Additional Information for

Hydrogen-bonding and hydrophobic interaction networks as structural determinants of microbial rhodopsin function

Éva Bertalan,[1,#] Masae Konno,[2,#] María del Carmen Marín,[2,§] Reza Bagherzadeh,[2,&] Takashi Nagata,[2] Leonid Brown,[3] Keiichi Inoue[2,*], and Ana-Nicoleta Bondar[4,5,*]

[1]RWTH Aachen University, Department of Mathematics and Natural Sciences, Aachen, Germany

[2]The Institute for Solid State Physics, The University of Tokyo, 5-1-5 Kashiwano-ha, Kashiwa, Chiba 277-8581 Japan

[3]University of Guelph, Department of Physics, Ontario N1G 2W1, Canada

[4]Forschungszentrum Jülich, Institute of Computational Biomedicine, IAS-5/INM-9, Wilhelm-Johnen Straße, 5428 Jülich, Germany

[5]University of Bucharest, Faculty of Physics, Atomiștilor 405, Măgurele 077125, Romania

[#]Equal contributions

This additional information contains the protocol to generate the NS-mrho numbers for microbial pump rhodopsins. The protocol, which was prepared by Eva Bertalan and Masae Konno, is provided on an as-is basis. Please report any bugs to eva0bertalan@gmail.com and masaek@issp.u-tokyo.ac.jp.

Sequence alignment

1. Align the sequences using either of the methods shown in A or B.

A. Sequence alignment using the structural data with PROMALS3D

A-1. Access the website of PROMALS3D (http://prodata.swmed.edu/promals3d/promals3d.php)

A-2. Enter a sequence of the target protein to the text box in FASTA format. Another way, upload the sequence file using the “Choose File” button.

A-3. If structural data is available, upload the structure file or enter the PDB ID and the chain ID.

A-4. After all of the sequences are entered, move to “DATA SUBMIT” box and click “Submit” button. If a specific job name is needed, enter a job name in the text box. Enter email address to receive the result if you need it.

A-5. Check the alignment results. Alignment data can be output in three different formats.

- COLORED alignment. The data is output as a html format. Amino acid residues predicted to constitute α-helices are marked as “h” in the “Consensus_ss” row.
- CLUSTAL format alignment. The data is output as an aln format.
- FASTA format alignment. The data is output as a fa format. Note: The data of "COLORED alignment" only includes information on the position of the predicted α-helices.

B. Sequence alignment using CLUSTALW with MEGA software

B-1. Download and install MEGA6 from the MEGA website (https://www.megasoftware.net/)

B-2. Start MEGA6 software.

B-3. Go to “File” > “Open A file/session” to load an already aligned sequence file (fasta format)

B-4. When the pop-up message "Analyze or Align file?" appears, select 'Align' to open the alignment file.

B-5. Select the amino acid sequence of interest (fasta or txt format) in Edit > Insert sequence from file by selecting the bottom line, the sequence of interest will be added to the bottom line of the alignment.

B-6. Select Alignment > Align by ClustalW to align all sequences. If no rows are selected, a pop-up message will ask if you want to select all sequences and click OK.

B-7. A pop-up window for setting ClustalW parameters appears, change the parameters as required and click OK.

2. Manually optimize the sequence alignments obtained by methods A or B as follows.

2-1. Manually remove any unnecessary gaps in the TM region if they exist, based on data that have already been aligned. If necessary, delete low-conserved N- and C-terminal sequences and hydrophilic loops between TMs; if the C-terminal side is deleted, ensure that at least 14 residues remain from the amino acid corresponding to the lysine residue (BR-K216) that forms the Schiff base linkage with the retinal chromophore.

2-2. If the sequence was edited in step 2-1, repeat the alignment using methods A or B.

2-3. Save the aligned data in CSV format.

2-3-1. Create data in Data > Phylogenic analysis

2-3-2. Open the created phylogenetic analysis data in the main window of MEGA.

2-3-3. Uncheck Use Identical Symbol in Display so that all amino acids are displayed in one letter way.

2-3-4. Select File > Export data, as the “Format” can be selected in the Exporting Sequence Data pop-up window. Select “CSV (Excel Importable)” from the dropdown list to save the data as a CSV file in comma-separated format. Also, set the value of “Sites per line” to be larger than the value of the total length of the alignment including gaps.

2-3-5. Delete the “Domain: Data” comment at the top of the file. Add a row at the top of the data showing the position corresponding to the TM region as “H” and the rest as “- (hyphen)” for the entire length of the alignment including the gap. Save the file in CSV format. If the file is opened in Excel, each residue is entered in each cell, making it easy to mark the TM region.

3. Load the CSV file containing aligned sequences into a Pandas DataFrame. Each row represents an aligned sequence, with the first column containing the protein name. Missing amino acids in a sequence are denoted by "-" (gap characters). The header row denotes structural features, with columns labeled as 'H' for helical regions and '-' for non-helical regions.

4. Call the `assign_conserved_numbers()` function from the script, which executes the following steps:
   4.1. Identify helical regions by selecting columns labeled 'H' in the header row.
   4.2. Compute the metrics for each column:
      - Most frequent amino acid: Represents the amino acid occurring most frequently at a given location across the aligned proteins.
      - Occurrence: Denotes how many the most common amino acid occurs at that position within the aligned sequences
      - Conservation: Reflects the ratio of the most frequent amino acid's occurrence at a specific position to the total number of aligned sequences.
      - Amino acid variability: Indicates the diversity of amino acids at a given position in proteins.
   4.3. Determine the most conserved amino acid within each helix: select from each helix the amino acid with the highest conservation. If multiple amino acids have the same conservation value in the helix, select the one closest to the middle of the helix.
   4.4. Assign Location IDs to amino acids within helices, with the most conserved amino acid receiving ID 50 and numbering ranging in both directions. This means the amino acids after the most conserved one get the numbers 51, 52, 53… and the amino acids before the most conserved one get 49, 48, 47…
   4.5. Generate a DataFrame where each column corresponds to a protein sequence, with the protein names as column headers. Then seven additional columns are appended:
      - Helix number: Identifies the transmembrane helix, ranging from 1 to 7 in a 7-transmembrane (TM) protein.
      - Amino acid variability
      - Most frequent value
      - Occurrence
      - Conservation

- Location ID: Assigns numerical identifiers in both positive and negative directions relative to the most conserved amino acid in the helix.
- Assigned conserved number: Combines the helix number and Location ID to create a final label (e.g., the most conserved amino acid residue on helix 2 receives the label 2.50).

These values are computed for each amino acid in the sequence, resulting in a table that can be saved as a CSV file.

5. Renumber the PDB files:

Subsequently, the script can renumber PDB files based on the assigned conserved numbers.

5.1. Call the renumber_pdb() function, which is using the table generated in the assign_conserved_numbers() step. In the function arguments a mapping has to be provided, which is matching the PDB ID to the protein name in the sequence alignment.

5.2. The script performs the following steps:

5.2.1. The PDB structure is read with MD Analysis.

5.2.2. The script identifies the first part of the sequence which matches the residues from the PDB structure, as the PDB structure often begins at a later amino acid than the full protein sequence.

5.2.3. The script then goes through the amino acids in the PDB structure, replacing original amino acid types with placeholders and overwriting the residue IDs with the assigned conserved numbers. For example, the most conserved amino acid in helix two would have a residue ID of 250 in the renumbered PDB file.

5.2.4. The data is saved as a CSV file as well, containing the original sequence, the assigned conserved numbers, and properties calculated in the `assign_conserved_numbers()` step (amino acid variability, occurrence, conservation).

This script facilitates the renumbering of protein sequences included in sequence alignments or structures with point mutations, where the original sequence was part of the alignment.