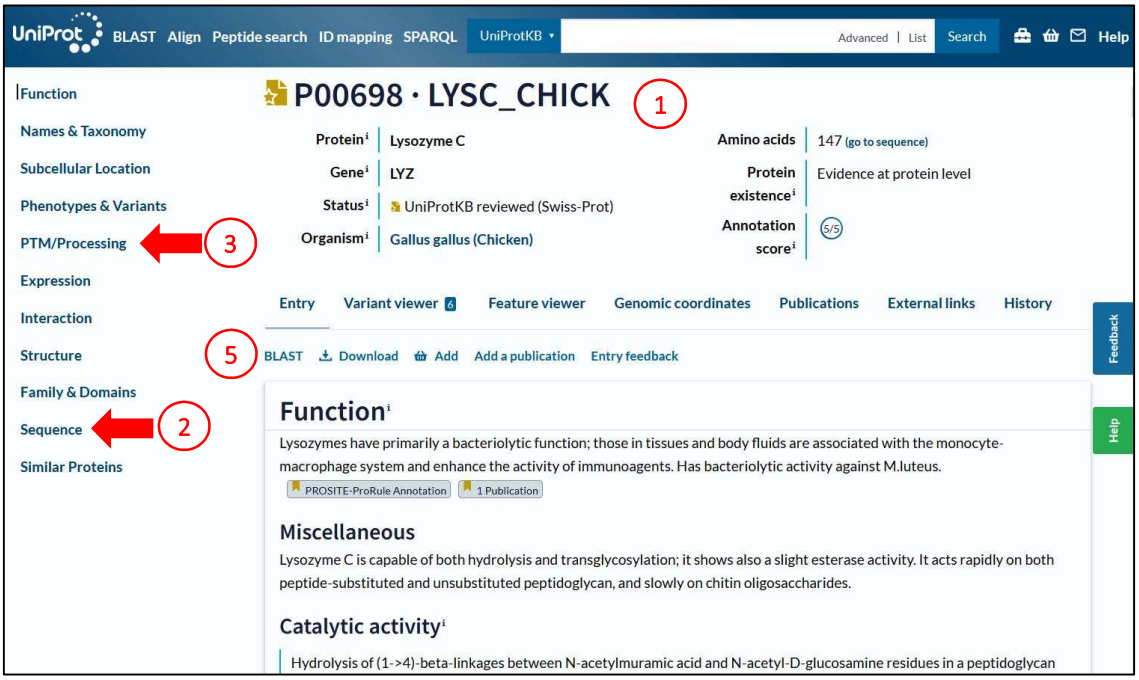
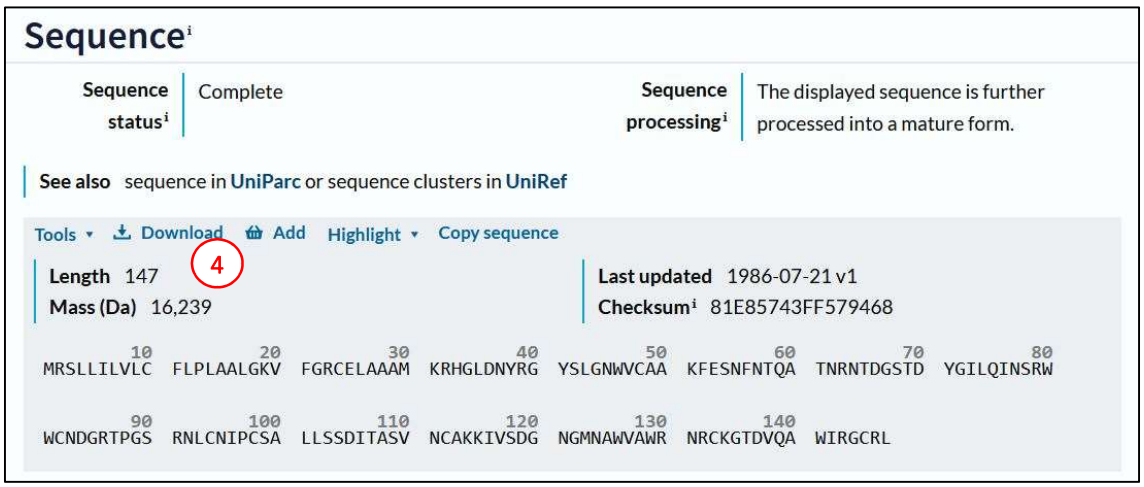
PHASING WITH THE MOLECULAR REPLACEMENT METHOD.

For phasing, crystallographic data should be used after scaling and merging, in order to have a list of unique reflections with respective intensities. To apply the Molecular Replacement method (MR), a model protein with known structure should be identified. Various resources are available to determine the best protein model, among which: (a) the UniProt online data bank [1] to obtain the primary sequence of the protein; (b) the online software Blast [2] to search for proteins with similar sequence and known structure; (c) the Protein Data Bank [3] to download the 3D structure of the selected protein; (d) the MOLREP software [4] of the crystallographic suite CCP4 [5] to solve the phase problem with MR; (e) the software Refmac [6] for the rigid body refinement and the determination of R_{work} e R_{free} values for the MR solution; (f) the Coot software [7] to visualize model and electron density obtained by Fourier transform using phases determined through MR.

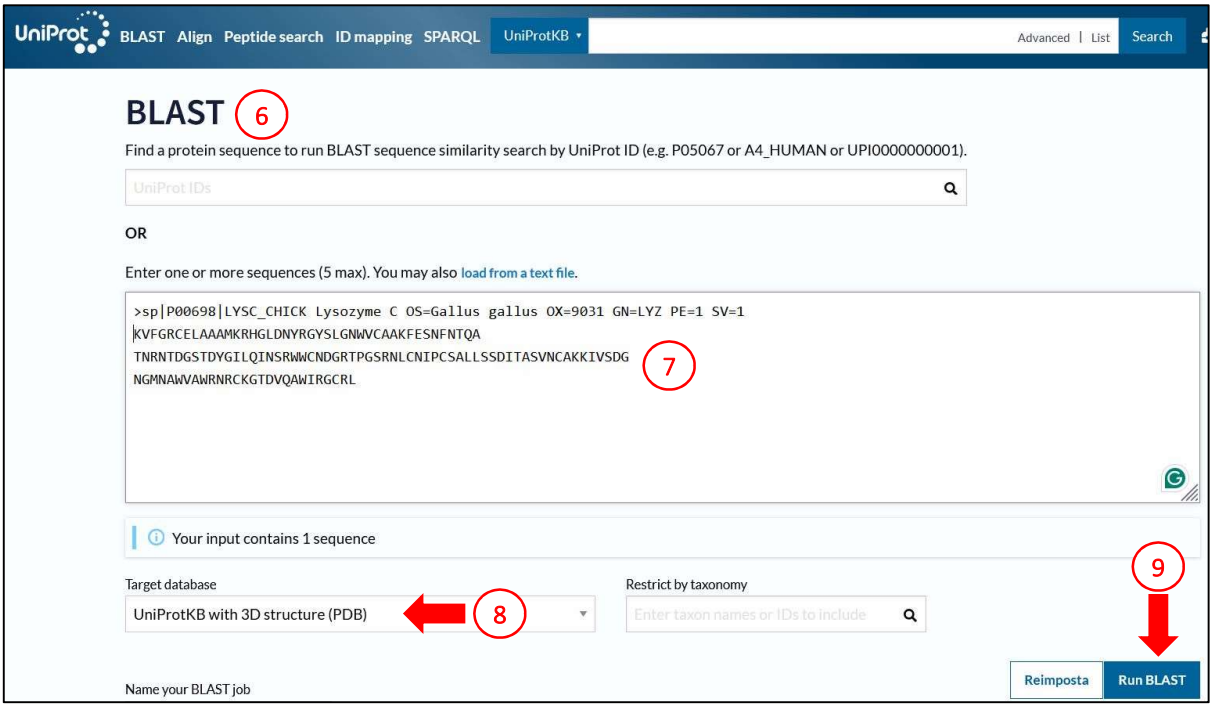
Selection and preparation of the model structure.

If the primary sequence of the protein is now yet available, it can be obtained from the UniProt data bank. In the example, we search for the Hen Egg White Lysozyme (https://www.uniprot.org/uniprot/P00698) (**1**). The databank contains a lot of information, among which the sequence (**2**) and the post-translational modifications (PTM/processing, **3**).

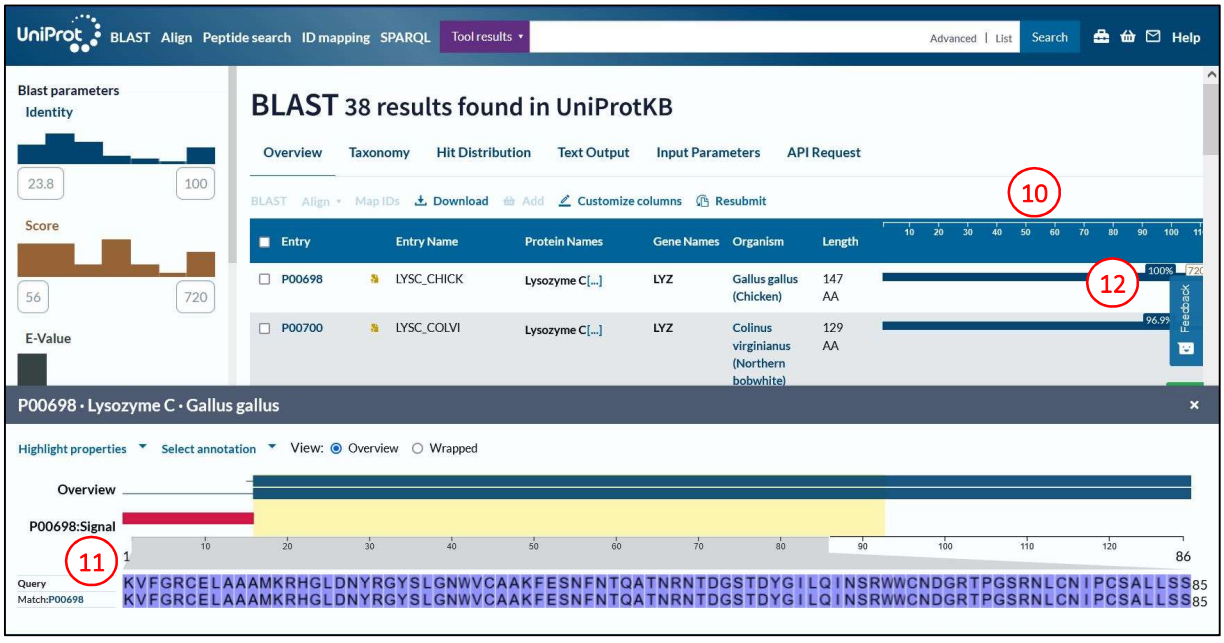


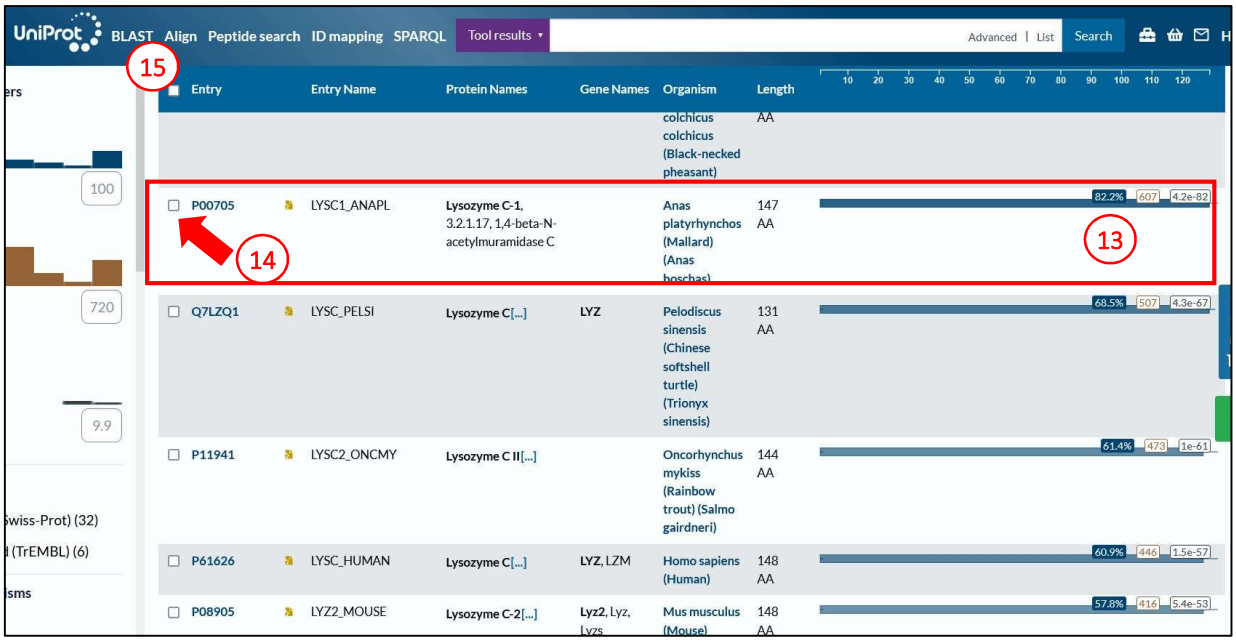


Considering the information available on the databank, the mature form of the protein lacks the first 18 residues of the sequence, which constitute the signal peptide and which are removed by proteolysis during protein maturation. The protein sequence in *Fasta* format can be obtained with the suitable button (**4**), it can be copied and the first 18 residues can be manually removed. From the UniProt webpage, the software Blast (**5**) can be opened to search for proteins with similar primary sequence.
A new Blast window opens (**6**) and the primary sequence of the protein in *Fasta* format can be pasted after removing the signal peptide residues (**7**). Among the options in the lower part of the window, it is advisable to select only the Protein Data Bank as target database (**8**), so that the proteins identified will have a known structure. The database search is started with the button "Run BLAST" (**9**).

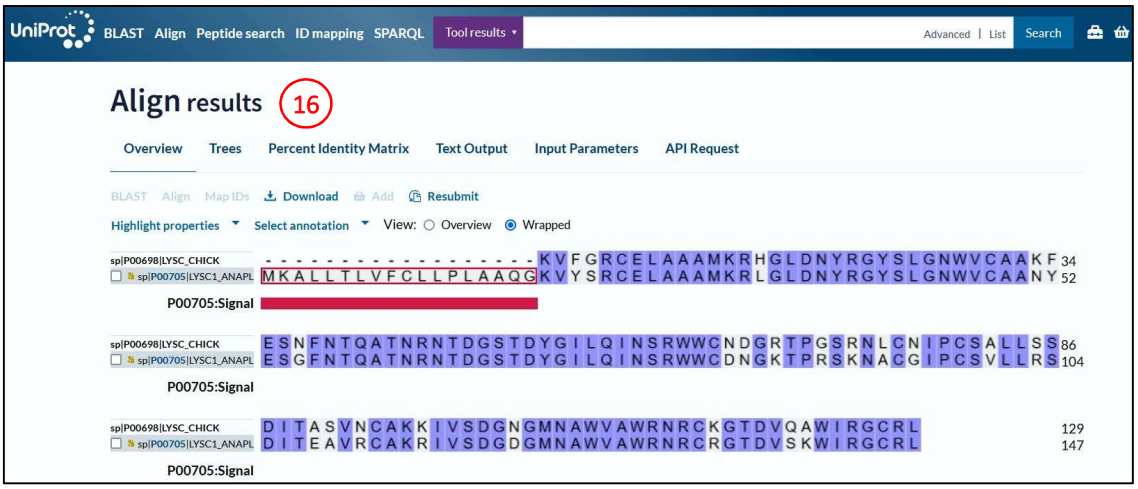


At the end of the search, the software shows a list of proteins with a sequence similar to the query (**10**), together with protein alignments (**11**) and identity percentage (**12**) between the query and the identified protein.

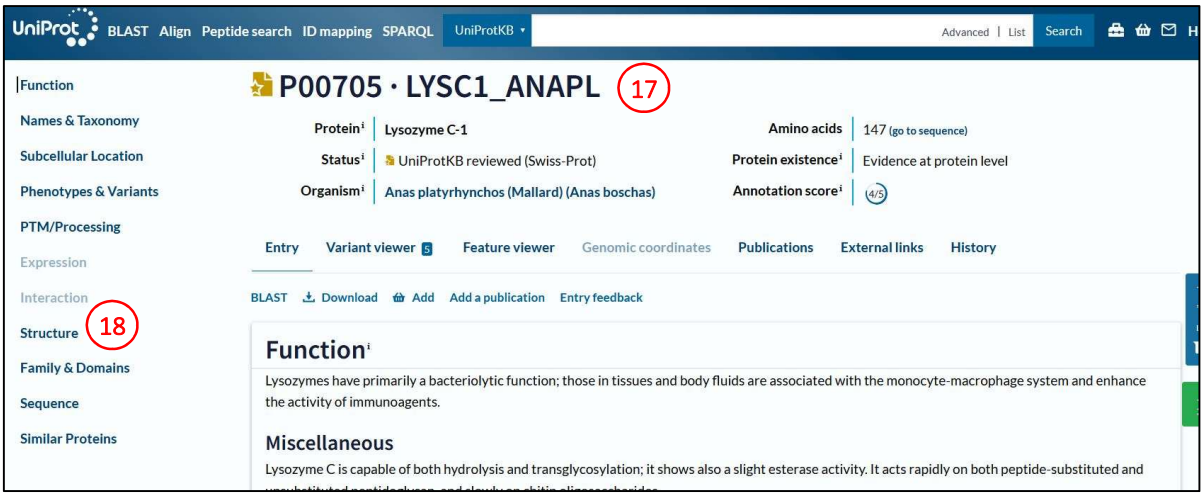


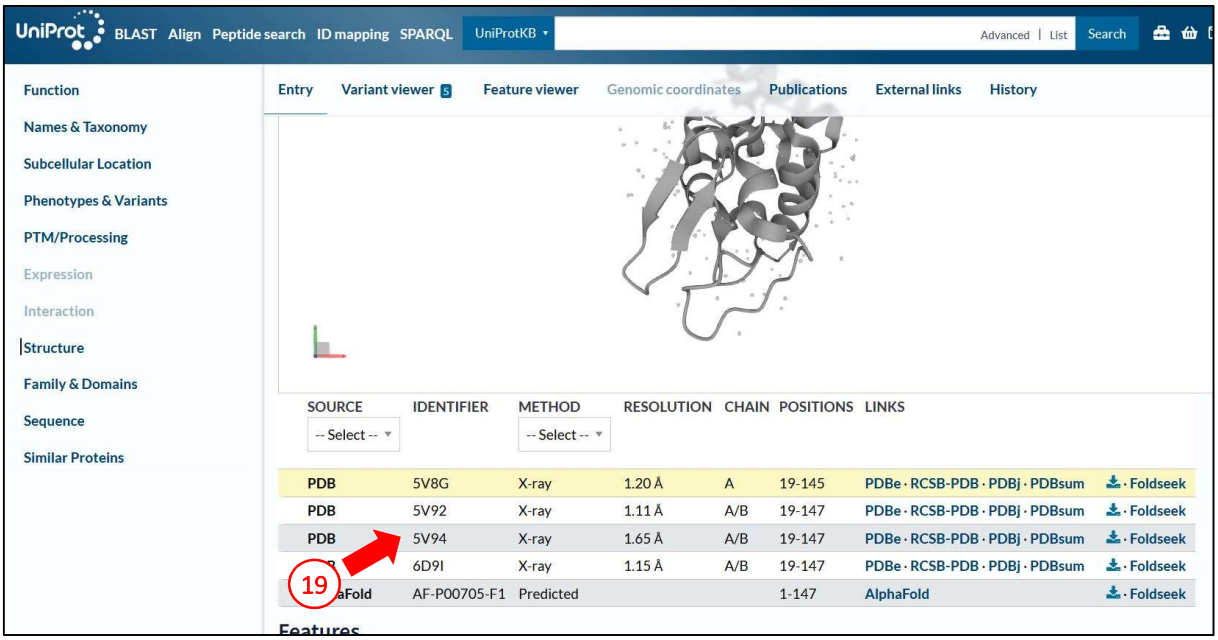


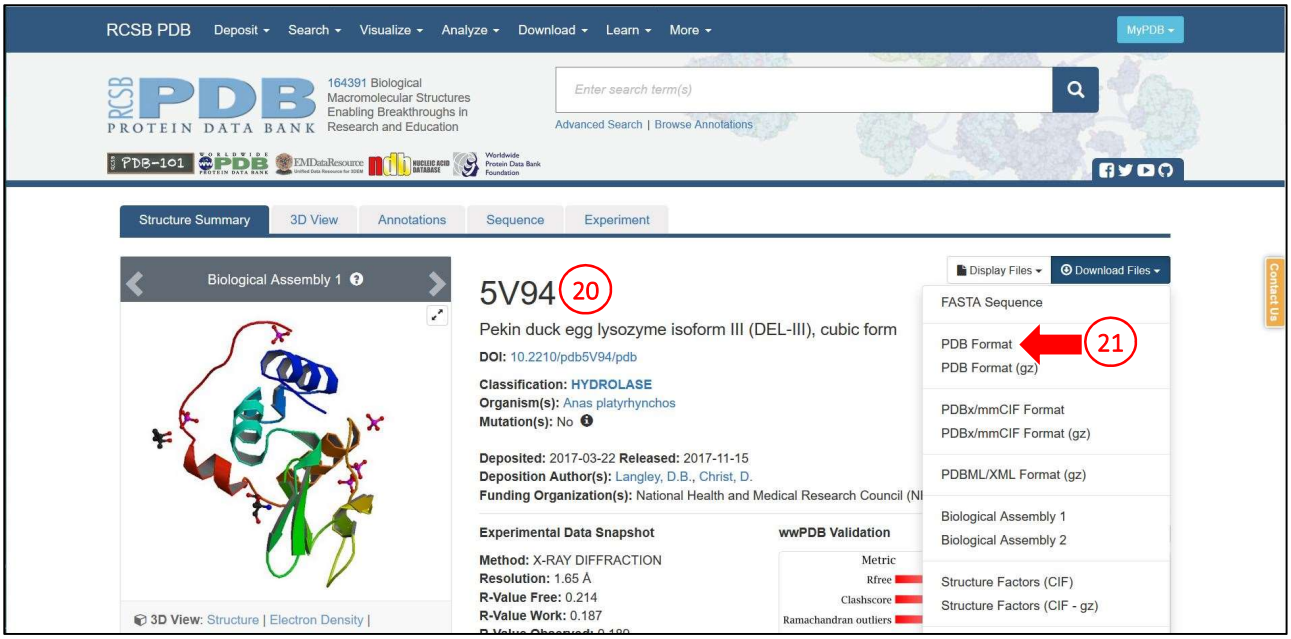
Among the protein sequences, we select a protein with about 80% identity (**13**) (in a real case, using the sequence with the highest identity value ensures a higher success probability for the MR search). For the selected sequence, the alignment details can be analyzed in a separate browser window by selecting the entry (**14**) and running the Align tool (**15**). The result of the alignment in Figure **16**. By clicking on the identification code of the sequence (in the example P00705), the user can open the UniProt page of the protein identified as probe (**17**).



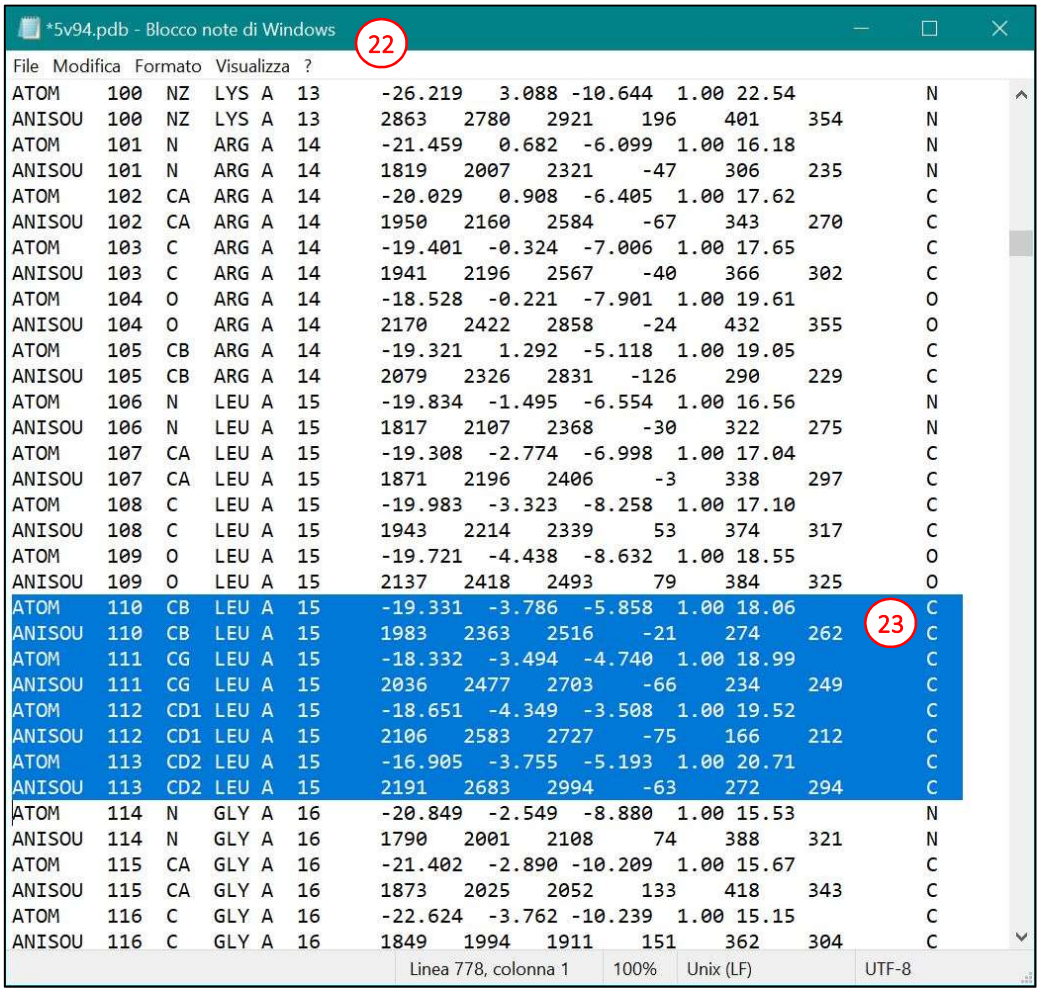
In this UniProt page, the Structure tab can be selected from the left menu (**18**), showing all the 3D structures corresponding to the primary sequence of the model protein, including structures determined using computational software such as AlphaFold. Among these structures, the user will choose the preferred for the MR step (for example the structure with PDB code 5V94). The second link on the right (**19**) opens the Protein Data Bank page of the structure (**20**).


UniProt
BLAST Align Peptide search ID mapping SPARQL
UniProtKB
Advanced | List
Search
Function
Names & Taxonomy
Subcellular Location
Phenotypes & Variants
PTM/Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar Proteins
P00705 · LYSC1_ANAPL
17
Protein
Lysozyme C-1
Amino acids
147 (go to sequence)
Status
UniProtKB reviewed (Swiss-Prot)
Protein existence
Evidence at protein level
Organism
Anas platyrhynchos (Mallard) (Anas boschas)
Annotation score
4/5
Entry
Variant viewer 5
Feature viewer
Genomic coordinates
Publications
External links
History
BLAST
Download
Add
Add a publication
Entry feedback
18
Function
Lysozymes have primarily a bacteriolytic function; those in tissues and body fluids are associated with the monocyte-macrophage system and enhance the activity of immunoagents.
Miscellaneous
Lysozyme C is capable of both hydrolysis and transglycosylation; it shows also a slight esterase activity. It acts rapidly on both peptide-substituted and


UniProt
BLAST Align Peptide search ID mapping SPARQL
UniProtKB
Advanced | List
Search
Function
Names & Taxonomy
Subcellular Location
Phenotypes & Variants
PTM/Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar Proteins
Entry
Variant viewer 5
Feature viewer
Genomic coordinates
Publications
External links
History
SOURCE
IDENTIFIER
METHOD
RESOLUTION
CHAIN
POSITIONS
LINKS
-- Select --
-- Select --
PDB 5V8G X-ray 1.20 Å A 19-145 PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB 5V92 X-ray 1.11 Å A/B 19-147 PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB 5V94 X-ray 1.65 Å A/B 19-147 PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
6D9I X-ray 1.15 Å A/B 19-147 PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
19
AF-P00705-F1 Predicted 1-147 AlphaFold · Foldseek
Features


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
164391 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Enter search term(s)
Advanced Search | Browse Annotations
PDB-101
Structure Summary
3D View
Annotations
Sequence
Experiment
Biological Assembly 1
5V94
20
Pekin duck egg lysozyme isoform III (DEL-III), cubic form
DOI: 10.2210/pdb5V94/pdb
Classification: HYDROLASE
Organism(s): Anas platyrhynchos
Mutation(s): No
Deposited: 2017-03-22 Released: 2017-11-15
Deposition Author(s): Langley, D.B., Christ, D.
Funding Organization(s): National Health and Medical Research Council (NI
Experimental Data Snapshot
Method: X-RAY DIFFRACTION
Resolution: 1.65 Å
R-Value Free: 0.214
R-Value Work: 0.187
wwPDB Validation
Metric
Rfree
Clashscore
Ramachandran outliers
3D View: Structure | Electron Density |
Display Files
Download Files
FASTA Sequence
PDB Format
PDB Format (gz)
21
PDBx/mmCIF Format
PDBx/mmCIF Format (gz)
PDBML/XML Format (gz)
Biological Assembly 1
Biological Assembly 2
Structure Factors (CIF)
Structure Factors (CIF - gz)
Contact Us

On the PDB page, the user can download the file containing atomic coordinates of the selected structure (21). Such file should be opened with a text editor (22) to manually edit the residues following the previous alignment (15). In particular, a common modification that yields good results is the removal of side chains (23) of residues that differ between the two protein sequences. In addition, only a single protein sequence should be included in the model file, while multiple chain should be removed together with water molecules, ions, ligands or other molecular species present in the 3D structure. The edited file is saved in *pdb* format. The edited structure can be visualized using a suitable software, such as PyMOL.



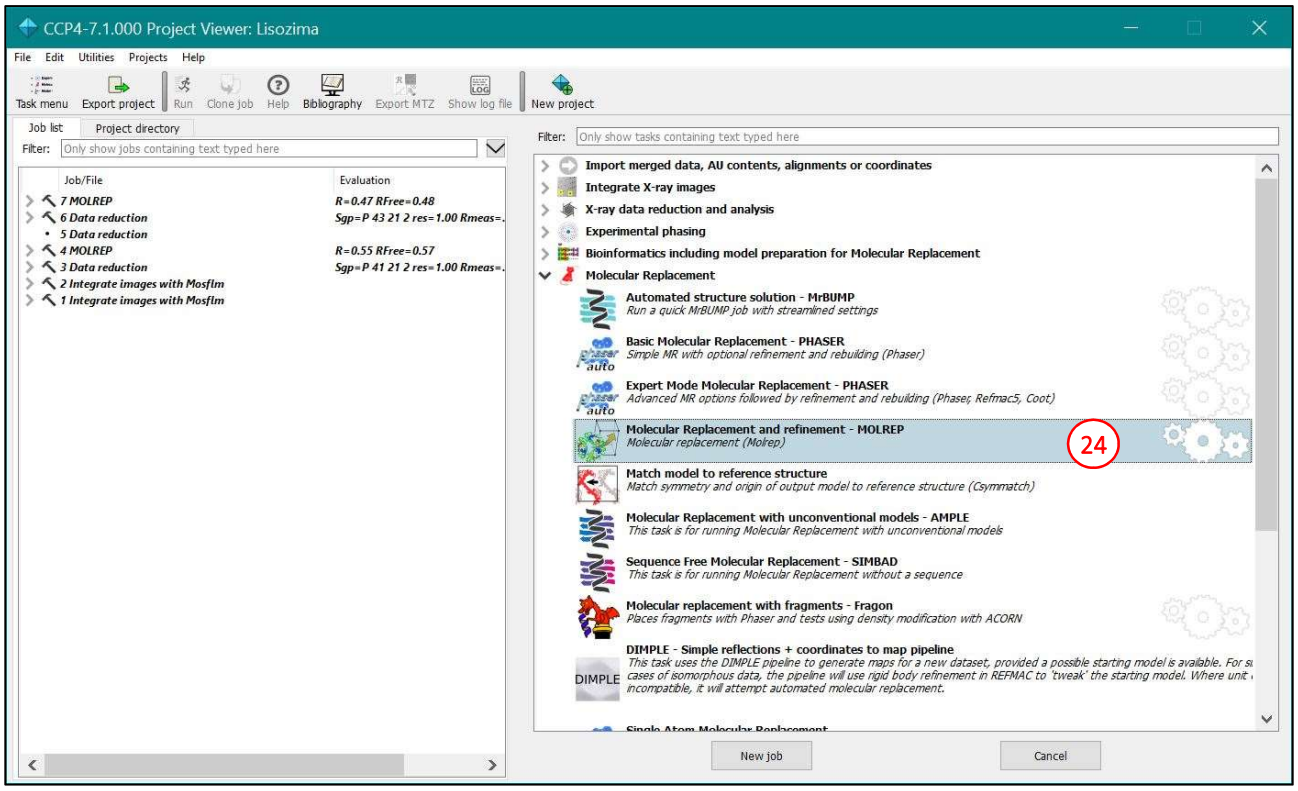
Phasing using the Molecular Replacement method with the MOLREP software.

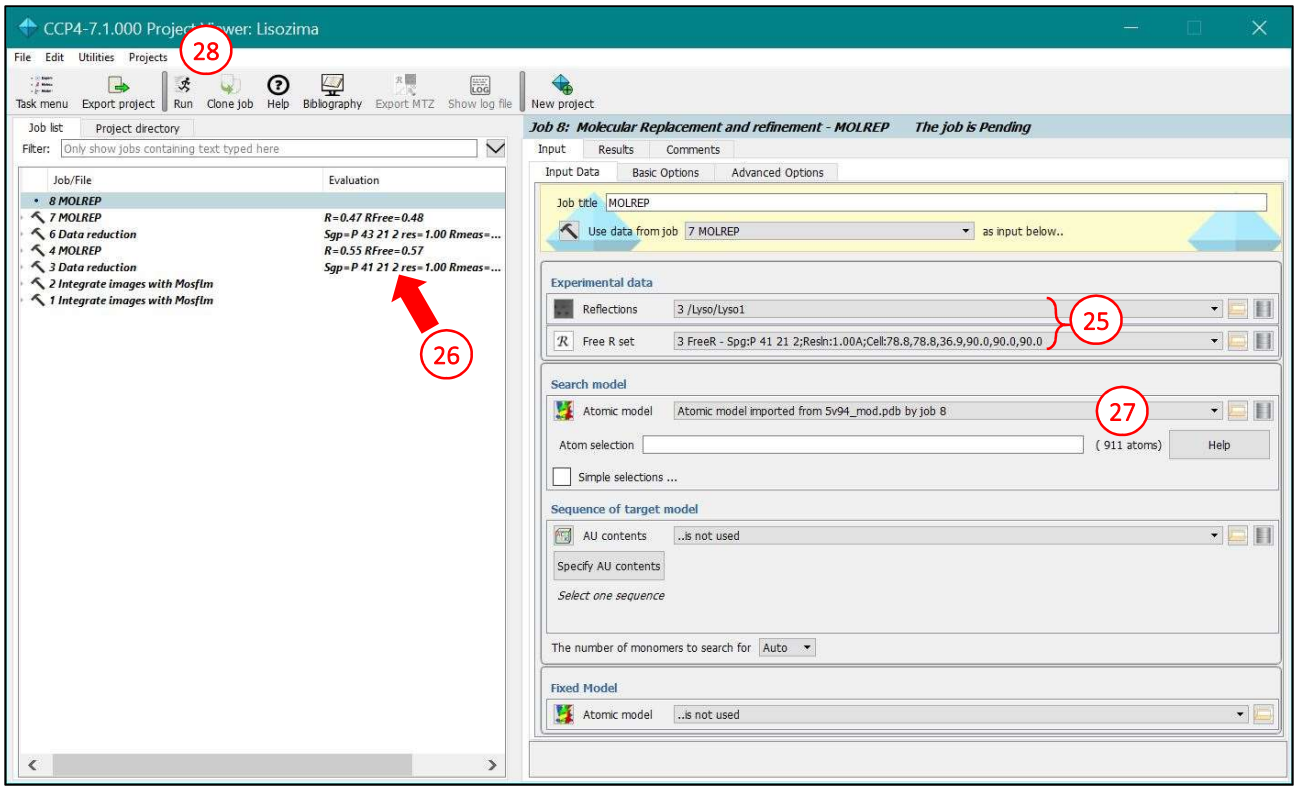
The MOLREP software can be started from the CCP4i2 interface (24). In the window that opens after selecting the program, the input diffraction data, i.e. the *mtz* file containing scaled intensities, can be selected.

Considering the enantiomorphism of the space group, the phasing protocol should be tested for both possible solutions. The first test is conducted with intensities scaled in the *P 41 21 2* space group (25) during the previous "job 3" (26). The menu on the right reports the space group, allowing the user to check the correctness of the desired scaling procedure.

In order to perform the MR search, a second input is required, namely the model previously prepared. The edited *pdb* file is selected in the appropriate space (27).

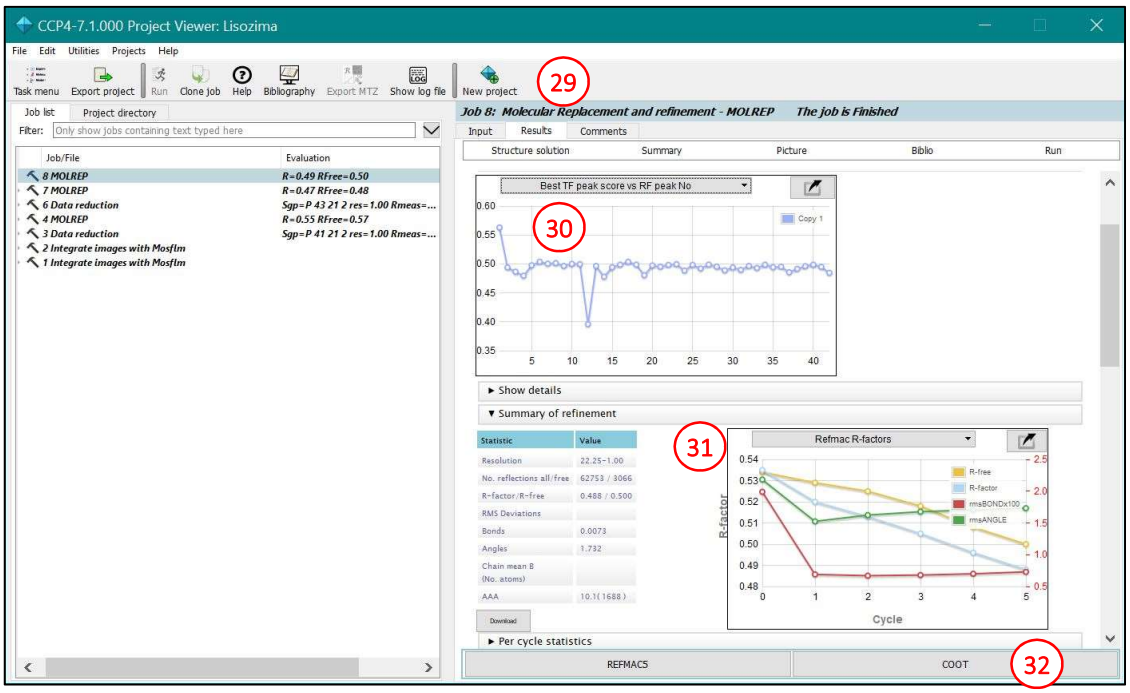
The MR search, corresponding to a rotation matrix and a translation vector search, can be can be started with the "Run" button (28).



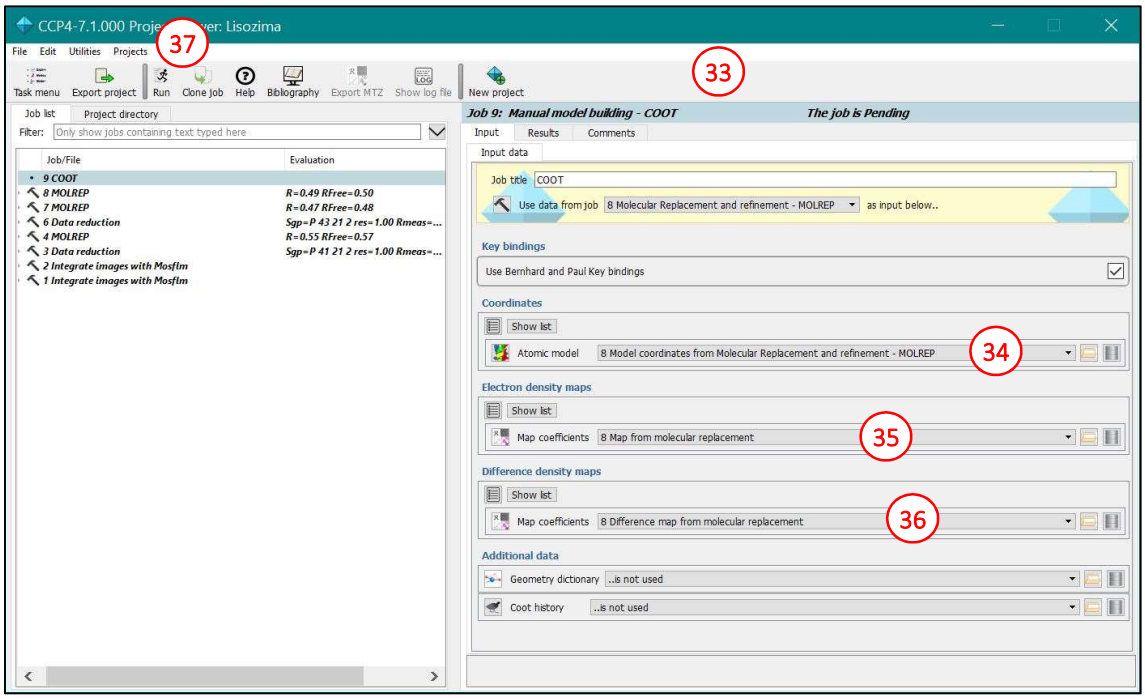


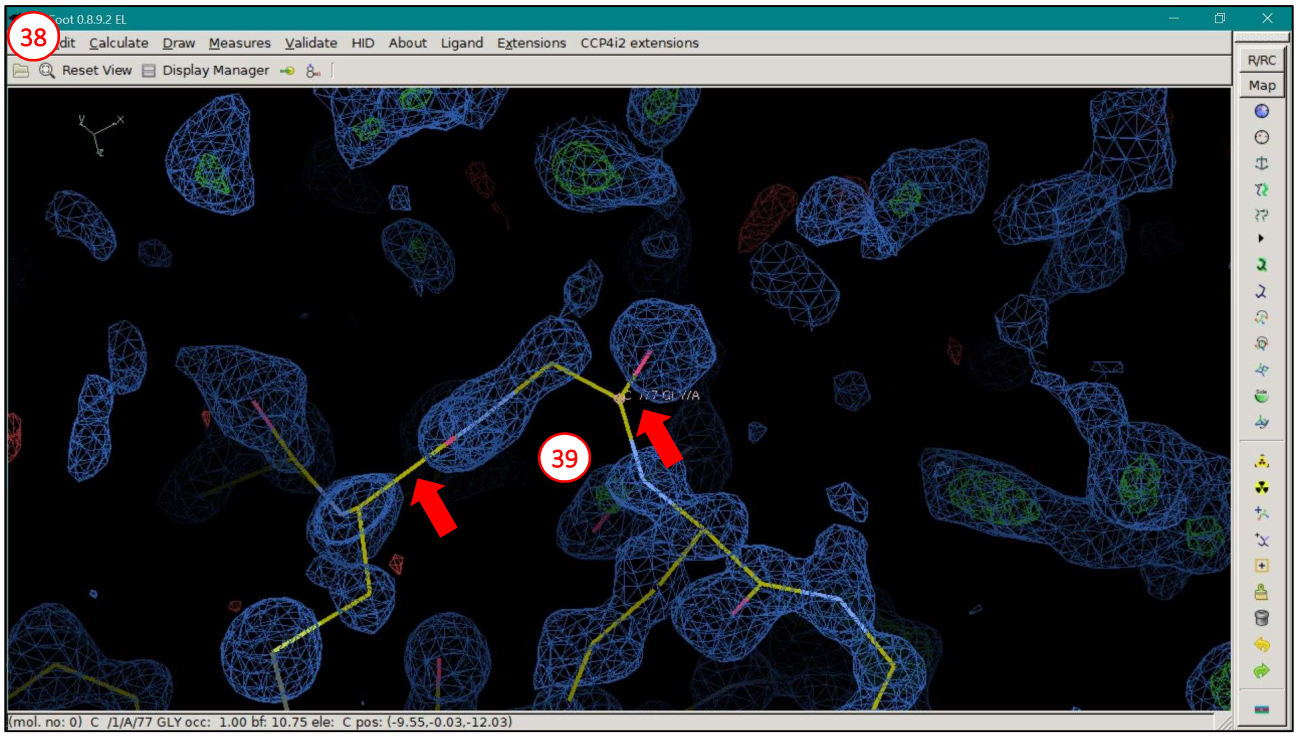
At the end of the calculation, the software provides the best solution identified for the model positioning in the unit cell (**29**). The graph on the right side of the window show the quality of the MR solution. In particular, graph **30** reports the quality of the best translational solution for each of the rotational solutions and, in the example, shows no optimal solution. The software automatically performs a rigid body refinement of the best solution (i.e., refining only the position of the whole protein structure, with no modification allowed on reciprocal positions of the atoms and residues), by recalling the Refmac software. This program yields also

values for the R_{work} e R_{free} factors after refinement. The graph **31** shows the variation of these indexes in the refinement cycles. In this case, the MR solution in the *P* 4_1 2_1 *2* space group yields an R_{work} value of 0.49 and an R_{free} value of 0.50 at the end of the 5th rigid body refinement cycle. These unsatisfactory values are indicative of a possible mistake in the space group choice, but this hypothesis can be confirmed only by testing the MR solution in the other enantiomorphic space group.



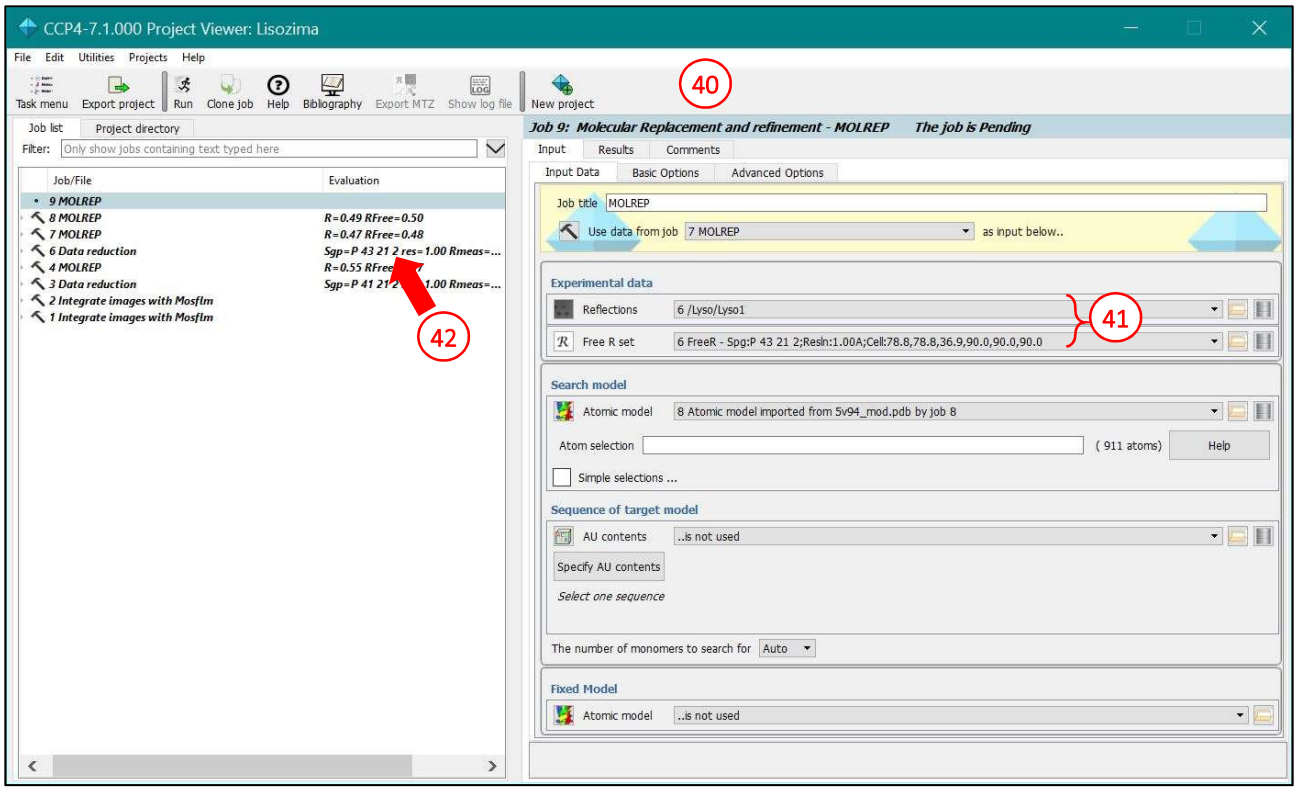
A further indication that the solution obtained is wrong comes from the direct observation of the calculated electron density, compared with the protein model used in the MR. The Coot software, that allows for the electron density inspection, can be started from the CCP4i2 interface (**32**). In the right window (**33**), the user can select both the model protein (**34**) and the data from which electron density (**35**) and difference electron density (**36**) are obtained. The "Run" button (**37**) opens the Coot window (**38**).



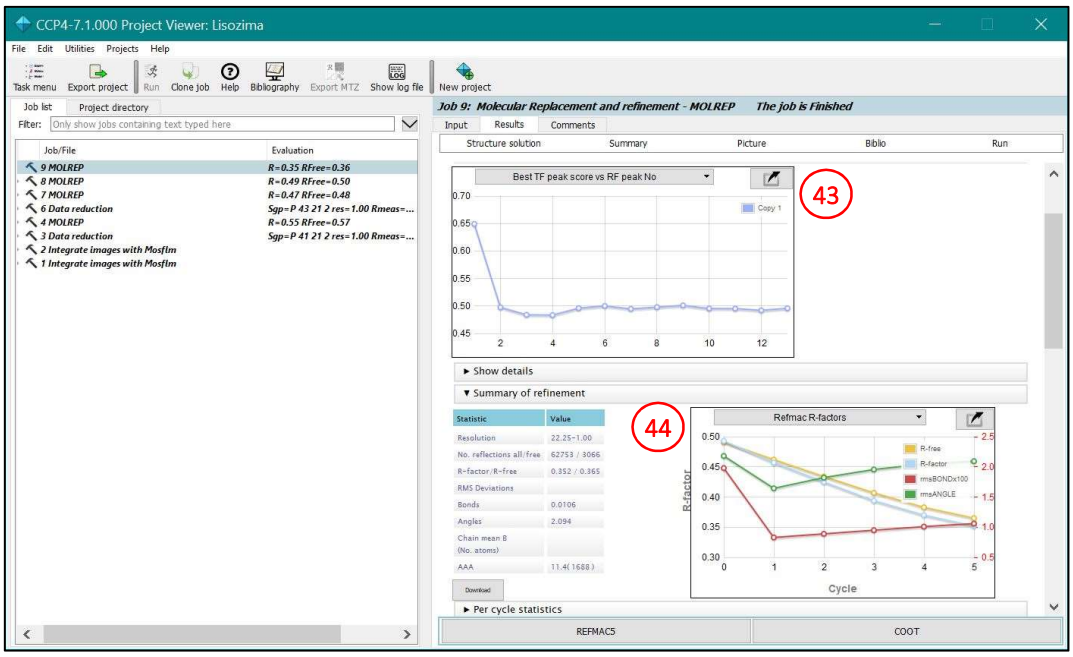


Despite the fact that the *P* 4_1 2_1 *2* is the wrong space group, the user can mistakenly think that there is a similarity between the electron density and the model. This apparent similarity is due to model bias. However, a more careful inspection shows that the density is not continuous, particularly in the main chain (**39**).

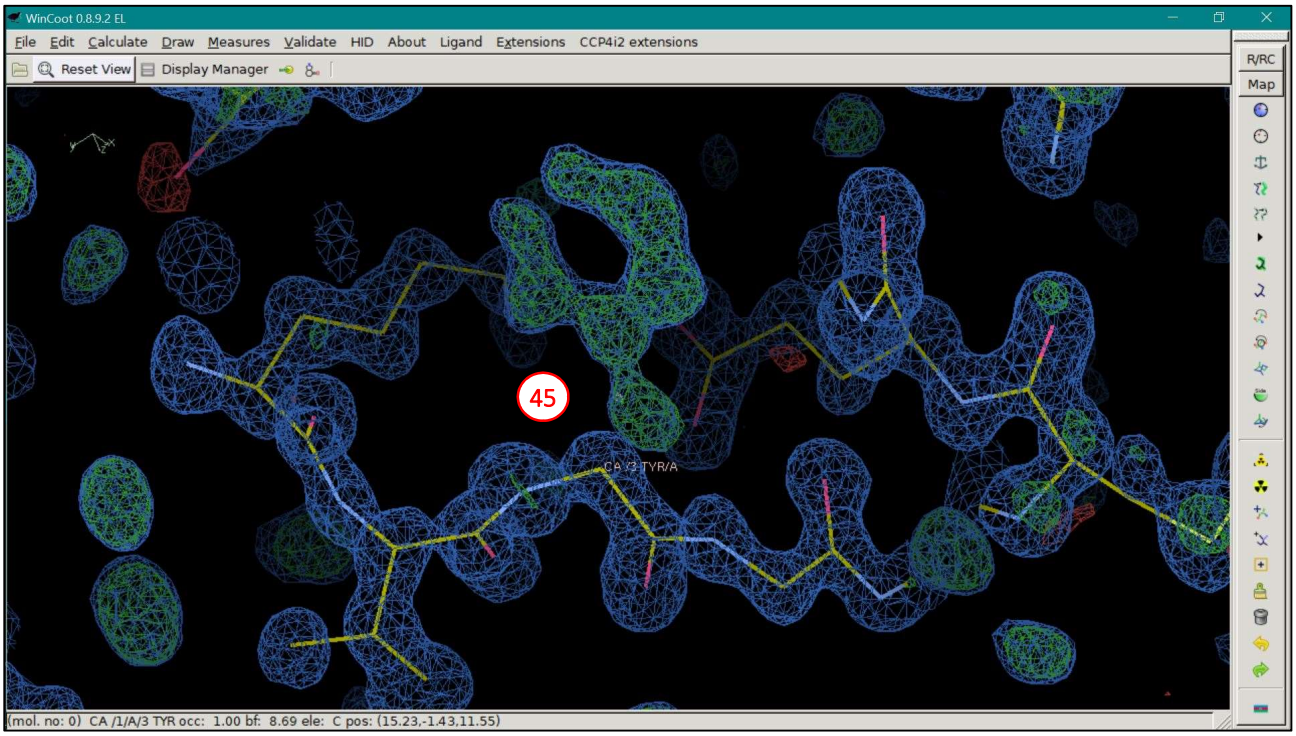
The same steps performed for the *P* 4_1 2_1 *2* solution will be repeated using data scaled in the *P* 4_3 2_1 *2* space group (**40**). In this case, data selected are those obtained after scaling in "job 6" (**41** e **42**).



The MR solution obtained by the MOLREP software for this space group is of higher quality. The graph **43** relative to the best translational solution shows that the best solution stands out among the other, with a significant difference. The presence of a clear optimal solution is an indication that the model has been correctly positioned in the unit cell. In the graph **44**, values of R_{work} and R_{free} (0.35 and 0.36, respectively) indicate a good fitting between model and experimental data, confirming the correct space group choice.



The analysis of the electron density with the Coot software shows a continuous electron density in the main chain. In addition, the calculated electron density predicts the mutation of some residues that differ from the model probe used in MR. For example, a tyrosine in position 3 was removed from the model, due to its mutation to phenylalanine in the analyzed protein. The electron density, **45**, shows the features of the aromatic ring, predicting the correct mutation.



References.

[1] The UniProt Consortium, *"UniProt: a worldwide hub of protein knowledge"*. Nucleic Acids Res. 2019; 47, D506-515.

[2] S. McGinnis, and T.L. Madden, *"BLAST: at the core of a powerful and diverse set of sequence analysis tools"*. Nucleic Acids Res. 2004; 32, W20-W25.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *"The Protein Data Bank"*. Nucleic Acids Res. 2000; 28, 235-242.

[4] A.Vagin and A.Teplyakov, *"MOLREP: an automated program for molecular replacement"*. J Appl Cryst. 1997; 30, 1022-1025.

[5] M. D. Winn et al., *"Overview of the CCP4 suite and current developments"*. Acta Cryst. 2011; D67, 235-242.

[6] G.N. Murshudov, A.A. Vagin, and E.J. Dodson, *"Refinement of Macromolecular Structures by the Maximum-Likelihood method"*. Acta Cryst. 1997; D53, 240-255.

[7] P. Emsley, B. Lohkamp, W.G. Scott, and K. Cowtan, *"Features and Development of Coot"*. Acta Cryst. 2010; D66(Pt 4), 486-501.