3QMA
3QPA	3QPC	3S6E	3SOJ	3TEU	3TRV	3U2C	3U7C	3UI4	3UI6	3V1A	3VHG	3VIF
3VIG	3VII	3VJQ	3VN3	3VOR	3WCQ	3WGE	3WGX	3WL2	3WOU	3WVM	3X2G	3X2M
3X32	3X33	3X34	3X35	3ZR8	3ZSJ	3ZSK	3ZUC	3ZZP	4A02	4ACJ	4AQO	4AWS
4AWT	4BCT	4BJ0	4BM8	4BVM	4DP9	4DPB	4DRQ	4E3Y	4EIC	4EKF	4F18	4FPT
4FRC	4FU5	4G78	4G9S	4GA2	4GCA	4GNR	4HGU	4HVU	4HVW	4I8G	4I8H	4I8J
4I8K	4I8L	4IAU	4IGS	4J5E	4JP6	4K8Y	4KQP	4LAU	4LAZ	4LB3	4LB4	4LBR
4LBS	4LFS	4M7G	4MTY	4MZC	4N1I	4NPD	4O6Q	4O6U	4O8H	4OO4	4PRT	4PSS
4PSY	4PTH	4Q4G	4Q78	4Q9W	4QB3	4QBX	4QXI	4R5R	4REK	4RTZ	4TJZ	4TKB
4TKH	4TKJ	4UA6	4UA7	4UA9	4UAA	4UYR	4WJX	4WKA	4WPK	4X5P	4X6H	4XDX
4XOJ	4XZH	4Y5L	4Y9W	4YXI	4Z8J	4ZC9	4ZGF	4ZM7	5A8C	5AVD	5AVG	5B28
5B5H	5CMT	5CTM	5D66	5DGJ	5DJ7	5DK1	5DKM	5DP2	5DZE	5E1K	5E1N	5E9N
5EMB	5F82	5FLK	5HB7	5I5B	5IG6	5II8	5IQN	5IVN	5JDK	5JDT	5JIG	5JUG
5KWM	5KXV	5LJT	5LP9	5LXW	5MAE	5MAJ	5MB5	5MEH	5MK9	5MN1	5MNC	5MNG
5MNK	5MNN	5MTU	5NFK	5NFM	5NW3	5O2X	5O99	5OAV	5OBK	5OGO	5OME	5OTN
5OU0	5OUJ	5OUK	5SV5	5TDA	5TIF	5U3A	5VLE	5WS7	5X9L	5X9M	5XBU	5XP6
5XQU	5XQV	5XR0	5Y2R	5Y2S	5YCE	5ZGE	5ZGI	5ZGW	5ZGX	5ZGY	5ZGZ	5ZIO
5ZJ1	5ZJ7	5ZJ8	5ZJC	6B00	6CNW	6EQE	6ETK	6ETM	6EUW	6F1O	6F7R	6F81
6FMC	6FO5	6FVI	6G1I	6HMD	6HMQ	6I74	6J93	6JJ1	6JJQ	6MU9	6NFR	6Q2Y
6Q49	6Q4G	6Q4H	6Q6T	6RGP	6RHH	6RHU	6RHX	6RI0	6RI6	6RI8	6RII	7A3H
8A3H												

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 crystallografic 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, 5XDQ, 5ZCP and 5ZCQ). 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 crystallografic water molecules in subunits I of CcO structures obtained by X-ray diffraction at 100 K (A and N subunits in the PDB

Table 2
The no-SHELX HR data set.

1EB6	1OAI	1OD3	1R2M	1SY2	1SY3	1T2H	1UG6	1V0L	1W0N	1ZUU	2ABB	2AT3
2AT8	2AYW	2CE2	2CNQ	2DDX	2GG2	2GGC	2H3L	2I4A	2IXT	2NLS	2NRL	2NX0
2OFR	2P5K	2PND	2PNE	2QCP	2V1M	2VHK	2VHR	2VI3	2VU6	2XFR	2XU3	2Z6W
2ZQ7	2ZQA	3A02	3A4R	3AJ4	3AKQ	3AKS	3AKT	3BOI	3C78	3E4G	3E6Z	3FSA
3I2Y	3I30	3I34	3I37	3IP0	3JUD	3LL1	3LL2	3NED	3NIR	3O5Q	3P8J	3PUC
3QL9	3QM5	3QM6	3QM8	3QM9	3QMA	3QPA	3QPC	3TEU	3UI4	3UI6	3V1A	3VIF
3VIG	3VII	3VN3	3WCQ	3WGE	3WGX	3WL2	3WVM	3X2G	3X2M	3ZR8	3ZUC	4ACJ
4AQO	4BCT	4BJ0	4BVM	4DP9	4DPB	4DRQ	4E3Y	4F18	4G9S	4GA2	4GCA	4GNR
4HVU	4HVW	4IGS	4J5E	4K8Y	4KQP	4LAU	4LAZ	4LB3	4LB4	4LBR	4LBS	4M7G
4MTY	4MZC	4O6Q	4O6U	4O8H	4OO4	4PRT	4Q4G	4Q78	4Q9W	4QB3	4QBX	4QXI
4R5R	4REK	4RTZ	4TJZ	4TKB	4TKH	4TKJ	4WJX	4WKA	4WPK	4X5P	4X6H	4XDX
4XOJ	4XZH	4Y5L	4Y9W	4YXI	4Z8J	4ZC9	4ZM7	5A8C	5AVD	5AVG	5B28	5B5H
5CMT	5CTM	5DGJ	5DJ7	5DK1	5DKM	5DP2	5DZE	5E1N	5E9N	5EMB	5F82	5FLK
5HB7	5I5B	5IG6	5II8	5IQN	5IVN	5JDK	5JDT	5JIG	5JUG	5KWM	5KXV	5LJT
5LP9	5LXW	5MAE	5MAJ	5MB5	5MEH	5MN1	5MNC	5MNG	5MNK	5MNN	5NFK	5NFM
5NW3	5O2X	5O99	5OAV	5OBK	5OGO	5OME	5OTN	5OU0	5OUJ	5OUK	5SV5	5TDA
5TIF	5U3A	5WS7	5XBU	5XP6	5XQU	5XQV	5XR0	5Y2R	5Y2S	5YCE	5ZGE	5ZGI
5ZGW	5ZGX	5ZGY	5ZGZ	5ZIO	5ZJ1	5ZJ7	5ZJ8	5ZJC	6B00	6CNW	6EQE	6EUW
6F1O	6FMC	6FO5	6FVI	6G1I	6HMD	6HMQ	6I74	6J93	6JJ1	6JJQ	6MU9	6NFR
6Q2Y	6Q49	6Q4G	6Q4H	6Q6T	6RGP	6RHH	6RHU	6RHX	6RI0	6RI6	6RI8	6RII
7A3H	8A3H											

Table 3
The sodium free HR data set.

1B0Y	1DY5	1EA7	1F94	1G4I	1I1W	1IQZ	1IR0	1J0P	1JFB	1K6U	1LNI	1LUG
1M1Q	1 MC2	1MJ5	1MUW	1N1P	1N4W	1N9B	1PWM	1R2M	1R6J	1SSX	1SY2	1SY3
1TG0	1TT8	1UG6	1V0L	1VB0	1VL9	1W0N	1X8P	1XMK	1YWA	1YWB	1YWC	1Z8A
1ZZK	2A6Z	2AGT	2AT3	2AT8	2AYW	2BF6	2CNQ	2EWI	2EWK	2FDN	2FMA	2FOU
2FVY	2GEW	2GG2	2GGC	2H3L	2H5C	2H5D	2JFR	2O9S	2OFR	2P74	2PEV	2PF8
2PFH	2PVE	2PZN	2WFI	2WFJ	2WUR	2XFR	2XU3	2Z6W	2ZPM	3A02	3A4R	3AKQ
3C78	3D1P	3D43	3DHA	3E4G	3EA6	3FSA	3FYM	3GHR	3GHS	3GOE	3I2Y	3I30
3I34	3I37	3JUD	3K34	3KFF	3LEP	3LZ5	3M4H	3NED	3NIR	3O5P	3O5Q	3PUC
3QL9	3QPA	3QPC	3SOJ	3TRV	3U2C	3U7C	3UI4	3UI6	3V1A	3VHG	3VIF	3VIG
3VII	3VJQ	3VOR	3WCQ	3WGE	3WGX	3WOU	3WVM	3X2G	3X2M	3ZR8	3ZUC	3ZZP
4ACJ	4AQO	4AWS	4AWT	4BCT	4BJ0	4DRQ	4E3Y	4EIC	4F18	4FU5	4G78	4G9S
4GA2	4GCA	4GNR	4IAU	4IGS	4J5E	4JP6	4LAU	4LAZ	4LB3	4LB4	4LBR	4LBS
4M7G	4O6U	4PRT	4PSS	4PSY	4PTH	4Q9W	4QB3	4QBX	4QXI	4RTZ	4TJZ	4TKB
4TKH	4TKJ	4UA6	4UA7	4UA9	4UAA	4XZH	4Y9W	4YXI	4Z8J	4ZC9	4ZGF	4ZM7
5A8C	5B28	5B5H	5DGJ	5DJ7	5DK1	5DP2	5E9N	5EMB	5F82	5FLK	5HB7	5IVN
5JDK	5JIG	5KWM	5MAE	5MAJ	5MEH	5MN1	5MNC	5MNG	5MNK	5MNN	5NFK	5NFM
5O2X	5O99	5OTN	5OU0	5OUJ	5OUK	5TDA	5TIF	5VLE	5XBU	5XP6	5YCE	5ZGE
5ZGY	5ZGZ	5ZIO	5ZJ1	5ZJ7	5ZJ8	5ZJC	6B00	6F7R	6F81	6FMC	6G1I	6MU9
6NFR	6Q2Y											

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 1–4; 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 $\leq$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 4
The sodium free, no-SHELX HR data set.

1R2M	1SY2	1SY3	1UG6	1V0L	1W0N	2AT3	2AT8	2AYW	2CNQ	2GG2	2GGC	2H3L
2OFR	2XFR	2XU3	2Z6W	3A02	3A4R	3AKQ	3C78	3E4G	3FSA	3I2Y	3I30	3I34
3I37	3JUD	3NED	3NIR	3O5Q	3PUC	3QL9	3QPA	3QPC	3UI4	3UI6	3V1A	3VIF
3VIG	3VII	3WCQ	3WGE	3WGX	3WVM	3X2G	3X2M	3ZR8	3ZUC	4ACJ	4AQO	4BCT
4BJ0	4DRQ	4E3Y	4F18	4G9S	4GA2	4GCA	4GNR	4IGS	4J5E	4LAU	4LAZ	4LB3
4LB4	4LBR	4LBS	4M7G	4O6U	4PRT	4Q9W	4QB3	4QBX	4QXI	4RTZ	4TJZ	4TKB
4TKH	4TKJ	4XZH	4Y9W	4YXI	4Z8J	4ZC9	4ZM7	5A8C	5B28	5B5H	5DGJ	5DJ7
5DK1	5DP2	5E9N	5EMB	5F82	5FLK	5HB7	5IVN	5JDK	5JIG	5KWM	5MAE	5MAJ
5MEH	5MN1	5MNC	5MNG	5MNK	5MNN	5NFK	5NFM	5O2X	5O99	5OTN	5OU0	5OUJ
5OUK	5TDA	5TIF	5XBU	5XP6	5YCE	5ZGE	5ZGY	5ZGZ	5ZIO	5ZJ1	5ZJ7	5ZJ8
5ZJC	6B00	6FMC	6G1I	6MU9	6NFR	6Q2Y						

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 (# O–O) and the number of oxygen-oxygen pairs considered as putative, charged (protonated or deprotonated), water clusters as defined in Ref. [1] (# OHO).

	4MTY	4Q78	4YXI	5LJT	5OGO	5Y2R	5Y2S	6B00
# H2O	398	214	264	439	333	441	429	409
# O–O	79003	22791	34716	96141	55278	97020	91806	83436
# OHO	3	3	1	1	9	12	17	10

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 (# O–O) and the number of oxygen-oxygen pairs considered as putative, charged (protonated or deprotonated), water clusters as defined in Ref. [1] (# OHO).

	5B1A_A	5B1A_N	5B1B_A	5B1B_N	5B3S_A	5B3S_N	5XDQ_A	5XDQ_N	5ZCP_A	5ZCP_N	5ZCQ_A	5ZCQ_N
# H2O	297	290	289	286	273	278	257	267	235	203	228	215
# O–O	43956	41905	41616	40755	37128	38503	32896	35511	27495	20503	25878	23005
# OHO	3	1	3	2	9	5	1	1	3	3	4	2

Table 7
The RDF pseudo-code.

```
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 $B = 8\pi^2U^2$.

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.bbagen.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.0910410107.

[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/BJ20140082.
[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.