Discovery of NOT-GPCRs: An Archaeal Membrane Protein Family Whose AlphaFold Predictions Structurally Resemble G Protein-coupled Receptors

Koya Sakuma[1,2]

1. Department of Basic Biology, Cellular and Structural Physiology Institute, Nagoya University, Nagoya, Japan
2. Department of Basic Medicinal Sciences, Graduate School of Pharmaceutical Sciences, Nagoya University, Nagoya, Japan

Corresponding author: K.S.

ksakuma {at} cespi.nagoya-u.ac.jp

Keywords and Abbreviations:

MGII; Marine Group II

GPCR; G protein-coupled receptor

bR; Bacteriorhodopsin

Data mining

Structure-function relationship

Abstract

G protein-coupled receptors (GPCRs) are a prominent class of membrane proteins recognized for their significance in human pharmaceutical research and central roles in eukaryotic cell communication. The author hypothesized that GPCR-like proteins might be distributed across domains other than eukaryotes and aimed to discover them via structure-based data mining. Comprehensive structure-based searches against the AlphaFold protein structure database identified a family of uncharacterized membrane proteins in *Methanobacteriati* (formerly *Euryarchaeota*) whose AlphaFold predictions structurally resemble the canonical GPCR fold. They were termed **N**ot **O**bjectively **T**rue GPCRs (NOT-GPCRs) because they show higher structural similarity to GPCRs than to archaerhodopsin and bacteriorhodopsin, while lacking evidence of coupling to G proteins. In addition to the GPCR-like domain, they possess putative eighth and ninth transmembrane helices on the C-terminus, and a subtype possesses EF-hand motifs in the probable cytosolic domain following them. The physiological and environmental significance of the genes remains entirely unknown. Therefore, NOT-GPCRs are structurally appealing, yet offer no biologically grounded interpretation at this stage. When seen from the viewpoint of metascience, this situation could illustrate *the Rorschach limit of structural biology* in the post-AlphaFold era. To ultimately reconnect the contextless atomic coordinate data to biology, experimental validation will be necessary to confirm the predictions, elucidate their functions, and uncover their evolutionary origin. Equally important will be the development of a novel interpretative framework in structural biology to reconsider the dogma *Structure Determines Function*.

Introduction

Structure comparison is a powerful tool for detecting distant relationships between proteins because tertiary structures tend to be more conserved than sequences, offering alternative ways to find hidden similarities among genes that are difficult to find by sequence similarities (1). Recent advancements in machine learning-based structure predictors could allow us to identify uncharacterized proteins distantly related to the query proteins of interest via structure-based similarity searches against databases or customized pools of predicted structure models (2–5). To demonstrate the effectiveness of this approach, the author composed a simple and straightforward question to be answered solely by structure-based similarity searches: Do proteins that structurally resemble G protein-coupled receptors (GPCRs) exist in domains other than the eukaryotes?

GPCRs are a class of membrane proteins well known for being distributed widely among eukaryotes. They play central roles in eukaryotic cell communication and are particularly prominent in pharmaceutical contexts, making them one of the most well-characterized classes of proteins (6–8). However, their evolutionary origin and phylogenetic distribution are of less research interest than their functions, and they are believed to be exclusively distributed in eukaryotes (9,10). Inspired by this gap and the evolutionary relationship between eukaryotes and archaea (11,12), the author hypothesized that the GPCR fold, or even GPCR ancestors, could be found in archaeal species. To test this, a dedicated protocol was designed to detect GPCR-like structures across the entire AlphaFold Protein Database (AFDB) (3).

Materials and Methods

Selection of structure-based search query

Structure-based search requires careful selection of query structure; the selection criteria can be arbitrary but should generally prioritize the resolution as a conventional metric of the quality of structures. The author investigated the Protein Data Bank to identify the highest-resolution GPCR structure eligible for a search query. While crystal structures of A2A adenosine receptor (PDB ID: 5IU4 and 5NM4) showed the highest resolution of 1.7 Å, these were chimeric proteins fused to thermostabilized apocytochrome b562 (so-

called BRIL). To avoid the unnecessary bias introduced by human interventions to define the structural boundary between GPCR and BRIL parts, the author avoided these two structures and instead selected a crystal structure of bovine rhodopsin (PDB ID: 8A6C), which is one of the second highest-resolution (1.8 Å) structures without fusion proteins.

Structure comparison iterated across the entire AlphaFold Protein Structure Database

AlphaFold Structure Database v4 UniProt (3) was downloaded from the Foldcomp database using Foldcomp version 0.0.2 installed via pip (13). An in-house Python script accessed the downloaded Foldcomp-format AFDB via the Foldcomp Python module to iterate structure comparisons over all the database entries. In each iteration, coordinate data were decompressed into the PDB format and input into MICAN (14) for structure alignment in the rewiring and reverse mode onto the representative bovine rhodopsin structure. Note that MICAN allows non-sequential structure comparison, where the order and orientations of secondary structure elements (SSEs) can be ignored in the rewiring and reverse mode; the author intended even to hit rewired versions, for example, circular permutations and reverse versions of GPCR folds, though they were not found. The entry-wise TM-score values were reported for downstream analysis. To keep the output file size smaller, the outputs from the reports were excluded if the TM-score normalized by the residue number of the reference rhodopsin structure was lower than 0.60 because they are considered in different folds (15). In actual computations, the Foldcomp database of AFDB was divided into 1,074 segments so that structure comparisons were parallelly executed on multiple nodes of clustered machines, and their respective reports were subsequently concatenated.

Initial identification of NOT-GPCR entries

Because the structure comparison detected many known GPCR-like structures, a list of GPCR entries was constructed if any of the following InterPro IDs were linked to the entry: IPR000276, IPR000832, IPR000337, IPR000366, IPR000848, IPR000539, and IPR022340 for classes A–F and putative plant GPCRs

(16), respectively. Using this as a negative list helped remove the entries recognized as GPCRs, reducing trivial data analysis. After removing trivial GPCR entries, the AFDB structures were clustered by MMseqs2 (17) applying the 35% sequence identity threshold. Each cluster was labeled by the NCBI Taxon ID at the superkingdom level most and second most frequently found in the cluster to indicate its taxonomic origins. Unless a cluster was assigned the taxonomic label of Archaea (NCBI taxonomic identifiers: 2157), the cluster was omitted from the following procedure. The resultant cluster representatives were again structure-aligned to the reference rhodopsin structure by MICAN (14) in the sequential mode. The cluster was removed if the representative structure showed the SSE match-weighted TM-score lower than 0.50 when normalized by the aligned length (18). Then, the author visually inspected all of the remaining cluster representatives to remove unwanted entries, such as eukaryotic GPCRs occurring in archaeal sequencing data; such genes are considered contamination for their lack of conservation.

Enhancing the comprehensiveness of NOT-GPCR search

Structure-based searches are generally sensitive to the reference (query) structure choice, and thus the initially identified NOT-GPCR structure could not be comprehensive and might miss possible NOT-GPCR entries. Therefore, the list of NOT-GPCR entries initially identified by structure comparisons was expanded to include evolutionarily related entries to ensure better coverage. First, the entries were mapped to the AFDB cluster (19) to identify structurally similar entries, and the cluster members possessing the same cluster representatives were considered NOT-GPCR candidates. This caused some clusters to include many eukaryotic GPCR entries because of their structure similarity, and the entries with eukaryotic taxonomic identifiers were removed. Second, the resultant list of AFDB IDs was mapped to the UniRef50 clusters to which each entry belongs (20). All of the members of these UniRef50 clusters were mapped to corresponding AFDB entries to obtain structure models, although some UniProt IDs referred to by the UniRef50 cluster could not be mapped to AFDB. All of the retrieved models were reviewed visually to remove contaminant structures and ensure the quality of the final entry list. The entry with AFDB ID A0A2U2RKL6 was the only contaminant originating

from the AFDB cluster represented by A0A1J8QVB6 and was removed. Third, these NOT-GPCR structures found in the AFDB were again mapped to the AFDB cluster to find entries sharing the same cluster representative with those appended via the UniRef50 search. The AFDB cluster members from domains other than the eukaryotes (NCBI taxonomic identifiers for eukaryotes: 2759) were retrieved from AFDB and again subjected to visual inspection to decide whether to include them in the final list, similarly to the previous cycle. Table S1 summarizes the resultant set of NOT-GPCR AFDB structures and UniProt-only candidates. The annotations, including phylogenetic information, were retrieved from UniProt for the downstream analysis (21).

Identification of an intact and non-redundant subset of NOT-GPCR, GPCR, and microbial rhodopsin structures

The expanded set of NOT-GPCRs identified by the method described above contained some fragmented structures that should be removed before detailed analysis. To find the representative NOT-GPCR structure, all structure models were sorted by their average Cα atom pLDDT scores and reviewed visually. The highest-score structure with all nine transmembrane helices intact was selected as the representative NOT-GPCR structure (AFDB ID: A0A327H4A3). Then, a subset of NOT-GPCR entries possessing an intact seven-transmembrane GPCR-like substructure was identified by structure alignment to residues 1–330, the GPCR-like substructure of the reference NOT-GPCR structure, by MICAN (14). The aligned structure was considered intact if it covered ≥85% of the reference structure. Among these intact structures, redundant entries were removed by sequence-based clustering using MMseqs2 (17) with a sequence identity threshold of 99%. Table S2 summarizes the resultant set of intact and non-redundant NOT-GPCR-predicted structures.

Microbial rhodopsin, referred to as bR hereafter, and GPCR structures were also curated using similar methods. As for GPCRs, the PDB IDs of available experimental GPCR structures were collected from GPCRdb (22) as of Jan 21, 2025. PDB IDs for bR structures, including archaerhodopsin and bacteriorhodopsin, were corrected using the UniProt search system and the RCSB-PDB advanced search as of Jan 21, 2025, by

using InterPro ID IPR001425 assigned for archaeal/bacterial/fungal rhodopsin as a query (16,21). All of the PDB entries were downloaded from PDB in mmCIF format for each class of proteins. All of the mmCIF files were parsed into chain-wise single PDB files. A bovine rhodopsin structure (PDB ID: 8A6C chain A) was taken as the reference structure for GPCRs, and a bacteriorhodopsin structure (PDB ID: 1Q5I chain A) for bRs, to which other structures were aligned. The aligned structure was considered intact if it covered ≥70% of the reference GPCR or bR structure's amino acid residues, excluding structures with many missing residues. Because GPCR structures are sometimes found to be fused to other proteins like T4-lysozyme to facilitate structure determination, all of the GPCR structures were parsed into domains using an in-house implementation of Protein Domain Parser (23,24) after first alignments. These domains were again structure-aligned to the reference bovine rhodopsin and identified as GPCR if they covered ≥70% of the reference structure's amino acid residues. For these respective sets of parsed GPCRs and intact bRs, redundant entries were removed by sequence-based clustering using MMseqs2 with a sequence identity threshold of 95%. Tables S2–S4 summarize the resultant set of intact and non-redundant GPCR and bR PDB-deposited structures.

All-against-all structure comparison among intact and non-redundant sets of NOT-GPCR, GPCR, and bR structures

The union of the intact and non-redundant subsets of NOT-GPCR, GPCR, and bR structures was used for the all-against-all structure comparison. The set contained 604 structures in total. The structure alignment was performed using MICAN (14) in sequential alignment mode. The TM-score normalized by the reference structure length and sequence identity based on the structure alignment were reported for all of the pairs of these structures. The TM-scores were used for inter-class and intra-class structure similarity analysis.

Subtype identification among NOT-GPCRs

The sequence identities based on the NOT-GPCR part of all-against-all structure alignment were used for subtype identification of NOT-GPCRs. The entities were clustered by hierarchical clustering using the

complete method with the dissimilarity matrix of sequences, whose matrix elements are pairwise sequence identities divided by 100 and then subtracted from 1. The cluster number with the maximum average silhouette value was considered optimal when the cluster number was scanned from 2 to 10. This ensures the cluster number of the hierarchical clustering result in maximum contrast between inter-cluster and intra-cluster distance, where the former needs to be maximized and the latter needs to be minimized.

Prediction of the orientation and position in the membrane

To infer the membrane orientation of NOT-GPCRs, TMbed (25) was used to predict the class probabilities (inside, outside, and transmembrane helix for each amino acid residue). The average probabilities over all predictions for intact and non-redundant structures were mapped onto the representative structure to infer average topology. Along with the membrane orientation, PPM 3.0 (26) was used to predict the position and pose of the representative NOT-GPCR structure within archaebacterial cell membranes.

Proteome-wide G protein identification

All UniProt entries in the proteomes containing NOT-GPCRs were downloaded to create a sequence database (21). A BLASTp search was performed using the amino acid sequence of the human G protein α subunit as a query, but no relevant hits were found. The UniProt IDs from the proteomes were mapped to 140,619 AFDB IDs to identify more distantly related proteins by structure comparison, and all AFDB entries were downloaded. These structures were aligned to the α subunit from a trimeric G protein crystal structure (PDB ID: 6CRK chain A), and their TM-scores were reported. Note that the α subunit subtypes (i/o, q/11, 12/13, and s) are structurally similar and can be queried using the same reference structure.

Figure Preparation

Images of molecular structures were rendered by open-source PyMOL 2.5.0 (27), where STRIDE was used to assign secondary structures (28).

Results

Database search results

Among more than 214 million entries in AlphaFold DB, 913,143 structures showed TM-scores of ≥0.60 against the reference bovine rhodopsin structure (15). The initial structure-based search was followed by additional structure- and sequence-based searches to improve coverage and robustness, since structure-based searches are inherently sensitive to query selection. After removing contaminating structures, 287 AFDB entries of non-eukaryotic proteins showing GPCR-like backbone conformations were identified and named Not Objectively True GPCRs (NOT-GPCRs). Among these 287 structures, 242 were considered structurally intact, and 173 out of these 242 entries were extracted to construct the intact and non-redundant subset (Table S1). The overall reliability of AlphaFold prediction was moderate for membrane proteins; the minimum, mean, and maximum values of the entity-wise Cα-averaged pLDDT among the non-redundant and intact NOT-GPCR structure subset were 75.85, 80.69, and 87.59, respectively, which were 79.99, 85.11, and 89.83 when focusing solely on the GPCR-like domains.

Table 1 summarizes the phylogenetic distribution of the NOT-GPCRs. Although the author had expected the Asgardarchaeota to be the primary source given its similarity to eukaryotes (29–31), the dominant organisms were archaeal species classified under *Euryarchaeota*, which has recently been proposed for renaming to *Methanobacteriati* (32). The source organisms included marine group II (MGII) archaea (33,34), whose proposed name is *Candidatus* Poseidoniales (35); MGII is one of the most abundant planktonic groups of archaea prominent in marine environmental and ecological studies, which remains enigmatic since the pioneering discovery of non-thermophilic classes of archaea widely distributed in the global ocean (33,36). To my knowledge, species in MGII remain uncultured even today (37). Consistently, NOT-GPCRs originated mainly from metagenome-assembled genomes (MAGs) (35,38–46) built on metagenomic sequence data from,

for example, *Tara* Oceans (47). These facts mean that almost no prior physiological and molecular biological knowledge can be employed to boost the inference on the biological functions of NOT-GPCRs.

The overall AlphaFold2-predicted structure of NOT-GPCRs

Typical NOT-GPCR structures possess nine transmembrane (TM) helices, where the arrangement of the seven N-terminal helices results in seven transmembrane (7TM) topologies (Figure 1A). Apart from the C-terminal eighth and ninth transmembrane helices and cytosolic domain, the predicted overall structure and the membrane orientation are reminiscent of the canonical 7TM GPCR fold and topology (Figure 1B). Therefore, the 7TM parts and the two C-terminal α-helices were named the "GPCR-like domain" (GLD) and the "auxiliary helices" (AuxHs) because these can be considered separate parts without tight atomic contacts.

To highlight typical features of NOT-GPCR structures, a predicted model (AFDB ID: A0A327H4A3) was selected on the basis of the intactness of the secondary structures and average Cα pLDDT values. The membrane orientation prediction of NOT-GPCRs by TMbed (25) reported that the N-terminus was exposed to the extracellular side (Figure 1C and D), consistent with the spatial position of the probable disulfide bond-forming cysteines conserved in the extracellular domain bridging the fourth and fifth TM helices. Interestingly, the extracellular domain of NOT-GPCRs serves as the junction between the fourth and fifth TM helices (Figure 1C), coinciding with the location of extracellular loop 2 (ECL2) of GPCRs, which is known to be structurally diverse.

Similarity between NOT-GPCR and GPCR/bR structures

Visual inspections of NOT-GPCR structures evoke impressions that the GLD part of NOT-GPCRs is reminiscent of the GPCR fold. The GLD not only shares the 7TM topology with the canonical GPCR fold, but also mimics GPCR's complicated packing pattern of TM helices (Figure 2A). To highlight these nuances in the arrangement of TM helices, it might be more informative and accessible to compare NOT-GPCRs with both GPCR and bR structures, contrasting their similarity and dissimilarity. Bacteriorhodopsin (bR) is another

prominent class of 7TM integral membrane proteins, which is no longer considered evolutionarily closely related to GPCRs (10). Upon visual inspection, it can easily be seen that GPCR and bR structures share 7TM topologies, but their spatial arrangement and packing pattern of TM helices differ markedly. For example, the pair of fourth and fifth TM helices overhang the third TM helix in NOT-GPCR and GPCR structures, with the fourth helix protruding from the plane defined by the third and fifth helices, but this back-and-forth trajectory of backbone trace is not observed in bR structures (Figure 1B). These observations highlight that these detailed structural characteristics, in addition to the shared basis of 7TM topology, make major contributions to the perceived resemblance between NOT-GPCR and GPCR structures.

Systematic structure comparisons were performed in silico to quantitatively evaluate the structural similarity between NOT-GPCR and GPCR. A set of non-redundant and intact GPCR and bR structures was constructed from available experimentally solved structures and compared with the predicted structures of NOT-GPCR. The average TM-score between NOT-GPCR and GPCR structures normalized by the amino acid numbers of referenced GPCR structures was 0.625 with a standard deviation of 0.035, indicating that the two structures are in the same fold. By contrast, bR structures showed an average TM-score of 0.532 with a standard deviation of 0.032. As shown in Figure 3 and Table 2, these statistical trends indicate a higher chance of NOT-GPCR structures being classified as the GPCR fold than the bacteriorhodopsin fold (15), while sequence identities of the aligned regions show the sequences of GPCRs and bRs are similarly distant from those of NOT-GPCRs (Table 3). In this sense, the GLD of NOT-GPCRs can be more structurally akin to GPCRs than bRs.

Classification of NOT-GPCR subtypes

The author noticed that NOT-GPCRs appear to have several structure subtypes by visually inspecting all of the curated NOT-GPCR models. Inspired by this observation, the non-redundant and intact subset of NOT-GPCR entries was clustered into five classes using sequence identities based on structure alignments as the similarity metric. The five types were named types 1–5; their populations, means and standard deviations

of sequence identities, and members for respective types are summarized in Tables 4, 5, and S5. The transmembrane-helix parts were quite similar among these classes, and the extracellular and intracellular domains more strongly determined the type (Figure 4A and 4B). The most dominant class, type 1, possesses two clear α-helices that form extracellular domains, while the other types show loop-like unstructured conformations in the corresponding regions (Figure 4A). The type 2, 3, 4, and 5 classes show loop-like conformations in place of the α-helices in the extracellular domains, whereas types 3 and 4 possess similar conformations in this region.

The intracellular domain of NOT-GPCR is located in the C-terminal region of the amino acid sequence, following the eighth and ninth AuxHs. In types 1, 2, 4, and 5, the conformation of the intracellular domain is α-helical with three α-helices forming an orthogonal bundle, sometimes followed by long loops and a four-stranded β-meander structure (Figure 4B and 4C). Some entries lack the β-meander or both domains, but this requires further investigation because it could be attributable to truncation in the original genomic sequence data, not structural variations.

Types 3 and 4 have similar extracellular domains; therefore, their structural distinctions depend on the conformation of intracellular domains. Among all of the classes, type 3 has a distinct structure of intracellular domains with more SSEs spatially arranged in complicated ways. Surprisingly, type 3 intracellular domains possess a pair of EF-hand motifs that form a typical calcium-binding substructure showing an α-EF-β-α-α-EF-β-α SSE sequence (Figure 4B, 4C) (48). This suggests this type of NOT-GPCR has something to do with divalent ions, but nothing biologically grounded can be said at this stage.

No α-subunit of trimeric G proteins found in the source organism genomes

A proteome-wide structure search using the human G protein α subunit (PDB ID: 6CRK chain A) as a query was performed to determine whether the source organisms have trimeric G proteins. The maximum value of the TM-scores observed was 0.454 for putative elongation factors (like AFDB ID: A0A358CVG8). This suggests that none of the organisms has G proteins that should couple with GPCRs. Of course, the absence

of the genes in the source database does not necessarily indicate their absence in these organisms; instead, it may simply reflect incomplete sequencing data. However, the author concludes here that NOT-GPCRs cannot be considered GPCRs; no evidence suggests they are coupled with G proteins. It is worth noting that, unlike bona fide GPCRs, the intracellular side of GLDs lacks pockets to interact with effector proteins, reinforcing this conclusion.

Conclusion

In this study, it was demonstrated that structure-based gene mining can detect distantly related proteins beyond the reach of conventional sequence searches. The results revealed that GPCR-like structures were distributed unexpectedly among certain types of archaea, but the author considers they are more likely to be structural analogs than GPCR homologs (49). Nevertheless, it would still be exciting to delve into the origin and history of GPCRs regarding the evolutionary trajectory between the last universal common ancestor and the last eukaryotic common ancestor. This would require further efforts to achieve high-quality sequencing of the genomes of diverse and uncharacterized archaea. Since the present study relies solely on predictions, experimental structure determination is essential to validate the discovery. Notably, the physiological roles of NOT-GPCRs remain entirely unknown, offering a fascinating opportunity to initiate entire biological investigations from structural bioinformatics hypotheses. Investigating the biology of MGII under natural conditions and developing cultivation strategies for MGII species are both critical for uncovering the biological significance of this enigmatic membrane protein family from the dark proteome of archaeal dark matter.

One might speculate that NOT-GPCRs participate in essential functions such as quorum sensing; however, such hypotheses remain entirely speculative. NOT-GPCR structures demonstrated that atomic coordinate data isolated from established biological contexts resist interpretation within existing narrative frameworks of structural biology; if structure truly determines function, why cannot NOT-GPCR structures directly tell their physiological significance? This could be one of the modern challenges in structural biology: the logic of structural biology becomes powerless when confronted with contextless structures, a modern

computational revival of the problem that challenged structural genomics in the late 1990s and early 2000s (50–55). This problem is well known, yet rarely acknowledged, and remains unresolved. Therefore, AlphaFold's most profound impact has been to reveal that coordinate data are, in themselves, devoid of meaning without contexts.

In the Rorschach limit of structural biology, where biological context reaches zero, interpretation becomes projection: subjected to a structural biology version of inkblot tests, we see only what we've internalized as structural biology, probably because we humans have not been trained to predict functions from coordinates. NOT-GPCR could be an extreme example of such structures, attractive enough to interpret but too isolated from biology. In the post-AlphaFold era, the real challenge facing structural biologists is not just the loss of supremacy in producing coordinate data, but the intrinsic non-obviousness of biological interpretation. To reconnect the contextless atomic coordinate data to biology, experimental validation will be necessary to confirm the predictions, elucidate their functions, and uncover their evolutionary origins. Equally important will be the development of a novel interpretative framework to reconsider what the core assumption in structural biology, *Structure Determines Function*, truly means.

Acknowledgments

The author acknowledges that this study was initiated in the previous affiliation, Graduate School of Informatics at Nagoya University, although it was excluded from the formal list of current affiliations for clarity. The author thanks Motonori Ota (Nagoya University) for providing computational resources that facilitated the recognition of NOT-GPCRs; Hideaki E. Kato (The University of Tokyo) for suggesting the identification of G proteins to assess whether NOT-GPCRs are true GPCRs; and Satomi Niwa (Osaka University) for discussions on structural features that make NOT-GPCRs resemble GPCRs. Some parts of the computations were executed at the Research Center for Computational Science, Okazaki, Japan (Project: 23-IMS-C188), and on the Type II subsystem of the supercomputer *Flow* at the Information Technology Center, Nagoya University, as part of the Nagoya University HPC Project. The author acknowledges the Program for

Young Researcher Units for the Advancement of New and Undeveloped Fields hosted by the Institute for Advanced Research at Nagoya University. Finally, the author thanks Edanz for editing a draft of this manuscript (https://jp.edanz.com/ac). However, the author should note that massive destructive changes were made before preprint submission, letting readers avoid unnecessary misgivings about their proofreading quality.

Declaration of Competing Interests

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Figures and Tables

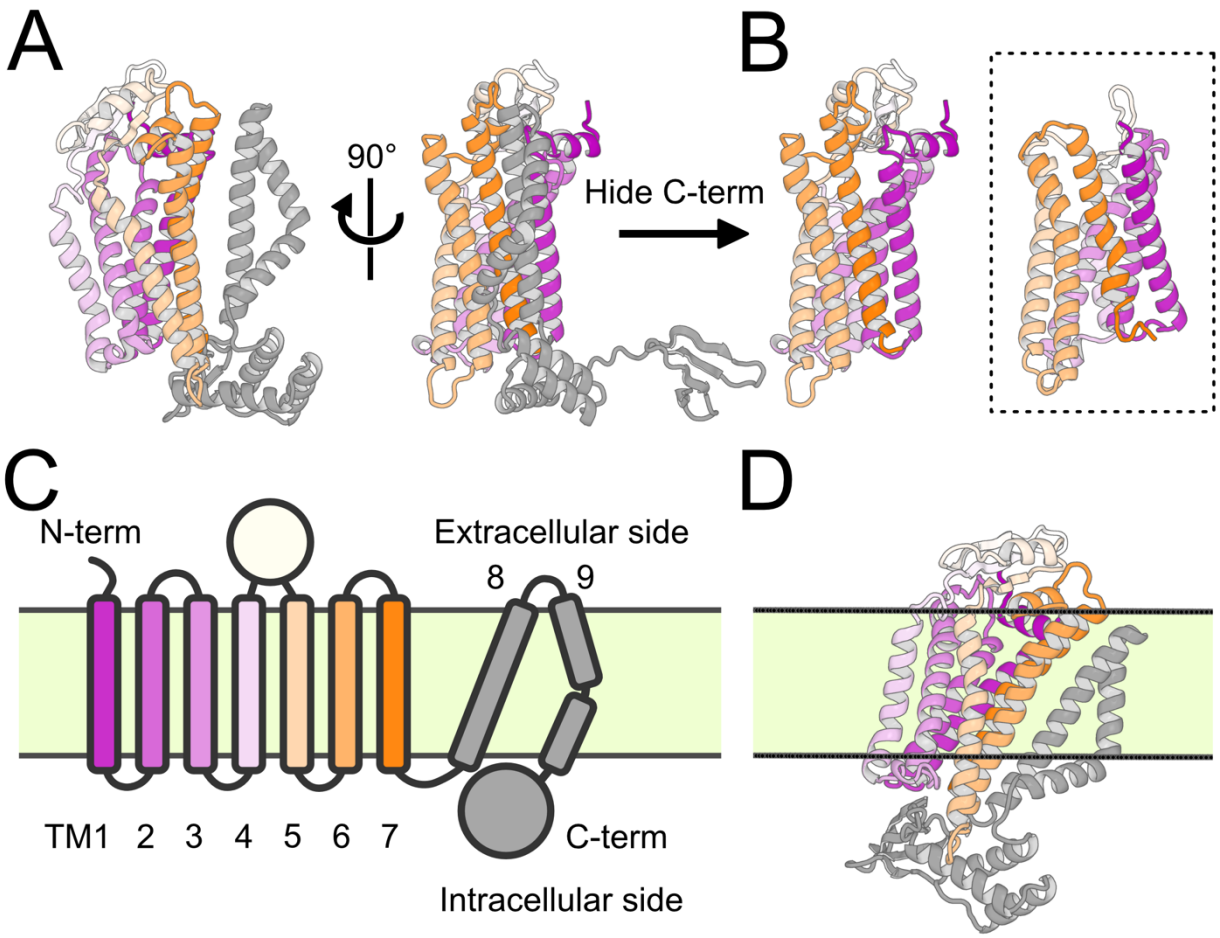

Figure 1: Overall structure of the representative NOT-GPCR

(A) Overall structure of the representative AlphaFold2-predicted NOT-GPCR (AFDB ID: A0A327H4A3). The GPCR-like domain (GLD) is colored in a purple-white-orange gradient from the N- to C-terminus. The auxiliary helices (AuxHs) and intracellular domains are colored gray. (B) GLD of the representative NOT-GPCR structure compared with a bona fide GPCR structure. The GPCR structure, mu-opioid receptor (PDB ID: 7T2H chain D), shown in a dashed box, was structure-aligned to GLD. Both are colored in a purple-white-orange gradient from the N- to C-terminus. (C) Schematic representation of NOT-GPCRs. The transmembrane helices are shown as rectangles, and extracellular/intracellular domains are shown as circles. The lipid bilayer region of the membrane is colored beige. (D) The predicted membrane topology and orientation of the

representative NOT-GPCR structure. The color scheme of the protein and membrane is the same as in (A) and (C).

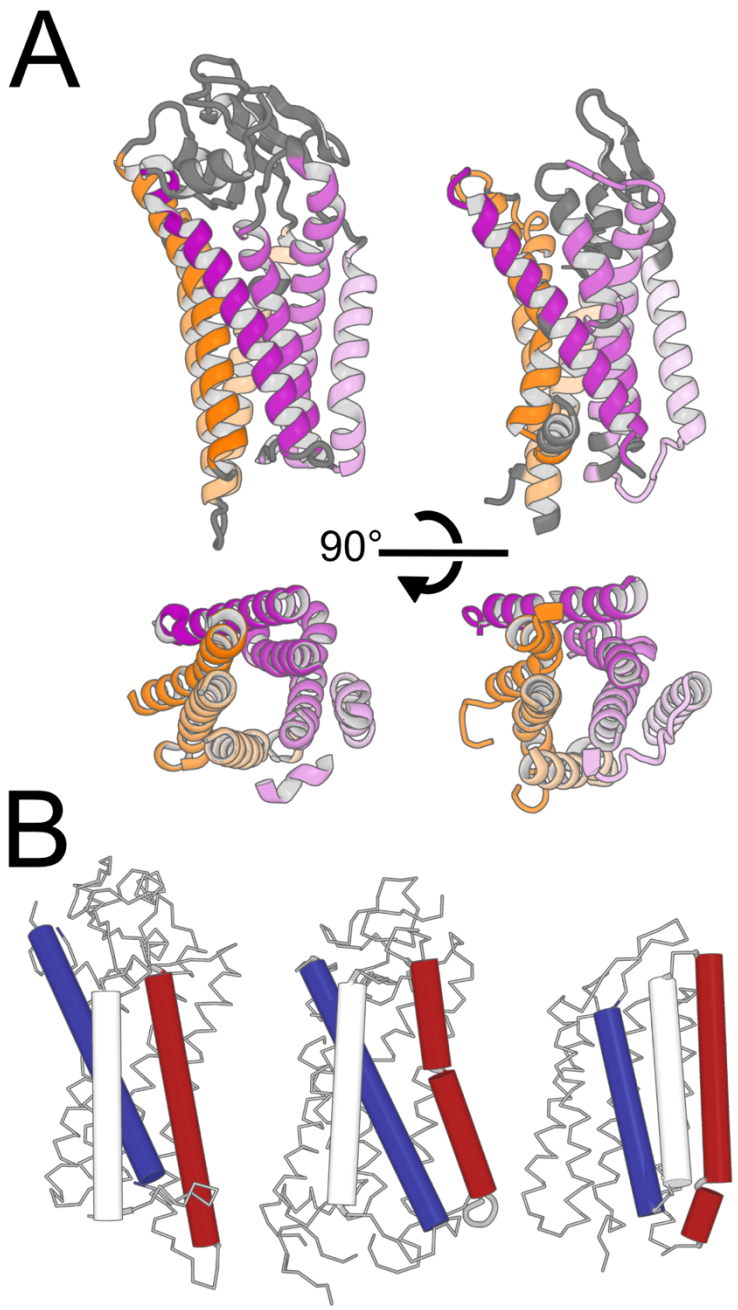

Figure 2: Similarity between NOT-GPCR and GPCR structures

(A) The representative NOT-GPCR structure aligned to a GPCR structure. The representative NOT-GPCR structure (AFDB ID: A0A327H4A3) was aligned to a bovine rhodopsin structure (PDB ID: 1F88), and the aligned regions are colored in a purple-white-orange gradient from the N- to C-terminus. The unaligned regions are colored gray. The bottom panel shows only the transmembrane (TM) helices to clarify the similarity of

their spatial arrangements and packing. (B) The key helix packing patterns to relate/distinguish NOT-GPCR, GPCR, and bR structures. The representative structures of NOT-GPCR (AFDB ID: A0A327H4A3), GPCR (PDB ID: 1F88), and bR (PDB ID: 1QHJ) were aligned and are shown as ribbons, and their third, fourth, and fifth TM helices are shown as cylinders and colored blue, white, and red, respectively. The rest of the structures are shown as thin ribbons.

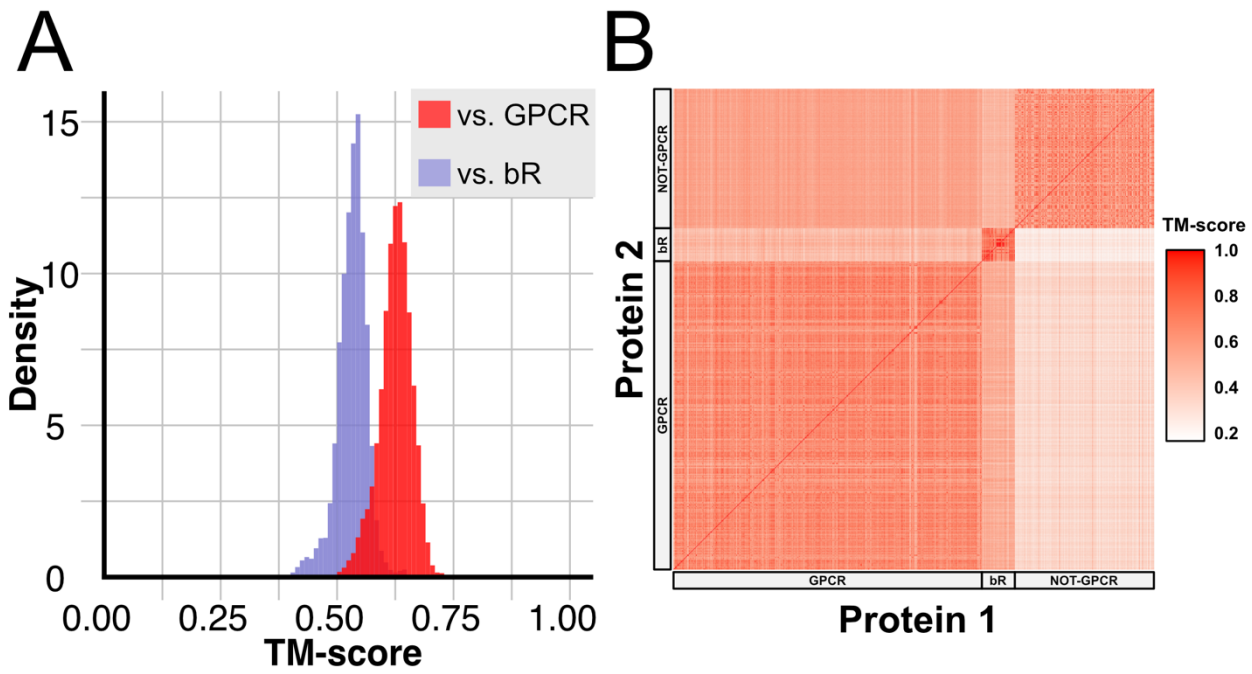

Figure 3: All-against-all structure comparisons among NOT-GPCR, GPCR, and bR structures

(A) The distribution of TM-score between NOT-GPCR and GPCR/bR structures. The horizontal axis represents the TM-score between NOT-GPCR and GPCR or bR structures. The vertical axis represents the density of the distributions. The red histogram shows the distribution of TM-scores between NOT-GPCR and GPCR structures when normalized by the amino acid sequence lengths of the respective GPCR structures. The blue histogram shows the distributions for TM-scores between NOT-GPCRs and bRs normalized by the amino acid sequence lengths of the respective bR structures. (B) A heatmap to show the all-against-all TM-score matrix among NOT-GPCR, GPCR, and bR structures. The horizontal axis (columns, Protein 1) corresponds to the reference structures to which the query structure (rows, Protein 2) was aligned. Note that TM-scores were normalized by the residue numbers of respective reference structures, resulting in an asymmetric matrix. The block in the top left corresponds to TM-scores between NOT-GPCRs and GPCRs, which was reduced into the red histogram in (A), and the block to the right of this corresponds to scores between NOT-GPCRs and bRs.

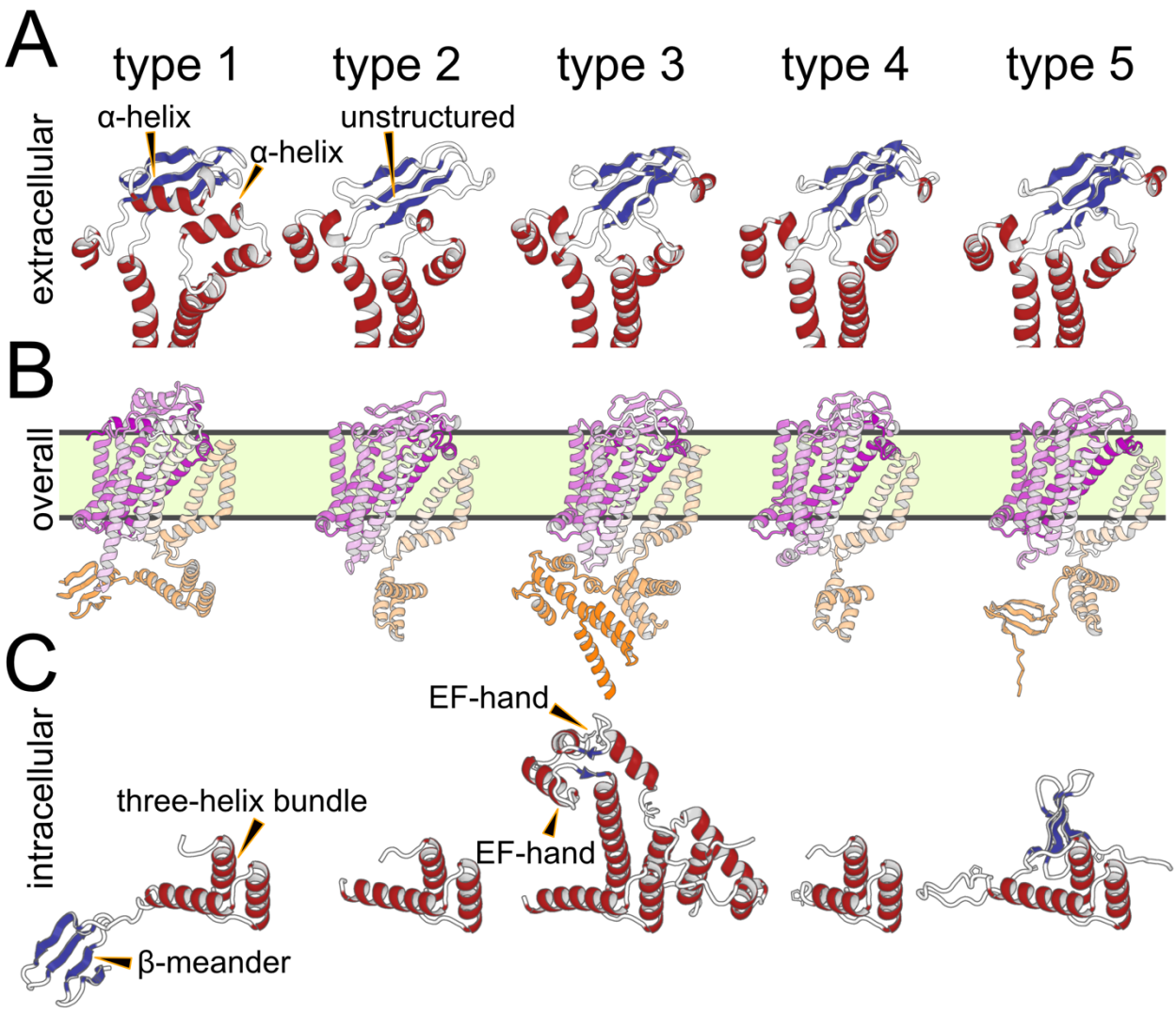

Figure 4: Five subtypes of NOT-GPCRs and their extracellular and intracellular domains

(A) The extracellular domains from all of the type-representative structures (AFDB ID: A0A2E6BZF1, A0A7J4T425, A0A2E4QIU6, A0A2E0BUJ4, and A0A7M3WXP5 for types 1–5, respectively). The continuous segments containing the extracellular domains are only shown as ribbons, where α-helices, β-strands, and loops are colored red, blue, and white, respectively. (B) The overall structure of the type-representatives. Each representative is colored in a purple-white-orange gradient from the N- to C-terminus, normalized by the type 3 structure having the largest number of residues. The beige region indicates the membrane. (C) The intracellular domains. The domains are aligned to the α-helical domain from the type 1 structure, and α-helices, β-strands, and loops are colored red, blue, and white, respectively.

Table 1: Phylogenetic distribution of NOT-GPCRs

Organism names are shown as they appeared during the phylogenetic analysis, regardless of whether the database naming convention has since been updated.

Organism	NCBI Taxon ID	Count
Euryarchaeota archaeon	2026739	198
Candidatus Poseidoniales archaeon	2163009	70
Candidatus Poseidoniaceae archaeon	2666346	26
Euryarchaeota archaeon TMED85	1986696	10
marine metagenome	408172	7
Candidatus Thalassarchaeaceae archaeon	2670411	7
Euryarchaeota archaeon TMED255	1986693	3
Neomarinimicrobiota bacterium	2026760	2
Marine Group II euryarchaeote MED-G33	2007294	2
Euryarchaeota archaeon TMED99	1986698	2
Euryarchaeota archaeon TMED141	1986686	2
Candidatus Woesearchaeota archaeon	2026803	2
Verrucomicrobiales bacterium	2026801	1
Thermoplasmata archaeon	1906666	1
Poseidonia sp.	2666344	1
Planctomycetaceae bacterium	2026779	1
Marine Group II euryarchaeote MED-G38	2007299	1
Legionellales bacterium	2026754	1
Euryarchaeota archaeon TMED97	1986697	1
Euryarchaeota archaeon TMED279	1986694	1
Euryarchaeota archaeon TMED103	1986682	1
Deltaproteobacteria bacterium	2026735	1
Acidimicrobiaceae bacterium	2024894	1

Table 2: TM-scores between/within NOT-GPCRs, GPCRs, and bR [presented as mean percentage (SD in parentheses)]

Note that the table is asymmetric because TM-scores differ depending on which structure is superposed onto the other, even for the same pair of structures. The columns correspond to the reference structure whose residue numbers are used for normalization of TM-scores.

Mean (SD)	NOT-GPCR	GPCR	bR
NOT-GPCR	0.703 (0.122)	0.625 (0.035)	0.532 (0.032)
GPCR	0.360 (0.047)	0.726 (0.083)	0.568 (0.046)
bR	0.272 (0.038)	0.504 (0.049)	0.814 (0.111)

Table 3: Structure-based sequence identities between/within NOT-GPCRs, GPCRs, and bR [presented as mean percentage (SD in parentheses)]

Mean (SD)	NOT-GPCR	GPCR	bR
NOT-GPCR	31.50 (15.77)	8.75 (1.78)	8.13 (1.98)
GPCR	-	19.04 (9.96)	9.18 (2.03)
bR	-	-	34.18 (26.31)

Table 4: Distribution of NOT-GPCR subtypes

Type	Count
1	111
2	55
3	62
4	7
5	7

Table 5: Structure-based sequence identities between/within NOT-GPCR subtypes [presented as mean percentage (SD in parentheses)]

Mean (SD)	type 1	type 2	type 3	type 4	type 5
type1	45.04 (12.85)	22.03 (2.71)	19.96 (2.67)	18.13 (2.66)	21.57 (1.86)
type2	-	55.83 (15.99)	28.42 (2.65)	25.03 (2.65)	26.36 (2.33)
type3	-	-	58.05 (16.01)	27.99 (3.12)	27.58 (2.96)
type4	-	-	-	80.26 (13.24)	30.17 (3.40)
type5	-	-	-	-	82.51 (13.72)

References

1. Rost B. Twilight zone of protein sequence alignments. Protein Engineering, Design and Selection. 1999 Feb;12(2):85–94.

2. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2022 Jan 7;50(D1):D439–44.

3. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Research. 2024 Jan 5;52(D1):D368–75.

4. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. 2023;

5. Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol [Internet]. 2023 May 8 [cited 2023 May 17]; Available from: https://www.nature.com/articles/s41587-023-01773-0

6. Gilman AG. G PROTEINS: TRANSDUCERS OF RECEPTOR-GENERATED SIGNALS. Vol. 56, Annual Review of Biochemistry. Annual Reviews; 1987. p. 615–49.

7. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, et al. High-Resolution Crystal Structure of an Engineered Human β_2 -Adrenergic G Protein–Coupled Receptor. Science. 2007 Nov 23;318(5854):1258–65.

8. Conflitti P, Lyman E, Sansom MSP, Hildebrand PW, Gutiérrez-de-Terán H, Carloni P, et al. Functional dynamics of G protein-coupled receptors reveal new routes for drug discovery. Nature Reviews Drug Discovery [Internet]. 2025 Jan 2; Available from: https://doi.org/10.1038/s41573-024-01083-3

9. De Mendoza A, Sebé-Pedrós A, Ruiz-Trillo I. The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity. Genome Biology and Evolution. 2014 Mar;6(3):606–19.

10. Kojima K, Sudo Y. Convergent evolution of animal and microbial rhodopsins. RSC Adv. 2023;13(8):5367–81.

11. Lake JA, Henderson E, Oakes M, Clark MW. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci USA. 1984 Jun;81(12):3786–90.

12. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences. 1990 Jun 1;87(12):4576–9.

13. Kim H, Mirdita M, Steinegger M. Foldcomp: a library and format for compressing and indexing large protein structure sets. Cowen L, editor. Bioinformatics. 2023 Apr 3;39(4):btad153.

14. Minami S, Sawada K, Ota M, Chikenji G. MICAN-SQ: a sequential protein structure alignment program that is applicable to monomers and all types of oligomers. Valencia A, editor. Bioinformatics. 2018 Oct 1;34(19):3324–31.

15. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010 Apr 1;26(7):889–95.

16. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, et al. InterPro: the protein sequence classification resource in 2025. Nucleic Acids Research. 2025 Jan 6;53(D1):D444–56.

17. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov;35(11):1026–8.

18. Minami S, Sawada K, Chikenji G. MICAN : a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, C α only models, Alternative alignments, and Non-sequential alignments. BMC Bioinformatics. 2013 Dec;14(1):24.

19. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, et al. Clustering predicted structures at the scale of the known protein universe. Nature [Internet]. 2023 Sep 13 [cited 2023 Oct 4]; Available from: https://www.nature.com/articles/s41586-023-06510-w

20. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015 Mar 15;31(6):926–32.

21. The UniProt Consortium, Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 2023 Jan 6;51(D1):D523–31.

22. Herrera LPT, Andreassen SN, Caroli J, Rodríguez-Espigares I, Kermani AA, Keserű GM, et al. GPCRdb in 2025: adding odorant receptors, data mapper, structure similarity search and models of physiological ligand complexes. Nucleic Acids Research. 2025 Jan 6;53(D1):D425–35.

23. Alexandrov N, Shindyalov I. PDP: protein domain parser. Bioinformatics. 2003 Feb 12;19(3):429–30.

24. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. Toward Consistent Assignment of Structural Domains in Proteins. Journal of Molecular Biology. 2004 Jun 4;339(3):647–78.

25. Bernhofer M, Rost B. TMbed: transmembrane proteins predicted through language model embeddings. BMC Bioinformatics. 2022 Aug 8;23(1):326.

26. Lomize AL, Todd SC, Pogozheva ID. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. Protein Science. 2022 Jan;31(1):209–20.

27. Schrodinger. The PyMOL Molecular Graphics System, Version 1.8. 2015.

28. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Bioinformatics. 1995 Dec 1;23(4):566–79.

29. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015 May;521(7551):173–9.

30. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017 Jan 19;541(7637):353–8.

31. Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. Phylogenomics provides robust support for a two-domains tree of life. Nature Ecology & Evolution. 2020 Jan 1;4(1):138–47.

32. Göker M, Oren A. Valid publication of names of two domains and seven kingdoms of prokaryotes. International Journal of Systematic and Evolutionary Microbiology [Internet]. 2024 Jan 22 [cited 2025 Jul 19];74(1). Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.006242

33. DeLong EF. Archaea in coastal marine environments. Proc Natl Acad Sci USA. 1992 Jun 15;89(12):5685–9.

34. DeLong EF, Wu KY, Prézelin BB, Jovine RVM. High abundance of Archaea in Antarctic marine picoplankton. Nature. 1994 Oct;371(6499):695–7.

35. Rinke C, Rubino F, Messer LF, Youssef N, Parks DH, Chuvochina M, et al. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (*Ca* . Poseidoniales ord. nov.). The ISME Journal. 2019 Mar 1;13(3):663–75.

36. Fuhrman JA, McCallum K, Davis AA. Novel major archaebacterial group from marine plankton. Nature. 1992 Mar;356(6365):148–9.

37. Santoro AE, Richter RA, Dupont CL. Planktonic Marine Archaea. Annu Rev Mar Sci. 2019 Jan 3;11(1):131–58.

38. Tully BJ, Sachdeva R, Graham ED, Heidelberg JF. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. PeerJ. 2017 Jul 10;5:e3558.

39. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018 Jan 16;5(1):170203.

40. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018 Nov;36(10):996–1004.

41. Haro-Moreno JM, López-Pérez M, De La Torre JR, Picazo A, Camacho A, Rodriguez-Valera F. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome. 2018 Dec;6(1):128.

42. Hiraoka S, Okazaki Y, Anda M, Toyoda A, Nakano S ichi, Iwasaki W. Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. Nat Commun. 2019 Jan 11;10(1):159.

43. Zhou Z, Tran PQ, Kieft K, Anantharaman K. Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation. The ISME Journal. 2020 Aug 1;14(8):2060–77.

44. Cabello-Yeves PJ, Callieri C, Picazo A, Mehrshad M, Haro-Moreno JM, Roda-Garcia JJ, et al. The microbiome of the Black Sea water column analyzed by shotgun and genome centric metagenomics. Environmental Microbiome. 2021 Mar 16;16(1):5.

45. Lin H, Ascher DB, Myung Y, Lamborg CH, Hallam SJ, Gionfriddo CM, et al. Mercury methylation by metabolically versatile and cosmopolitan marine bacteria. The ISME Journal. 2021 Jun 1;15(6):1810–25.

46. Rinke C, Chuvochina M, Mussig AJ, Chaumeil PA, Davín AA, Waite DW, et al. A standardized archaeal taxonomy for the Genome Taxonomy Database. Nat Microbiol. 2021 Jun 21;6(7):946–59.

47. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015 May 22;348(6237):1261359.

48. Ban C, Ramakrishnan B, Ling KY, Kung C, Sundaralingam M. Structure of the recombinant ıt Paramecium tetraurelia calmodulin at 1.68 Å resolution. Acta Crystallographica Section D. 1994 Jan;50(1):50–63.

49. Krishna SS, Grishin NV. Structurally Analogous Proteins Do Exist! Structure. 2004 Jul 1;12(7):1125–7.

50. Gaasterland T. Structural genomics taking shape. Trends in Genetics. 1998 Apr 1;14(4):135–135.

51. Gaasterland T. Structural genomics: Bioinformatics in the driver's seat. Nat Biotechnol. 1998 Jul;16(7):625–7.

52. Kim SH. Shining a light on structural genomics. Nature Structural Biology. 1998 Aug 1;5(8):643–5.

53. Williamson AR. Creating a structural genomics consortium. Nature Structural Biology. 2000 Nov 1;7(11):953–953.

54. Cyranoski D. "Big science" protein project under fire. Nature. 2006 Sep;443(7110):382–382.

55. Yokoyama S, Terwilliger TC, Kuramitsu S, Moras D, Sussman JL. RIKEN aids international structural genomics efforts. Nature. 2007 Jan 1;445(7123):21–21.