STRUCTURE OF MACROMOLECULAR
COMPOUNDS

Determination of the Near-Atomic Structure of Non-Purified Oligomeric *E. coli* Proteins

E. B. Pichkur[a,b], V. I. Mikirtumov[c], O. V. Tikhonova[d], N. I. Derkacheva[e],
L. P. Kurochkina[f], and O. S. Sokolova[c],*

[a] *National Research Centre "Kurchatov Institute," Moscow, 123182 Russia*

[b] *St. Petersburg Nuclear Physics Institute named by B.P. Konstantinov, National Research Centre "Kurchatov Institute,"*
Gatchina, Leningrad oblast, 188300 Russia

[c] *Moscow State University, Moscow, 119234 Russia*

[d] *Institute of Biomedical Chemistry, Russian Academy of Sciences, Moscow, 119121 Russia*

[e] *Yevdokimov Moscow State University of Medicine and Dentistry, Moscow, 127473 Russia*

[f] *Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119992 Russia*

***e-mail: sokolova@mail.bio.msu.ru**

Received May 4, 2020; revised May 4, 2020; accepted May 29, 2020

Abstract—The roughly purified extract of *E. coli* proteins has been studied by cryoelectron microscopy, the class-sums containing 2D projections of two proteins (β-galactosidase and 2-oxoglutarate dehydrogenase complex catalytic domain (ODC-CD)), identified in an extract by tandem mass spectrometry, have been distinguished. The structures of these proteins have been solved at near-atomic resolution. *De novo* simulation of the ODC-CD structure yielded an atomic model that revealed differences in the positions of some amino acid residues of the active center, in comparison with the known crystal structures.

INTRODUCTION

Within the previous decades, structural biology methods have been rapidly developed, resulting in the detection of new atomic structures of various proteins. Cryoelectron transmission microscopy (cryo-TEM) has recently been acknowledged as one of the main tools for high-resolution analysis of protein structures, along with X-ray diffraction analysis and nuclear magnetic resonance (NMR) spectroscopy [1]. This method became very popular in view of certain technological advances, consisting not only of new devices, direct electron detectors, and phase plates, but also leading software such as Relion [2], cryo-SPARC [3], Chimera [4, 5], and Coot [6]. All these factors provided impressive results to the study of proteins and their complexes. A striking example of cryo-TEM advantages is the structure of the coronavirus 2019-nCoV S-protein that was obtained in one month [7]. These data revealed fundamental differences in the structures of closely related viruses, which induce SARS and COVID-19 infections.

Currently, the biochemical purification of proteins remains a bottleneck for the acquisition of high-resolution data. This stage is difficult to automate, because to a great extent it is based on human expertise. There are a number of problems related to the choice of the protein expression system, its purification method, and concentration of the functional protein. The goals are to minimize the number of purification steps, reduce their influence on the protein, and obtain a sufficient amount of homogeneous specimen for further structural analysis.

In this study, this problem was considered from a technological point of view. A sample containing some cellular proteins along with the target protein was obtained by the ammonium sulfate precipitation of the recombinant protein from the extract of the productive *E. coli* cells. The sample was analyzed by cryo-TEM, and the structures of two *E. coli* proteins, identified by mass spectrometry: β-galactosidase and 2-oxoglutarate dehydrogenase complex catalytic domain (ODC-CD), were resolved with a near-atomic resolution. *De novo* simulation of the ODC-CD structure yielded an atomic model that revealed differences in the positions of some amino acid (a.a.) residues of the active center, in comparison with the known crystal structures [8, 9].

EXPERIMENTAL

Preparation of E. coli Proteins

Isolation and precipitation of *E. coli* proteins were performed according to the procedure described in [10]. *E. coli* BL21(DE3) cells, which produce recom-

binant chaperonin of bacteriophage OBP, were grown in a large volume of the 2xTY medium at 37°C, until optical density $A_{600} = 0.7$ was achieved. The protein expression was induced by adding isopropyl-β-D-thiogalactopyranoside until the final concentration of 1 mM was obtained, and the cell culture was grown for 3.5 hr at 25°C. Cells were precipitated by centrifugation, resuspended in 50 mM tris-HCl-buffer (pH 7.5), subjected to ultrasonic treatment in a Virsonic 100 desintegrator (Vertis, United States), and centrifuged (13 000g) to remove cellular debris. Nucleic acids were precipitated by the addition of 1/10 volume of 30 wt % streptomycin sulfate solution to the supernatant, followed by centrifugation. The proteins were precipitated from the supernatant by the addition of a saturated ammonium sulfate solution to the final concentration of 30 wt %. After centrifugation, the protein pellet was dissolved in 50 mM Tris-HCl buffer (pH 7.5), containing 100 mM NaCl, and washed with 30 volumes of the buffer (50 mM Tris-HCl (pH 7.5) with 10 mM $MgCl_2$ and 100 mM KCl) on Amicon filters (Millipore, United States). The protein composition was analyzed by tandem mass spectrometry.

Mass Spectrometry

Protein hydrolysis in solution was performed using trypsin (Sequencing Grade Modified, Promega, Madison, United States), as was described in [12]. The peptides obtained during the hydrolysis were analyzed using an Ultimate 3000 RSLCnano high-efficiency liquid chromatography system (Thermo Scientific, United States), connected to a Q Exactive™ HF mass spectrometer (Thermo Scientific, United States). Before analytical separation, the peptides were applied to an Accalaim μ-Precolumn enriching column (0.5 × 3 mm, 5 μm particle size; Thermo Scientific, United States), at a flow rate of 10 mL/min for 5 min in the isocratic mode of mobile phase B (2% acetonitrile, 0.1% formic acid). Then, the peptides were separated on an Acclaim Pepmap C18 column (75 μm × 150 mm, 2 μm particle size; Thermo Scientific, United States) in the gradient elution mode. The gradient was formed by mobile phases A (0.1% formic acid) and B (80% acetonitrile, 0.1% formic acid) at a flow rate of 0.3 mL/min. The column was washed with 2% mobile phase B for 5 min, after which the concentration of mobile phase B was linearly increased to 35% for 40 min and then to 99% for 5 min. After washing the column for 5 min at 99% phase B concentration, it was reduced to the initial conditions (2% phase B) for 5 min. The experiment was completed within 60 min.

The mass spectrometric analysis was performed at three technical repetitions on a Q Exactive™ HF hybrid quadrupole-orbitrap mass spectrometer (Thermo Scientific, United States) in positive ionization mode, using NESI source (Thermo Scientific, United States). The mass spectrometry was carried out at the following settings: emitter voltage of 2.1 kV and capillary temperature of 240°C. The scanning was performed in the mass range from 300 to 1500 m/z; the tandem scanning of fragment ions was performed from the lower boundary of 100 m/z to the upper boundary determined by the charge state of a precursor ion (but no larger than 2000 m/z). For tandem scanning, only ions with a charge state from $z = 2+$ to $z = 6+$ were taken into account. The maximum number of ions allowed for synchronous isolation in MS2 mode was set to be no more than 20. The maximum accumulation time did not exceed 50 ms for precursor ions and 110 ms for fragment ions.

The identification of proteins based on mass spectra was carried out in a MASCOT search engine [13] using the SwissProt database of amino acid sequences with the restriction of the species identity to the *E. coli*. To estimate the false discovery rate (FDR), a database of erroneous protein sequences was generated in the form of reverse sequences of *E. coli* proteome amino acids. The search parameters were set as follows: as the splitting enzyme—trypsin with a possibility of skipping one site of sequence splitting, the error in determining mass monoisotopic peaks of peptides: ±5 ppm, the error in determining masses in the MS/MS spectra: ±0.01 Da. Carbamide methylation of cysteine and oxidation of methionine were taken into account as necessary and possible modifications, respectively. To confirm the comparisons of the spectra and PSM peptides, the identification of peptides and proteins, the applied FDR value did not exceed 1.0%. The proteins were considered to be reliably identified if at least two unique validation-passed peptides were found for them. Marker-free quantitative estimation of the protein content was performed on the basis of the emPAI empirical factor [14].

Cryoelectron Microscopy

A 3.0 μL sample of *E. coli* protein cocktail was applied to a carbon-coated grid with circular holes (diameter 1.2 μm) and a 1.3 μm period (R1.2/1.3, Quantifoil). The grids were preliminarily treated by a glow discharge for 30 s in a PELCO easyGlow glow discharge cleaning system (Ted Pella, United States) with the chamber pressure of 0.26 mbar and a current strength of 25 mA. Grids were instantly frozen in liquid ethane on a Vitrobot Mark IV system (Thermo Fischer Scientific) at 100% humidity and a temperature of +4.5°C. Micrograph stacks, each containing 38 frames (4017 stacks), were recorded in a Titan Krios cryomicroscope (Thermo Fisher Scientific) with a Falcon II direct electron detector under the following conditions: accelerating voltage 300 kV, magnification 75000, pixel size 0.86 Å, and electron dose ~100 e/Å². The stacks were preliminarily processed in the Warp program [11] to estimate the parameters of the contrast transfer function (CTF) and to correct the drift in individual frames. Images of protein molecules were selected using the BoxNet neural network and cut

Table 1. List of reliably identified *E. coli* proteins

Uniprot ID	Protein	Score	MW, kDa	Sequence coverage, %	emPAI
P00722	β-Galactosidase (BGAL_ECOLI)	4240	117.35	39	3.06
P0A9L8	Pyrroline-5-carboxylate reductase (P5CR_ECOLI)	395	28.36	21	1.18
P0A9P0	Dihydrolipoyl dehydrogenase (DLDH_ECOLI)	484	50.94	18	0.99
P05055	Polyribonucleotide nucleotidyltransferase (PNP_ECOLI)	394	77.11	14	0.65
P0A6F5	60 kDa chaperonin (CH60_ECOLI)	396	57.46	15	0.65
P0A9H3	Inducible lysine decarboxylase (LDCI_ECOLI)	275	81.61	7	0.54
P0AFG7	Succinyltransferase component of 2-oxoglutarate dehydrogenase complex (ODO2_ECOLI)	311	43.98	6	0.34
P0AES4	DNA gyrase subunit A (GYRA_ECOLI)	41	97.13	3	0.07

from the micrographs via a 256 × 256 pixel frame, resulting in the collection of a set of 470 117 particles. Among them, more than 250 000 particles represent OBP chaperonin [10]. 2D classification was performed in the CryoSPARC program [3]. 2D classes corresponding to different proteins were chosen manually with the subsequent generation of an ab initio model for each class separately. Then, 3D refinement of the structure of each class was carried out using a preliminary model with no symmetry imposed. In the next stage, the obtained structures were used as references for further 3D classification of the initial set of particles. Final 3D protein structures were reconstructed with imposed symmetry: dihedral (D2) for β-galactosidase and octahedral (O) for ODC-CD. The final reconstructions included 99 246 particles for β-galactosidase and 13 770 particles for ODC-CD. Resolutions of the reconstructions were estimated at FSC = 0.143, 2.6 Å for β-galactosidase and 3.2 Å for ODC-CD.

De Novo Simulation of the Protein Structure

The simulation was carried out using the Coot [6] and Phenix [15] programs. The ODC-CD map resolution obtained by cryo-TEM was sufficient for the side chains identification and unambiguous determination of the protein sequence. A unique part of the ODC-CD map was extracted using the *phenix.mapbox* algorithm and NCS operators, obtained by means of the *phenix.mapsymmetry* algorithm, after setting the symmetry of the "O" point group; the atomic model of this part was constructed automatically using the *phenix.maptomodel* algorithm. The obtained model was verified manually in the Coot program (wide ranges of mismatch residues and empty regions were filled with alanine residues); new peptide bonds were formed (if necessary) so that the model would consist of one chain. Afterwards, the alanine side chains in the model were replaced by the side chains of amino acids corresponding to sequences of the C-terminal part of the ODC-CD protein (174−405 a.a.). All the residues were manually corrected using the Coot program for a maximally accurate correspondence to the electrostatic potential map with the correct geometry of residues retained; rotamers with optimal orientation were chosen for all of the side chains. NCS operators obtained using *phenix.mapsymmetry* algorithm were applied to the corrected model of the unique structural part.

RESULTS AND DISCUSSION

The purpose of this study was to demonstrate that cryo-TEM and image processing allow to obtain 3D structures without fine purification of target proteins. We used *E. coli* cells, heterologously expressing viral chaperonin (OBP) [10] to obtain a roughly purified protein cocktail. The recombinant chaperonin was precipitated from the cellular extract by 30% ammonium sulfate [16]; in this case, some cellular proteins appeared in the pellet. The mixture of precipitated *E. coli* proteins was identified by high-efficiency liquid chromatography in combination with tandem mass spectrometry. The results are listed in Table 1. The relative amounts of simultaneously precipitated proteins were estimated using the emPAI factor, calculated in the MASCOT program. The highest correspondence of the sequence (39%) was obtained for β-galactosidase (P00722); the other seven proteins are represented by several peptides each.

After precipitation and washing, the protein mixture was instantly frozen on the Vitrobot MARK IV system (Thermo Fisher Scientific) and examined by cryo-TEM using a standard protocol (described above). Two successive cycles of 2D classification were performed in CryoSPARC [3] to identify different proteins. The first classification was used to exclude "junk" classes from the set; after the second one (Fig. 1), 40 classes were distinguished and visually divided into three datasets, each corresponding to a separate protein.
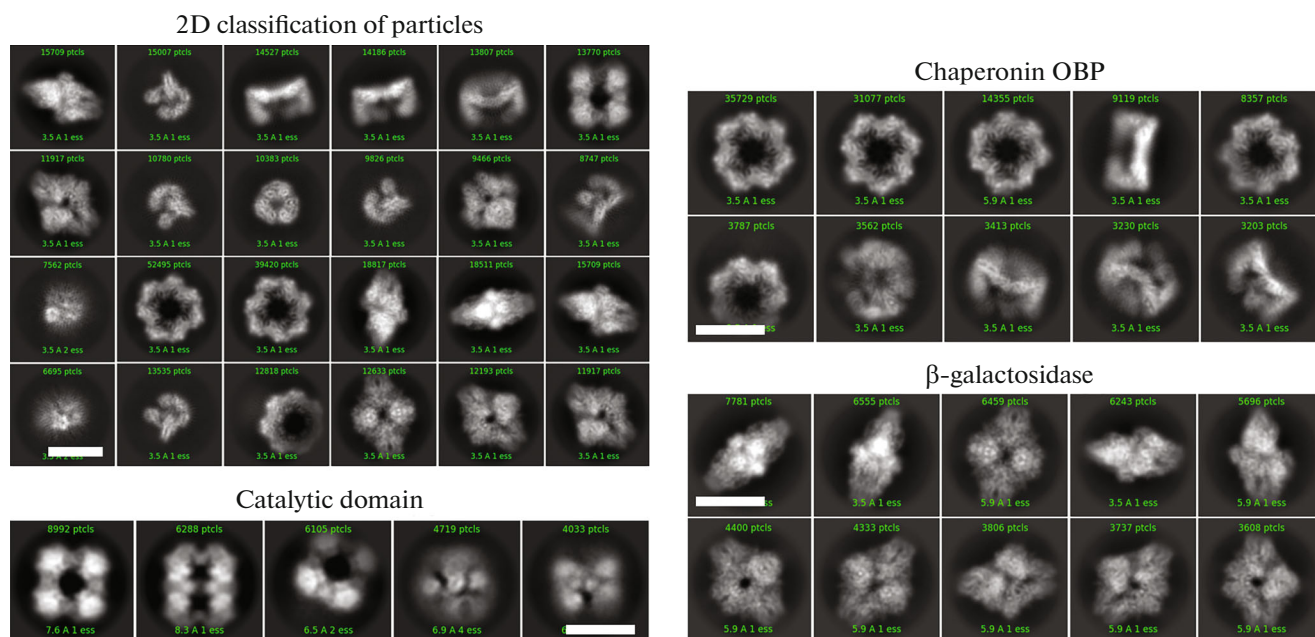
2D classification of particles

Chaperonin OBP

β-galactosidase

Catalytic domain

Fig. 1. Cryo-TEM analysis of the *E. coli* protein extract: 2D classification and distinguished classes of separate proteins. Scale intervals are equal to 10 nm.

Three oligomeric proteins (transiently expressed heptameric chaperonin of the *P. fluorescens* phage OBP [10] and two *E. coli* proteins (tetrameric β-galactosidase and 24-meric ODC-CD)) possess particles of similar sizes (12−15 nm in cross section; Fig. 1). In addition, classes corresponding to a smaller asymmetric protein of uncertain nature were observed. Other proteins, the peptides of which were found by mass spectrometry (Table 1), could not be identified on the cryo-TEM images. It is possible that these proteins do not form stable oligomeric complexes under the used conditions, or that their concentrations are too low for cryo-TEM analysis.

According to tandem mass spectrometry data (Table 1) and 2D classification (Fig. 1), the most widespread *E. coli* protein in the studied sample was β-galactosidase (99 246 particles). It is encoded by the lacZ gene and forms a stable homotetramer with a molecular weight of ~464 kDa [17]. It has a well-distinguishable shape in the cryo-TEM images and can be easily separated using 2D classification (Fig. 1). A standard approach to image analysis [18] yielded 3D reconstruction of β-galactosidase with D2 symmetry and a resolution of 2.6 Å (Figs. 2a−2c).

The 2-oxoglutarate dehydrogenase complex of *E. coli* incorporates three different enzymes: E1 (2-oxoglutarate decarboxylase), E2 (dihydrolipoyl transsuccinylase), and E3 (dihydrolipoyl dehydrogenase). Enzyme E2 catalyzes transport of the succinyl group from S-succinyl-dihydrolipoyl to coenzime A (CoA) [19]. Its full-size monomer with a length of 405 a.a. consists of three distinguished domains: N-terminal lipoyl-binding domain, E3-binding domain, and C-terminal catalytic domain. The molecular weight of the entire protein is ~44 kDa.

Only two significant peptides for enzyme E2 (one from the N-terminal lipoyl-binding domain and one from the C-terminal catalytic domain) were identified in the studied *E. coli* protein cocktail by tandem mass spectrometry. Nevertheless, in the obtained cryo-TEM images, the C-terminal domain of E2 (ODC-CD) was the second most widespread protein after the 2D classification (13 770 particles). The 2D classification revealed characteristic particles with octahedral symmetry (Fig. 1). Linear sizes of 2D classes are close to the sizes of previously identified ODC-CD crystal structures [8, 9]. A comparison of the data collected with the results of [8, 20] showed that the density for N-terminal E3-binding and lipoyl-binding domains is absent in the obtained cryoelectronic structure (as well as the density for linkers connected to these domains). It was previously suggested that these domains can be removed by endogenous proteases [9]. The structures of isolated N-terminal domains were resolved by NMR [21, 22].

Note that crystals of this domain were previously obtained accidentally when trying to crystallize *Arabidopsis thaliana* recombinant amidase expressed in the *E. coli* heterologous system [9]. It was suggested that histidine residues on the surface of the catalytic domain can facilitate its binding with nickel-containing resin [23]; however, we did not use Ni-affinity chromatography in this study.

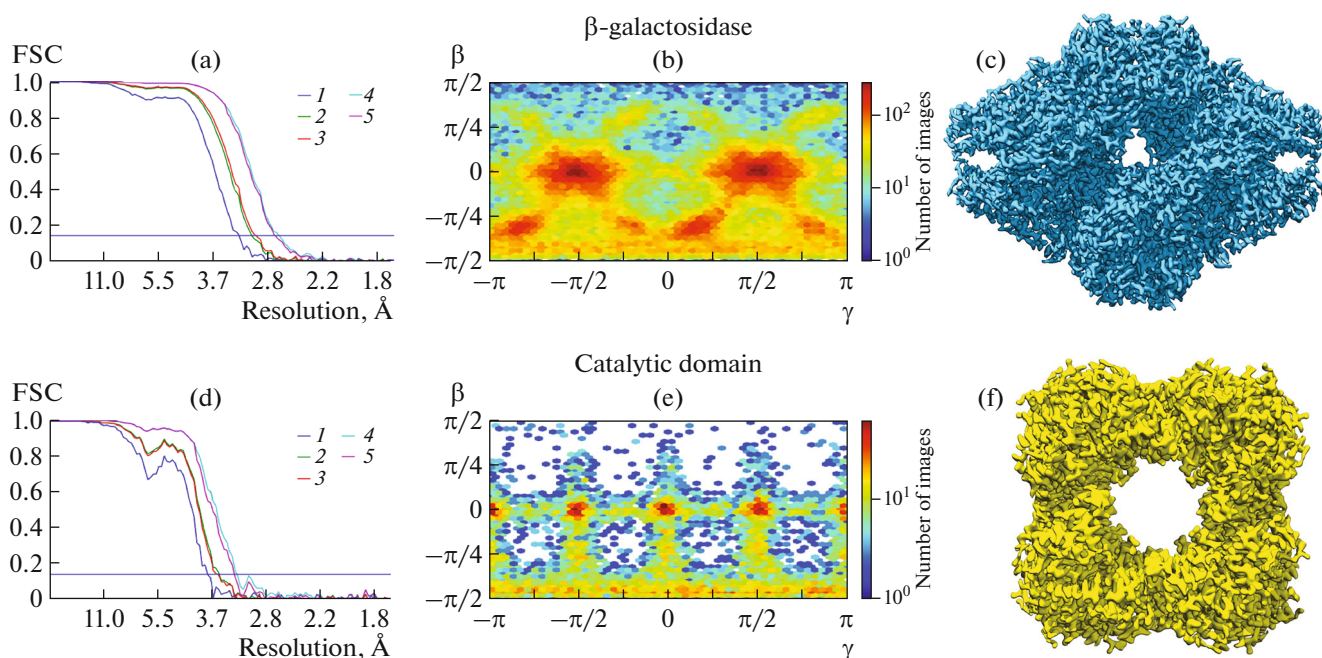Reconstruction of ODC-CD with octahedral symmetry, implemented in cryoSPARC, had a resolution

Fig. 2. Construction of 3D structures of *E. coli* proteins: (a−c) β-galactosidase and (d−f) ODC-CD. (a, d) Curves of Fourier shell correlation (FSC) for the corresponding structures: (*1*) without mask, (*2*) spherical mask, (*3*) free mask, (*4*) narrow mask, and (*5*) after correction. (b, e) Distributions of particle projections relative to the vertical *z* axis of an Euler sphere: β is the angle between the projection and the *z* axis and γ is the angle of rotation around the *z* axis; images are obtained in the cryoSPARC software [3]. (c, f) 3D reconstructions of β-galactosidase and ODC-CD.
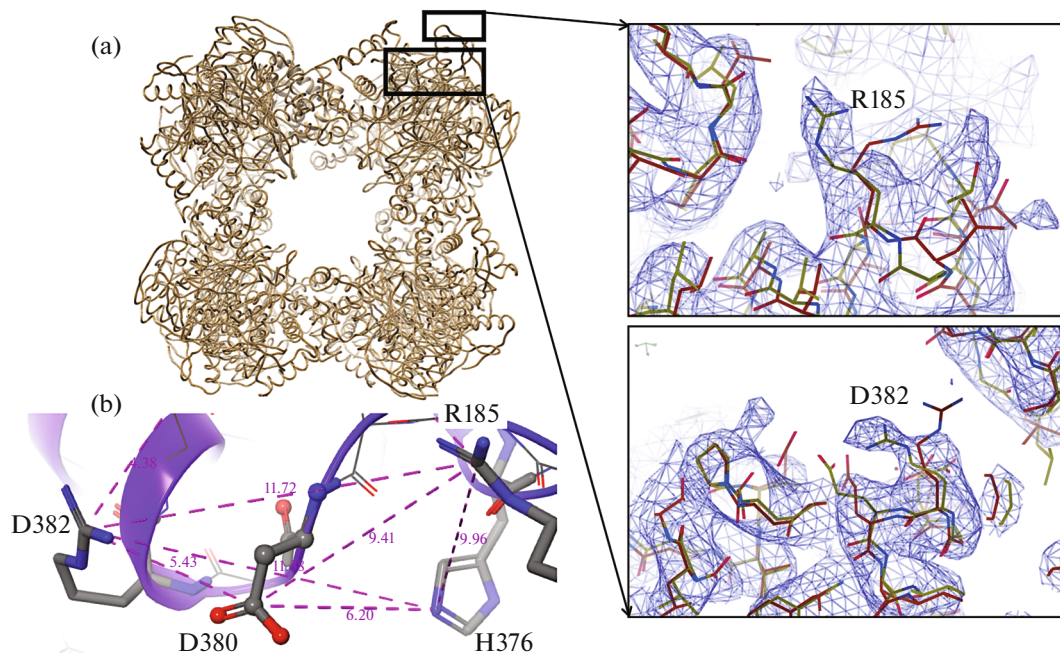
Fig. 3. Structure of ODC-CD. (a) *De novo* atomic model of ODC-CD; position of the active center is indicated by the frame; differences in the positions of arginines R185 and R382 of the active center are shown in the insets. The crystal structure from [9] and our structure are presented. The electrostatic potential map is shown at level 2σ. (b) Distances between key amino acids of the ODC-CD active center.

of 3.2 Å (Figs. 2d−2f). The high resolution allowed *de novo* structure determination of the catalytic domain. The structural model was obtained using the Coot program [6] and three-dimensional coordinates for each atom (Fig. 3a).

A comparison of our cryo-TEM structure with the X-ray structure (PDB ID 6prb) showed that there are some differences in the orientations of the side chains of key amino acids in the active center: D382 and D185 (Fig. 3a, inset). The main unsolved question

regarding the ODC-CD structure is the nature of the changes in the side-chain orientations, which occur in the active center upon substrate binding. It was previously suggested that the H375 residue initiates the first stage of catalysis by deprotonation of the CoA thiol group, which then attack the carbonyl atom of the succinyl residue bound to the dihydrolipoamide. The destruction of the intermediate compound leads to the formation of succinyl-CoA and a protonated dihydrolipoyl group [8]. It was also suggested that, upon the binding of CoA in the active center, asparagine D380 forms salt bridges with histidine H376 and with one of the two arginine residues: R185 or R382 [20]. For that, D380 should change its orientation, depending on the substrate. The structure obtained by cryo-TEM allowed us to accurately localize the R185 and R382 side chains; these side chains were not ordered in the early X-ray study (PDB ID 1e2o); however, they were revealed later (PDB ID 6prb). A comparison of the PDB:6prb model with the one obtained in this study (Fig. 3a, inset) showed that in the *de novo* model Arg residues are rotated. The measured distances between the key amino acid residues in the active center were as follows: D380−H376 = 6.2 Å, D380−R382 = 5.4 Å, and D380−R185 = 9.4 Å (Fig. 3b). Thus, asparagine D380 is located approximately in the middle between histidine H376 and arginine R380, and its interactions with these residues apparently possess an ionic nature. According to our model, asparagine D380 does not interact with arginine R185. To determine the structure of the enzyme active center more accurately, the reconstruction of the ODC-CD with substrate is needed.

CONCLUSIONS

Cryo-EM becomes an irreplaceable tool for structural research of biological macromolecules. Until recently, acquiring structures with high resolution was possible only for highly purified protein preparations. However, it has been shown that 3D classification of heterogeneous particles implemented in the Relion [2] and cryoSPARC [3] programs, in combination with mass spectrometry, allows to bypass special purification [24]. In this study, cryo-TEM and tandem mass spectrometry were used to identify protein molecules in a roughly purified protein sample. Using image analysis, we distinguished two particle samples, corresponding to two *E. coli* proteins, and obtained two 3D structures (β-galactosidase and ODC-CD). Both proteins were in the native state and had no recombinant tags used for purification. The close-to-atomic resolution made it possible to perform *de novo* simulation of the ODC-CD atomic model. A comparison of the obtained structure with a previously published crystal structure of the same domain [9] showed that the active site in the solution is in a somewhat different conformation, in comparison with the crystal.

REFERENCES

1. O. S. Sokolova, M. P. Kirpichnikov, K. V. Shaitan, et al., *Modern Methods for Studying the Structure and Functions of Ionic Channels* (Tovarishchestvo Nauchnykh Izdanii KMK, Moscow, 2020) [in Russian].

2. S. H. Scheres, Methods Enzymol. **579**, 125 (2016). https://doi.org/10.1016/bs.mie.2016.04.012

3. A. Punjani, J. L. Rubinstein, D. J. Fleet, et al., Nat. Methods **14**, 290 (2017). https://doi.org/10.1038/nmeth.4169

4. Z. Yang, K. Lasker, D. Schneidman-Duhovny, et al., J. Struct. Biol. **179**, 269 (2012). https://doi.org/10.1016/j.jsb.2011.09.006

5. E. F. Pettersen, T. D. Goddard, C. C. Huang, et al., J. Comput. Chem. **25**, 1605 (2004). https://doi.org/10.1002/jcc.20084

6. P. Emsley, B. Lohkamp, W. G. Scott, et al., Acta Crystallogr. D **66**, 486 (2010). https://doi.org/10.1107/S0907444904019158

7. D. Wrapp, N. Wang, K. S. Corbett, et al., Science **367**, 1260 (2020). https://doi.org/10.1126/science.abb2507

8. J. E. Knapp, D. Carroll, J. E. Lawson, et al., Protein Sci. **9**, 37 (2000). https://doi.org/10.1110/ps.9.1.37

9. B. Andi, A. S. Soares, W. Shi, et al., Acta Cryst. F. **75**, 616 (2019). https://doi.org/10.1107/S2053230X19011488

10. T. B. Stanishneva-Konovalova, P. I. Semenyuk, L. P. Kurochkina, et al., J. Struct. Biol. **209**, 107439 (2020). https://doi.org/10.1016/j.jsb.2019.107439

11. D. Tegunov and P. Cramer, Nat. Methods **16**, 1146 (2019).
https://doi.org/10.1038/s41592-019-0580-y

12. V. Anselm, S. Novikova, and V. Zgoda, Int. J. Mol. Sci. **18** (2017).
https://doi.org/10.3390/ijms18081763

13. D. N. Perkins, D. J. Pappin, D. M. Creasy, et al., Electrophoresis **20**, 3551 (1999).
https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

14. Y. Ishihama, Y. Oda, T. Tabata, et al., Mol. Cell. Proteomics **4**, 1265 (2005).
https://doi.org/10.1074/mcp.M500061-MCP200

15. P. V. Afonine, B. K. Poon, R. J. Read, et al., Acta Crystallogr. D **74**, 531 (2018).
https://doi.org/10.1107/S2059798318006551

16. P. Wingfield, Curr. Protoc. Protein Sci., Appendix **3** (2001).
https://doi.org/10.1002/0471140864.psa03fs13

17. R. H. Jacobson, X. J. Zhang, R. F. DuBose, et al., Nature **369**, 761 (1994).
https://doi.org/10.1038/369761a0

18. J. Zivanov, T. Nakane, B. O. Forsberg, et al., Elife **7**, e42166 (2018).
https://doi.org/10.7554/eLife.42166

19. R. N. Perham, Biochem. **30**, 8501 (1991).
https://doi.org/10.1021/bi00099a001

20. J. E. Knapp, D. T. Mitchell, M. A. Yazdi, et al., J. Mol. Biol. **280**, 655 (1998).
https://doi.org/10.1110/ps.9.1.37

21. P. M. Ricaud, M. J. Howard, E. L. Roberts, et al., J. Mol. Biol. **264**, 179 (1996).
https://doi.org/10.1006/jmbi.1996.0632

22. M. A. Robien, G. M. Clore, J. G. Omichinski, et al., Biochemistry **31**, 3463 (1992).
https://doi.org/10.1021/bi00128a021

23. V. M. Bolanos-Garcia and O. R. Davies, Biochim. Biophys. Acta **1760**, 1304 (2006).
https://doi.org/10.1016/j.bbagen.2006.03.027

24. E. J. Verbeke, A. L. Mallam, K. Drew, et al., Cell. Rep. **24**, 259 (2018).
https://doi.org/10.1016/j.celrep.2018.06.022

Translated by A. Sin'kov