# Illuminating environmental adaptation of proteorhodopsins based on their distribution and ecophysiological traits in the Yangtze River

Tianyi Chen1,2,3, Qi Chen1, Liyu Sun2,3, Guoping Hu4, Malan Yi5, Zongzhi Wu2,3, Bingxue Wang2,3, Chenyuan Dang6, Tang Liu1,3✉ and Jinren Ni3

1Environmental Microbiome Engineering and Innovative Genomics Laboratory, College of Chemistry and Environmental Engineering, Shenzhen University, Shenzhen 518060, China.

2School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen 518055, China.

3College of Environmental Sciences and Engineering, Key Laboratory of Water and Sediment Sciences, Ministry of Education, Peking University, Beijing 100871, China.

4Ganjiang Innovation Academy, Chinese Academy of Sciences, Ganzhou 341000, China.

5Tianjin Research Institute for Water Transport Engineering, M. O. T, Tianjin, 300456, China.

6School of Environmental Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.

✉Correspondence to: Tang Liu, Chemistry and Environmental Engineering, Shenzhen University, Shenzhen 518060, China

E-mail addresses: liutang@szu.edu.cn (T. Liu)

**Supplementary Data**

Supplementary Data 1. Raw protein sequences of identified 816 proteorhodopsins in the Yangtze River.

Supplementary Data 2. Raw protein sequences of 200 non-redundant reference rhodopsins of nine different types.

Supplementary Data 3. Raw protein sequences of 120 used reference proteorhodopsins in the present study.

Supplementary Data 4. Phylogenetic tree file of 816 putative proteorhodopsins in the Yangtze River and 200 reference rhodopsins with different types.

Supplementary Data 5. Phylogenetic tree file of 816 putative proteorhodopsins in the Yangtze River and 120 used reference proteorhodopsins.

**Supplementary Tables**

Table S1. Sequencing depth for each of the 51 sequenced samples in the Yangtze River.

Table S2. Motif summary of identified PR genes in the Yangtze River.

Table S3. Completeness, contamination and total scaffold size of PR+ and PR- genomes.

Table S4. Calculated Z-score of selected gene annotations (GO/eggNOG) within comparison genomes.

Table S5. Calculated reporter scores of selected KEGG modules/pathways within comparison genomes.

**Supplementary Figures**



**Fig. S1. Phylogenetic tree of rhodopsins.** Maximum-likelihood method with 1000 bootstrap replications is applied for the construction of the phylogenetic tree. Reference sequences (marked in grey strips) cover 200 non-redundant rhodopsins (clustered at 70% average amino acid identity), including used publicly-reported PRs (see Methods) and other types of rhodopsins downloaded from the NCBI protein database. The black strips represent the 816 putative PR sequences identified in the Yangtze River.



**Fig. S2. Phylogenetic tree of identified proteorhodopsins in the Yangtze River.** The branch length displays the relative genetic distance within 120 publicly reported PR references and 816 putative PRs in the present study.Corresponding ion pumping, spectral tuning and phylum classification are marked in colored strip for each sequence.



**Fig. S3. Averaged copy number of PRs with different spectral tuning variants.**



**Fig. S4. Linear correlation between averaged gene copies of PRs and gradients of environmental factors.** Correlation coefficient and *p*-values are denoted in the chart.



**Fig. S5. Mean genome relative abundance of PR+ and PR-(CR) genomes across landforms.** The filled dot represents the averaged abundance value within each landform type. Statistical significanceis estimated by Bonferroni-adjusted Wilcoxon test, marked in asterisks (\*\*\*\*: ≤0.0001; \*\*\*: ≤0.001; \*\*: ≤0.01).



**Fig. S6. Differences in amino acid composition between PR+, PR-(CR) and PR-(NCR) genomes.** Statistical significanceis estimated by Bonferroni-adjusted Wilcoxon test for multiple-range comparisons, marked in asterisks (\*\*\*\*: ≤0.0001; \*\*\*: ≤0.001; \*\*: ≤0.01; \*: ≤0.05; ns: >0.05).