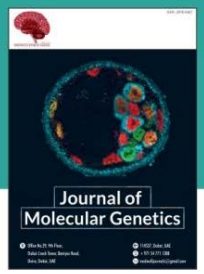
Journal of Molecular Genetics



A Consensus Sequence among Natural Transmembrane Linkers for Single-Polypeptide Chain (SC) Connection of Two G-Protein-Coupled Receptors in Tandem

[1, 2, 3, 5]Toshio Kamiya, [3]Takashi Masuko, [4]Dasiel Oscar Borroto-Escuela, [5]Haruo Okado and [1]Hiroyasu Nakata

[1]*Department of Molecular Cell Signaling, 2-6 Musashidai, Fuchu, Tokyo 183-8526, Japan*
[2]*Tokyo Metropolitan Institute for Neuroscience, Department of Neurology, 2-6 Musashidai, Fuchu, Tokyo 183-8526, Japan*
[3]*Cell Biology Laboratory, School of Pharmaceutical Sciences, Kinki University, 3-4-1 Kowakae, Higashi-Osaka, Osaka 577-8502, Japan*
[4]*Department of Neuroscience, Karolinska Institutet, Retzius väg8, 17177 Stockholm, Sweden*
[5]*Neural Development Project, Tokyo Metropolitan Institute of Medical Science, 2-1-6 Kamikitazawa, Setagaya-ku, Tokyo 156-8506, Japan*

Key words: Transmembrane single polypeptide chain linker, oligomerization, adenosine A_{2A} receptor, dopamine D_2 receptor; receptor allostery, fusion protein, striatum, supramolecular protein assembly, consensus sequence of natural scGPCR TM linker

Corresponding Author:
Toshio Kamiya
Division of Gene Regulation, Institute for Advanced Medical Research, Keio University School of Medicine, Keio University, 35 Shinano-Machi, Shinjuku-ku, Tokyo 160-8582, Japan

Page No.: 1-45
Volume: 13, Issue 1, 2021
ISSN: 2070-4267
Journal of Molecular Genetics

Abstract: A transmembrane (TM) single-polypeptide chain (sc) linker that can connect any two class A G-protein-coupled receptors (GPCRs) in tandem as a gene-fusion strategy has been demonstrated. Here, we report that a natural TM linker for sc exists that can connect two GPCRs in tandem. *In silico* analysis indicated that the central intervening portion (amino acid residues 311~416) of *Exaiptasia diaphana* (LOC110241027) GPCR fusion protein contains a single probable TM domain and that a consensus sequence, T(I/A/P) (A/S) (L/N) (I/W/L) (I/A/V) GL (L/G) (A/T) (S/L/G) (I/L) is conserved between this TM and previously characterized TM-linkers deriving from the type II TM proteins with a cytoplasmic N-ter segment, single TM and extracellular C-ter tail, *Caenorhabditis elegans odr4* and human CD23. In addition, to the invertebrate TM-linker, it was found that in the central intervening region of tropical bird *Opisthocomus hoazin* protein LOC104327099 (XP_009930279.1), TM-linker contains the consensus sequence. Here, we directed for the first time, our attention to the natural TM sc linker as a rare evolutionary event among all mammalian and avian proteins.

INTRODUCTION

Dopamine (DA)[1] fills a role that is important to striatal function[2, 3]. In the striatum, the contribution of the G-protein-coupled receptor (GPCR) for adenosine, subtype A_{2A} ($A_{2A}R$)[4, 5] to DA neurotransmission in the widespread [, i.e., non-(narrow/fast),] "volume transmission"/dual-transmission model[6,7] has been

Table 1: Natural GPCR fusions

Number	Organisms	Protein product	TM-linker	Resulting GPCR	
First analyzed:					
XP_020902518.1[1]	*Exaiptasia pallida*[1]	LOC110241027	consensus	dimer	Fig. 1b
XP_020602879.1	*Orbicella faveolata*	LOC110041892	no TM	dimer?	Fig. 1a
XP_022805553.1	*Stylophora pistillata*	LOC111342710	1st: no TM 2nd: N-ter signal peptide	dimer? (GPCR 2: TM2~TM6) monomer	Fig. 1c
Later analyzed					
TM-linker					
XP_009930279.1	*Opisthocomus hoazin*	LOC104327099	consensus	dimer	Fig. 3a
OBS78147.1	*Neotoma lepida*	A6R68_19462, partial	TM1 (amino acids 10~61) of Taar 8	dimer	Fig. 3b
No TM-linker					
KPP73650.1	*Scleropages formosus*	alpha-1B adrenergic receptor-like		dimer?	Fig. S6A
RMX53929.1	*Pocillopora damicornis*	pdam_00010707		pentamer?	Fig. S6B
XP_011663779.1[2]	*Strongylocentrotus purpuratus*	octopamine receptor beta-2R-like			
XP_020603616.1	*Orbicella faveolata*	octopamine receptor beta-1R-like		dimer?	Fig. S6C
ERE85073.1	*Cricetulus griseus*	trace amine-associated receptor 8a-like		dimer	Fig. S6D
XP_020629325.1	*Orbicella faveolata*	LOC110066440		dimer?	Fig. S6E
XP_001631773.1[2]	*Nematostella vectensis*	predicted			
KPP58082	*Scleropages formosus*	trace amine-associated receptor 1-like, partial		trimer? (GPCR 1: TM1, 2, 5~TM7; GPCR 2: TM1~TM6; GPCR 3: TM1~TM5)	Fig. S6F

[1]This record 'XP_020902518.1' was changed to 'XP_028515471.1' [All the amino acid sequences remain unchanged except that Arg of the *E. pallida* (that is at present, classified as *E. diaphana*) XP_020902518.1 is replaced by Phe of the XP_028515471.1 at amino acid residue 360. See text for details]. [2]These records 'XP_011663779.1' and 'XP_001631773.1' that are at present obsolete were removed as a result of standard genome annotation processing (NCBI)

studied extensively[8, 9]. $A_{2A}R$ and the GPCR for DA, subtype D_2[10], bind each other, leading to reciprocal and allosteric regulation of their functions. However, the dimers are not fully formed but depend on the equilibrium between monomer and dimer. Moreover, in order to stimulate the heteromerization, the prototypical single-polypeptide chain (sc) heterodimeric $A_{2A}R/D_2R$ complex[11], a fusion receptor and several other types were created, for example by joining the C-terminus of the $A_{2A}R$ via. the transmembrane (TM) of a type II TM protein to the N-terminus of the D_2R in tandem and tested experimentally[12].

Although, in addition to the fusion receptors, to further elucidate this *in vivo*, we designed a new molecular tool, namely, the self-assembling supermolecule-type $scA_{2A}R/D_2R$[12], based on a relationship between nanoscale surface curvature and surface-bound protein oligomerization[14-16], we have not yet constructed or tested it. However, here, the purpose of this study is that we relate a new computational finding of a natural invertebrate TM sc linker (, i.e., an additional TM helix between the annotated GPCR pairs) to previous experimental works on $scA_{2A}R/D_2R$. As a result, first, three class A GPCR fusion proteins homologous to the human $A_{2A}R$ were obtained using the BLAST search: *Orbicella faveolata* (LOC110041892), *Exaiptasia pallida* (that is at present, classified as *Exaiptasia diaphana*; LOC110241027) and *Stylophora pistillata* un characterized proteins (LOC111342710). *In silico* analysis indicated that among these, only the central intervening portion (amino acid residues 311~416) of the LOC110241027 protein contains a single probable TM domain. Unexpectedly, we found out a consensus sequence in these TM-linkers that can connect any two class A GPCRs in tandem, resulting in a 15-TM membrane protein, similar to early evolution events[17-19] and in vertebrate, both this type and another type TM sc linkers exist (Fig. 2, Table 1). Also, our results (Table 2 and 3) indicate that natural TM-linked GPCR occurred only as very rare evolutionary events when it is compared with in some degree many cases of no existence of TM-linkers between natural GPCR fusions. While transient class A GPCR nanoclusters have not been identified at the cell surface membrane, so far such a TM linker including a consensus sequence to continually connect the nonobligate dimer deserves discussion through the use of $scA_{2A}R/D_2R$ as a tool for the study of transient protein-protein interactions, focusing on sc GPCR oligomeric complexes that exist really at present in the natural world.

Table 2: Natural GPCR fusions from mammals

Number	Organisms	Protein product	TM-linker in intervening region	Resulting GPCR	
Primates[1]:					
ACA53481.1	*Plecturocebus moloch*	hypothetical	no TM	dimer	Fig. S7A
ABZ80295.1	*Callithrix jacchus*	hypothetical	no TM	dimer	Fig. S7B
Euarchontoglires[2]:					
OBS78147.1	*Neotoma lepida*	A6R68_19462, partial	TM1 (amino acids 10~61) of Taar 8	dimer	Fig.3b, T. 1[4]
OBS80343.1	*Neotoma lepida*	A6R68_21460, partial	0~2 TM? similar to itself	dimer	Fig. S8A, T. S1[4]
OBS75582.1	*Neotoma lepida*	A6R68_17966, partial	0 or 1 TM? similar to itself	dimer	Fig. S8B, T. S2
OBS82940.1	*Neotoma lepida*	A6R68_23065[2]	no TM	dimer	Fig. S8C, T. S3
OBS74663.1	*Neotoma lepida*	A6R68_14817	no TM	trimer [(GPCR 1)$_3$]	Fig. S8D, T. S4
OBS78080.1	*Neotoma lepida*	A6R68_19529	no TM?	tetramer	Fig. S8E, T. S5
VTJ79832.1	*Marmota monax*	hypothetical, partial	0 or 1 TM? similar to itself in each	pentamer	Fig. S8F, T. S6
VTJ82433.1	*Marmota monax*	hypothetical	no TM	trimer	Fig. S8G, T. S7
VTJ70100.1	*Marmota monax*	predicted	2 TM[3]	dimer[3]	Fig. S8H
Laurasiatheria:					
XP_010960385.1	*Camelus bactrianus*	LOC105074509	no TM	dimer	Fig. S8I
ACE79115.1	*Sorex araneus*	hypothetical	no TM	dimer	Fig. S8J
XP_006073590.2	*Bubalus bubalis*	LOC102415980	no TM	dimer	Fig. S8K
XP_004613594.2	*Sorex araneus*	LOC101542683	no TM	dimer	Fig. S8L
MXQ95485.1	*Bos mutus*	E5288_WYG003327	no TM	tetramer	Fig. S8M
ACE79123.1	*Sorex araneus*	hypothetical	no TM	trimer	Fig. S8N
MXQ95489.1	*Bos mutus*	E5288_WYG003309	no TM	tetramer	Fig. S8O
MXQ88753.1	*Bos mutus*	E5288_WYG003391	no TM	dimer	Fig. S8P
Mammalia (except for the above):					
XP_028909644.1	*Ornithorhynchus anatinus*	LOC100087811	no TM	trimer	Fig. S9

[1] Human genome has no GPCR fusion; [2]All natural GPCR fusions with 700 or more amino acids in length from Euarchontoglires are here shown, except for A6R68_23065 (OBS82940.1) (with 685 amino acids in length): 17 additional *Neotoma lepida* and *Marmota monax* GPCR fusions with 600~699 amino acids in length are shown in Table 11; [3][GPCR 1 (TM helices 1~7)]-(the central intervening region = TM helices 1 and 2 of GPCR 1)-{GPCR 2 [(TM helices 1~7) = (TM helices 3-4-6-2-3-4 of GPCR 1)]}; [4] T. 1: Table 1; T. S1~ T. S7: Table 4~10

Table 3: Natural GPCR fusions from Aves

Number	Organisms	Protein product	TM-linker in intervening region		Resulting GPCR
Galloanserae[1]:					
OXB57691.1	*Callipepla squamata*	ASZ78_008252	no TM	dimer	Fig. S10A
OXB61274.1	*Callipepla squamata*	ASZ78_010190	no TM	dimer	Fig. S10B
OXB75603.1	*Colinus virginianus*	H355_015719	no TM	dimer	Fig. S10C
POI30912.1	*Bambusicola thoracicus*	CIB84_005337	no TM	dimer	Fig. S10D
Aves (except for the above):					
XP_009930279.1	*Opisthocomus hoazin*	LOC104327099	consensus	dimer	Fig. 3a, T. 1[2]
XP_009819412.1	*Gavia stellata*	LOC104264164	TM 1 (amino acids 25~47) of olf 52B2-like	dimer	Fig. S10E
XP_008947395.1	*Merops nubicus*	LOC103781429	no TM	dimer	Fig. S10F
XP_019149630.2	*Corvus cornix cornix*	LOC104698100	no TM	dimer	Fig. S10G
KAF2980960.1	*Melospiza melodia maxima*	EK904_014712	no TM	dimer	Fig. S10H
OPJ89483.1	*Patagioenas fasciata monilis*	adhesion GPCR G4	no TM	dimer	Fig. S10I
PKU37004.1	*Limosa lapponica baueri*	llap _12689	no TM	dimer	Fig. S10J
PKU42245.1	*Limosa lapponica baueri*	histamine h3 receptor-like	no TM	dimer	Fig. S10K
RLW13346.1	*Erythrura gouldiae*	DV515_00000209	0 or 1 TM? similar to TM (amino acids 244-266) of RCC_04814 related to metalloreductase	dimer	Fig. S10L
RMC04025.1	*Hirundo rustica rustica*	DUI87_19362	no TM	dimer	Fig. S10M
TRZ15306.1	*Zosterops borbonicus*	HGM15179_011826	no TM	dimer	Fig. S10N

[1] In order to obtain avian GPCR fusion proteins, blastp search of the full-length of NP_000666 human $A_{2A}R$ protein was done, separately, against either Aves proteins, excluding Galloa nserae proteins or only Galloanserae proteins; 2 T. 1: Table 1

Table 4: *Neotoma lepida* (desert woodrat) hypothetical protein A6R68_21460, partial [with 775 amino acids in length; GenBank: OBS80343.1]

	Analysis tool		
Annotation	TMHMM	MEMSTAT3	Phobius
GPCR 1:			
TM helix 1	(203-225)	TM 1 (197-223)	TM 1 (197-222)
TM helix 2	(240-262)	TM 2 (234-256)	TM 2 (234-256)
TM helix 3	(275-294)	TM 3 (273-294)	TM 3 (276-294)

Table 4: Continue

Annotation	Analysis tool		
	TMHMM	MEMSTAT3	Phobius
TM helix 4	(314-336)	TM 4 (315-337)	TM 4 (315-337)
TM helix 5	(356-374)	TM 5 (369	TM 5 (370
TM helix 6	(378-400)	-396)	-399)
TM helix 7	(413-435)	TM 6 (416-431)	TM 6 (411-434)
		TM 7 (447-462) = TM 7 of GPCR 2	TM 7 (446-465)
Central intervening region:	(436-546)	no TM (463-504)	(466-503)
TM helix 1	(445-467) = TM 7 of GPCR 2		TM 1 (477-498)[1]
TM helix 2	(510-532) = TM 1 of GPCR 1		
GPCR 2:		TM 1 (505-531) = TM 1 of GPCR 1	TM 1 (504-529)
TM helix 1	(547-569) = TM 2 of GPCR 1	TM 2 (541-556) = TM 2 of GPCR 1	TM 2 (541-559)
TM helix 2	(582-604) = TM 3 of GPCR 1	TM 3 (577-602) = TM 3 of GPCR 1	TM 3 (579-602)
TM helix 3	(614-636) = TM 4 of GPCR 1	TM 4 (614-640) = TM 4 of GPCR 1	TM 4 (614-636)
TM helix 4	(657-679) = TM 5 of GPCR 1	TM 5 (666	TM 5 (669
TM helix 5	(684-701) = TM 6 of GPCR 1	-696) = TM 5 of GPCR 1	-698)
TM helix 6	(713-735) = TM 7 of GPCR 1	TM 6 (715-730) = TM 6 of GPCR 1	TM 6 (710-733)
TM helix 7	(745-767)	TM 7 (746-761)	TM 7 (745-767)

[1]Analysis by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/), an Artificial Neural Network program (MEMSAT3: http://bioinf.cs.ucl.ac.uk/?id = 756)[20] and Phobius indicated different annotation, i.e., TMHMM: the central intervening portion (amino acid residues 436~546) of this clone (OBS80343.1) does contain two transmembrane (TM) helices domain segments; MEMSTAT3: no TM; Phobius: a single TM [TM 1 (477-498)] of central intervening region shows homology to a sequence that precedes TM1 (197-222) of GPCR 1: 477-ILLLFPKWSHYNPGPFLLVGIP-498 * : : : * ** *** * *** 171-ISMAMYNTSSYNTGPFTLSGIP-192

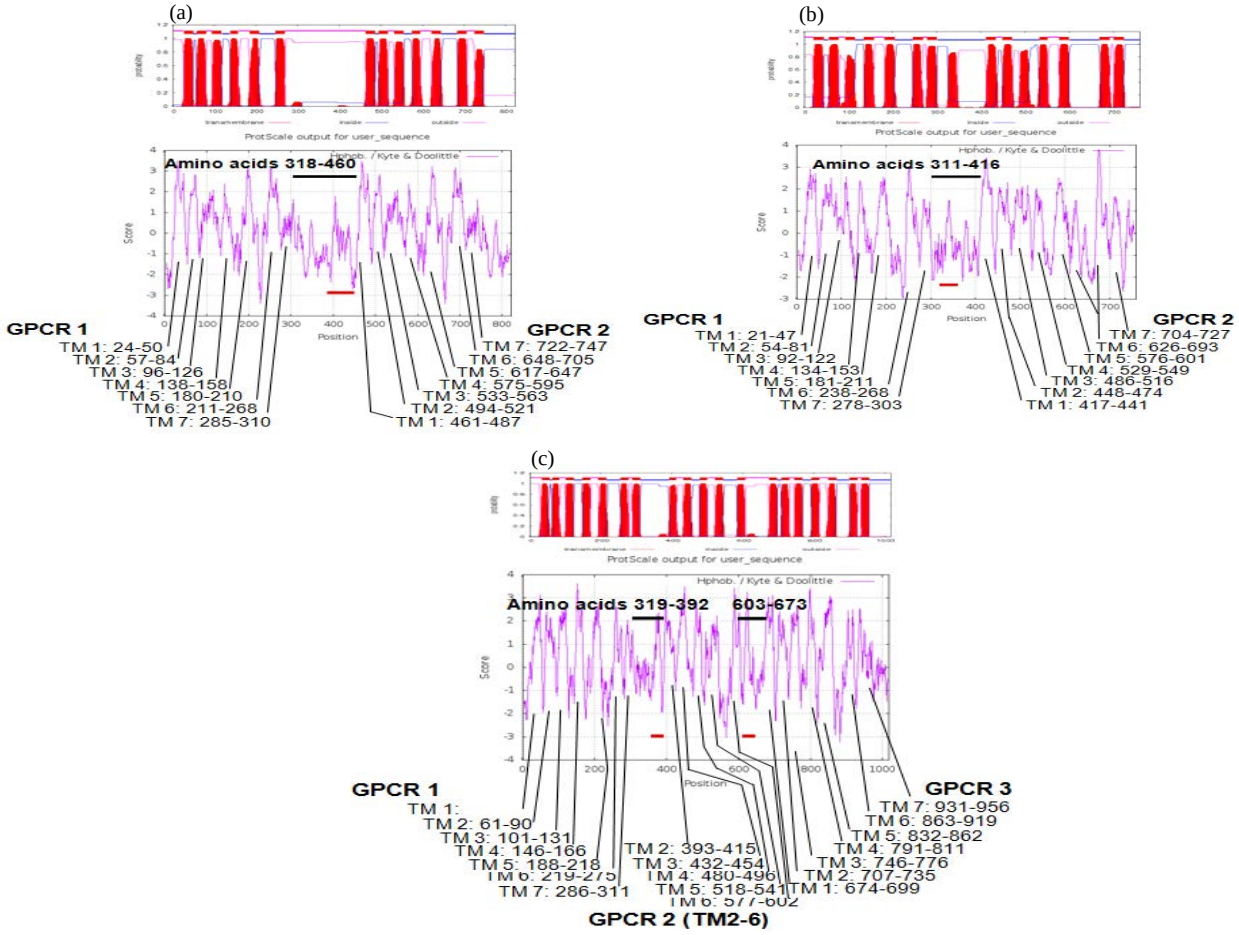


Fig. 1(a-c): (a) *Orbicella faveolata* protein LOC110041892 (XP_020602879.1), (b) *Exaiptasia pallida* protein LOC110241027 (XP_020902518.1) and (c) *Stylophora pistillata* protein LOC111342710 (XP_022805553.1)

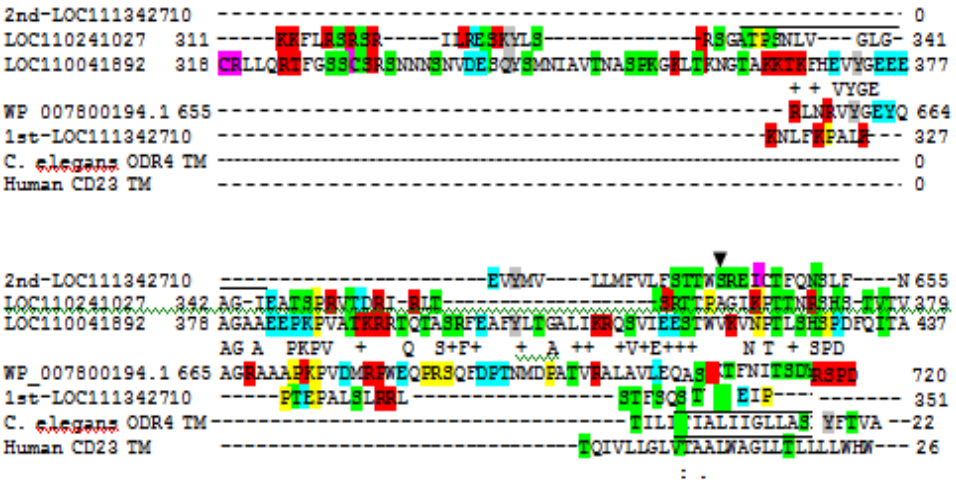


Fig. 2: Sequence of natural scGPCR TM linker evolution events

MATERIALS AND METHODS

Homology study of the proteins: The homologous sequences of the proteins were identified from successive iterations using the Basic Local Alignment Search Tool (BLAST) (BLASTP 2.8.0+), Position-Specific Iterated (PSI)-BLAST or Smart Blast. Because a lot of $A_{2A}R$ homologs among 20000 hits obtained in 2018-search (3. Results) were found to be replaced in 2019-search that is a BLAST search of the gene databases {blastp against all deposited organisms [nonredundant protein sequences (nr)]} with the full-length sequence of human $A_{2A}R$ (NCBI Reference Sequence: NP_000666.2) carried out on March 18, 2019, we made up our mind that we would better introduce more rational methods of identifying natural TM-linkers in GPCR fusions, rather than making a comparison between hits of 2018- and 2019-search and we have not completed the comparison [Supplemental Experimental procedures (S2)]. Multiple sequences were aligned using ClustalW (1.81) (http://align.genome.jp/) (This site was not found at present) (Fig. 3) or Clustal O (1.2.4) (http://www.ebi.ac.uk/) (Fig. 4-6). The secondary structure of the predicted protein was analyzed using the program PSIPRED (http://www.expasy.org/tools/)(http://bioinf.cs.ucl.ac.uk/psipred/) and using the hydropathicity scale (Kyte and Doolittle) (https://web.expasy.org/protscale/). A transmembrane (TM) domain for GPCR proteins was confirmed by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/) and/or an Artificial Neural Network program (MEMSAT3: http://bioinf.cs.ucl.ac.uk/?id = 756)[20]. In case of invertebrates [*Orbicella faveolata* (LOC110041892), *Exaiptasia diaphana* (LOC110241027) and *Stylophora pistillata* (LOC111342710) uncharacterized proteins] and vertebrates *Opisthocomus hoazin* (uncharacterized, LOC104327099) and *Neotoma lepida* (hypothetical A6R68_19462, partial) proteins and other cases of GPCR fusions without TM linker (Table 1, membrane topology was also reassessed by using Phobius[21, 22]. Analysis by Phobius indicated that *Caenorhabditis elegans odr-4* and human CD23 are single membrane-spanning type II TM proteins with a cytoplasmic N-ter but that human ODR-4 is the two-membrane-spanning protein with an extracellular N-ter and a conserved tail-anchored TM of the same configuration as that of *Caenorhabditis elegans odr-4*, although, TM1 (amino acid residues 82~102: RLLRMLPGGIHVVCIAWFSDK) of human ODR-4 is highly similar to an amino acid sequence [78-qVsRMLPGGIlVLGVfiiTtl-98 (Regions representing identity at that position or conserved amino acids to ODR4 TM1 are indicated in capital letters)[23] of *Caenorhabditis elegans odr-4*. The repeated pattern motifs were analyzed using the program Rapid Automatic Detection and Alignment of Repeats RADAR (http://www.ebi.ac.uk/Tools/Radar/index.html). ESyPred3D, an automated homology modeling program using neural networks (http://www.expasy.org/tools/) where the final three-dimensional structure is built using the modeling package MODELLER, was also used. The human $A_{2A}R$ structure PDB 3EML is shown by RasMol molecular graphics (Windows ver. 2.7.4.2) (http://www.rasmol.org/) (Fig. 3). An automated neural-network-based protein modeling CPH Models 3.2 server (http://www.cbs.dtu.dk/services/CPHmodels/) was also used. The coiled-coil propensity of proteins such as possible coiled-coil periodicities, i.e., the numbers of residues per turn, the sequence repeat length and the number of helical turns per repeat[24], was not

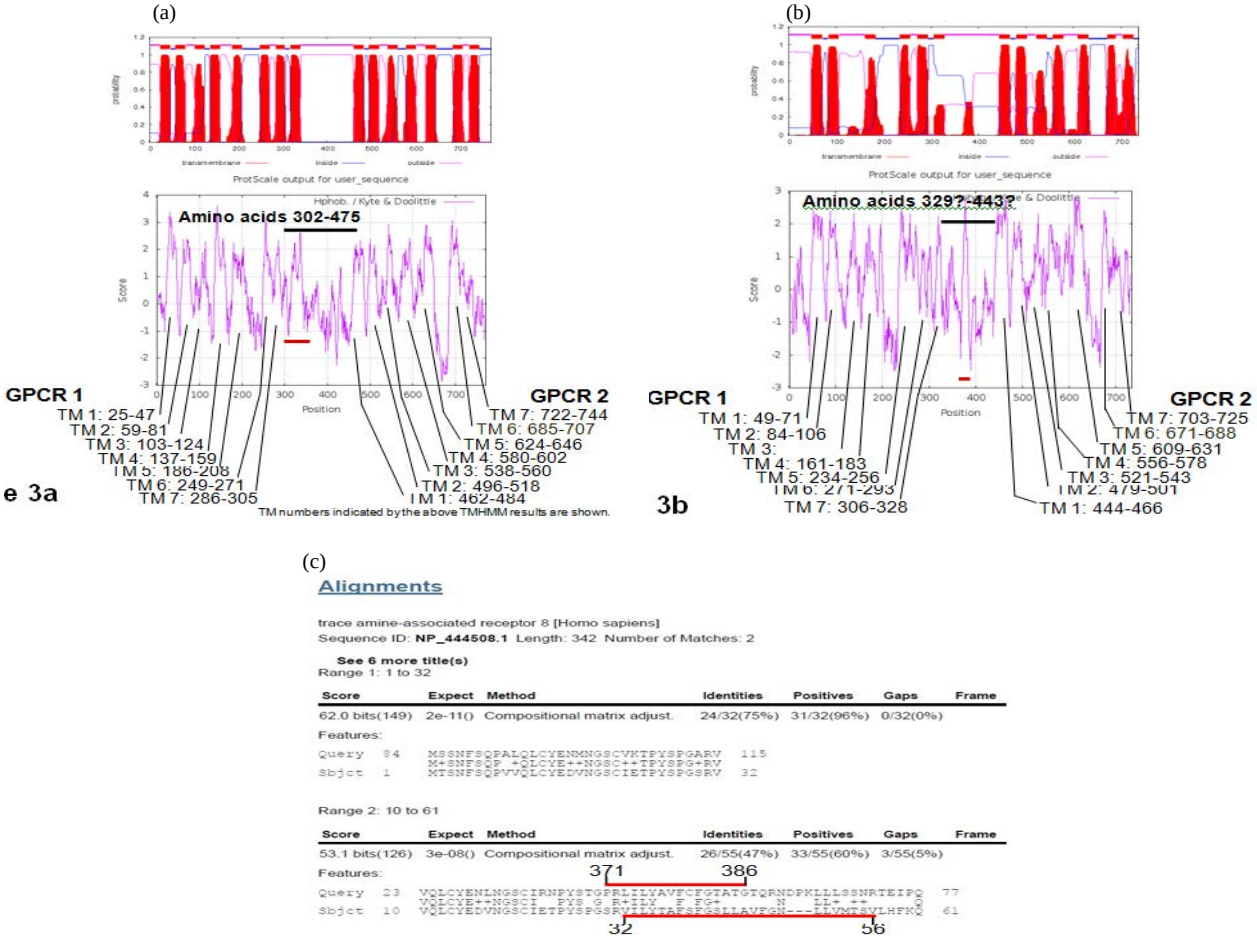


Fig. 3: A sequence that was found to exist in the natural vertebrate *Opisthocomus hoazin* protein LOC104327099 (NCBI Reference Sequence: XP_009930279.1) TM-linker (318~340)(VSI WGLAVLSVPS PSFLCILSLL)

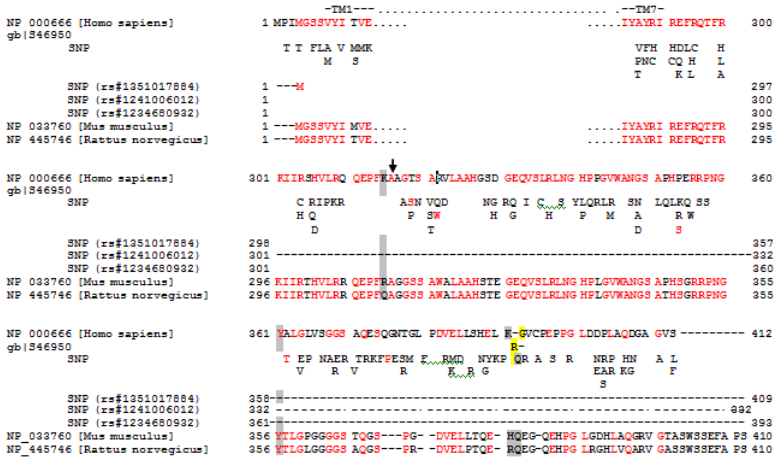


Fig. 4: Cytoplasmic

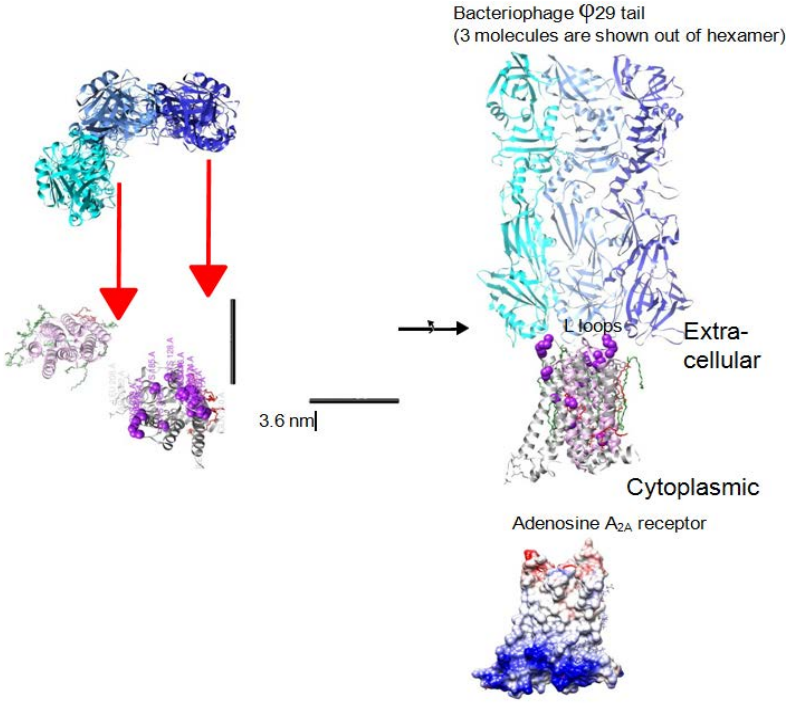


Fig. 5: Bacteriophage φ29 tail (3 molecules are shown out of hexamer)

Fig. 6: Supplementary (Fig. 3A/B)

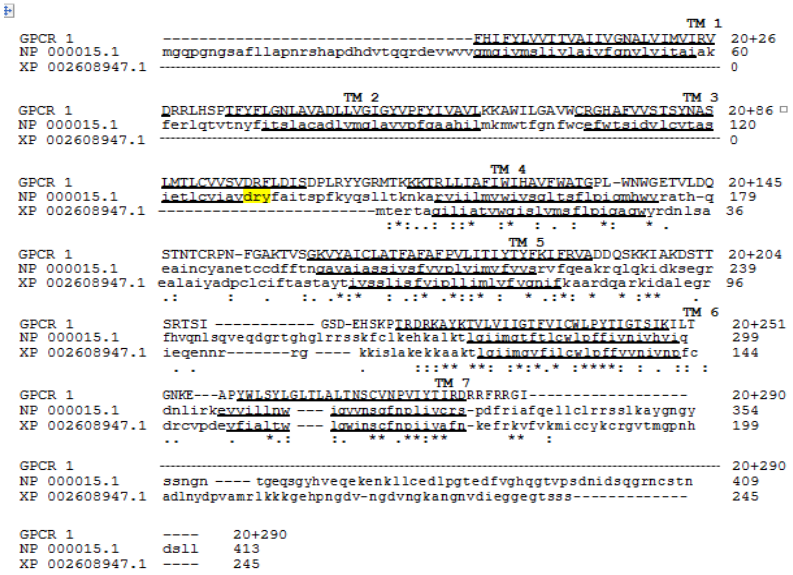


Fig. 7(a, b): Supplementary

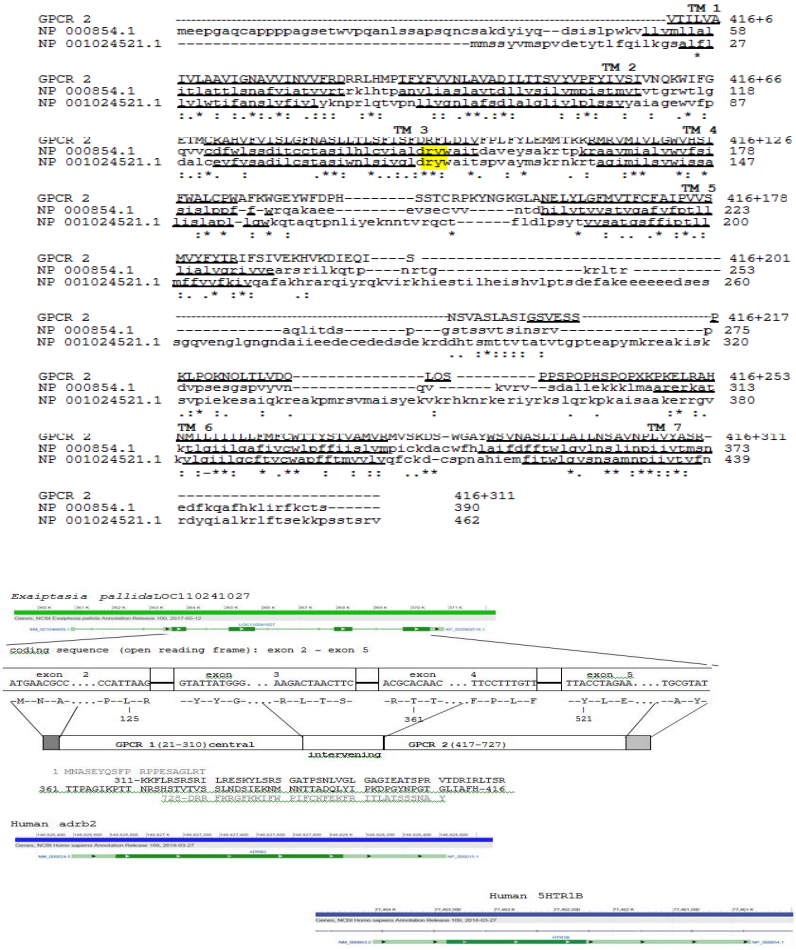


Fig. 8(a, b): Supplementary (Fig. 4A/B)

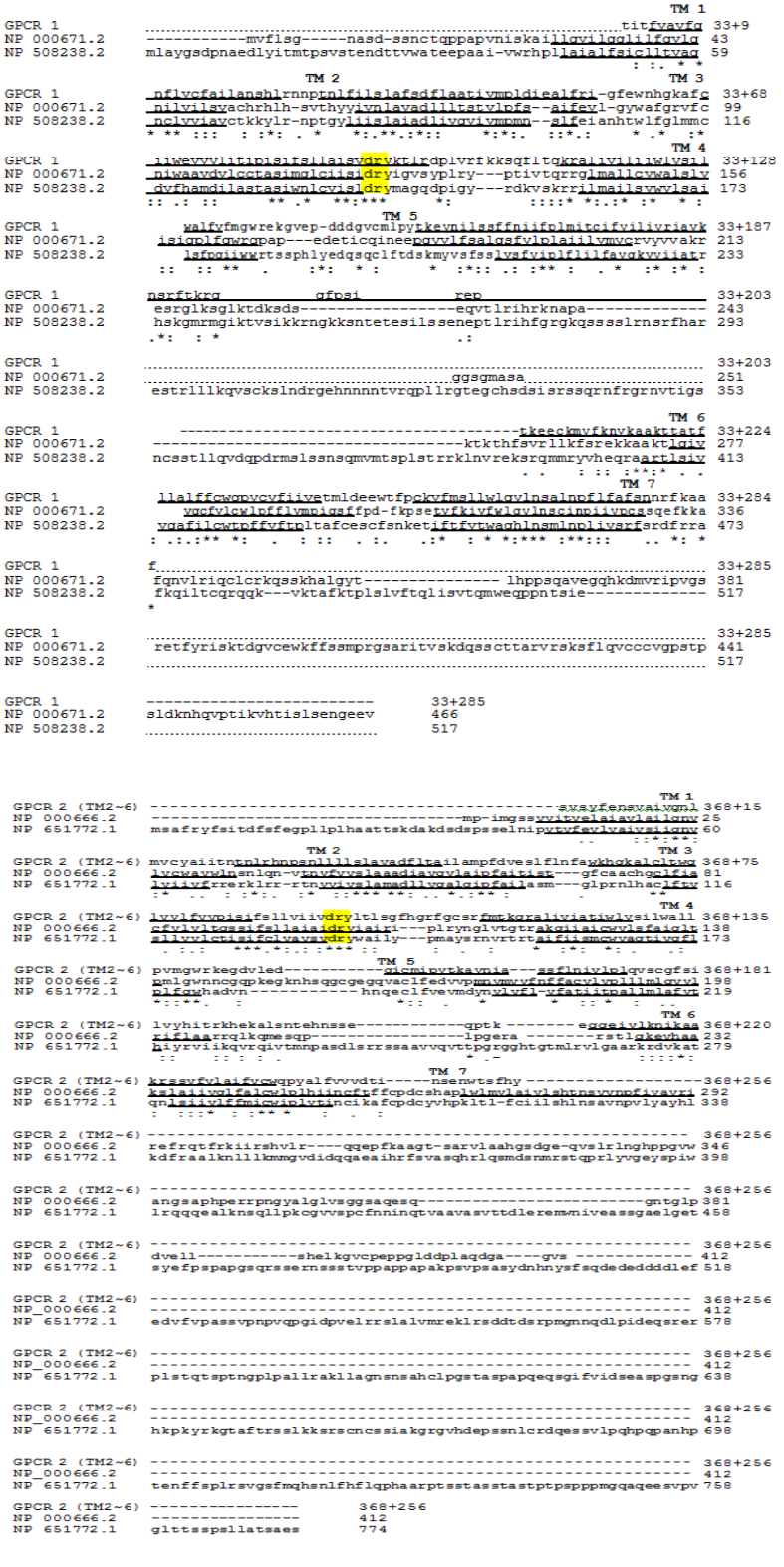


Fig. 9(a, b): Supplementary (Fig. 5A/B)

Fig. 10: Supplementary (Fig. S5A/B)

determined. Unless otherwise indicated, all *in silico* experiments were carried out according to procedures essentially from resources[25, 26].

RESULTS AND DISCUSSION

Identification of natural TM-linkers: A BLAST search of the gene databases {blastp against all deposited organisms [nonredundant protein sequences (nr)]} with the full-length sequence of human $A_{2A}R$ (NCBI Reference Sequence: NP_000666.2).

Figure SI as the query sequence to screen for the presence of sc GPCR oligomeric complexes revealed obvious homologs (as of April 10, 2018) [among 20000 hits of maximal target sequences, i.e., the maximum number of aligned sequences to display (though the actual number of alignments may be greater than this)], while three proteins of interest were selected (Table 1): 1, *Orbicella faveolata* uncharacterized protein LOC110041892 [with 824 amino acids in length; NCBI Reference Sequence: XP_020602879.1. It encodes two identical class A GPCRs in tandem (amino acid residues 24~317 and 461~754)] [A coding sequence (open reading frame) is on exon 5 and another is on exon 7] (Fig. 1a, Fig. S2B): 2, *Exaiptasia pallida* uncharacterized protein LOC110241027 [with 761 amino acids in length; NCBI Reference Sequence:

XP_020902518.1. It encodes two class A GPCRs in tandem (amino acid residues 21~310 and 417~727)] {This record 'XP_020902518.1' was later

Table 5: *Neotoma lepida* (desert woodrat) hypothetical protein A6R68_17966, partial [with 765 amino acids in length; GenBank: OBS75582.1]

	Analysis tool		
Annotation	TMHMM	MEMSTAT3	Phobius
GPCR 1:			
		TM 1 (5-27)	TM 1 (6-27)
TM helix 1	(42-64)	TM 2 (40-62)	TM 2 (39-61)
TM helix 2	(101-123)	TM 3 (95-122)	TM 3 (98-123)
TM helix 3	(135-157)	TM 4 (139-155)	TM 4 (135-157)
TM helix 4	(167-189)	TM 5 (170-186)	TM 5 (169-189)
	(187~216) = (690~719)		
		TM 6 (259-284) = TM 1 of GPCR 1	
TM helix 5	(304-326)	TM 7 (306-328) = TM 2 of GPCR 1	TM 6 (304-321)
TM helix 6	(358-380)		TM 7 (357-382)
TM helix 7	(not identified: probability around 0.7)		
central intervening region:	no TM (?-425)	(329-426)	(383-426)
TM helix 1		(360-389) = TM 3 of GPCR 1	
		= TM 5 of GPCR 2	TM 1 (403-421)[1]
GPCR 2:			
TM helix 1	(426-448)	TM 1 (427-451)	TM 1 (427-450)
TM helix 2	(463-485)	TM 2 (461-483) [(477~) = (240~)]	(not identified: probability around 0.9)
TM helix 3	(not identified: probability around 0.7)	TM 3 (496-521) = TM 1 of GPCR 1	(not identified: probability 0.6~0.8)
TM helix 4	(541-563)	TM 4 (544-565) = TM 2 of GPCR 1	TM 4 (541-564)
TM helix 5	(606-628)	TM 5 (597-625) = TM 3 of GPCR 1	TM 5 (601-626)
TM helix 6	(641-663)	TM 6 (642-659) [(~654) = (~417)]	TM 6 (638-660)
TM helix 7	(673-692)	TM 7 (674-689) = TM 5 of GPCR 1	TM 7 (672-692)

Table 6: *Neotoma lepida* (desert woodrat) hypothetical protein A6R68_23065 [with 685 amino acids in length; GenBank: OBS82940.1]

	Analysis tool		
Annotation	TMHMM	MEMSTAT3	Phobius
GPCR 1:			
TM helix 1	(24-46)	TM 1 (20-44)	TM 2 (39-61)
TM helix 2	(59-81)	TM 2 (59-86)	TM 3 (98-123)
TM helix 3	(96-118)	TM 3 (90-120)	TM 4 (135-157)
TM helix 4	(141-163)	TM 4 (139-158)	TM 5 (169-189)
TM helix 5	(186-208)	TM 5 (181-200)	TM 6 (304-321)
TM helix 6	(221-243)	TM 6 (226-246)	TM 7 (357-382)
TM helix 7	(Not identified: probability around 0.7)		
Central intervening region:	No TM (300?-357)	(-)	(383-426)
GPCR 2:			
TM helix 1	(358-380)	TM 1 (359-381)	TM 1 (427-450)
TM helix 2	(393-415)	TM 2 (392-415)	(not identified: probability around 0.9)
TM helix 3	(430-452)	TM 3 (426-456)	(not identified: probability 0.6~0.8)
TM helix 4	(472-494)	TM 4 (474-493) = TM 2 of GPCR 1	TM 4 (541-564)
TM helix 5	(518-540)	TM 5 (520-541) = TM 3 of GPCR 1	TM 5 (601-626)
TM helix 6	(561-579)	TM 6 (562-582)	TM 6 (638-660)
TM helix 7	(609-631)	TM 7 (not identified: probability)	TM 7 (672-692)

[1]Analysis by the TM-HMM 2.0 algorithm an Artificial Neural Network program (MEMSAT3) and Phobius indicated different annotation, i.e., TMHMM: the central intervening portion (amino acid residues ?~425) of this clone does contain no transmembrane (TM) helices domain segment; MEMSTAT3: a single TM [TM 1 (360~389)] of central intervening region (329~426) shows homology to sequences {93~122 [GPCR 1-TM helix 3 (MEMSTAT3: 95-122)]/[GPCR 1-TM helix 2 (TMHMM: 101-123)] and 596~625 [GPCR 2-TM helix 5 (TMHMM: 606-628)(MEMSTAT3: 597-625)]} of this clone; Phobius: a single TM [TM 1 (403~421)] of central intervening region (383~426) shows no homology to any other sequences of this clone. Highly homologous sequences are as follows: 240 ~ 417 ≈ 477 ~ 654; 260 ~ 423 ≈ 3~168; 187~216 ≈ 690 ~719

removed as a result of standard genome annotation processing (NCBI). At present, NCBI Reference Sequence: XP_028515471.1; Taxonomically within the genus *Aiptasia*, *Exaiptasia pallida* is at present classified as *Exaiptasia diaphana*[27, 28] [NCBI: Taxid 1720309 (*Exaiptasia pallida*) was merged into taxid 2652724 (*Exaiptasia diaphana*) on July 11, 2020]}[A coding sequence (open reading frame) is on exons 2 and 3 and another is on exons 4 and 5] (Fig. 1b: 7a) and 3, *Stylophora pistillata* uncharacterized protein LOC111342710 [with 1018 amino acids in length; NCBI Reference Sequence: XP_022805553.1. It encodes three class A GPCRs in tandem, of which the second GPCR is annotated as a deletion-type only with transmembrane (TM) helix domain 2 (TM2)~TM6 (amino acid residues 34~318, 369~626 and 674~963)] (Fig. 8a, b). Among some higher scores of known full-length class A GPCR proteins such the above three proteins were obtained, of

Table 7: *Neotoma lepida* (desert woodrat) hypothetical protein A6R68_21460 [with 1135 amino acids in length; GenBank: OBS74663.1]

	Analysis tool		
Annotation	TMHMM	MEMSTAT3	Phobius
GPCR 1:			
TM helix 1	(42-64)	TM 1 (42-66)	TM 1 (42-63)
TM helix 2	(79-101)	TM 2 (76-99)	TM 2 (75-101)
TM helix 3	(114-136)	TM 3 (111-137)	TM 3 (113-136)
TM helix 4	(156-178)	TM 4 (157-182)	TM 4 (156-179)
TM helix 5	(Not identified: probability around 0.2)	TM 5 (213-239)	(Not identified: probability around 0.2)
TM helix 6	(259-281)	TM 6 (256-274)	TM 6 (253-271)
TM helix 7	(291-310)	TM 7 (290-306)	TM 7 (291-310)
Central intervening region, 1st:	No TM (311-483)	(307-483) TM 1 (388-403)	no TM (311-483)
GPCR 2:			
TM helix 1	(484-506)	TM 1 (484-508)	TM 1 (484-505)
TM helix 2	(521-543)	TM 2 (518-541)	TM 2 (517-543)
TM helix 3	(556-578)	TM 3 (553-579)	TM 3 (555-578)
TM helix 4	(598-620)	TM 4 (599-623)	TM 4 (598-621)
TM helix 5	(Not identified: probability around 0.1)	TM 5 (656-683)	(not identified: probability around 0.4)
TM helix 6	(Not identified: probability around 0.9)	TM 6 (698-716)	TM 6 (702-723)
TM helix 7	(730-752)	TM 7 (732-750)	TM 7 (735-752)
Central intervening region, 2nd:	(753-803) (Not identified: probability around 0.7)	No TM (751-803)	(753-805) (not identified: probability around 0.1
GPCR 3:			
TM helix 1	(804-826)	TM 1 (804-830)	TM 1 (806-827)
TM helix 2	(839-858)	TM 2 (840-865)	TM 2 (839-857)
TM helix 3	(878-900)	TM 3 (875-900)	TM 3 (878-900)
TM helix 4	(920-942)	TM 4 (921-944)	TM 4 (920-941)
TM helix 5	(980-1002)	TM 5 (977-1006)	TM 6 (980-1003)[1]
TM helix 6	(1023-1045)	TM 6 (1022-1040)	TM 7 (1024-1042)
TM helix 7	(1055-1072)		TM 8 (1054-1072)

[1]Annotation resulted from BLAST search of this clone. Unlike analysis by the TM-HMM 2.0 algorithm and the MEMSAT3 program, Phobius analysis indicated GPCR 3-TM 6 (probability ≈ 0.9); -TM 5 (953-974: probability ≈ 0.75)[, similar to which an amino acid sequence of Mus musculus olf543 (AAI47343.1) is not TM]

Table 8: *Neotoma lepida* (desert woodrat) hypothetical protein A6R68_19529 [with 1229 amino acids in length; GenBank: OBS78080.1]

	Analysis tool		
Annotation	TMHMM	MEMSTAT3	Phobius
GPCR 1:			
TM helix 1	(69-91)	TM 1 (69-92)	TM 1 (68-90)
TM helix 2	(not identified)	TM 2 (not identified)	TM 2 (102-123)
TM helix 3	(140-162)	TM 3 (137-161)	TM 3 (143-162)
TM helix 4	(182-204)	TM 4 (184-206)	TM 7 (182-200)
Central intervening region, 1st:	no TM (205-304)	(207-305) TM 1 (236-251)	no TM (201-304)
GPCR 2:			
TM helix 1	(305-327)	TM 1 (306-329)	TM 1 (305-328)
TM helix 2	(340-362)	TM 2 (not identified)	TM 2 (340-361)
TM helix 3	(377-399)	TM 3 (374-399)	TM 3 (381-399)
TM helix 4	(419-441)	TM 4 (420-443)	TM 4 (419-441)
TM helix 5	(not identified)	TM 5 (not identified)	(not identified)
TM helix 6	(504-526)	TM 6 (500-526)	TM 6 (507-526)
TM helix 7	(546-568)	TM 7 (545-570)	TM 7 (546-564)
Central intervening region, 2nd:	No TM (569-610)	No TM (571-603)	(565-?) (Not identified: probability around 0.6)
GPCR 3:			
TM helix 1	(611-633)	TM 1 (604-628)	TM 1 (not identified)
TM helix 2	(663-685)	TM 2 (663-686)	TM 2 (662-684) TM 3 (696-717: probability around 0.4)
TM helix 3	(734-756)	TM 3 (731-755)	TM 4 (737-756)
TM helix 4	(776-798)	TM 4 (777-800)	TM 5 (776-794)
TM helix 5	(835-857)	TM 5 (833-861)	TM 6 (833-857)
TM helix 6	(869-891)	TM 6 (874-889)	TM 7 (873-891)
Central intervening region, 3rd:	No TM (892-919)	No TM (890-920)	no TM (892-919)

Table 8: Continue

	Analysis tool		
Annotation	TMHMM[1]	MEMSTAT3	Phobius
GPCR 4:			
TM helix 1	(920-942)	TM 1 (921-944)	TM 1 (920-943)
TM helix 2	(955-977)	TM 2 (954-975)	TM 2 (955-976)
TM helix 3	(992-1014)	TM 3 (989-1014)	TM 3 (996-1014)
TM helix 4	(1034-1056)	TM 4 (1035-1060)	TM 4 (1034-1056)
TM helix 5	(1093-1115)	TM 5 (1092-1117)	TM 5 (1093-1120)
TM helix 6	(1136-1155)	TM 6 (1136-1151)	TM 6 (1132-1154)
TM helix 7	(1165-1187)	TM 7 (1167-1182)	TM 7 (1166-1185)

[1]Annotation resulted from BLAST search of this clone. Cobalt multiple alignment of GPCRs 1~4 (Fig. S9) indicated [GPCR 1 (TM1~TM4)] ≈ [GPCR 2 (TM1~TM4)] ≈ [GPCR 4 (TM1~TM4)] ≈ [GPCR 3 (TM2~TM4)] where a sequence of GPCR 3 (697-FFLANLSSVDISAPSVTVPKALV-719, double-underlined) similar to TM2 of GPCRs 1/2/4 was identified as a TM and where GPCR 1-TM2 (double-underlined) was identified (in Phobius analysis unlike analyses by the TM-HMM 2.0 algorithm and the MEMSAT3 program)

Table 9: *Marmota monax* (woodchuck) hypothetical predicted protein, partial [with 1503 amino acids in length; GenBank: VTJ79832.1]

	Analysis tool		
Annotation	TMHMM[1]	MEMSTAT3	Phobius
GPCR 1:			
TM helix 1	(28-50)	TM 1 (29-52)	TM 1 (27-52)
TM helix 2	(63-85)	TM 2 (65-84)	TM 2 (64-84)
TM helix 3	(100-122)	TM 3 (97-120)	TM 3 (104-122)
TM helix 4	(142-164)	TM 4 (145-166)	TM 4 (142-164)
TM helix 5	(202-224)	TM 5 (198-225)	TM 5 (199-224)
TM helix 6	(245-262)	TM 6 (245-261)	TM 6 (245-263)
TM helix 7	(Not identified)	TM 7 (275-290)	TM 7 (275-294: probability around 0.8)
Central intervening region, 1st:	No TM (?-338)	No TM (291-338)	(295-336) TM 1 (315-331: probability around 0.8)
GPCR 2:			
TM helix 1	(339-361)	TM 1 (339-363)	TM 1 (337-364)
TM helix 2	(374-396)	TM 2 (376-395)	TM 2 (376-399)
TM helix 3	(411-433)	TM 3 (408-432)	TM 3 (411-433)
TM helix 4	(453-475)	TM 4 (456-476)	TM 4 (453-470)
TM helix 5	(510-532)	TM 5 (509-536)	TM 5 (511-534)
TM helix 6	(551-573)	TM 6 (556-571)	TM 6 (554-573)
TM helix 7	(Not identified)	TM 7 (586-601)	(not identified: probability around 0.5)
Central intervening region, 2nd:	No TM (?-641)	(602-638) TM 1 (612-627)	(?-644) (not identified: probability around 0.5)
GPCR 3:			
TM helix 3	(642-664)	TM 3 (639-663)	TM 3 (645-664)
TM helix 4	(684-706)	TM 4 (687-707)	TM 4 (684-708)
TM helix 5	(744-766)	TM 5 (740-767)	TM 5 (744-765)
TM helix 6	(787-804)	TM 6 (787-803)	TM 6 (786-805)
TM helix 7	(Not identified)	TM 7 (817-832)	TM 7 (817-836: probability around 0.9)
Central intervening region, 3rd:	No TM (?-884)	(833-881) TM 1 (855-870)	(837-887) TM 1 (848-868: probability around 0.9)
GPCR 4:			
TM helix 3	(885-907)	TM 3 (882-906)	TM 3 (888-907)
TM helix 4	(927-949)	TM 4 (930-950)	TM 4 (927-948)
TM helix 5	(987-1009)	TM 5 (983-1010)	TM 5 (984-1008)
TM helix 6	(1030-1047)	TM 6 (1030-1045)	TM 6 (1028-1047)
Cntral intervening region, 4th:	No TM (1048-1214)	(1046-1214)	No TM (1048-1213) TM 1 (1062-1077)
GPCR 5:			
TM helix 1	(1215-1237)	TM 1 (1215-1239)	TM 1 (1214-1239)
TM helix 2	(1250-1272)	TM 2 (1252-1271)	TM 2 (1251-1270)
TM helix 3	(1287-1309)	TM 3 (1284-1307)	TM 3 (1290-1309)
TM helix 4	(1329-1351)	TM 4 (1332-1352)	TM 4 (1329-1346)
TM helix 5	(1389-1411)	TM 5 (1385-1412)	TM 5 (1387-1410)
TM helix 6	(1432-1449)	TM 6 (1432-1447)	TM 6 (1430-1449)
TM helix 7	(1464-1481)	TM 7 (1462-1477)	TM 7 (1461-1481: probability around 0.8)

[1]Annotation resulted from BLAST search of this clone. Cobalt multiple alignment of GPCRs 1~5 (Fig. S9) indicated [GPCR 1 (TM3~TM6)] ≈ [GPCR 2 (TM3~TM6)] ≈ [GPCR 3 (TM3~TM6)] ≈ [GPCR 4 (TM3~TM6)] ≈ [GPCR 5 (TM3~TM6)] and [GPCR 1 (TM1/TM2)] ≈ [GPCR 2 (TM1/TM2)] ≈ [GPCR 5 (TM1/TM2)] where sequences of GPCRs 1/2/3 [respectively, 275/586/817-GVA(V/I/I)LNTSVAP(LL/ML/MM)NPF-290/601/832, double-underlined] similar to TM7 of GPCR 5 were identified as single TMs in the MEMSAT3 program analysis unlike analysis by the TMHMM 2.0 algorithm

Table 10: *Marmota monax* (woodchuck) hypothetical predicted protein [with 850 amino acids in length; GenBank: VTJ82433.1]

	Analysis tool		
Annotation	TMHMM[1]	MEMSTAT3	Phobius
GPCR 1:			
TM helix 1	(27-49)	TM 1 (27-49)	TM 1 (26-47)
TM helix 2	(59-78: probability around 0.7)	TM 2 (not identified)	TM 2 (59-78: probability around 0.9)
TM helix 3	(98-120)	TM 3 (96-119)	TM 3 (98-120)
TM helix 4	(130-152)	TM 7 (140-165)	TM 7 (132-151)
Central intervening region, 1st:	No TM (153-240)	No TM (166-239)	(152-239) TM 1 (157-178: probability around 0.5)
GPCR 2:			
TM helix 1	(241-263)	TM 1 (240-264)	TM 1 (240-263)
TM helix 2	(278-300)	TM 2 (not identified)	TM 2 (275-297: probability around 0.9)
TM helix 4	(320-342)	TM 4 (328-353)	TM 4 (318-342)
TM helix 5	(385-407)	TM 5 (383-410)	TM 5 (385-409)
TM helix 6	(428-450)	TM 6 (430-447)	TM 6 (430-448)
TM helix 7	(460-482)	TM 7 (not identified)	TM 7 (460-480: probability around 0.9)
Central intervening region, 2nd:	No TM (483-560)	No TM (?-560)	(481-559) (Not identified: probability around 0.2)
GPCR 3:			
TM helix 1	(561-583)	TM 1 (561-583)	TM 1 (560-583)
TM helix 2	(593-612: probability around 0.3)	TM 2 (not identified)	(not identified: probability around 0.85)
TM helix 3	(632-654)	TM 3 (629-653)	TM 3 (635-654)
TM helix 4	(664-686)	TM 4 (674-699)	TM 4 (666-688) (not identified: probability around 0.2)
TM helix 5	(730-752)	TM 5 (729-754)	TM 5 (731-750)[1]
TM helix 6	(772-794)	TM 6 (776-792)	TM 6 (771-794)
TM helix 7	(not identified)	TM 7 (not identified)	TM 7 (806-826)

[1]Annotation resulted from BLAST search of this clone. [GPCR 1 (TM1/TM2)] ≈ [GPCR 2 (TM1/TM2)] ≈ [GPCR 3 (TM1/TM2)]; [GPCR 1 (TM3)] ≈ [GPCR 3 (TM3)]; [GPCR 1 (TM4)] ≈ [GPCR 2 (TM4)] ≈ [GPCR 3 (TM4)]; [GPCR 2 (TM5~TM7)] ≈ [GPCR 3 (TM5~TM7)]. Unlike analysis by the TM-HMM 2.0 algorithm and the MEMSAT3 program, Phobius analysis indicated a TM-like part between GPCR 3-TM 4 and TM 5 [, similar to which an amino acid sequence of GPCR 2 is not TM]

sequences producing significant alignments with expectation values (E-values) of less than 10^{-16} (better than the threshold). To more clarify the above clones, we then carried out blastp search using *a*, the amino-terminal (N-ter) portion [, i.e., the GPCR 1/2; the GPCR 1; the GPCR 1 and the central GPCR 2 (*Δ*DTM1/*Δ*DTM7) without the central intervening portion] of the above three, respectively, 1, *O. faveolata* protein LOC110041892, 2, *E. diaphana* protein LOC110241027 and 3, *S. pistillata* protein LOC111342710; b, the carboxy-terminal (C-ter) portion (, i.e., none; the GPCR 2; the GPCR 3) and c, the central intervening portion (LOC110041892: amino acid residues 318-460; LOC110241027: 311~416 LOC111342710: 319~368 and 627~673) (which are encoded on exons 5~7, 3/4, and 1/2 and 2/3, respectively), containing partly single cell membrane-penetrating/transverse amino acid sequences that are predicted to form sc and to be necessary for its function. [The LOC111342710 central intervening portion excluding GPCR 2 (*Δ*TM1/*Δ*TM7) was used]. The blastp search using the central portion (c) found no similarity to human proteins. The blastp search using the N-ter portion (a) identified the following three groups of proteins with the respective E-values of (1) 0.5×10^{-46} [identity = 35% (114/327, identical amino acids out of the selected sequence); similarity = 52% (172/327, conserved amino acids out of the selected sequence)], (2) 0.1×10^{-40} [identity = 32% (95/294); similarity = 50% (147/294)] and (3) 0.1×10^{-33} [identity = 28% (86/307); similarity = 46% (144/307)] and 0.2×10^{-33} [identity = 31% (95/304); similarity = 48% (146/304)] and 0.8×10^{-19} [identity = 30% (73/246); similarity = 47% (118/246)]: 1, human dopamine D_1 receptor (D_1R) [with 446 amino acids in length (The open reading frame is encoded on a single exon); NCBI Reference Sequence: NP_000785.1]; 2, human b2 adrenergic receptor (b2AR) [with 342 amino acids in length (The open reading frame is encoded on a single exon); NCBI Reference Sequence: NP_000015.1] and 3 [the GPCR 1 and the central GPCR 2 (*Δ*DTM1/*Δ*DTM7) without the central intervening portion], human α1a adrenergic receptor (α1aAR) [with 466 amino acids in length (11 exons); GenBank: AAK77197.1; NCBI Reference Sequence: NP_000671.2] and human $A_{2A}R$ [with 412 amino acids in length (6 exons) (Fig. 11-15).

PDB: 3PWH_A; NCBI Reference Sequence: NP_000666.2] and 1~3, other proteins. Additionally, the search using the C-ter portion (b) resulted in the identification of 2, 5-hydroxytryptamine (serotonin) receptor 1B (5-HT_1B) [with 390 amino acids in length (a single exon); NCBI Reference Sequence: NP_000854.1], with an E-value of 0.4×10^{-25} [identity = 28% (91/326, identical amino acids out of the selected sequence);

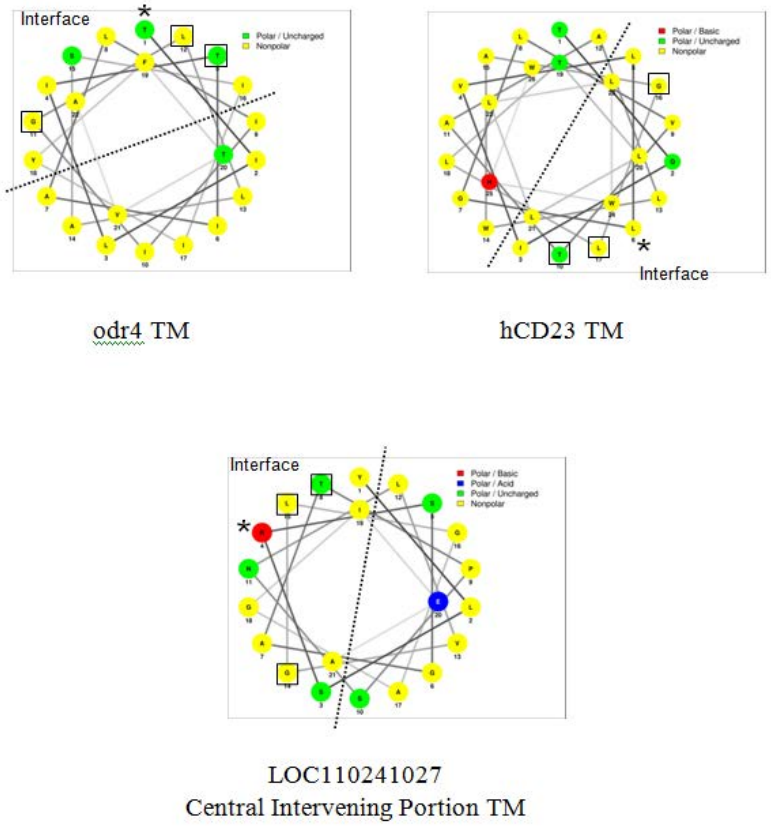


Fig. 11: Supplementary (Fig. 5A)

Table 11: Natural GPCR fusions with 600~699 amino acids in length from Euarchontoglires (excluding primates)[1]

Number	Organisms	Protein product	TM-linker in intervening region		Resulting GPCR
OBS82940.1	*Neotoma lepida*	A6R68_23065	No TM	dimer	Fig. S8C, T. 2, T. S3
OBS57425.1	*Neotoma lepida*	A6R68_11448	No TM	dimer	Fig. S11A
OBS69813.1	*Neotoma lepida*	A6R68_01647	No TM	dimer	Fig. S11B
OBS75694.1	*Neotoma lepida*	A6R68_17852, partial	No TM	dimer	Fig. S11C
OBS80342.1	*Neotoma lepida*	A6R68_21459, partial	No TM	dimer	Fig. S11D
OBS83474.1[2]	*Neotoma lepida*	A6R68_22502	No TM	dimer	Fig. S11E
OBS83645.1	*Neotoma lepida*	A6R68_22360, partial[3]	1 TM[4]	dimer	Fig. S11F
OBS71544.1	*Neotoma lepida*	A6R68_13879, partial	No TM	dimer	Fig. S11G
OBS69033.1	*Neotoma lepida*	A6R68_02426, partial	No TM	dimer	Fig. S11H
OBS59748.1	*Neotoma lepida*	A6R68_09130, partial	No TM	dimer	Fig. S11I
OBS59291.1	*Neotoma lepida*	A6R68_09584	No TM	dimer	Fig. S11J
OBS70632.1	*Neotoma lepida*	A6R68_00824	No TM	dimer	Fig. S11K
OBS77767.1	*Neotoma lepida*	A6R68_19843, partial	No TM	trimer	Fig. S11L
OBS57429.1	*Neotoma lepida*	A6R68_11439, partial	No TM	dimer	Fig. S11M
VTJ83394.1	*Marmota monax*	hypothetical, partial	No TM	trimer	Fig. S11N
VTJ85523.1	*Marmota monax*	hypothetical predicted	No TM	trimer	Fig. S11O
VTJ52678.1[2]	*Marmota monax*	hypothetical predicted	No TM	dimer	Fig. S11P
VTJ88622.1[5]	*Marmota monax*	hypothetical predicted	1 TM[5]	dimer	Fig. S11Q

[1]All 17 natural GPCR fusions with 600~699 amino acids in length from Euarchontoglires are here shown; A6R68_23065 (OBS82940.1)(with 685 amino acids in length) and *Neotoma lepida* and *Marmota monax* GPCR fusions with 700 or more amino acids in length are shown in Table 2. Both clones of *N. lepida* and *M. monax* proteins, respectively, with 674 and 605 amino acids in length, are highly similar each to each.3 [GPCR 1 (TM helices 1~7)]-(the central intervening region)-{GPCR 2 [(TM helices 1~7) ≈ (TM helices 1~7 of GPCR 1)]}4 The TM is similar to the latter half (84~94) of a TM helix (TMHMM result: 73~95) of *Panthera pardus fusca* cytochrome b, partial (with 267 amino acids in length; GenBank: ABN79993.1). 5 Also, (S3.10), this clone (with 613 amino acids in length; GenBank: VTJ88622.1) that had been regarded as a GPCR monomer at the screen [, i.e., by manual review of suspicious clones among GPCR fusions with 600 amino acids or more in length (S2.1),] was found to be a TM-linked GPCR dimer whose intervening region contains a TM (TMHMM: 331~353), which is similar to TM1 (TMHMM: 29~51) of *Rattus norvegicus* PREDICTED LOW QUALITY olfactory receptor 8D2 (with 309 amino acids in length; NCBI Reference Sequence: XP_008756332.1) and TM1 (TMHMM: 26~48) of *N. lepida* hypothetical protein A6R68_23906, partial (with 339 amino acids in length; GenBank: OBS82101.1)

Table 12: GPCR fusions with <600 amino acids in length and with low similarity to the human $A_{2A}R$ [1]

Number	Organisms	Protein product	TM-linker in intervening region	Resulting GPCR	
OBS81854.1	*Neotoma lepida*	A6R68_24156, partial[2]	No TM	dimer	Fig. S12A
OBS75898.1	*Neotoma lepida*	A6R68_17650, partial[3]	No TM	dimer	Fig. S12D
OBS69909.1	*Neotoma lepida*	A6R68_01549, partial[3]	No TM	dimer	Fig. S12E
OBS63836.1	*Neotoma lepida*	A6R68_07625, partial[4]	No TM	dimer	Fig. S12F
OBS68388.1	*Neotoma lepida*	A6R68_03066, partial[4]	No TM	dimer	Fig. S12G
VTJ67479.1	*Marmota monax*	hypothetical, predicted[2]	No TM	dimer	Fig. S12B
VTJ87989.1	*Marmota monaxs*	hypothetical predicted[2]	No TM	dimer	Fig. S12C
VTJ86726.1	*Marmota monax*	hypothetical predicted[4]	No TM	dimer	Fig. S12H

[1]During analyzing the clone (XP_033818019.1), one of Amphibia proteins, *Marmota monax* hypothetical predicted protein (with 613 amino acids in length; GenBank: VTJ88622.1) (Table 11, Fig. S11Q) and the 3 clones, respectively, were found to be present at and absent from the file obtained from the blastp search [of the full length of NP_000666 human $A_{2A}R$ protein against Euarchontoglires proteins, excluding primates (done on December 11, 2019), considering "Lineage"] (S3.6): The presence or absence of identity of 25% or more with regard to similarity to the query $A_{2A}R$, distinguished likely the hit-one from the other 3 non-hits. These 3 proteins (OBS81854.1, VTJ67479.1 and VTJ87989.1) are respectively with 645, 633 and 618 amino acids in length. [3]These 2 proteins (OBS75898.1 and OBS69909.1; respectively with 789 and 516 amino acids in length) (both of which have not been found in the above file) were identified as ones of clones that are similar to the clone (VTJ88622.1) 4 These 3 proteins (OBS63836.1, OBS68388.1 and VTJ86726.1) are respectively with 626, 631 and 627 amino acids in length

```
C. elegans odr4              TILITIALIIGLLASIIYFTVA----------  22
human ODR4                   ---------IgVia---AftVAVLAagIsFhyf  21
                                      **::*    ****


O. hoazin LOC104327099       ------VSIWGLAVLSVPSPSFLCILSLL------     23
Human ODR4                   ---------IgVia---AftVAV-LAagIsFhyf     21
hCD23                        ----tqivllglvt---aalwagltlllllwhw-     26
mCD23                        ----tqlmlvglls---tamwagllallllwhw-     26
rCD23                        ----tqlvlvgllt---tvmwagllallllwhw-     26
C. elegans odr4              TILITTALIIGLLA----SIIYFTVA---------     22
E. pallida LOC110241027      ----TPSNLVGLGA----GI---------------     12
                                      *:
consensus                        TIALIIGLLA    SI
                                  A NWA  GT    LL
                                  PS LV        G
```

Fig. 12: Supplementary S12

similarity = 46% (153/326, conserved amino acids out of the selected sequence)] and 3, 5-hydroxytryptamine (serotonin) receptor 4 ($5\text{-}HT_4$) isoform d [with 360 amino acids in length (12 exons); NCBI Reference Sequence: NP_001035262.2], with an E-value of 0.3×10^{-29} [identity = 29% (87/302); similarity = 46% (140/302)].

Next, the central intervening portions (LOC110041892: amino acid residues 318~460; LOC110241027: 311~416; LOC111342710: 319~368 and 627~673) were analyzed for their single cell membrane-penetrating/transverse amino acid sequences. Only the central intervening portion (amino acid residues 311~416) of the LOC110241027 protein contains a single probable TM domain (Fig. 1a~1c and Fig. 2), although, the TM-HMM 2.0 algorithm indicates that none of the three contain an additional TM helix between the annotated GPCR pairs. Also, (Fig. 2) shows little similarity of the amino acid sequence of the newly identified natural TM sc linkers and previously characterized *Caenorhabditis elegans odr4* TM-and human CD23 TM-linkers. However, while considering this alignment and little relationship between *C. elegans odr4* TM-and human CD23 TM[13], a consensus sequence (Fig. 10), T(I/A/P)(A/S)(L/N) (I/W/L)-(I/A/V) GL(L/G)-(A/T)(S/L/G) (I/L) was identified to be conserved, respectively, among *odr4* TM, CD23 TM and a single probable TM domain that shows probability comparable to those (around 0.8~0.9) of TM helices 3 of GPCR 1 and 2 (Fig. 1b) in the LOC110241027-central intervening portion (Fig. 2, overlined). BLAST search using *odr4* TM, CD23 TM and the single probable TM domain in the LOC110241027-central intervening portion, respectively, indicated that *odr4* TM has, for example, a high similarity to TM helix 2 (amino acid residues 82~104) of *Rhodococcus* sp. 29MFTsu3.1 ABC transporter permease (NCBI Reference Sequence: WP_019666760.1) (The TM-HMM 2.0 algorithm showed it to be a 6-TM protein.) and that the single probable TM domain in the LOC110241027-central intervening portion has, for example, a high similarity to TM helix 10 (amino acid residues 357~379) of *Rhodococcus imtechensis*

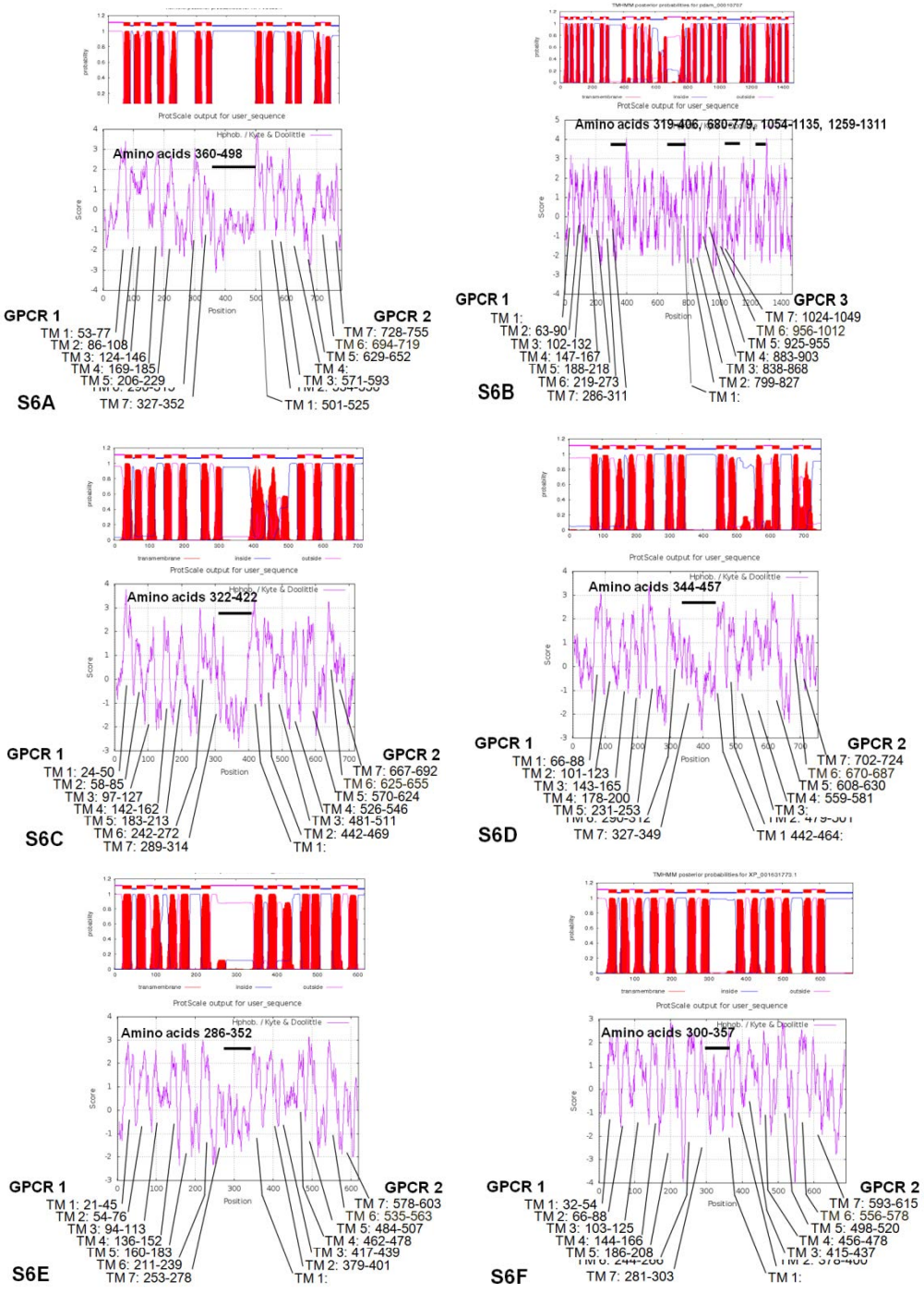


Fig. 13(a-f): Supplementary S13

Major facilitator superfamily (MFS) transporter (NCBI Reference Sequence: WP_039953125.1) (The TM-HMM 2.0 algorithm showed it to be a 11-TM protein.) but that CD23 TM has only similarity to itself.

Moreover, in addition to the invertebrate sea anemone *E. diaphana* protein LOC110241027-central intervening portion, it was found that in the central intervening region (amino acid residues 302~475) of tropical bird *Opisthocomus hoazin* protein LOC104327099 (NCBI Reference Sequence: XP_009930279.1), TM-linker (amino acid residues 318~340) [VSIWGLAVLS VPSPS FLCI(-)L(-)SLL]

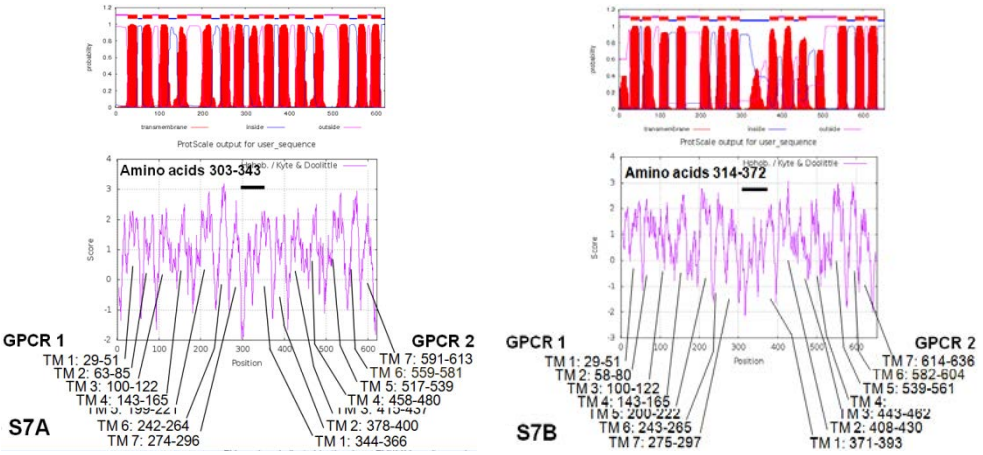


Fig. 14(a, b): Supplementary S14

contains the same consensus sequence (underlined) (Fig. 3a, Table 1): The TM within this central intervening region and TM 7 of GPCRs 1 and 2 were not determined by MEMSTAT3. Also, an alignment between the central intervening regions of invertebrates [*O. faveolata* (LOC110041892), *E. diaphana* (LOC110241027) and *S. pistillata* (LOC111342710) uncharacterized proteins] (Fig. 2) and this vertebrate *Opisthocomus hoazin* (LOC104327099) uncharacterized proteins suggested no similarity of these amino acid sequences. On the other hand in another natural vertebrate TM-linker of the desert woodrat *Neotoma lepida* hypothetical protein A6R68_19462, partial (GenBank: OBS78147.1), the major part of the central intervening region was found to consist of the N-terminus and TM helix 1 (amino acids 10~61), followed by the N-terminus (amino acids 1~32) of trace amine-associated receptor 8 (Fig. 3b, Table 1).

In order to more identify natural TM-linkers in GPCR fusions, we then narrowed "Lineage" such as order/family/genus/species in the class of Mammalia in web-based blastp search by the organism limit [S2. Supplemental Experimental procedures (S2.1. Rational methods of identifying natural TM-linkers in GPCR fusions)]. Among all mammalian proteins searched so far, 37 GPCR fusions were found to exist, although, not conclusively from lack of genome annotations, in which 29 GPCR fusions lack TM-linkers and 8 GPCR fusions of Euarchontoglires (, excluding Primates,) do not contain the consensus sequence in their central intervening portions (Table 2) [Supplemental Data (S3. Results)]. *Neotoma lepida* hypothetical protein A6R68_22360, partial (with 637 amino acids in length; GenBank: OBS83645.1) is a GPCR dimer fused through the third type of TM-linker that is similar to the latter half of a TM helix (TMHMM result: 73~95) of *Panthera pardus fusca* cytochrome b, partial (with 267 amino acids in length; GenBank: ABN79993.1) (Table 11, Fig. S11F). This is unrelated to GPCR itself or its receptor-interacting proteins[11]. *Marmota monax* hypothetical predicted protein (with 613 amino acids in length;

GenBank: VTJ88622.1) is a TM-linked GPCR dimer, whose intervening region contains a TM (TMHMM: 331~353) (the second type of TM-linker) which is similar to TM1 (TMHMM: 29~51) of *Rattus norvegicus* PREDICTED LOW QUALITY olfactory receptor 8D2 (with 309 amino acids in length; NCBI Reference Sequence: XP_008756332.1) and TM1 (TMHMM: 26~48) of *N. lepida* hypothetical protein A6R68_23906, partial (with 339 amino acids in length; GenBank: OBS82101.1) (Table 11, Fig. S11Q). Also, while among all avian proteins searched so far, 15 GPCR fusions were found to exist, all of these are dimers in which 12 GPCR fusions lack TM-linkers and *Opisthocomus hoazin* protein LOC104327099 (XP_009930279.1) does contain the consensus sequence in its central intervening portion but other 2 GPCR fusions of Aves do not (Table 3) [Supplemental Data (S3. Results)]. For the sake of clarity in presenting the above results, the details concerning GPCR fusions with <600 amino acids in length and with low similarity to the human $A_{2A}R$ which have been obtained during analyzing the clone (XP_033818019.1), one of Amphibia proteins, are provided in Supplemental Information (S3.10).

The evolutionary view of natural TM-linked GPCR fusions: A plant *Arabidopsis thaliana* fusion protein of a GPCR and the regulator of G-protein signaling protein 1 (AtRGS1) functions as a cell surface membrane sensor for D-glucose[29, 30]. Although, endogenous ligands of the

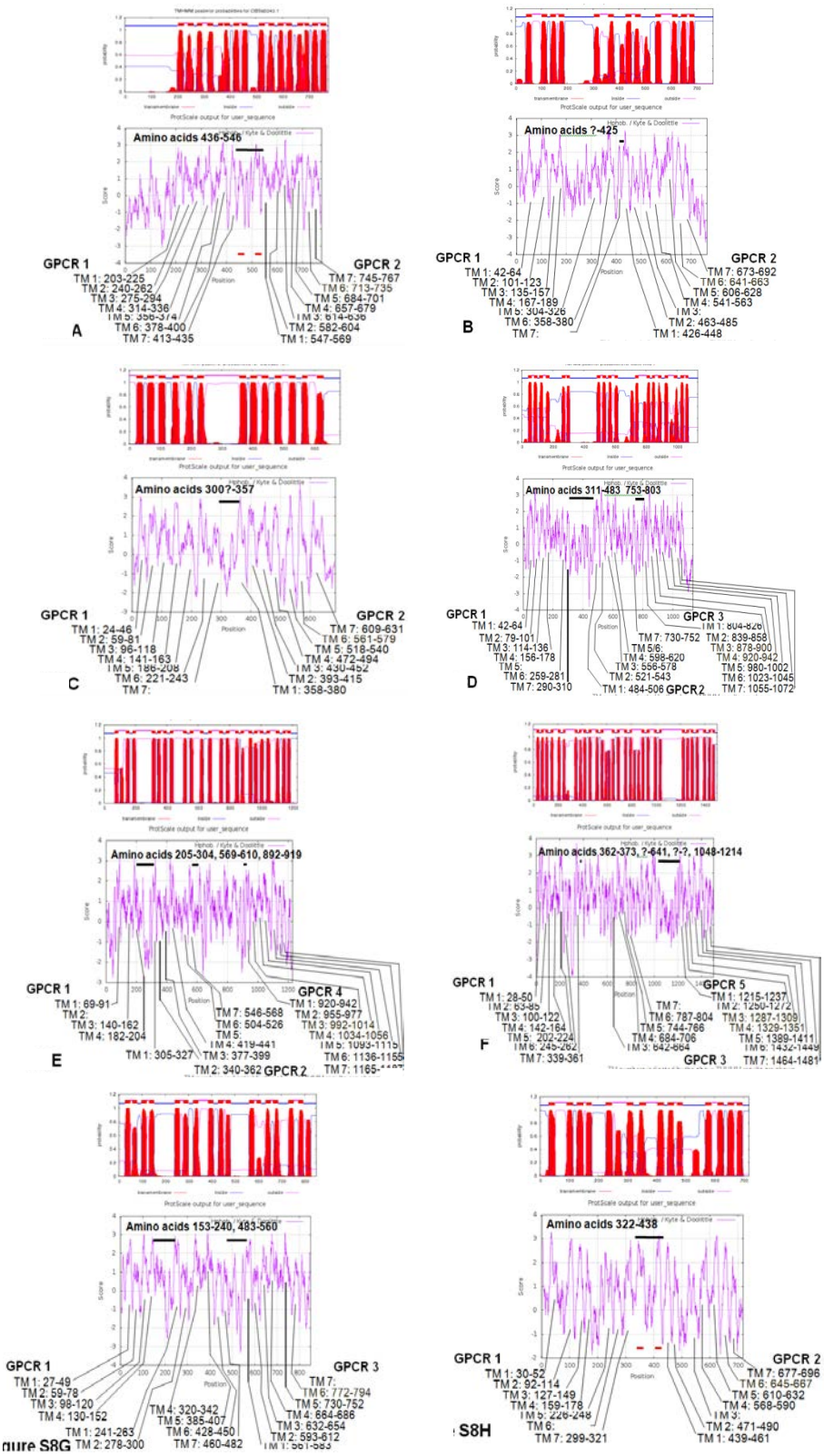


Fig. 15(a-p): Continue

\

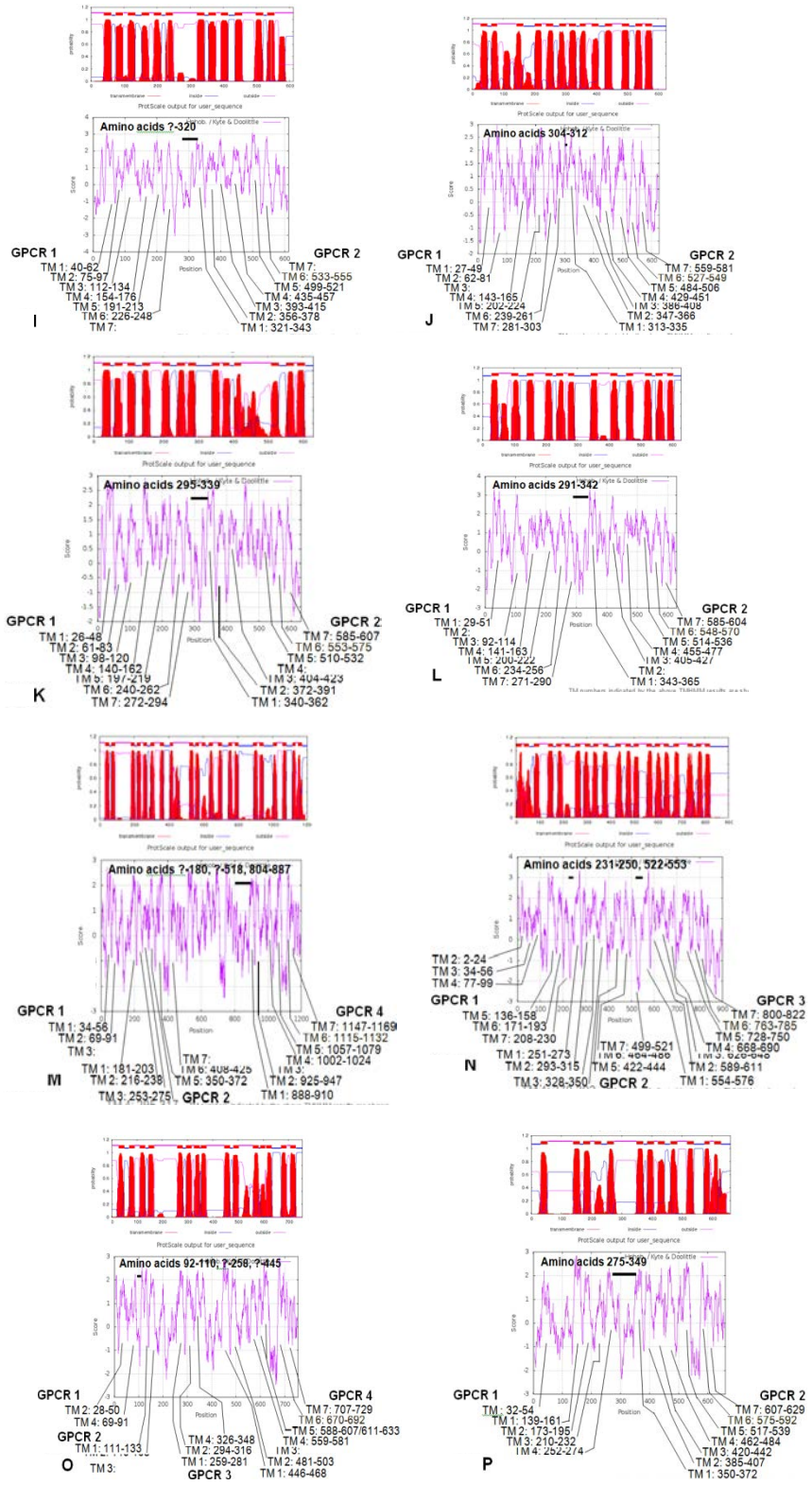


Fig. 15(a-p): Supplementary S15

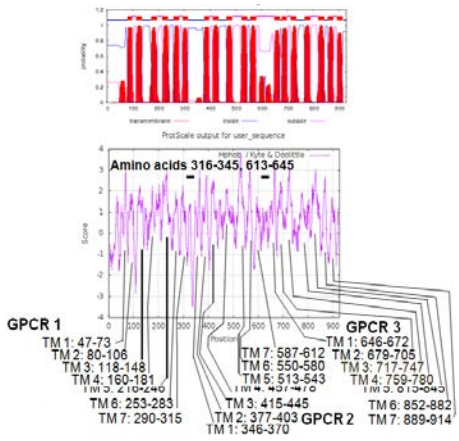


Fig. 16: Supplementary S16

above three GPCR oligomers that have been reported at NCBI, respectively, O. faveolata protein LOC110041892, *E. diaphana* protein LOC110241027 and *S. pistillata* protein LOC111342710, remain unknown, these results {on the amino acid sequence similarity and the genomic structures (Fig. S2A/B, S3A/B and S4A/B) of these obtained class A GPCR fusion proteins [homologous to the human $A_{2A}R$ (Fig. S1)] (Fig. 1a~1c) with their homologene(s) and on the central intervening portions analyzed for their single cell membrane-penetrating/transverse amino acid sequences (Fig. 2)} suggest that such a single-polypeptide chain GPCR dimeric (or trimeric) complex exists really in the natural world in agreement with the following evolutionary view[17-19]: "One of the most notable events is a vast expansion of the rhodopsin family at the origin of metazoans"[19] (Fig. 16). However, in the protein evolution of TM transport systems, i.e., intragenic duplication or gene fusion events from genes encoding simple pore-forming peptides with just 1~3 TM a-helical segments [30], the concept of TM linker was unnecessary. Additionally, a-helical TM proteins evolved by the strategy of "non-covalent oligomeric associations" 20 rather than fusing evolutionarily unrelated TM domains through a gene duplication and fusion event[31, 32]. Accordingly, in rare evolutionary cases, not only the invertebrate sea anemone *Exaiptasia diaphana* protein LOC110241027 but also vertebrate tropical bird *Opisthocomus hoazin* protein LOC104327099 became a GPCR dimer concatenated with a TM-linker that contains the consensus sequence. Two partial mRNA species (GenBank: JV084338.1, JV121130.1) are indeed expressed among transcripts of the LOC110241027 gene in the whole animal body[33] and the full-length mature mRNA (NCBI Reference Sequence: XM_028659670.1) of the TM-linked GPCR dimer is suggested with dataset on exon coverage and coverage of introns (intron-spanning reads)/intron features that are derived from alignments of sequences of spliced RNA (NCBI) while its expression has not yet been reported. Although, the consensus sequence mutation experiment has not been done, the consensus sequence may make a contribution to the important role of the receptors because $scA_{2A}R/D_2R$ of both $A_{2A}R$ -D2LR (DTM1) and D2LR-$A_{2A}R$ (DTM1) which have specific radioligand binding of only N-ter halves, i.e., respectively, $A_{2A}R$ and D_2R of the fusions, lack a TM-linker as well as TM1 of C-ter halves[11, 12] [This novel consensus sequence is discussed in Supplemental Information in more detail (S3.3.; Fig. S5: Helical-wheel representations)]. On the other hand the second and third types of TM-linkers were suggested, respectively, related and unrelated to GPCR itself or its receptor-interacting proteins[11], except for odr4TM that contains the consensus sequence (as described above). In the cases of natural vertebrate TM-linkers of the desert woodrat *Neotoma lepida* hypothetical proteins A6R68_19462 (with 737 amino acids in length; GenBank: OBS78147.1) (Fig. 3b, Table 1) and A6R68_22360, partial (with 637 amino acids in length; GenBank: OBS83645.1) (Table 11, Fig. S11F), the major part of the central intervening region of the former consists of the N-terminus and TM helix 1 [in a Nout-Cin orientation (that is with an extracellular N ter and a intracellular C ter)], followed by the N-terminus of trace amine-associated receptor 8 and the TM-linker of the latter is similar to the latter half of a TM helix (TMHMM result: 73~95) of Panthera pardus fusca cytochrome b, partial (with 267 amino acids in length; GenBank: ABN79993.1) while studies of membrane protein topology remains to be done in this and other cases of GPCR fusions without TM-linker (Table 1). In *Marmota monax* hypothetical predicted protein (with 613 amino acids in length; GenBank: VTJ88622.1), the second type of TM-linker is also suggested (Table 11, Fig. S11Q). Taken together, this suggests varied GPCR fusion patterns including a case of no TM-linker (Fig. 17-19).

The use of the central intervening portion (amino acids 311~416) of the *Exaiptasia diaphana* uncharacterized protein LOC110241027 for connection between the N-ter receptor half ($A_{2A}R$) and the C-ter receptor half (D2LR) of the $scA_{2A}R$ /D2LR Taking advantage of the rarity of TM-linker in the TM protein evolution, our previous reports[11,12] describe 'a type II TM protein with a cytoplasmic N-ter segment, single TM and extracellular C-ter tail', i.e., the Caenorhabditis elegans accessory protein of odorant receptor (odr4)[34] was first selected for connection between the N-ter receptor half ($A_{2A}R$) and the C-ter receptor half (D2LR, the long form of D_2R) of the $scA_{2A}R$ /D2LR. Then, it was demonstrated that an insertion of some other TM sequence instead of the *odr4* TM sequence (TILITIALIIGLLASIIYFTVA)

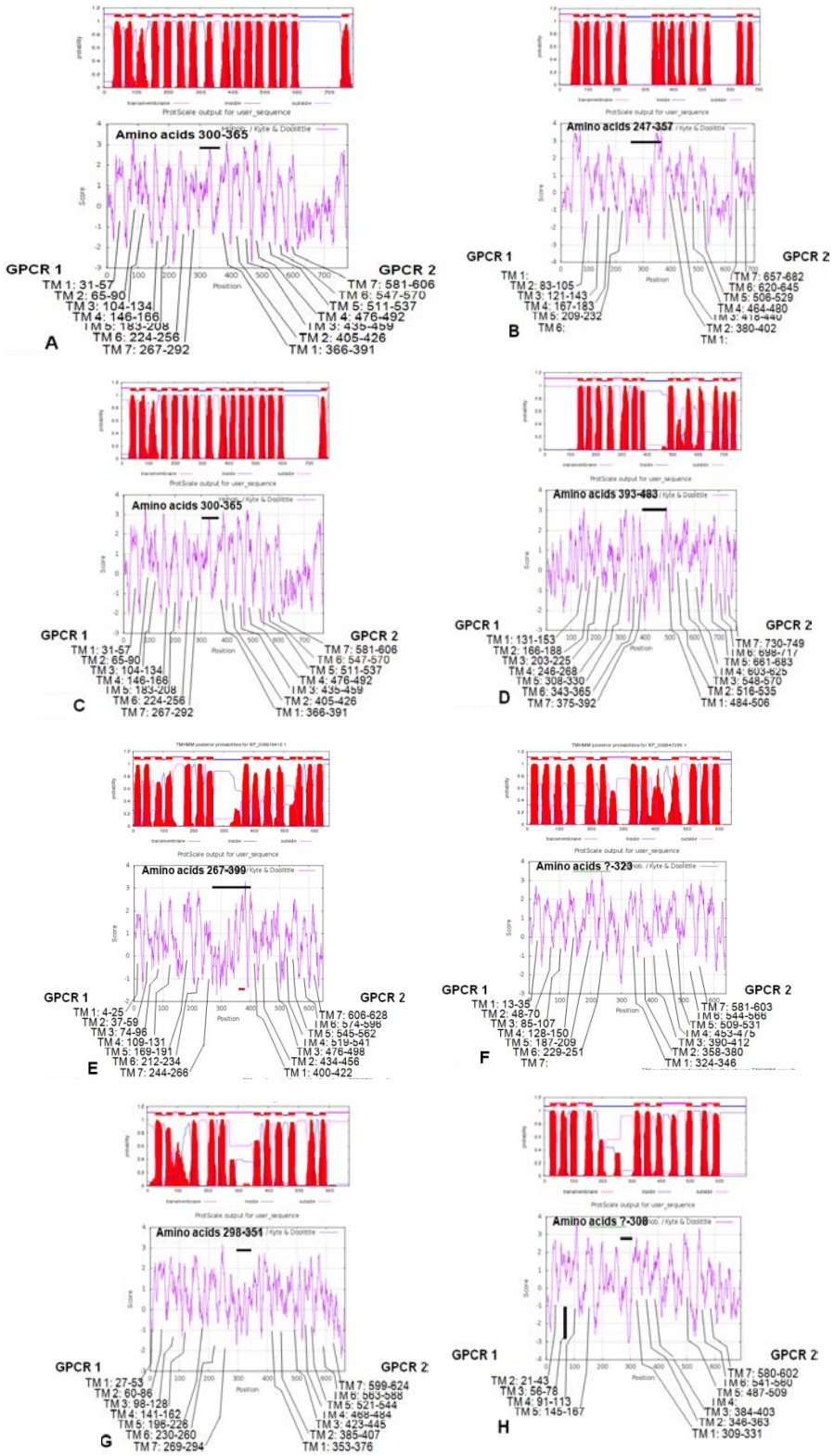


Fig. 17(a-p): Continue

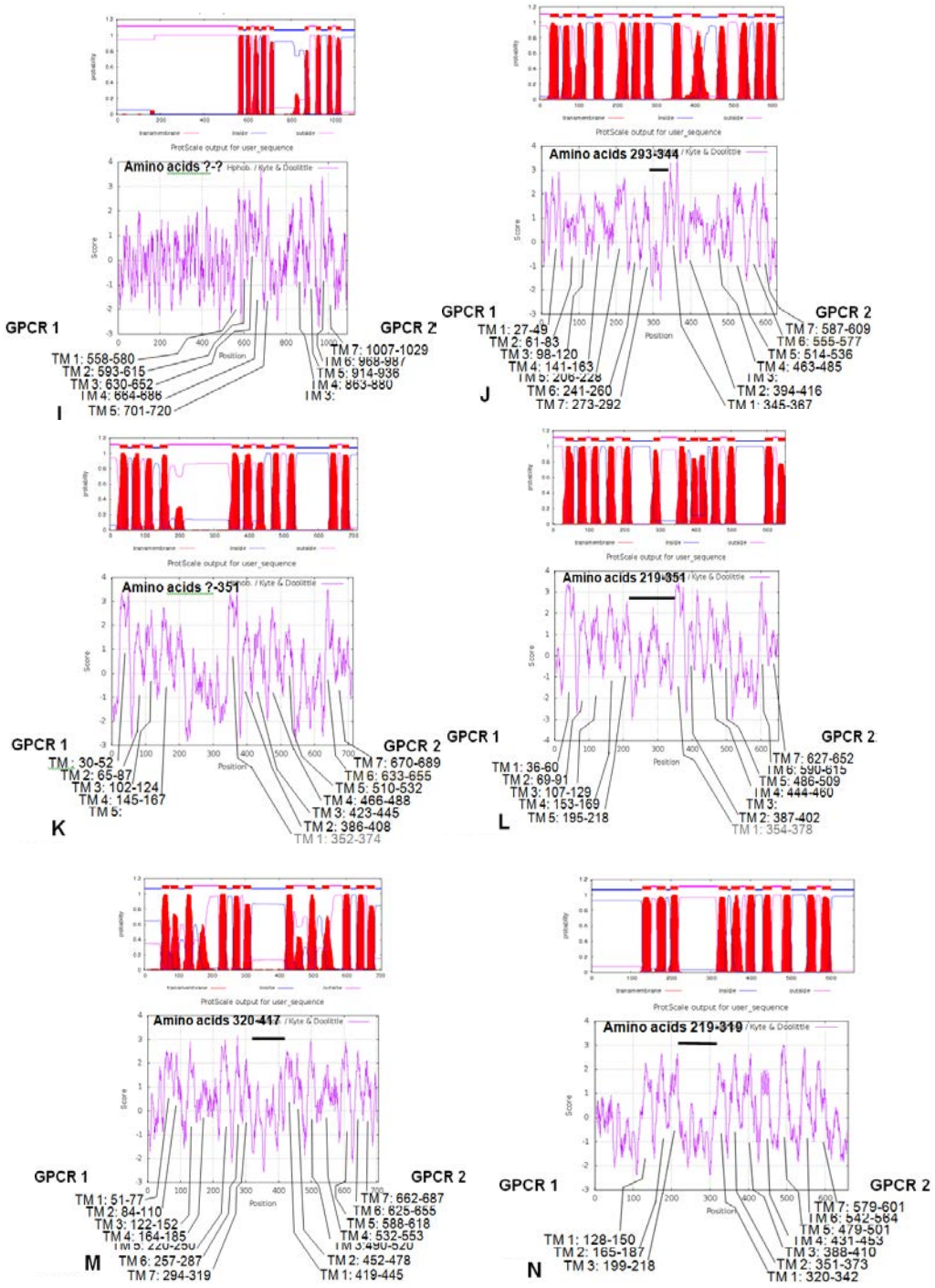


Fig. 17(a-n): Supplementary S17

works similarly, thus, it does not have to be odr4TM to onsensus sequence of natural scGPCR TM linker work, using the $scA_{2A}R$ /D2LR designed with another TM of a type II TM protein, the human low affinity receptor for IgE designated CD23. These show just retrospectively the usefulness of the consensus sequence that was found to be present not only in odr4TM and CD23 TM but also in natural scGPCRs in this study, although, natural scGPCR TM linkers themselves have not been tested experimentally: To investigate whether newly identified

TM sc linker can connect any two class A GPCRs in tandem, for example, the central intervening portion (amino acid residues 311~416) of *Exaiptasia diaphana* uncharacterized protein LOC110241027 including a single probable TM domain will be used for connection between the N-ter receptor half ($A_{2A}R$) and the C-ter receptor half (D2LR) of the sc$A_{2A}R$ /D2LR. With regard to hydropathicity of a single probable TM domain in the LOC110241027-central intervening portion (the prior section), the resulting natural TM-linked sc$A_{2A}R$/D2LR will be elusive. However, a long reach by the LOC110241027-central intervening portion could separate both from each other, i.e., the N-ter receptor half ($A_{2A}R$) and the C-ter receptor half (D2LR) of the sc$A_{2A}R$ /D2LR. As a result while the apparent ratio of $A_{2A}R$ to D_2R binding ites was shown to be approximately 4 in prototype sc$A_{2A}R$/D2LR possibly forced to form oligomers[12], the ratio in this natural TM-linked sc$A_{2A}R$/ D2LR could be 1 because of no or false interaction between the two receptors ($A_{2A}R$ and D2LR) within a single sc$A_{2A}R$/D2LR molecule. By using single-domain antibody (sdab: the antigen-binding moiety of heavy chain antibodies occurring in camelid species and cartilaginous fishes; nanobody), structural and functional analyses of GPCR indicate that 'GPCRs adopt multiple conformational states that can be differentially stabilized by molecules binding to topographically disparate sites, either alone or in combination.'[35] In our previous report[13], the 'exclusive' monomeric GPCRs with a rigid Ce2 domain of IgE-Fc, i.e., the C-ter of the odr4TM of the prototype sc$A_{2A}R$/D2LR fused to the N-ter of Ce2 and its C-ter fused to the N-ter of the D2LR, were designed, given that unlike IgG {[Fab (= VH+CH1)]-hinge-[Fc (= CH_2+CH_3)]}, IgE lacks the hinge region and IgE-Fc consists of Cε2-3-4 domains. Unlike the anti-IgE antibody so far, 026 sdab has been shown to induce the closed conformation of IgE-Fc (in which IgE cannot bind FceRI), by binding the site of IgE to its low-affinity receptor CD23 and through a half-bent conformation of IgE-Fc, not by directly binding the site of IgE to its high-affinity receptor FceRI, thus to inhibit binding to both FceRI and CD23 and to exhibit antiallergic activity[36]: Similar to the 026 sdab that can abrogate the two exclusive interacting sites at the same time, the central intervening portion of the LOC110241027 protein, i.e., the native sc(GPCR 1)/(GPCR 2), may confer the function of intra-molecular interactions between both GPCRs. Additionally, given that the central intervening portion of the LOC110241027 protein, i.e., the native sc(GPCR 1)/(GPCR 2), may or may not inhibit inter-molecular interactions such as higher order oligomerization or aggregation, we will take advantage of a monoclonal antibody against its long central intervening portion, so that, this natural TM-linked sc$A_{2A}R$/D2LR can be an exclusive dimer under an artificial condition with the addition of such an antibody. In any way, the Rosetta program and other algorithms for de novo protein structure prediction have been used successfully so far, and even though "evolutionary processes are not a good guide for its exploration"[37] of protein design[37-43], the abovementioned central intervening portion of the native sc(GPCR 1)/(GPCR 2) LOC110241027.

Protein born in "protein universe" is worth attempting. Here, we showed, for the first time, the capability of the natural TM sc linker and we have recently reviewed the TM sc linker deriving from the type II TM protein with a cytoplasmic N-ter segment, single TM and extracellular C-ter tail[13] while others have reviewed soluble linkers and fusion proteins[44-46].

Figure 1a *Orbicella faveolata* uncharacterized protein LOC110041892 (824 amino acids in length; NCBI Reference Sequence: XP_020602879.1). It encodes two identical class A GPCRs in tandem (amino acid residues 24~317 and 461~754) (NCBI annotation). Secondary structure of the predicted protein was analyzed using the program PSIPRED (http://www.expasy.org/tools/) (http://bioinf.cs.ucl.ac.uk/psipred/) and also using the scale hydropathicity (Kyte and Doolittle) (https:// web.expasy. org/cgi-bin/protscale/protscale.pl) (lower). A transmembrane (TM) domain for GPCR proteins was confirmed by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/) (upper).

Figure 1b *Exaiptasia pallida* uncharacterized protein LOC110241027 (761 amino acids in length; NCBI Reference Sequence: XP_020902518.1). It encodes two class A GPCRs in tandem (amino acid residues 21~310 and 417~727) (NCBI annotation). The scale hydropathicity (Kyte and Doolittle) (lower) and TM domains for GPCR proteins confirmed by the TM-HMM 2.0 algorithm (upper) are shown. The central intervening portions (LOC110041892: amino acid residues 318-460; LOC110241027: 311~416; LOC111342710: 319~368 and 627~673) were analyzed for their single cell membrane-penetrating/transverse amino acid sequences: only the central intervening portion (amino acids 311~416) of LOC110241027 contains a single probable TM domain (Fig. 1a~1c), although, the TM-HMM 2.0 algorithm indicates that none of the three contain an additional TM helix between the respective annotated GPCR pairs.

Figure 1c *Stylophora pistillata* uncharacterized protein LOC111342710 (1018 amino acids in length; NCBI Reference Sequence: XP_022805553.1). This encodes three class A GPCRs in tandem, of which the second GPCR is annotated as a deletion-type only with TM helix domain 2 (TM2)~TM6 (amino acid residues 34~318, 369~626 and 674~963). The scale hydropathicity (Kyte and Doolittle) (lower) and TM domains for GPCR proteins confirmed by the TM-HMM 2.0 algorithm (upper) are shown. Although, according to the NCBI

annotation with regard to the GPCR 2 (*Δ*TM1/*Δ*TM7) of the *S. pistillata* LOC111342710 protein [amino acid residues 369~626 (Section 3.1: p. 5, line 4 from the bottom)], TM 2 and TM 6 are amino acid residues 393~415 and 577~602, respectively and the central intervening portions at both ends of the GPCR 2 are amino acid residues 319~368 and 627~673. In this analysis, they were defined as amino acid residues 319~392 and 603~673, respectively, just before and after (neighboring) TM 2 and TM 6 of GPCR 2.

Figure 2 Little similarity of the amino acid sequence of the newly identified natural TM single-polypeptide chain (sc) linkers and previously characterized *Caenorhabditis elegans odr4* TM- and human CD23 TM-linkers. First, multiple sequence alignment was done among the indicated RC protein TM helices [of the thermophilic bacterium Thermo chromatium tepidum (Ref. 46: Extended Data Fig. 3)[47] and the purple bacterium Rhodopseudomonas viridis[48] photosynthetic reaction center (RC) subunits L (with 5 TM helices; GenBank: X03915.1), M (with 5 TM helices; GenBank: X03915.1) and H (with a TM helix; GenBank: X02659.1).], the Caenorhabditis elegans accessory protein of odorant receptor odr4 TM, human CD23 TM and a hit of Smart blast search of odr4TM [NCBI Reference Sequence: YP_009056868.1: PSII RC protein ycf12 (Bathycoccus prasinos)][13]. Here, a comprehensive comparison of a large number of membrane proteins with a single TM domain from the major secretory organelles from both fungi and vertebrates[49] was not used. Then, the set of odr4 TM and CD23 TM was manually aligned with the probable (parts of) TMs identified in the other set that had been aligned between the central intervening portions of *Orbicella faveolata* LOC110041892 and *Exaiptasia pallida* LOC110241027 and the 1st and the 2nd ones of *Stylophora pistillata* LOC111342710, showing little similarity of these amino acid sequences. The asterisk indicates an amino acid identical to odr4 TM and CD23 TM among the proteins. ":" means that conserved substitutions have been observed, according to the COLOR table: P in one letter code: yellow, among AVFPMILW including small [small+hydrophobic (including aromatic, except for Y)] amino acids; DE: sky blue, acidic amino acids; RK: red, basic amino acids; ST/Y: green/gray, among STYHCNGQ including hydroxyl (S/T/Y), uncharged polar (N/Q), basic (H) and nonpolar (C/G) amino acids. "." means that semi-conserved substitutions are observed. Also, aconsensus sequence, T(I/A/P)- (A/S) (L/N) (I/W/L) (I/A/V) GL (L/G) (A/T) (S/L/G)(I/L) is overlined, respectively, conserved among odr4TM, CD23 TM and a single probable TM domain that shows probability comparable to those (around 0.8~0.9) of TM helices 3 of GPCR 1 and 2 (Fig. 1b) in the LOC110241027-central intervening portion.

Bioinformatic analysis using Signal P[50] predicts that an amino acid sequence (627-EVYM VLLMFVLFST TWSR...) will be cleaved off (arrowhead) between position 626+17 Ser (S) and 626+18 Arg (R) as the signal sequence within the 2nd central intervening portion of *S. pistillata* LOC111342710, but that none of the central intervening portion of O. faveolata LOC110041892, *E. pallida* LOC110241027 and the 1st one of *S. pistillata* LOC111342710 have the signal sequence. This suggests *S. pistillata* LOC111342710 consisting of a dimer of the a1aAR-like GPCR 1 and the central A_{2A}R -like GPCR 2 (DTM1/DTM7) and a monomer of 5-HT4-like GPCR 3. On the othe hand, BLAST search showed that only the central intervening portion of *O. faveolata* LOC110041892 has a sequence producing an alignment of some significance (E = 3.2) (This row displays identities; plus (+) sign for similar matches) with a phage tape measure transmembrane protein (NCBI Reference Sequence: WP_007800194.1) from Salipiger bermudensis, respectively, shown just below the LOC110041892 sequence. This result recalls a common evolutionary origin of a major part of the type-6 secretion system machine and membrane-penetrating tails of bacteriophages within six types of secretion systems in Gram-negative bacteria that have been characterized, so far, including both the N-ter signal peptide-dependent (type-2 and -5 secretion systems) and -independent ones (type-1, -3 and usually also type-4 systems)[51-53]. Also, while in process of our forming a plan with the non-type II TM linker, a use of the Human Immunodefiency Virus (HIV)-1-derived, cell-permeable TAT peptide (GRKKRRQRRR)[54-56] has been come up with there are no such sequence.

Figure 3a *Opisthocomus hoazin* uncharacterized protein LOC104327099 (773 amino acids in length; NCBI Reference Sequence: XP_009930279.1). It encodes two class A GPCRs in tandem (amino acid residues 39~301 and 476~737) [NCBI annotation; However, in its item "FEATURES Region," TM helices are not assigned unlike LOC110041892, LOC110241027 and LOC111342710 (Fig. 1a~1c; their amino acid sequence similarities and their genomic structures: Fig. S2A/B, S3A/B and S4A/B) and not further analyzed for the amino acid sequence similarities and the genomic structures.]. The scale hydropathicity (Kyte and Doolittle) (lower) and TM domains for GPCR proteins confirmed by the TM-HMM 2.0 algorithm (upper) are shown. The central intervening portion (amino acid residues 302~475) was analyzed for its single cell membrane-penetrating/transverse amino acid sequence: A consensus sequence was found to exist in this natural vertebrate TM-linker (318~340) (VSI WGLAVLSVPS PSFLCILSLL).

The TM within this central intervening region and TM 7 of GPCRs 1 and 2 were not determined by

MEMSTAT3. Also, an alignment between the central intervening regions of invertebrates [*O. faveolata* (LOC110041892), *E. pallida* (LOC110241027) and *S. pistillata* (LOC111342710) uncharacterized proteins] (Fig. 2) and this vertebrate *Opisthocomus hoazin* (LOC104327099) uncharacterized proteins suggested no similarity of these amino acid sequences.

Figure 3b *Neotoma lepida* hypothetical protein A6R68_19462, partial (737 amino acids in length; GenBank: OBS78147.1). It encodes two class A GPCRs in tandem (amino acid residues 1~328 and 444~737)(TM numbers indicated by the TM-HMM results are shown.) [On NCBI annotation in its item "FEATURES" TM helices are not assigned unlike LOC110041892, LOC110241027 and LOC111342710.

Figure 1a-c; their amino acid sequence similarities and their genomic structures Fig. S2A/B, S3A/B and S4A/B and not further analyzed for the amino acid sequence similarities and the genomic structures.]. The scale hydropathicity (Kyte and Doolittle) (lower) and TM domains for GPCR proteins confirmed by the TM HMM 2.0 algorithm (upper) are shown. The central intervening portion (amino acid residues 329~443) was analyzed for its single cell membrane penetrating/transverse amino acid sequence:

Analysis by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/) indicated that the central intervening portion (amino acid residues 329?~443?) of this clone (OBS78147.1) does not contain an additional TM helix between the GPCR pairs, although it does contain a single probable transmembrane helix domain segment [that shows probability comparable to that (around 0.4) of TM helix 7 of GPCR 1] and that by an Artificial Neural Network program (MEMSAT3: http://bioinf.cs.ucl.ac.uk/?id=756)[20] did that it does contain a single probable transmembrane helix domain segment (amino acids 371~386: prlilyavfc fgtatg). The TM within this central intervening region and TM 3 of GPCR 1 and TM 7 of GPCR 1 were not determined, respectively, by the TM-HMM 2.0 algorithm and MEMSTAT3. The consensus sequence was found to exist in a natural vertebrate TM-linker (amino acids 318~340)(VSI WGLAVLSVPS.

PSFLCILSLL) in the central intervening region (amino acids 302~475) of tropical bird *Opisthocomus hoazin* protein LOC104327099 (XP_009930279.1) (Fig. 3a). In another natural vertebrate TM-linker (amino acids 371~386) identified between the putative dual GPCR of the desert woodrat *Neotoma lepida* protein hypothetical A6R68_19462, partial (OBS78147.1), the major part of the central intervening region (amino acids 329?~443?) was found to consist of the N-terminus and TM helix 1 (amino acids 10~61), followed by the N-terminus (amino acids 1~32) of trace amine-associated receptor 8. The latter case may be an artefact while its genome annotation has not been done due to the above reason. Also, an alignment between the central intervening regions of invertebrates [*Orbicella faveolata* (LOC110041892), *Exaiptasia pallida* (LOC110241027) and *Stylophora pistillata* (LOC111342710) uncharacterized proteins] (Fig. 2) and this vertebrate *Neotoma lepida* hypothetical (A6R68_19462) proteins suggested no similarity of these amino acid sequences.

A sequence that was found to exist in the natural vertebrate *Opisthocomus hoazin* protein LOC104327099 (NCBI Reference Sequence: XP_009930279.1) TM-linker (318~340) (VSI WGLAVLSVPS PSFLCILSLL) S-----P-S-F-L------C----I-L-S L T(I/A/P)(A/S)-(L/N)(I/W/L)(I/A/V)GL (L/G)(A/T)(S/L/G) (I/L) A consensus sequence in previously characterized *Caenorhabditis elegans odr4* TM- and human CD23 TM-linkers and the natural invertebrate *Exaiptasia pallida* protein LOC110241027 (NCBI Reference Sequence: XP_020902518.1) TM-linker.

Supplemental information

A consensus sequence among natural transmembrane linkers for single-polypeptide chain (sc) connection of two G-protein-coupled receptors in tandem: Toshio Kamiya, Takashi Masuko, Dasiel Oscar Borroto-Escuela, Haruo Okado and Hiroyasu Nakata:

- S1. Contents
- Supplemental Experimental procedures (S2), Supplemental Data (S3. Results and discussion and
- Supplemental Tables), Supplemental Fig. S4 and Supplemental References (S5)

Figure S1 related to Table 1; Fig. S2A/B, S3A/B and S4A/B, respectively, related to Fig. 1a-c and also to Fig. 2; Fig. S5A/B related to Fig. 2; Fig. S6 related to Table 1; Fig. S7A/B, S8A~S8P and S9 related to Table 2; Fig. S10A~S10N related to Table 3; Fig. S11A~S11Q related to Table 11; Fig. S12A~S12H related to Table 12.

S2. Supplemental Experimental procedures

S2.1. Rational methods of identifying natural TM-linkers in GPCR fusions: We empirically set the lower limit of the length of GPCR fusion protein to 600 amino acids while considering that a GPCR dimer is twice as long as GPCR monomer and that one of the longest and shortest class A GPCR monomer, respectively are for example, Ceratina calcarata uncharacterized protein LOC108631058 [with 1040 amino acids in length; NCBI Reference Sequence: XP_017890244.1. It encodes a class A GPCR (amino acid residues 42~321)] and human trace amine-associated receptor 8 (with 342 amino acids in length; NCBI Reference Sequence: NP_444508.1). In

order to limit searches to sequence lengths, we conducted the Entrez search independently from the BLAST search by using the E- utilities, one of the following: Web search in the Protein database; Entrez Programming Utilities (ESearch); EDirect. Next, we have continued to try to make the step [Find out the E-utilities equivalent to "Entrez Direct (E-utilities on the UNIX Command Line) for a work of homology search (= blastp)"] forward in our study, just such as Example: Find protein UIDs that are rat Reference Sequences and that are sequence similar to GI 15718680 https://eutils.ncbi.nlm.nih. gov/entrez/eutils/elink.fcgi?dbfrom=protein&db=protein&id=15718680&term=rat[orgn]+AND+srcdb+refseq[prop]&cmd=neighbor_history.

However, microbial organisms, including pathogen, may have GPCRs such as cytomegalovirus-encoded homologs of GPCRs in the co-evolution with the host cells, although GPCR fusions have not been reported so far. Accordingly, considering a reply e-Mail letter from NCBI, we made up our mind that we would better select web-based blastp search by the organism limit.

By this reason, first, blastp search of the full-length of NP_000666 human $A_{2A}R$ protein against human proteins resulted in the identification of a clone with the longest full-length among the total 1057 hits, follicle-stimulating hormone receptor isoform X1 (with 729 amino acids in length; NCBI Reference Sequence: XP_011531035.1). Next, blastp search of the full-length of NP_000666 human $A_{2A}R$ protein against mammalian proteins excluding humans resulted in the 20,000 hits, among which a clone with the longest full-length is Cricetulus griseus (Chinese hamster) DNA (cytosine-5)-methyltransferase 1-like protein (with 2024 amino acids in length; GenBank: ERE75644.1). Thus, we made up our mind that we would better select narrower "Lineage" than "mammalia" in web-based blastp search by the organism limit (S3.5).

S3.1. The human brain $A_{2A}R$ phenotype: With regard to the full length sequence of human $A_{2A}R$ (NCBI Reference Sequence: NP_000666.2) used in this study, a relevant brain-type Arg392 of the human $A_{2A}R$ (gb|S46950) is suggested (Fig. S1). Indeed, the missense (Gly392Arg) mutation of the human $A_{2A}R$ (dbSNP rs#1277013918), highlighted in yellow in Fig. S1 of our report [13] has been reported at NCBI recently. The human $A_{2A}R$ can be connected with the central intervening portion (amino acid residues 311~416) (not shown here) of *Exaiptasia pallida* uncharacterized protein LOC110241027 including a single probable TM domain.

S3.2. GPCR fusions to identify natural TM-linkers: As GPCR fusions, invertebrates *Orbicella faveolata* (LOC110041892; NCBI Reference Sequence: XP_020602879.1), *Exaiptasia pallida* (LOC110241027; NCBI Reference Sequence: XP_020902518.1) and *Stylophora pistillata* (LOC111342710; NCBI Reference Sequence: XP_022805553.1) uncharacterized proteins were first selected and analyzed 0 (Fig. S2A/B, S3A/B and S4A/B). The record 'XP_020902518.1' was changed to 'XP_028515471.1'(NCBI)[All the amino acid sequences remain unchanged except that Arg of the *Exaiptasia pallida* (which is at present, classified as E. diaphana,) uncharacterized protein LOC110241027 (NCBI Reference Sequence: XP_020902518.1) is replaced by Phe of the protein (NCBI Reference Sequence: XP_028515471.1) at amino acid residue 360 (3.1), in the extracellular loop between the TM sequence (amino acid residues 326~346) in the central intervening portion (amino acid residues 311~416) and GPCR 2 (417~727) (NCBI annotation). Therefore, we here described the sequence of events to correctly show results obtained using the record 'XP_020902518.1' in the main text (Figures) and Supplemental Information].

S3.3. Novel consensus sequence found in TM helices: Helical-wheel representations of the TM sequences from *Caenorhabditis elegans odr4*, human CD23 and the central intervening portion (amino acid residues 311~416). [The TM sequence (amino acid residues 326~346) is numbered 1~21 here] of *Exaiptasia pallida* uncharacterized protein LOC110241027 were generated with NetWheels at a website (http://lbqp.unb.br/NetWheels/) (done in December, 2018 but later this site was not found)] (Fig. S5A). Thr and Gly-Leu residues 5 and 11-12 (odr4), 10 and 16-17 (hCD23) and 8 and 14-15 (LOC110241027-central intervening portion) in the square boxes in TMs of these three proteins are parts of the consensus sequence, T(I/A/P)(A/S)(L/N)(I/W/L)(I/A/V)GL(L/G)(A/T)(S/L/G)(I/L). Also, Ser and (−)-Leu residues 13 and (−)- 20 are kept as parts of this consensus sequence in the natural vertebrate *Opisthocomus hoazin* protein LOC104327099 (NCBI Reference Sequence: XP_009930279.1) TM-linker (, i.e., an additional TM helix between the annotated GPCR pairs) (318-VSI WGLAVLSVPS PSFLCILSLL-340) (Fig. 3a, Table 1, Fig. S5B)(S3.4). This novel consensus sequence differs from the putative GXXXG motif commonly found in TM helices known to dimerize or the Rossmann fold (the GXGXXG motif: a super-secondary structure called a bab fold) which underlie inter-or intramolecular interactions of proteins. Note that between *C. elegans* ODR-4 {with 475 amino acids in length (11 exons); NCBI Reference Sequence: NP_001022814.1; (with 445 amino acids in length; NCBI Reference Sequence: NP_001367826.1) [This record was removed as a result of standard genome annotation processing (NCBI) during this study.]} and the human orthologous

protein [with 454 amino acids in length (15 exons); NCBI Reference Sequence: NP_060317.3], homology is across the entire regions including the TM helix segments, based on its sequence alignment[13, 23]; Also, not only an alignment between these two ODR4 TMs alone, i.e., *C. elegans* odr4 TM (422/452-TILITIALIIGLLASIIYFTVA-443/473) and human ODR4 TM [amino acid residues 432~452: 432 IgViaAftVAVLAagIsFhyf 452 (Regions representing identity at that position or conserved amino acids to *C. elegans* ODR4 TM are indicated in capital letters)] (Fig. S5B, upper) but also another one among *C. elegans*/human ODR-4 TMs, human/mouse/rat CD23 TMs and TM-linkers in the central intervening portions of the *Exaiptasia pallida* LOC110241027 and *Opisthocomus hoazin* LOC104327099 (XP_009930279.1) proteins indicated different alignments, but with all the same minimal consensus Gly-(Leu/Val) residues (Fig. S5B, lower). Accordingly, in human CD23 TM (and also in mouse and rat CD23 TMs), two minimal consensus Gly-(Leu/Val) residues [or (Thr/Ser)-(5 amino acids)-Gly-(Leu/Val)], both of which are encoded on exon 3 (human gene:[57, 58]; mouse gene:[59]; rat gene: NCBI) are included.

S3.4. Other GPCR fusions without TM-linker: In addition to 3 GPCR fusions of the above invertebrates [*Orbicella faveolata* (LOC110041892), *Exaiptasia pallida* (LOC110241027) and *Stylophora pistillata* (LOC111342710) uncharacterized proteins], 2 GPCR fusions of the vertebrates [*Opisthocomus hoazin* (LOC104327099) uncharacterized protein and *Neotoma lepida* hypothetical protein (A6R68_19462, partial)] (Table 1) were analyzed. Eight class A GPCR fusion proteins homologous to the human $A_{2A}R$ were obtained additionally and analyzed using the hydropathicity scale (Kyte and Doolittle) (https://web.expasy.org/protscale/), the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/) and/or MEMSAT3 (http://bioinf.cs.ucl.ac.uk/?id=756). The results indicate that all these single-chain (sc) GPCRs, i.e., the GPCR fusions (each consisting of either a GPCR dimer, trimer, or pentamer), lack likely natural TM-linker: Scleropages formosus (KPP73650.1) (Fig. S6A); *Pocillopora damicornis* (RMX53929.1) (Fig. S6B); Strongylocentrotus purpuratus (XP_011663779.1); *Orbicella faveolata* (XP_020603616.1) (Fig. S6C); Cricetulus griseus (ERE85073.1) (Fig. S6D); *Orbicella faveolata* (XP_020629325.1) (Fig. S6E); Nematostella vectensis (XP_001631773.1); Scleropages formosus (KPP58082) (Fig. S6F). Among these, the records 'XP_011663779.1' and 'XP_001631773.1' that are at present obsolete were removed as a result of standard genome annotation processing (NCBI) [It gives GPCR monomers, respectively, 'XP_030853119.1' (LOC582575 dopamine D2-like receptor) and 'XP_032236613.1' (LOC5511358 octopamine receptor)] during this study. Because we knew a case of the human $A_{2A}R$ SNP (S4. Figure legends-Fig. S1) and removals of records are expected to occur, we here described the sequence of events (Table 1).

Primates: In order to more identify natural TM-linkers in GPCR fusions, we then narrowed "Lineage" such as order/family/genus/species in the class of Mammalia in web-based blastp search by the organism limit (S2.1): Blastp search of the full-length of NP_000666 human $A_{2A}R$ protein against primates proteins excluding humans (done on November 25, 2019) resulted in the 7296 hits, among which a clone with the longest full-length is Otolemur garnettii (small-eared galago) mediator of RNA polymerase II transcription subunit 14 (predicted) [with 1677 amino acids in length; GenBank: ACG63590.1]. Also, among them, two GPCR fusions (consisting of a GPCR dimer) exist at the NCBI BLAST search whereas human genome has no GPCR fusion. One clone, Plecturocebus moloch (red-bellied titi) hypothetical protein (with 622 amino acids in length; GenBank: ACA53481.1) (Its genome annotation was not found), contains likely no natural TM-linker (Fig. S7A). Analysis by the TM-HMM 2.0 algorithm indicated that the central intervening portion (amino acid residues 303~343) of this clone (ACA53481.1) does contain a transmembrane helix domain segment but an Artificial Neural Network program (MEMSAT3) indicated that it does not. Against all organisms proteins, we then carried out blastp search using the full length of the above central intervening portion (amino acid residues 303~343) of this clone (ACA53481.1). The search resulted in the identification of Ceratotherium simum simum (southern white rhinoceros) protein (PREDICTED: LOW QUALITY PROTEIN: olfactory receptor 51I2-like)[with 341 amino acids in length (2 exons); NCBI Reference Sequence: XP_014645641.1] with an E-value of 1×10^{-11} [identity= 78% (32/41, identical amino acids out of the selected sequence); similarity = 80% (33/41, conserved amino acids out of the selected sequence)]: This suggests that the central intervening region consists of a protein sequence of Plecturocebus moloch (red-bellied titi) which is similar to the N-terminus (amino acids 12~52) that precedes/excludes TM1 (amino acids 51~73) of Ceratotherium simum simum (southern white rhinoceros) protein (PREDICTED: LOW QUALITY PROTEIN: olfactory receptor 51I2-like)(XP_014645641.1). The other clone, Callithrix jacchus (white-tufted-ear marmoset) hypothetical protein (with 654 amino acids in length; GenBank: ABZ80295.1) shows no natural TM-linker existence between the putative dual GPCR (Fig. S7B).

Euarchontoglires: Blastp search of the full-length of NP_000666 human $A_{2A}R$ protein against Euarchontoglires proteins excluding primates (done on December 11, 2019), considering "Lineage" (Mammalia; Eutheria; Euarchontoglires; Primates), resulted in the 15645 hits, among which a clone with the longest full-length is *Cricetulus griseus* (Chinese hamster) DNA (cytosine-5)-methyltransferase 1-like protein (with 2024 amino acids in length; GenBank: ERE75644.1). Also, among them, 9 GPCR fusions exist at the NCBI BLAST search, only of either *Neotoma lepida* (desert woodrat)[, including A6R68_19462 (GenBank: OBS78147.1) (Fig. 3b, Table 1)] or *Marmota monax* (woodchuck) hypothetical proteins: Genome annotations of all these clones were not found at the databases. Thus, such critical analyses of all these clones as done in invertebrates were suggested to need their genome and transcriptome annotations in order to obtain conclusive evidence. Whereas *N. lepida* hypothetical proteins A6R68_23065 (with 685 amino acids in length; GenBank: OBS82940.1) and A6R68_14817 (with 1135 amino acids in length; GenBank: OBS74663.1) and *M. monax* hypothetical predicted protein (with 850 amino acids in length; GenBank: VTJ82433.1) lack likely natural TM-linker, other scGPCRs, i.e., the GPCR fusions (each consisting of either a GPCR dimer, trimer, tetramer or pentamer) do not contain the consensus sequence in their central intervening portions (Table 2): *Neotoma lepida* (OBS80343.1) (Fig. S8A; Table 4), (OBS75582.1) (Fig. S8B; Table 5), (OBS82940.1) (Fig. S8C; Table 6), (OBS74663.1) (Fig. S8D; Table 7), (OBS78080.1) (Fig. S8E; Table 8); *Marmota monax* (VTJ79832.1) (Fig. S8F; Table 9), (VTJ82433.1) (Fig. S8G; Table 10), (VTJ70100.1) (Fig. S8H).

In the middle of July, 2020, during analyzing a clone of Amphibia proteins, in the file obtained from the above blastp search [of the full length of NP_000666 human $A_{2A}R$ protein against Euarchontoglires proteins, excluding primates (done on December 11, 2019), considering "Lineage"], additional 16 GPCR fusions, either 13 dimers or 3 trimers (Among which *Marmota monax* has a dimer and two trimers and *Neotoma lepida* has all others) were found to exist. All the GPCR fusions lack likely natural TM linker, except for the TM-linked GPCR dimer, i.e., *Neotoma lepida* hypothetical protein A6R68_22360, partial (with 637 amino acids in length; GenBank: OBS83645.1). This clone (OBS83645.1) is [GPCR1 (TM1~TM7)]-[central intervening region (TMHMM: 296~352)]-[GPCR 2 (TM1~TM7)] where GPCR1 (TM1~TM7) is almost identical to GPCR2 (TM1~TM7) and to Mus musculus olfactory receptor 464 (with 313 amino acids in length; Sequence ID: NP_666524.2) and where this clone (OBS83645.1) is highly similar to *Neotoma lepida* hypothetical protein A6R68_03066, partial (with 631 amino acids in length; GenBank: OBS68388.1) (S3.10) and where the central intervening region [(TMHMM: 296~352)(MEMSAT3: 293-352) (Phobius: 296~351)]-TM helix 1 [(TMHMM: 307~324: probability around 0.875) (Phobius: 307~324): 307-lakrlvsmsficflymhg-324] is rather than a consensus sequence, T(I/A/P) (A/S) (L/N) (I/W/L) (I/A/V)GL (L/G) (A/T)(S/L/G)(I/L), similar to the latter half (84-smfficlymhg-94) of a TM helix (TMHMM result: 73~95) of Panthera pardus fusca cytochrome b, partial (with 267 amino acids in length; GenBank: ABN79993.1).: *Neotoma lepida* (OBS57425.1) (Fig. S11A) which (OBS69813.1) (Fig. S11B), (OBS75694.1) (Fig. S11C), OBS80342.1) (Fig. S11D), (OBS83474.1) (Fig. S11E), (OBS83645.1) (Fig. S11F), (OBS71544.1) (Fig. S11G), (OBS69033.1) (Fig. S11H), (OBS59748.1) (Fig. S11I), (OBS59291.1) (Fig. S11J), (OBS70632.1) (Fig. S11K), (OBS77767.1) (Fig. S11L), (OBS57429.1) (Fig. S11M); *Marmota monax* (VTJ83394.1) (Fig. S11N), (VTJ85523.1) (Fig. S11O), (VTJ52678.1) (Fig. S11P).

Mammalia: Blastp search of the full length of NP_000666 human $A_{2A}R$ protein against Mammalia proteins excluding two superordinal clades Euarchontoglires and Laurasiatheria (done on January 27, 2020) resulted in the 6122 hits, among which a clone with the longest full length is Trichechus manatus latirostris (Florida manatee) LOW QUALITY PROTEIN: PDZ domain-containing protein 8-like [with 1178 amino acids in length; NCBI Reference Sequence: XP_012411926.2]. Also, among them, only one fusion exists at the NCBI BLAST search: Ornithorhynchus anatinus (platypus) uncharacterized protein LOC100087811 [LOW QUALITY PROTEIN: with 928 amino acids in length (6 exons); NCBI Reference Sequence: XP_028909644.1] (Fig. S9). This sc GPCR consisting of a GPCR trimer lacks natural TM-linker in its central intervening portions.

Laurasiatheria: Blastp search of the full length of NP_000666 human $A_{2A}R$ protein against Laurasiatheria proteins (done on January 28, 2020) resulted in the 18970 hits, among which a clone with the longest full length is Bos mutus (wild yak) hypothetical protein [with 2194 amino acids in length; MXQ95491.1]. Also, among them, 8 fusions, either dimer (total 5 dimers), trimer (total one), or tetramer (total two), exist at the NCBI BLAST search. All these scGPCRs lack natural TM linker in their central intervening portions (Table 2): Camelus bactrianus (XP_010960385.1) (Fig. S8I); Bubalus bubalis (XP_006073590.2) (Fig. S8J); Sorex araneus (ACE79115.1) (Fig. S8K), (XP_004613594.2) (Fig. S8L), (ACE79123.1) (Fig. S8M); Bos mutus (MXQ95485.1) (Fig. S8N), (MXQ95489.1) (Fig. S8O), (MXQ88753.1) (Fig. S8P).

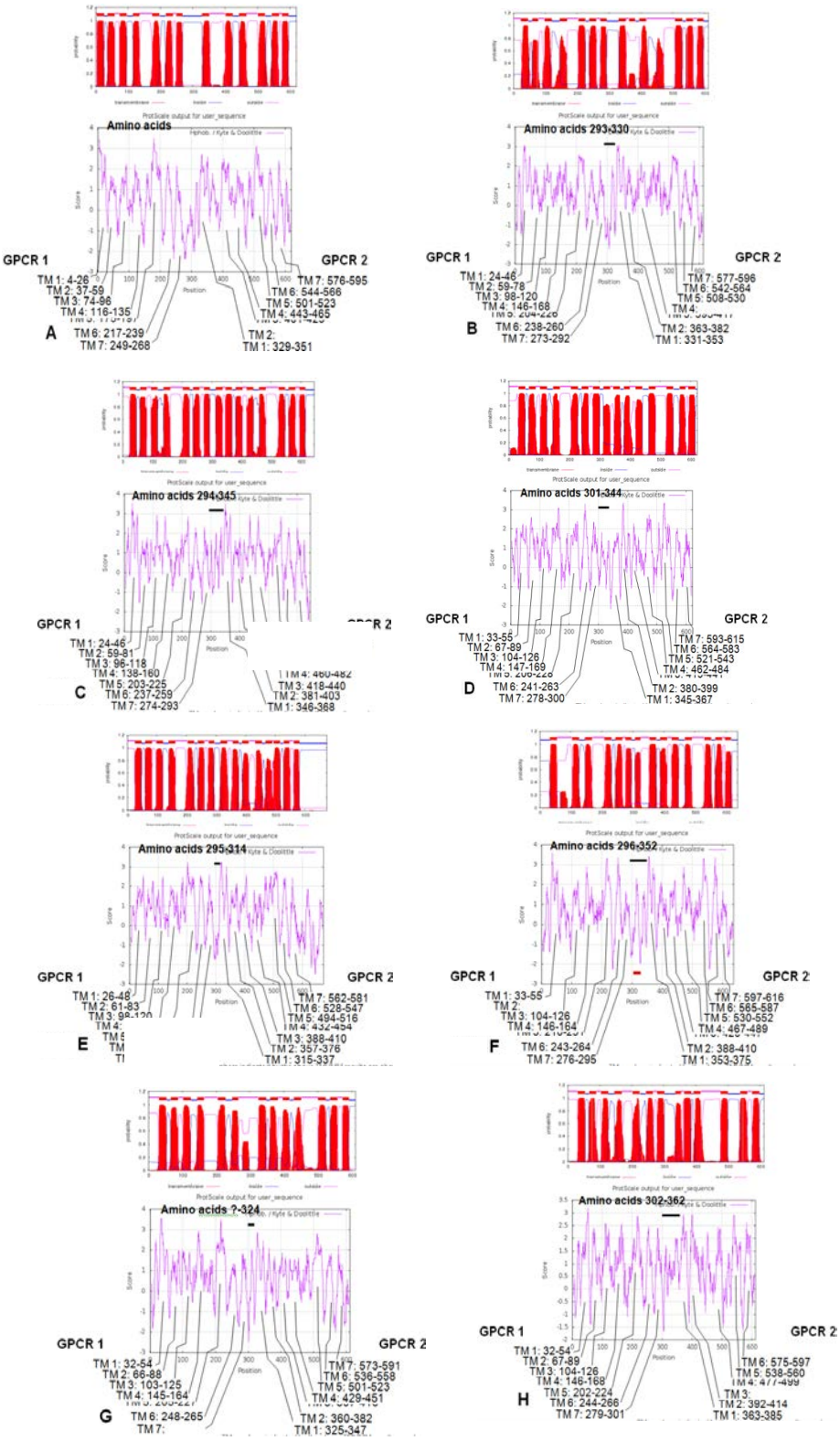


Fig. 18(a-q): Continue

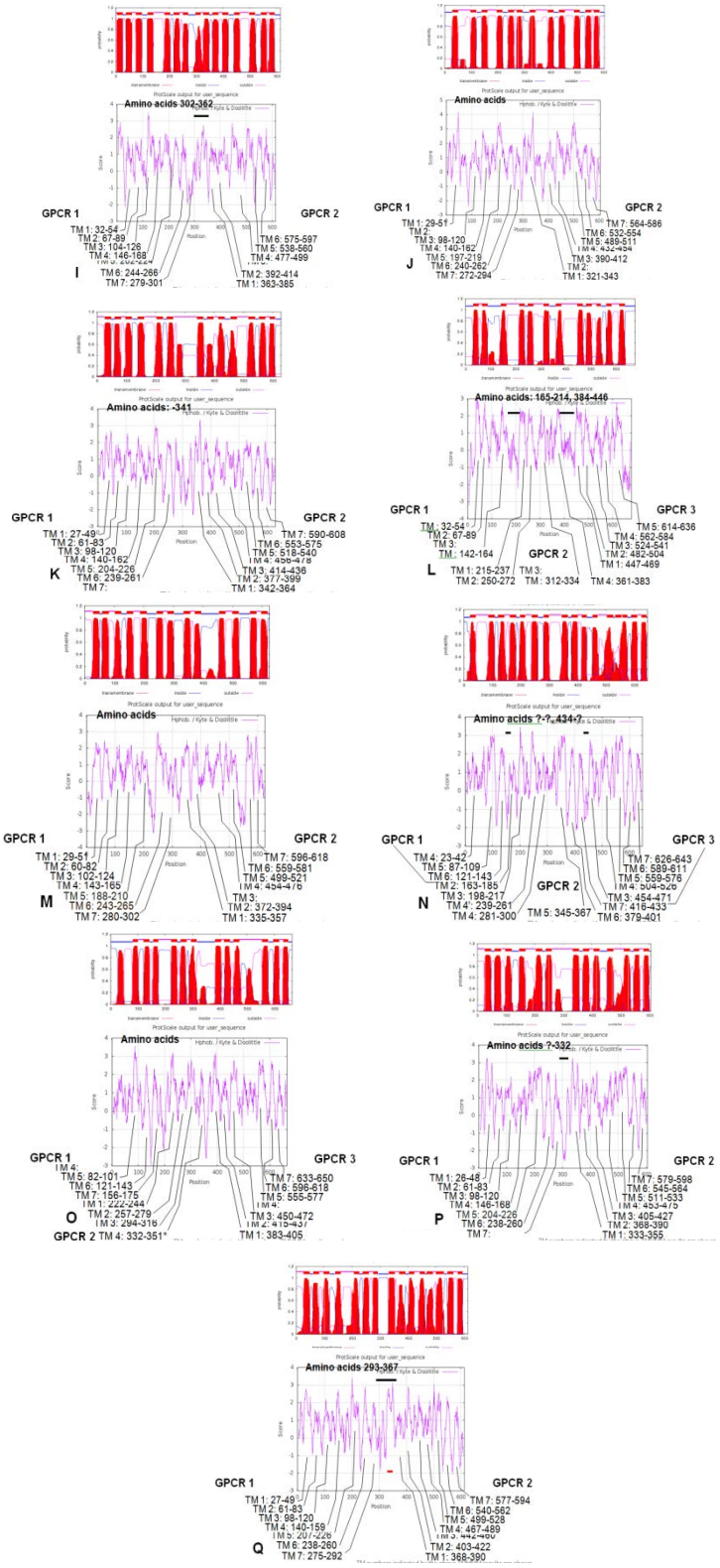


Fig. 18(a-q): Supplementary S18

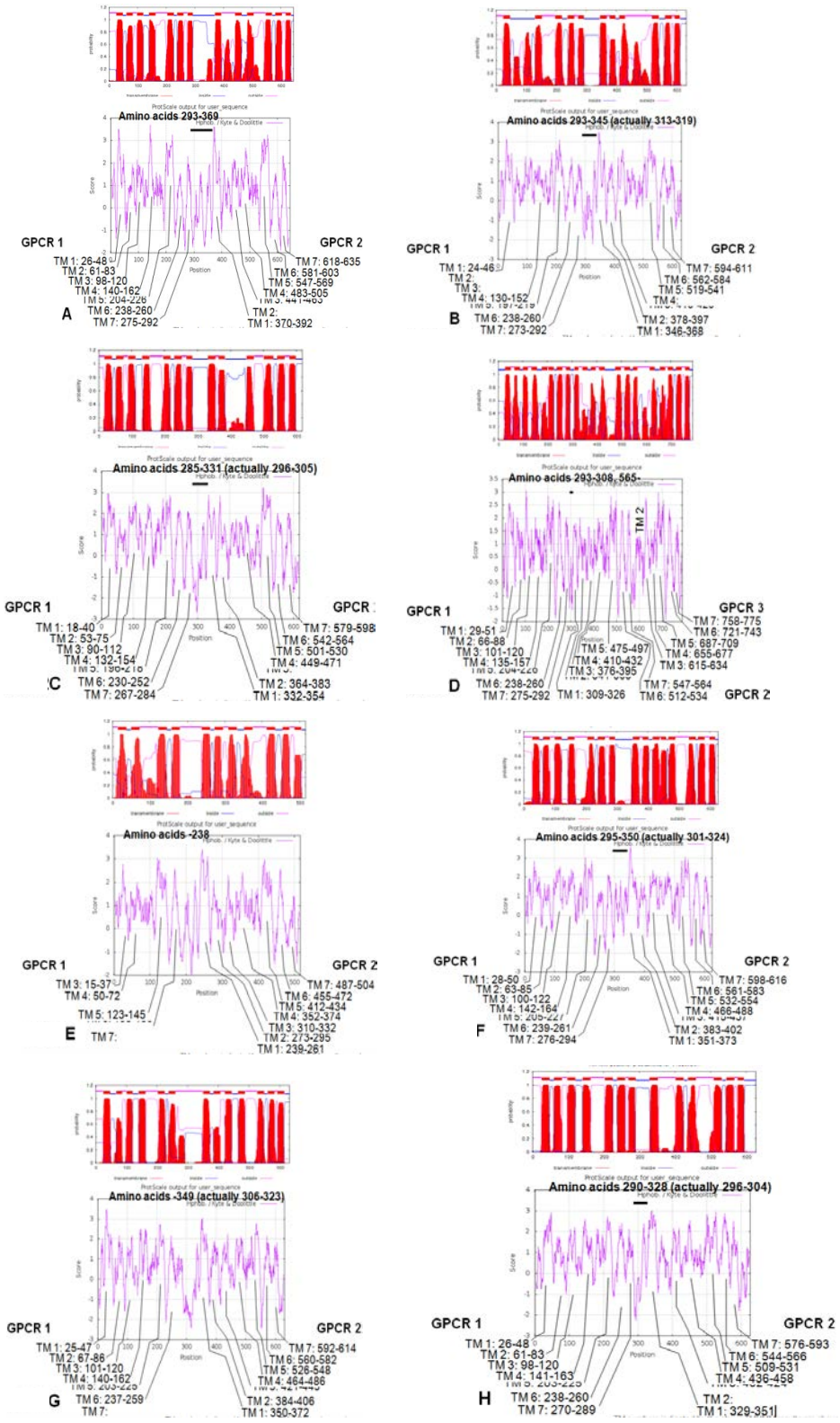


Fig. 19(a-h): Supplementary S19

Aves: Blastp search of the full length of NP_000666 human $A_{2A}R$ protein against Aves proteins, excluding Galloanserae proteins, (done on March 13, 2020) resulted in the 17266 hits, among which a clone with the longest full length is Columba livia (rock pigeon) DNA (cytosine-5)-methyltransferase 1 (with 1905 amino acids in length; XP_021157054.1). Also, blastp search of the full length of NP_000666 human $A_{2A}R$ protein only against Galloanserae proteins, out of Aves proteins (done on April 9, 2020) resulted in the 2334 hits, among which a clone with the longest full length is Bambusicola thoracicus hypothetical protein CIB84_009651, partial LOC110041892 (with 1286 amino acids in length; GenBank: POI26599.1) that is homologous to Gallus gallus (chicken) canalicular multispecific organic anion transporter 1 isoform X2 (with 1505 amino acids in length; NCBI Reference Sequence: XP_025007572.1) (ABCC2 ATP binding cassette subfamily C member 2). Among all these hits, 15 GPCR fusions exist at the NCBI BLAST search in which 12 fusions lack likely natural TM-linker and *Opisthocomus hoazin* protein LOC104327099 (XP_009930279.1) does contain the consensus sequence in its central intervening portion but other 2 GPCR fusions of Aves do not (Table 3): In Gavia stellata (red-throated loon) PREDICTED/LOW QUALITY uncharacterized protein LOC104264164 [with 650 amino acids in length (3 exons); NCBI Reference Sequence: XP_009819412.1] that is a dimer, [GPCR 1 (TM1~TM7)]-[central intervening region (TMHMM: 267~399)]-[GPCR 2 (TM1~TM7)], an amino acid sequence (338~ 399) in the central intervening region (TMHMM: 267~399; MEMSAT3: 261~396; Phobius: 267~396){containing TM helix 1 [intracellular (^{in}) 363~385 extracellular (^{out})](MEMSAT3: ^{out}355~385^{in}; Phobius: ^{in}362~385^{out})} has similarity to a sequence (1~62), consisting of the extracellular N-terminus, TM helix 1 (TMHMM result: 25~47) and the 1st intracellular region, of *Opisthocomus hoazin* olfactory receptor 52B2, partial (with 310 amino acids in length; GenBank: KFR05239.1), i.e., a related sequence of *O. hoazin* LOC104339471 olfactory receptor 52B2-like [with 316 amino acids in length (a single exon); NCBI Reference Sequence: XP_009943929.1](Fig. S10E); In Erythrura gouldiae hypothetical protein DV515_00000209, partial (with 652 amino acids in length; GenBank: RLW13346.1) that is a TM-linked? GPCR dimer, i.e., |GPCR 1 (TM1~TM5){(TM1/TM2, TM4/TM5) = [(TM1/TM2, TM4/TM5) of GPCR 2]}|-[central intervening region (NCBI: 219~351)]-[GPCR 2 (TM1/TM2, TM4~TM7)], in the central intervening region (NCBI: 219~351; TMHMM: >302~350; MEMSAT3:-; Phobius: >302~345), amino acid sequences (219~272; 319~351) are identical to those 510~563 [just between TM helix 5 (486~509) and TM helix 6 (590~615)] and 1~33 [just before TM helix 1 (36~60)], respectively and TM helix 1 (TMHMM: 282~301; 282- mlvtwcamv lqsvsatamw l-301) shows some similarity to a TM (TMHMM: 244~266: probability around 0.85; 244-imvlesv mhtamwaana isaggy-266) of {Erythrura gouldiae DV515_00000209 RLW13346- [central intervening region (NCBI: 219-351)-homologous} RCC_04814 related to metalloreductase of Ramularia collo-cygni (with 721 amino acids in length; NCBI Reference Sequence: XP_023625859.1) (Fig. S10L). All 12 fusions that lack likely natural TM-linker are as follows: Callipepla squamata (OXB57691.1) (Fig. S10A), (OXB61274.1) (Fig. S10B); Colinus virginianus (OXB75603.1) (Fig. S10C); Bambusicola thoracicus (POI30912.1) (Fig. S10D); Merops nubicus (XP_008947395.1) (Fig. S10F); Corvus cornix (XP_019149630.2) (Fig. S10G); Melospiza melodia maxima (KAF2980960.1) (Fig. S10H); Patagioenas fasciata monilis (OPJ89483.1) (Fig. S 10I); Limosa lapponica baueri (PKU37004.1) (Fig. S10J), (PKU42245.1) (Fig. S10K); Hirundo rustica rustica (RMC04025.1) (Fig. S10M); Zosterops borbonicus (TRZ15306.1) (Fig. S10N).

Notes are as follows: Lonchura striata domestica (Bengalese finch) C-C chemokine receptor type 5 (with 1110 amino acids in length: GenBank: OWK55386.1) is not a GPCR trimer, because this clone was derived from the genomic sequence: Nucleotide (Accession and Version: MUZQ01000204.1) and at present has been assigned to the record, LOC110471067 C-C chemokine receptor type 8 of L. striata domestica (with 351 amino acids in length: NCBI Reference Sequence: XP_031360030.1) that is derived from the genomic sequence [Annotation release: 101; Status: current; Assembly: lonStrDom2 (GCF_005870125.1); Chr: 2; Location: NC_042567.1 (44335887..44336951, complement)]; Limosa lapponica baueri (the Bar-tailed Godwit) d dopamine receptor (with 1890 amino acids in length; GenBank: PKU33311.1) includes unidentified amino acid sequences (506~1418); Fig. S10A: Callipepla squamata hypothetical protein ASZ78_008252, partial (with 777 amino acids in length; GenBank: OXB57691.1) is a TM-linker-lacking GPCR dimer because of being highly homologous to hypothetical protein H355_015719, partial of Colinus virginianus (with 775 amino acids in length; GenBank: OXB75603.1) that is a TM-linker-lacking GPCR dimer as described below. Note that Colinus virginianus hypothetical protein H355_006548 (with 737 amino acids in length; GenBank: OXB73539.1) is considered not to be a GPCR dimer because this clone contains a unidentified amino acid sequence (376-xxxxx xxxxxxxxxx xx-392) in the central intervening region (NCBI: 299~449) whereas Egretta garzetta (little egret) LOC104124944 G-protein coupled receptor family C group 5 member C-like [PREDICTED: with 611 amino acids in length (5 exons); NCBI Reference Sequence:

XP_009635631.1] is a TM-linker-lacking GPCR dimer and contains a unidentified amino acid sequence (442-xxxxxxxxx-450) outside its central intervening region (TMHMM: 291~325); Fig. S10B: Callipepla squamata hypothetical protein ASZ78_010190 (with 706 amino acids in length; GenBank: OXB61274.1) (Genome annotation of this clone was unavailable) that is a TM-linker-lacking GPCR dimer, i.e., [GPCR 1 (TM1?, TM2~TM5, TM6?)]-[central intervening region (NCBI: 247~357; TMHMM: 349~351; MEMSAT3: 371~379; Phobius: 348~352)]-[GPCR 2 (TM2~TM7)] where a sequence (247~333) of its central intervening region (NCBI: 247~357) shows a high similarity to the sequence (247~334) that encompasses a part of the 3rd intracellular region and a part of TM 6, of Numida meleagris histamine H3 receptor-like (with 414 amino acids in length; NCBI Reference Sequence: XP_021244296.1) and where also, (GPCR 1)-{TM helix 6 (TMHMM: 326~348)[(MEMSAT3: 342~370; Phobius: 326~347)]}(326-IAKSLAVITSTPQFSLGVLVLLA-348) or TM1? of central intervening region (NCBI: 247-357): (TMHMM: 349~351)(MEMSAT3: 371~379)(Phobius: 348~352) of *C. squamata* hypothetical protein ASZ78_010190 is not a TM helix because this sequence is homologous to regions, i.e., amino acid sequences 37~51 [that encompasses a part of TM helix 1 (TMHMM result: 50~72; MEMSAT3: 44~73; Phobius: 42~73) and that precedes TM helix 2 (NCBI: 83~105; TMHMM: 85~107; MEMSAT3: 84~106; Phobius: 85~105) of GPCR 1] and 622~629 [that is a part of TM helix 6 (NCBI: 620~645) of GPCR 2] of this clone itself and to regions, i.e., amino acid sequences 13~27 [that precedes TM helix 2 (NCBI: 59~81)] and 301~308 [that is a part of TM helix 6 (NCBI: 299~324)] of Numida meleagris (helmeted guineafowl) LOC110394665 histamine H3 receptor-like [with 385 amino acids in length (3 exons); NCBI Reference Sequence: XP_021244303.1]; Fig. S10C: Colinus virginianus hypothetical protein H355_015719, partial (with 775 amino acids in length; GenBank: OXB75603.1) is a TM-linker- lacking GPCR dimer, i.e., [GPCR 1 (TM1~TM7)]-[central intervening region (NCBI: 300~365)]-[GPCR 2 (TM1~TM7)], because Gallus gallus GPCR family C group 5 member C (with 542 amino acids in length; NCBI Reference Sequence: XP_004946305.1) has a sequence (106~163) which precedes TM helix 1 (TMHMM: 172~194), similar to the central intervening region of Colinus virginianus hypothetical protein H355_015719 partial (with 775 amino acids in length; GenBank: OXB75603.1); Fig. S10D: Bambusicola thoracicus hypothetical protein CIB84_005337, partial (with 771 amino acids in length; GenBank: POI30912.1) is a TM-linker-lacking GPCR dimer, i.e., [GPCR 1 (TM1~TM7)]-[central intervening region (TMHMM: 393~483)]-[GPCR 2 (TM1~TM7)], because the central intervening region (TMHMM: 393~483; MEMSAT3: –; Phobius: 395~482) of this clone (POI30912.1)[containing TM helix 1 (TMHMM: not identified; Phobius: 460~477: probability around 0.55)] has a sequence (amino acid residues 458~483) similar to the N-terminus (amino acid residues 1~26) which precedes just TM helix 1 (26~52) of Gallus gallus (chicken) OR1F2P olfactory receptor 5V1-like (with 322 amino acids in length; NCBI Reference Sequence: XP_025002272.1) while this clone (POI30912.1) shows similarity (458~757 and 110~398, respectively, higher and lower) to the sequences (1~314 and 5~296) of G. gallus OR1F2P (XP_025002272.1); Fig. S10F: Merops nubicus (carmine bee-eater) PREDICTED uncharacterized protein LOC103781429 [with 639 amino acids in length (3 exons); NCBI Reference Sequence: XP_008947395.1] is a TM-linker-lacking GPCR dimer, i.e., [GPCR 1 (TM1~TM7)]-[central intervening region (TMHMM: ?~323)(MEMSAT3: 277~321)(Phobius: ?~322)]-[GPCR 2 (TM1~TM7)]; Fig. S10G: Corvus cornix cornix LOW QUALITY uncharacterized protein LOC104698100 (with 666 amino acids in length; NCBI Reference Sequence: XP_019149630.2) is a TM-linker-lacking GPCR dimer and has similarity to Corvus brachyrhynchos LOC103624072 olfactory receptor 52K2-like (with 342 amino acids in length; XP_017600043.1), while the central intervening region (NCBI: 298~351) of this clone (XP_019149630.2) has similarity to a sequence (1~27), i.e., non-TM, of LOC103624072 (XP_017600043.1); Fig. S10H: Melospiza melodia maxima (the North American Song Sparrow) hypothetical protein EK904_014712 (with 693 amino acids in length; GenBank: KAF2980960.1) is a TM-linker-lacking GPCR dimer, i.e., [GPCR 1 (TM2~TM5, TM6?, TM7?)]-[central intervening region (TMHMM: ?~308; MEMSAT3: ?~309; Phobius: ?~312)]-[GPCR 2 (TM1~TM7)]; Fig. S10I: Patagioenas fasciata monilis (the band-tailed pigeon) adhesion G-protein coupled receptor G4 [with 1092 amino acids in length (Its genome annotation is unavailable); GenBank: OPJ89483.1] is a TM-linker-lacking GPCR dimer, Laminin G-like domain containing [GPCR 1 (TM1~TM4, TM5?)]-[GPCR 2 (TM 3~TM7)] in which although Colius striatus (speckled mousebird) LOC104562211 probable G-protein coupled receptor 112 [with 1026 amino acids in length (20 exons); NCBI Reference Sequence: XP_010206488.1] with a sequence (821~988) has a similarity to a sequence (687~816), i.e., [just after TM4 (TMHMM: 664~686),] consisting of [TM5 (TMHMM: 701~720)]-[central intervening region (TMHMM: ?~?; MEMSAT3: –; Phobius: ?~816)], of this clone (OPJ89483.1), within the sequence (821~988) of C. striatus GPCR 112 (XP_010206488.1), {GPCR [TM1 (TMHMM result: 693~715)]-[TM2 (727~749)]-[TM3 (762~784)]-[TM4 (797~819)]-[TM5 (851~873)]}, TM helix 5 (851-VAAYIFLIFLMNIAMFITVLLQI-873) of *C. striatus* GPCR 112 (XP_010206488.1) shows little

similarity to the relevant sequence within the sequence (687~816) of *P. fasciata* monilis adhesion GPCR G4 (OPJ89483.1) and in which the central intervening region of *P. fasciata* monilis adhesion GPCR G4 (OPJ89483.1) does contain no TM helix; Fig. S10J: Limosa lapponica baueri hypothetical protein llap_12689 [with 631 amino acids in length (Genome annotation of this clone was unavailable); GenBank: PKU37004.1) is a TM-linker-lacking GPCR dimer in which the N-terminus, i.e., an amino acid sequence (8-NETAVVEFVIL GFQSIPEVQFLL-30), of Calidris pugnax (ruff) LOC106890090 PREDICTED LOW QUALITY PROTEIN olfactory receptor 11A1-like [with 323 amino acids in length (a single exon); NCBI Reference Sequence: XP_014801159.1], {similar to the central intervening region (TMHMM: 293~344) of *L.lapponica* baueri protein llap_12689 (PKU37004.1)} is not a TM helix while this clone-homologous LOC112060720 olfactory receptor 9K2-like of Chrysemys picta bellii (western painted turtle) (with 630 amino acids in length; NCBI Reference Sequence: XP_023968235.1) that is derived from the genomic sequence (NW_007281696 Unplaced Scaffold Reference Chrysemys_picta_bellii-3.0.3 Primary Assembly) contains about a 50-base-pairs (bp) sequence that has not been determined (194.15 K~194.2 K), consisting of a region, the intron between the two exons of the gene of the clone (XP_023968235.1); Fig. S10K: Limosa lapponica baueri (the Bar-tailed Godwit) histamine h3 receptor-like (with 714 amino acids in length; GenBank: PKU42245.1) is a TM-linker-lacking GPCR dimer, i.e., {GPCR 1 [= (TM1~TM5) of GPCR 2]}-[GPCR 2 (TM1~TM7)]; Fig. S10 M: Hirundo rustica rustica (the barn swallow) hypothetical protein DUI87_19362 [with 705 amino acids in length (It genome annotation is unavailable); GenBank: RMC04025.1] is a TM-linker-lacking GPCR dimer, i.e., [GPCR1 (TM1~TM7)]-(central intervening region)-{GPCR 2 [= GPCR 1 (TM1~TM7)]}, where a sequence (390~417) of the central intervening region (NCBI: 320~417; TMHMM: 318~420; MEMSAT3: 313~418; Phobius: 318~417) has a high similarity to the sequence (22~49), i.e., parts of its own TM helix 1 (51~77) and TM helix 2 (84~110); Fig. S10N: Zosterops borbonicus hypothetical protein HGM15179_011826 (with 660 amino acids in length; GenBank: TRZ15306.1) is a TM-linker-lacking GPCR dimer, i.e., [GPCR 1 (TM1~TM3)]-[GPCR 2 (TM1~TM7)]; Note that Lonchura striata domestica (Bengalese finch) G-protein coupled receptor 35 (with 611 amino acids in length; GenBank: OWK57606.1) is according to the present genome annotation, LOC110484860 LOW QUALITY PROTEIN G-protein coupled receptor 35 of L. striata domestica (Gene ID: 110484860, updated on 6-Nov-2019) [with 362 amino acids in length (2 exons); NCBI Reference Sequence: XP_031361553.1]; Lonchura striata domestica P2X purinoceptor 3 (with 757 amino acids in length; GenBank: OWK57554.1) is not a (P2X 3)-fused monomer GPCR, because this clone was derived from the genomic sequence (Accession and Version: MUZQ01000123.1) and at present has been assigned to the record, P2X purinoceptor 3 of *L. striata* domestica (with 403 amino acids in length; NCBI Reference Sequence: XP_021411674.1)(updated on VRT 05-NOV-2019), that is derived from the genomic sequence [Annotation release: 101; Status: current; Assembly: lonStrDom2 (GCF_005870125.1); Chr: 5; Location: NC_042571.1 (61066629..61070804, complement)]; Nipponia nippon PREDICTED uncharacterized protein LOC104014851 (with 1048 amino acids in length; NCBI Reference Sequence: XP_009465805.1) is (Voltage-dependent potassium channel)-[(Collagen triple helix repeat)(20 copies)]-(GPCR monomer)-(metallophosphatase domain); Limosa lapponica baueri olfactory receptor 6f1-like [with 716 amino acids in length (Its genome annotation is unavailable); GenBank: PKU32848.1] is RVT_1 reverse transcriptase fused monomer GPCR; Haliaeetus leucocephalus PREDICTED sodium-dependent phosphate transport protein 2A (with 1016 amino acids in length; NCBI Reference Sequence: XP_010577535.1) is SLC34A1 (solute carrier family 34 member 1) fused GPCR (TM2~TM7) monomer; Charadrius vociferus PREDICTED uncharacterized protein LOC104286740 (with 784 amino acids in length; NCBI Reference Sequence: XP_009883672.1) that is derived from the genomic sequence (NW_009648402.1 Unplaced Scaffold Reference ASM70802v2) contains about a 3.5 kbase-pairs (bp) sequence that has not been determined (23.1 K~26.6 K), consisting of a region (the intron between the 2nd and 3rd exons of the gene of the clone (NCBI Reference Sequence: XP_009883672.1); Phaethon lepturus (White-tailed tropicbird) PREDICTED suppressor of SWI4 1 homolog, partial (with 736 amino acids in length; NCBI Reference Sequence: XP_010280749.1) that is derived from the genomic sequence (NW_010524399.1 Unplaced Scaffold Reference ASM68728v1) contains about a 0.5 kbase-pairs (bp) sequence that has not been determined (3.7 K~7.9 K), consisting of regions (the introns upstream of the 3rd exon and between the 11 and 13th exons of the gene of the clone (NCBI Reference Sequence: XP_010280749.1).

S3.10. GPCR fusions with <600 amino acids in length and with low similarity to the human $A_{2A}R$ In the middle of August, 2020, during analyzing the clone (XP_033818019.1), one of Amphibia proteins, the below one clone (1) (Table 11) and 3 clones (2~4:TM-linker-lacking GPCR dimers) (Table 12), respectively, (1) *Marmota monax* hypothetical predicted protein (with 613 amino acids in length; GenBank: VTJ88622.1) (Fig. S11Q) and (2) *Neotoma lepida* hypothetical protein A6R68_24156, partial (with 645 amino acids in length;

GenBank: OBS81854.1) (Fig. S12A), (3) *M. monax* hypothetical predicted protein (with 633 amino acids in length; GenBank: VTJ67479.1) (Fig. S12B) and (4) *M. monax* hypothetical predicted protein (with 618 amino acids in length; GenBank: VTJ87989.1) (Fig. S12C), all of which are similar to Geotrypetes seraphini LOW QUALITY uncharacterized protein LOC117368452 (with 648 amino acids in length; NCBI Reference Sequence: XP_033818019.1), i.e., a TM-linker-lacking GPCR dimer, were found to be present at and absent from the file obtained from the above blastp search [of the full length of NP_000666 human $A_{2A}R$ protein against Euarchontoglires proteins, excluding primates (done on December 11, 2019), considering "Lineage"] (S3.6): The presence or absence of identity of 25% or more, with regard to similarity to the query $A_{2A}R$, distinguished likely the hit-one (1) from the other 3 non-hits (2~4). Also, in this time, *M. monax* hypothetical predicted protein (with 613 amino acids in length; GenBank: VTJ88622.1)(1) that had been regarded as a GPCR monomer at the screen [, i.e., by manual review of suspicious clones among GPCR fusions with 600 amino acids or more in length (S2.1)] was found to be a TM-linked GPCR dimer (Table 11, Fig. S11Q), [GPCR1 (TM1~TM7)]-[central intervening region (TMHMM: 293~367)(MEMSAT3: 290~) (Phobius: 293~360)]-[GPCR 2 (TM1~TM7)] where this clone (VTJ88622.1) is similar to the query clone (XP_033818019.1) and to *N. lepida* hypothetical proteins A6R68_17650, partial (with 789 amino acids in leength; GenBank: OBS75898.1)(, i.e., a TM-linker-lacking GPCR trimer) (Fig. S12D) and A6R68_01549, partial (with 516 amino acids in length; GenBank: OBS69909.1) (i.e., a TM-linker-lacking GPCR dimer) (Fig. S12E) (both of which have not been found in the above file) where the central intervening region of this clone (VTJ88622.1) does contain a TM helix segment (TMHMM: 331~353) (MEMSAT3: 328~352)(Phobius: 329~355) which is similar to TM1 (TMHMM: 29~51) of Rattus norvegicus PREDICTED LOW QUALITY olfactory receptor 8D2 (with 309 amino acids in length; NCBI Reference Sequence: XP_008756332.1) and TM1 (TMHMM: 26~48) of *N. lepida* hypothetical protein A6R68_23906, partial (with 339 amino acids in length; GenBank: OBS82101.1) where GPCRs 1 and 2 of this clone (VTJ88622.1) are similar to Marmota marmota marmota PREDICTED olfactory receptor 8D2-like (with 311 amino acids in length; NCBI Reference Sequence: XP_015354588.1) and where GPCR1 (TM1~TM7) is with identity of 60.26%, similar to GPCR2 (TM1~TM7) of this clone (VTJ88622.1). These indicate that in the NCBI database, data of GPCR fusions are maintained other than ones that we have analyzed here so far [Tables 1~3 and S8 and Fig. 1a-c, 3a-b, S6A~S6F, S7A, S7B, S8A (Table 4)~S8G (Table 10)~S8P, S9, S10A~10N and S11A~S11Q]. However, here, we directed, for the first time, our attention to the natural TM sc linker and we set the purpose of this study such that we relate a new computational finding of a natural invertebrate TM sc linker to previous experimental works on sc$A_{2A}R/D_2R$ and limit our analyses to all mammalian and avian proteins, except for GPCR fusions with <600 amino acids in length and with low similarity to the human $A_{2A}R$.

As described above, *Neotoma lepida* hypothetical protein A6R68_17650, partial (with 789 amino acids in leength; GenBank: OBS75898.1) (Fig. S12D) is a GPCR trimer lacking TM-linkers and was identified as one of clones that are similar to the clone (VTJ88622.1)(1) and also has been found to be absent from the above file: This protein consists of [GPCR1 (TM1~TM7)]-[1st central intervening region (TMHMM: 293~308)(MEMSAT3: 290~310)(Phobius: 290~308)]-[GPCR 2 (TM1~TM7)]-[2nd central intervening region (TMHMM: 565~) (MEMSAT3: 563~591)(Phobius: 565~583)]-[GPCR 3 (TM2~TM7)] where the central intervening region of this clone (OBS75898.1) does contain no TM helix segment where GPCRs 1, 2 and 3 of this clone (OBS75898.1) are similar to Mus musculus olfactory receptor 967 (with 310 amino acids in length; NCBI Reference Sequence: NP_001011826.1) and where GPCR1 (TM1~TM7), GPCR2 (TM1~TM7) and GPCR 3 (TM2~TM7) of this clone (OBS75898.1) are similar each to each with identity (GPCR 1 vs. GPCR 2: 87.73%; GPCR 1 vs. GPCR 3: 80.69%; GPCR 2 vs. GPCR 3: 78.17%). This clone (OBS75898.1) is, with identity of 22.89%, similar to the human $A_{2A}R$ (NCBI Reference Sequence: NP_000666.2), resulting in non-hit at the blastp search [of the full length of NP_000666 human $A_{2A}R$ protein against Euarchontoglires proteins, excluding primates (done on December 11, 2019)] and thus is absent from the file. Other 4 fusions, all of which are TM-linker-lacking GPCR dimers are as follows: *N. lepida* hypothetical protein A6R68_01549, partial (with 516 amino acids in length; GenBank: OBS69909.1) (Fig. S12E) consisting of [GPCR1 (TM3~TM7)]-[central intervening region (TMHMM: ~238)(MEMSAT3: 209~237)(Phobius: ~238)]-[GPCR 2 (TM1~TM7)] was also identified as one of clones that are similar to the clone (VTJ88622.1)(1) and has not been found in the above file. In addition, *N. lepida* hypothetical protein A6R68_07625, partial (with 626 amino acids in length; GenBank: OBS63836.1) (Fig. S12F) consisting of [GPCR1 (TM1~TM7)]-[central intervening region (TMHMM: 295~350)(MEMSAT3: 291~350)(Phobius: ~346)]-[GPCR 2 (TM1~TM7)] was identified as one of clones that are similar to the clone (VTJ67479.1)(3) and also has not been found in the above file; *N. lepida* hypothetical protein A6R68_03066, partial (with 631 amino acids in length; GenBank: OBS68388.1) (Fig. S12G) and *M. monax* hypothetical predicted protein (with 627 amino acids in length; GenBank: VTJ86726.1) (Fig. S12H), respectively, [GPCR1 (TM1~TM7)]-[central

intervening region (TMHMM: ~349)(MEMSAT3: 288~347) (Phobius: ~348)]-[GPCR 2 (TM1~TM7)] and [GPCR1 (TM1~TM7)]-[central intervening region (TMHMM: 290~328)(MEMSAT3: 287~328)(Phobius: 290~328)]-[GPCR 2 (TM1~TM7)], were identified as ones of clones that are similar to an Amphibia protein [Microcaecilia unicolor uncharacterized protein LOC115457245 (with 648 amino acids in length; NCBI Reference Sequence: XP_030042527.1)] and also have not been found in the above file.

The human brain $A_{2A}R$ phenotype: A multiple alignment using ClustalW 1.8 of the human $A_{2A}R$ and D2LR together with other subtypes was done[60]. Here, the extracellular N-terminal and cytoplasmic C-terminal portion encompassing parts of the Transmembrane (TM) domains 1 and 7, respectively, of the human, mouse and rat $A_{2A}R$ (412, 410 and 410 amino acids in length, respectively; NCBI Reference Sequence: NP_000666, NP_033760 and NP_445746) were aligned manually, showing the amino acid sequence similarity of these proteins. Amino acids in red denote those being identical and conserved among these proteins. A missense (Gly392Arg) and a synonymous mutation (Tyr361) of the human $A_{2A}R$ are highlighted in yellow and gray, respectively. An arrow indicates the site where the human $A_{2A}R$ C-terminus was reportedly truncated for X-ray crystallography[61]. Missense SNPs are shown together in one line or two, meaning SNPs with a single mutation as in either Glu294Asp or Phe295Leu but no single variant clone with all the mutations (dbSNP rs#17650937 with a missense mutation of Phe295Leu by G at codon position 3 was reported which at least for once deleted, this information was accessed on January 29, 2016; This sequence of event was not applied to this figure). A similarity of the amino acid sequences between the two parts highlighted in gray, (Lys315/Arg310/Gln310) and [(Lys391-Gly392) or (Lys391-Arg392)]/(His380-Gln381)/(Arg380- Gln381)], respectively, human, mouse and rat $A_{2A}R$, suggests a relevant brain-type Arg392 of the human $A_{2A}R$ (gb|S46950) but which is not necessarily derived from errors in the sequencing library preparation; errors in the sequencing process were repeatedly denied. In fact, this relevancy is also supported by that the matched genome and RNA sequencing data so far were obtained from samples of mostly lymphoblastoid cell lines which were used in the 1000 Genomes Project but not from brain tissues[62-64]. Indeed, the missense (Gly392Arg) mutation of the human $A_{2A}R$ (dbSNP rs#1277013918), highlighted in yellow in Fig. S1 of our report (13) has been reported at NCBI recently.

The human $A_{2A}R$ [its structure (PDB ID: 3EML) determined by T4-lysozyme fusion Strategy (S2) where most of the 3rd intracellular loop (Leu20 95.70~Ala221 6.23: The Ballesteros-Weinstein numbering scheme is shown in superscript)[65] was replaced with lysozyme from T4 bacteriophage (not shown here) and the C-ter tail (Ala317~Ser412) was deleted to improve the likelihood of crystallization is shown with a bundle width diameter of about 36 Å] (cysteines as balls in purple) which was drawn by Chimera molecular graphics (University of California San Francisco) (Fig. S1: lower left, a view from extracellular upper side perpendicular to the cell surface membrane; lower right, a sideview parallel to the cell surface membrane) can be connected with the central intervening portion (amino acid residues 311~416) (not shown here) of *Exaiptasia pallida* uncharacterized protein LOC110241027 including a single probable TM domain. Lines for 36 Å are shown in black. Additionally, for comparison, the bacteriorhodopsin lipid complex (PDB ID: 1C3W;[66]) (Fig. S1: lower left, a view from extracellular upper side perpendicular to cell surface membrane) and the human $A_{2A}R$ complexed with lipids and a chemical ligand and the electrostatic surface coloring of the human $A_{2A}R$ which was obtained by the Chimera (Fig. S1: lower right, a sideview parallel to the cell surface membrane) and the bacteriophage f29 tail (PDB ID: 3fb4; Three molecules are shown, respectively, in distinct blue, out of hexamer;[67]) are shown.

Supplementary (Fig. 2). The amino acid sequence similarity of two identical GPCRs of *Orbicella faveolata* uncharacterized LOC110041892 protein (with 824 amino acids in length; NCBI Reference Sequence: XP_020602879.1), human DRD1 (NP_000785.1) and putative Aplysia californica dop1 (NP_001191631.1) D1 dopamine receptor (D1R) (A) and the genomic structures of the first two (B). A, A multiple alignment using ClustalO 1.2.4 shows the sequence similarity of these proteins with an overall length homology approximately 35% as a rule of thumb for protein sequences[68] (aligned score of *O. faveolata* GPCR 1/2 vs. human DRD1 = 36%; *O. faveolata* GPCR 1/2 vs. putative A. californica dop1 = 34%; human DRD1 vs. putative A. californica dop1 = 35%). A tripeptide motif (DRY in one letter code), together with Transmembrane (TM) domains (underlined) is shown in yellow, confirmed by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/ services/TMHMM/). The *Orbicella faveolata* LOC110041892 protein was found to have two identical class A GPCRs in tandem [named here as GPCR 1 (amino acids 24~317) and GPCR 2 (amino acid residues 461~754)] (numbering from XP_020602879.1) across the central intervening portion (amino acids 318~460) (which is encoded on exon 5 through exon 6 to exon 7: B), containing the N-ter sequence corresponding to 70 amino acid residues attached to the C-ter end of GPCR 2,55-CRLLQRTFGSSCSRSNNNSNVDESQYSMNI AVTNASPKGKLTKNGTAKKTKFHEVYGEEEAGAA EEPKPV-824 (middle: shown as light gray boxes; lower: shown as dotted underlined sequences) and the C-ter

sequence corresponding to 23 amino acid residues ascending the N-ter end of GPCR 1, 1-MKMSNSTNQTQDRCQQEIQRDFI-23 (middle: shown as dark gray boxes; lower: shown as underlined sequences). The asterisk indicates an identical amino acid among the three proteins. ":" means that conserved substitutions have been observed. "." means that semiconserved substitutions are observed. B, *O. faveolata* LOC110041892 and human DRD1 have 7 (only exons 5~7 are shown in the middle column as white boxes) and 2 exons, respectively [5'- and 3'-untranslated regions are shown as light green boxes and a coding sequence (open reading frame) is shown as a dark green box: from NCBI, USA]; for example, exons 1~4 (with a 5'-untranslated region) of *O. faveolata* LOC110041892 do not necessarily agree with exon 1 (with a 5'-untranslated region) of human DRD1, suggesting the widespread loss of introns during the evolution of the mammalian class A GPCR family[17] while exons 5 and 7 of *O. faveolata* LOC110041892 and exon 2 of human DRD1 encode the respective GPCR protein. Thin lines (in black) between the respective exons of *O. faveolata* LOC110041892 indicate introns, not drawn to scale. The corresponding amino acid residues (in one letter code) just below the schematic illustration of the genomic organization and the below boxes drawn to scale, for example numbered 1 or 2 or other-mentioned, represent protein products (middle: shown as dark or light gray boxes; lower: shown as underlined or dotted underlined sequences) of the genes. The thin line (in green) indicates an intron of *O. faveolata* LOC110041892 and human DRD1 with the scale bar in base pairs above the schematic illustration of the genomic organization.

Supplementary Fig. 3 the amino acid sequence similarity of two GPCRs of *Exaiptasia pallida* uncharacterized LOC110241027 protein (761 amino acids in length; NCBI Reference Sequence: XP_020902518.1), human ADRB2 (NP_000015.1) and a fish-like marine chordate Branchiostoma floridae hypothetical protein BRAFLDRAFT_85478 (XP_002608947.1) b2 adrenergic receptor (b2AR) (vs. GPCR 1) or human HTR1B (NP_000854.1) and Caenorhabditis elegans tyra-2 tyramine receptor (NP_001024521.1) 5-hydroxytryptamine receptor 1B (5-HT1B) (vs. GPCR 2) (A) and the genomic structures of the three homologenes (B). A, A multiple alignment using Clustal O 1.2.4 shows the sequence similarity of these proteins with an overall length homology approximately 30% as a rule of thumb for protein sequences (S10) (aligned score of *E. pallida* GPCR 1 vs. human ADRB2 = 31%; *E. pallida* GPCR 1 vs. hypothetical B. floridae BRAFLDRAFT_85478 = 26%; human ADRB2 vs. hypothetical B. floridae BRAFLDRAFT_85478 = 40%; *E. pallida* GPCR 2 vs. human HTR1B = 28%; *E. pallida* GPCR 2 vs. *C. elegans* tyra-2 = 24%; human HTR1B vs. C. elegans tyra-2 = 37%). A tripeptide motif (DRY in one letter code), together with transmembrane (TM) domains (underlined), is shown in yellow, confirmed by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/). The *Exaiptasia pallida* LOC110241027 protein has two class A GPCRs in tandem, named here as GPCR 1 (amino acids 21~310) and GPCR 2 (amino acids 417~727) (numbering from XP_020902518.1), across the central intervening portion (amino acids 311~416) which is encoded on exon 4 and exon 5 (B), consisting of the 106 amino acid residue sequence. The asterisk indicates an identical amino acid among the three proteins. ":" means that conserved substitutions have been observed. "." means that semi conserved substitutions are observed. B, *E. pallida* LOC110241027, human ADRB2 and 5-HTR1B have 5 (only exons 2~5 are shown in the middle column as white boxes) and single exons [5'- and 3'-untranslated regions are shown as light green boxes and a coding sequence (open reading frame) is shown as a dark green box: from NCBI, USA], respectively; for example, exons 1 and 2 (with 5'-untranslated region) of E. pallida LOC110241027 do not necessarily agree with exon 1 (with 5'-untranslated region) of human ADRB2 and 5-HTR1B, suggesting the widespread loss of introns during the evolution of the mammalian class A GPCR family[17] while exons 2/3 and 4/5 of *E. pallida* LOC110241027 and the single exons of human ADRB2 and 5-HTR1B encode the respective GPCR proteins. Thin lines (in black) between the respective exons of *E. pallida* LOC110241027 indicate introns, not drawn to scale; The corresponding amino acid residues (in one letter code) just below the schematic illustration of the genomic organization and the below boxes drawn to scale, for example numbered 1 or 2 or other-mentioned, represent protein products (middle: shown as dark or light gray boxes; lower: shown as underlined or dotted underlined sequences) of the genes. The sequence corresponding to 20 amino acid residues (amino acids 1~20) ascending the N-ter end of GPCR 1 and the sequence (middle: shown as dark and light gray boxes, respectively) corresponding to 34 amino acid residues (amino acid residues 728~761) attached to the C-ter end of GPCR 2 are also shown in gray. The thin line (in green) indicates an intron of *E. pallida* LOC110241027, human ADRB2 and 5-HTR1B with the scale bar in base pairs above the schematic illustration of the genomic organization.

Supplementary Fig. 4, the amino acid sequence similarity of three GPCRs of *Stylophora pistillata* uncharacterized LOC111342710 protein (with 1018 amino acids in length; NCBI Reference Sequence: XP_022805553.1), human ADRA1A a1a adrenergic receptor (a1aAR) (with 466 amino acids in length; GenBank: AAK77197.1; NCBI Reference Sequence: NP_000671.2), Caenorhabditis elegans ortholog DOP-4 dopamine receptor 4 (with 517 amino acids in length;

NCBI Reference Sequence: NP_508238.2) (vs. GPCR 1), human ADORA2A $A_{2A}R$ (with 412 amino acids in length; NCBI Reference Sequence: NP_000666.2), Drosophila melanogaster AdoR adenosine receptor isoform A (with 774 amino acids in length; NCBI Reference Sequence: NP_651772.1) (vs. GPCR 2) or human HTR4 5-hydroxytryptamine (serotonin) receptor 4 (5-HT4) isoform i (with 428 amino acids in length; NCBI Reference Sequence: NP_001035263.1) and Aplysia californica 5-HT4 (with 384 amino acids in length; NCBI Reference Sequence: NP_001240691.1) (vs. GPCR 3)(A) and the genomic structures of the four homologenes (B). A, A multiple alignment using ClustalO 1.2.4 shows the sequence similarity of these proteins, with an overall length homology approximately 30% as a rule of thumb for protein sequences (S10). Aligned score of S. pistillata GPCR 1 vs. human ADRA1A = 30%; S. pistillata GPCR 1 vs. C. elegans DOP-4 = 30%; human ADRA1A vs. C. elegans DOP-4 = 38%; S. pistillata GPCR 2 (DTM1/DTM7) vs. human ADORA2A = 29%; S. pistillata GPCR 2 (DTM1/DTM7) vs. D. melanogaster AdoR = 24%; human ADORA2A vs. D. melanogaster AdoR = 33%; S. pistillata GPCR 3 vs. human HTR4 = 31%; S. pistillata GPCR 3 vs. A. californica 5-HT4 = 28%; human HTR4 vs. A. californica 5-HT4 = 41%. A tripeptide motif (DRY in one letter code), together with transmembrane (TM) domains (underlined), is shown in yellow, confirmed by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/). The Stylophora pistillata LOC111342710 protein has three class A GPCRs in tandem, named here as GPCR 1 (amino acid residues 34~318), GPCR 2 (DTM1/DTM7) (amino acid residues 369~626) and GPCR 3 (amino acid residues 674~963) (numbering from XP_022805553.1) across the central intervening portions, consisting of 70 and 47 amino acid residue sequences [amino acid residues 319~368 and 627~673 are encoded on exons 1/2 (mainly on exon 1) and exons 2/3, respectively: B]. *C. elegans* DOP-4 shows high sequence homology with D_1-like dopamine receptors unique to invertebrates which are distinct from mammalian D1-like dopamine receptors and relatively high sequence identity to invertebrate octopamine receptors and human α1aAR[69]. The human 5-hydroxytryptamine (serotonin) receptor 4 isoform i protein which is larger than other isoforms b (with 388 amino acids in length; NCBI Reference Sequence: NP_000861.1), a (with 387 amino acids in length; NP_001035259.1), d (with 360 amino acids in length; NP_001035262.2), c (with 411 amino acids in length; NP_001273339.1) and g (with 378 amino acids in length; NP_955525.1)[70] was used here. Hydropathic analyses have been developed in the process of predicting antigenic determinants which is important for antibody preparation in a protein by analyzing its hydropathicity, flexibility or secondary structure [77, 72]. The asterisk indicates an identical amino acid among the three proteins. ":" means that conserved substitutions have been observed. "." means that semiconserved substitutions are observed. B, S. pistillata LOC111342710, human ADRA1A, human ADORA2A and human HTR4 have 3 (all three exons are shown in the middle column as white boxes) and 11, 1 and 12 exons [5'- and 3'-untranslated regions are shown as light green boxes and a coding sequence (open reading frame) is shown as a dark green box: from NCBI, USA], respectively. Other 11 human GPCRs for 5-hydroxytryptamine (serotonin), i.e., subtypes 1A (HTR1A: a single exon), 1B (HTR1B: a single exon), 1D (HTR1D: a single exon), 1E (HTR1E: 2 exons), 1F (HTR1F: 6 exons), 2A (HTR2A : 4 exons), 2B (HTR2B: 4 exons), 2C (HTR2C: 7 exons), 5A (HTR5A: 2 exons), 6 (HTR6: 3 exons) and 7 (HTR7: 5 exons); HTR3A 5-hydroxytryptamine (serotonin) receptor 3A: ionotropic[73]. Thin lines (in black) between the respective exons of *S. pistillata* LOC111342710 indicate introns, not drawn to scale. The corresponding amino acid residues (in one letter code) just below the schematic illustration of the genomic organization and the below boxes drawn to scale, for example numbered 1 or 2 or other-mentioned, represent protein products (middle: shown as dark or light gray boxes; lower: shown as underlined or dotted underlined sequences) of the genes. The sequence corresponding to 33-amino acid residues (amino acid residues 1~33) ascending the N-ter end of GPCR 1 and the sequence (middle: shown as dark and light gray boxes, respectively) corresponding to 55-amino acid residues (amino acid residues 964~1018) attached to the C-ter end of GPCR 3 are also shown in gray. Thin line (in green) indicates an intron of S. pistillata LOC111342710 and human ADRA1A, human ADORA2A and human HTR4 with the scale bar in base pairs above the schematic illustration of the genomic organization.

Supplementary Fig. 5 A, Helical-wheel representations of the TM sequences from *Caenorhabditis elegans odr4*, human CD23 and the central intervening portion (amino acid residues 311~416) [The TM sequence (amino acid residues 326~346) is numbered 1~21 here] of *Exaiptasia pallida* uncharacterized protein LOC110241027 are shown. Consensus sequence is an amino acid sequence of proteins which underlies inter-or intramolecular interactions including motifs necessary for posttranslational modification such as phosphorylation, ubiquitination, acetylation, methylation and N-linked glycosylation (the N-X-S/T motif) by enzymes[77-76] (http://www.uniprot.org/docs/ptmlist). With about 40 to hundreds posttranslational modifications through molecular interactions as molecular switch, cells regulate their biological functions which depend on protein functions including enzymatic activity, protein turnover, conformation, localization and posttranslational modification crosstalk to each other. Consensus sequence

is also found in the Rossmann fold[77] (the GXGXXG - motif) (a super-secondary structure called a bab fold) or nucleotide binding fold at the binding site of the nucleotide, the leucine zipper (every seventh amino acid residue is a leucine) (dimerizing a-helices in DNA-binding proteins). With regard to a common feature of the TM domain helices involved in protein interactions, the "noncovalent forces that stabilize protein structures are not fully understood" Focusing on "these interactions experimentally using designed peptides", Baker *et al*.[78] found "that both terminal backbone-side chain and certain side chain-side chain interactions (which include both local effects between proximal charges and interatomic contacts) contribute much more to helix stability than side chain-helix macrodipole electrostatics which are believed to operate at larger distances. "[78] A BLAST search of the gene databases {blastp against all deposited organisms [nonredundant protein sequences (nr)]} with the 94 Designed Heterodimers (DHDs) ["well-expressed in E. coli with both monomers co-purifying by Ni-affinity chromatography (only one monomer contains a hexahistidine tag; Supplementary Table 3)"][79]: p. 108, The left colun-line 14; Supplementary Table 2), i.e., different 188 polypeptide chains [DHD9_a (### 1) ~ DHD147_b (### 188)](The proteins are here referred to by serial numbers (###)}[S25], out of 97 DHDs that had been selected among de novo designs of millions of four-helix backbones with varying degrees of supercoiling around a central axis, did not reveal similarity to the amino acid sequences of the previously characterized *Caenorhabditis elegans odr4* TM-and human CD23TM-sc linkers (https://osf.io/h 7ewu/?view_only = e18fc8ebe 66443d895b098280 3758eb3) and also did not to those of single consensus-containing TM-sc linkers in the central intervening region of the invertebrate *Exaiptasia pallida* LOC110241027 and the vertebrate *Opisthocomus hoazin* LOC104327099 proteins. Known TM domain–TM domain interfaces from bitopic proteins often contain common interaction motifs such as GxxxG, polar amino acids including Asn, Gln, Asp, His, or Trp, or more complex patterns such as Ser/Thr- clusters and QxxS-motifs[80, 81]. In the helical-wheel representations of the TM sequences from Caenorhabditis elegans odr4, human CD23 and the central intervening portion (amino acid residues 311~416) [The TM sequence (amino acid residues 326~346) is numbered 1~21 here] of *Exaiptasia pallida* uncharacterized protein LOC110241027, Thr and Gly-Leu residues 5 and 11-12 (odr4), 10 and 16-17 (hCD23) and 8 and 14-15 (LOC110241027 central intervening portion) in the square boxes in TMs are parts of the consensus sequence, T(I/A/P)(A/S)(L/N)(I/W/L)(I/A/V)GL(L/G)(A/T)(S/L/G)(I/L). The TM interface which contains the consensus sequence, is also shown by being split up with a dotted line. Asterisks mark nonconservative amino acid differences among odr4, hCD23 and LOC110241027 central intervening portion. The putative GXXXG motif commonly found in transmembrane helices known to dimerize is absent in these three sequences and human $A_{2A}R$ TM4 (121-akgiiaicwv lsfaigltpm lgw-143) and TM5 domains (174-mnymvyf 181-nffacvlvpl llmlgvyl-198). B, Between C. elegans ODR-4 [with 475 amino acids in length (11 exons); NCBI Reference Sequence: NP_001022814.1] and the human orthologous protein [with 454 amino acids in length (15 exons); NCBI Reference Sequence: NP_060317.3], homology is across the entire regions, including the TM helix segments, based on its sequence alignment[13,23]. Also, not only an alignment between these two ODR4 TMs alone, i.e., C. elegans odr4 TM (452-TILITIALIIGLLASIIYFTVA-473) and human ODR4 TM [amino acid residues 432~452: 432 IgViaAftVAVLAagIsFhyf 452 (Regions representing identity at that position or conserved amino acids to *C. elegans* ODR4 TM are indicated in capital letters)] (B, upper) but also another one among C. elegans/human ODR-4 TMs, human/mouse/rat CD23 TMs and TM-linkers in the central intervening portions of the *Exaiptasia pallida* LOC110241027 and *Opisthocomus hoazin* LOC104327099 (XP_009930279.1) proteins indicated different alignments but with all the same minimal consensus Gly-(Leu/Val) residues (B, lower). Accordingly, in human CD23 TM (and also in mouse and rat CD23 TMs), two minimal consensus Gly-(Leu/Val) residues [or (Thr/Ser)-(5 amino acids)-Gly-(Leu/Val)] are included. Also, Ser and (–)-Leu residues 13 and (–)-20 are kept as parts of this consensus sequence in the *O. hoazin* protein LOC104327099 TM-linker (Fig. 3a, Table 1).

Supplementary (Fig. 6) a The clone, Scleropages formosus protein alpha-1B adrenergic receptor-like [with 782 amino acids in length (Its genome annotation was not found); GenBank: KPP73650.1](a freshwater fish, Asian arowana), consisting of the amino-terminal (N-ter) portion (i.e., the GPCR 1: amino acid residues 51~359), the central intervening portion (amino acid residues 360~498) and the Carboxy-terminal (Cter) portion (, i.e., the GPCR 2: amino acid residues 499~762), lacks likely natural TM-linker. This clone (KPP73650.1) differs from LOC108934154, Scleropages formosus protein [PREDICTED: alpha-1B adrenergic receptor- like; with 430 amino acids in length; NCBI Reference Sequence: XP_018607202.1)(It encodes a class A GPCR monomer)[A coding sequence (open reading frame) is on exons 1~6]. Analysis by the TM-HMM 2.0 algorithm (http://www.cbs.dtu.dk/services/TMHMM/) and an Artificial Neural Network program (MEMSAT3: http://bioinf.cs.ucl.ac.uk/?id=756)[20] indicated that the central intervening portion (amino acid residues 360~498) of this clone (KPP73650.1) does not contain any TM helix domain segment. B: The clones, *Pocillopora*

damicornis hypothetical protein pdam_00010707 [with 1473 amino acids in length (Its genome annotation was not found); GenBank: RMX53929.1] (the cauliflower coral) (B) and Strongylocentrotus purpuratus protein octopamine receptor beta-2R-like [with 703 amino acids in length (Its genome annotation was updated on 8-Apr-2017; Gene symbol: LOC582575; 2 exons); NCBI Reference Sequence: XP_011663779.1] (purple sea urchin), lack likely natural TM-linker. Analysis by the TM-HMM 2.0 and MEMSAT3 indicated that the central intervening portion (RMX53929.1: amino acid residues 360~498; XP_011663779.1: amino acid residues 247~367) of these clones do not contain any TM helix domain segment. C and D: The clones, *Orbicella faveolata* protein octopamine receptor beta-1R-like [with 717 amino acids in length (Its genome annotation was updated on 8 Sep 2017; Gene symbol: LOC110042595; 2 exons); NCBI Reference Sequence: XP_020603616.1] (commonly known as mount ainous star coral) and Cricetulus griseus protein trace amine associated receptor 8a like [with 754 amino acids in length (Its genome annotation was not open); GenBank: ERE85073.1] (Chinese hamster), lack likely natural TM linker. Analysis by the TM-HMM 2.0 and MEMSAT3 indicated that the central intervening portion (XP_011663779.1: amino acid residues 322~422; ERE85073.1: amino acid residues 344~457) of these clones do not contain any TM helix domain segment. E and F: The clones, *Orbicella faveolata* protein uncharacterized LOC110066440 [with 619 amino acids in length (Its genome annotation was updated on 8-Sep-2017; Gene symbol: LOC110066440; 3 exons); NCBI Reference Sequence: XP_020629325.1] (commonly known as mountainous star coral) and Nematostella vectensis protein predicted [with 690 amino acids in length (Its genome annotation was updated on 24-Sep-2016; Gene symbol: NEMVEDRAFT_v1g208740; 2 exons); NCBI Reference Sequence: XP_001631773.1] (starlet sea anemone), lack likely natural TM-linker. Analysis by the TM-HMM 2.0 and MEMSAT3 indicated that the central intervening portion (XP_020629325.1: amino acid residues 286~352; XP_001631773.1: amino acid residues 300~357) of these clones do not contain any TM helix domain segment. The clone, Scleropages formosus protein trace amine-associated receptor 1-like, partial [with 696 amino acids in length (Its genome annotation was not open); GenBank: KPP58082.1] (Asian arowana) lacks likely natural TM-linker. Analysis by the TM-HMM 2.0 and MEMSAT3 and an alignment with human trace amine-associated receptor 1 (NP_612200.1) {that is one with the most significant E-value of 0.3×10^{-82} [identity = 54% (129/238, identical amino acids out of the selected sequence); similarity = 70% (168/238, conserved amino acids out of the selected sequence)], obtained by blastp search using this clone (KPP58082.1) against human genes, indicated that the central intervening portion (amino acid residues 235~255)(determined by NCBI annotation) of this clone (KPP58082.1) does not contain an additional TM helix domain segment as a linker.

Supplementary (Fig. 7) in order to more identify natural TM linkers in GPCR fusions, we then narrowed "Lineage" such as order/family/genus/species in the class of Mammalia in web-based blastp search by the organism limit (S2.1; S3.5). Two primates GPCR fusions (consisting of a GPCR dimer) exist at the NCBI BLAST search whereas human genome has no GPCR fusion. One clone, Plecturocebus moloch (red bellied titi) hypothetical protein (with 622 amino acids in length; GenBank: ACA53481.1) (Its genome annotation was not found), contains likely no natural TM linker (A). The other clone, Callithrix jacchus (white tufted ear marmoset) hypothetical protein (with 654 amino acids in length; GenBank: ABZ80295.1) shows no natural TM linker existence between the putative dual GPCR (B).

Supplementary (Fig. 8) in two superordinal clades Euarchontoglires (excluding primates) and Laurasiatheria, at the NCBI BLAST search, respectively, 9 GPCR fusions, only of either *Neotoma lepida* (desert woodrat) [including A6R68_19462 (GenBank: OBS78147.1) (Fig. 3b, Table 1)] (total 4 dimers/one trimer/one tetramer) or *Marmota monax* (woodchuck) hypothetical proteins (total one dimer/one trimer/one pentamer) and 8 fusions, either dimer (total 5 dimers), trimer (total one) or tetramer (total two), exist: *Neotoma lepida* (OBS80343.1) (A; Table 4), (OBS75582.1) (B; Table 5), (OBS82940.1) (C; Table 6), (OBS74663.1) (D; Table 7), (OBS78080.1)(E; Table 8); *Marmota monax* (VTJ79832.1) (F; Table S6), (VTJ82433.1) (G; Table 10), (VTJ70100.1) (H); Camelus bactrianus (XP_010960385.1) (I); Bubalus bubalis (XP_006073590.2) (J); Sorex araneus (ACE79115.1) (K), (XP_004613594.2) (L), (ACE79123.1) (M); Bos mutus (MXQ95485.1)(N), (MXQ95489.1) (O), (MXQ88753.1) (P). Whereas *N. lepida* hypothetical proteins [A6R68_23065 (with 685 amino acids in length; GenBank: OBS82940.1) and A6R68_14817 (with 1135 amino acids in length; GenBank: OBS74663.1)] and *M. monax* hypothetical predicted protein (with 850 amino acids in length; GenBank: VTJ82433.1) and all 8 fusions of Laurasiatheria lack likely natural TM-linker, other *N. lepida*/*M. monax* scGPCRs, i.e., the GPCR fusions (each consisting of either a GPCR dimer, trimer, tetramer or pentamer), do not contain the consensus sequence in their central intervening portions (Table 2).

Supplementary (Fig. 9) in the class of Mammalia, excluding two superordinal clades Euarchontoglires and Laurasiatheria, only one fusion exists at the NCBI BLAST search: Ornithorhynchus anatinus (platypus) uncharacterized protein LOC100087811 [LOW QUALITY PROTEIN: with 928 amino acids in length (6 exons); NCBI Reference Sequence:

XP_028909644.1]. This scGPCR consisting of a GPCR trimer lacks natural TM-linker in its central intervening portions.

Supplementary (Fig. 10), in the class of Aves, out of total 15 GPCR fusions obtained at the NCBI BLAST search (Table 3), three TM-linked GPCR dimers exist: Only *Opisthocomus hoazin* protein LOC104327099 (XP_009930279.1) does contain the consensus sequence in its central intervening portion but other 2 GPCR dimers of Aves do not: Gavia stellata (XP_009819412.1) (E); RLW13346.1 (L). The remaining 12 GPCR dimers lack TM-linkers: *Callipepla squamata* (OXB57691.1) (A), (OXB61274.1)(B); *Colinus virginianus* (OXB75603.1) (C); *Bambusicola thoracicus* (POI30912.1) (D); *Merops nubicus* (XP_008947395.1) (F); Corvus cornix (XP_019149630.2) (G); *Melospiza melodia* maxima (KAF2980960.1) (H); *Patagioenas fasciata* monilis (OPJ89483.1) (I); *Limosa lapponica* baueri (PKU37004.1) (J), (PKU42245.1) (K); Hirundo rustica rustica (RMC04025.1) (M); *Zosterops borbonicus* (TRZ15306.1) (N).

Supplementary (Fig. 11) in a superordinal clade Euarchontoglires (excluding primates) at the NCBI BLAST search while 8 natural GPCR fusions with 700 or more amino acids in length, plus A6R68_23065 (OBS82940.1) (with 685 amino acids in length) are shown in Table 2, i.e., only of either *Neotoma lepida* (desert woodrat)[, including A6R68_19462 (GenBank: OBS78147.1) (Fig. 3b, Table 1)] (total 4 dimers/one trimer/one tetramer) or *Marmota monax* (woodchuck) hypothetical proteins (total one dimer/one trimer/one pentamer), 17 additional GPCR fusions with 600~699 amino acids in length exist as shown here, either 14 dimers or 3 trimers (among which *M. monax* has two dimers and two trimers and *N. lepida* has all others): *Neotoma lepida* (OBS57425.1) (A), (OBS69813.1) (B), (OBS75694.1) (C), (OBS80342.1) (D), (OBS83474.1) (E), (OBS83645.1)(F), (OBS71544.1) (G), (OBS69033.1) (H), (OBS59748.1) (I), (OBS59291.1) (J), (OBS70632.1) (K), (OBS77767.1) (L), (OBS57429.1) (M) *Marmota monax* (VTJ83394.1) (N), (VTJ85523.1) (O), (VTJ52678.1) (P), (VTJ88622.1) (Q).

Supplementary (Fig. 12) during analyzing the clone (XP_033818019.1), one of Amphibia proteins, *Marmota monax* hypothetical predicted protein (with 613 amino acids in length; GenBank: VTJ88622.1) (Table S8, Fig. S11Q) and the 3 clones, respectively, were found to be present and absent in the file obtained from the blastp search [of the full length of NP_000666 human $A_{2A}R$ protein against Euarchontoglires proteins, excluding primates (done on December 11, 2019), considering "Lineage"] (S3.6): The presence or absence of identity of 25% or more with regard to similarity to the query $A_{2A}R$, distinguished likely the hit-one from the other 3 non-hits. The latter 3 clones, i.e., GPCR fusions with <600 amino acids in length and with low similarity to the human $A_{2A}R$ are TM-linker-lacking GPCR dimers as follows: *Neotoma lepida* hypothetical protein A6R68_24156, partial (with 645 amino acids in length; GenBank: OBS81854.1) (A) and *M. monax* hypothetical predicted proteins {[with 633 amino acids in length; GenBank: VTJ67479.1 (B)] and [with 618 amino acids in length; GenBank: VTJ87989.1 (C)]}. The clone (VTJ88622.1) is similar to the query clone (XP_033818019.1) and to *N. lepida* hypothetical proteins A6R68_17650, partial (with 789 amino acids in leength; GenBank: OBS75898.1) (i.e., a TM-linker-lacking GPCR trimer) (D) and A6R68_01549, partial (with 516 amino acids in length; GenBank: OBS69909.1) (i.e., a TM-linker-lacking GPCR dimer) (E) (both of which have not been found in the above file) where the central intervening region of this clone (VTJ88622.1) does contain a TM helix segment (TMHMM: 331~353)(MEMSAT3: 328~352) (Phobius: 329~355). Other 3 fusions, all of which are TM-linker-lacking GPCR dimers are as follows: *N. lepida* (OBS63836.1) (F), (OBS68388.1) (G); *M. monax* (VTJ86726.1) (H).

CONCLUSION

The design of transmembrane (TM)-linked sc$A_{2A}R$/D2LR 'exclusive' monomers and dimers using the newly identified natural TM linkers that accompany long amino acid sequences possibly allows us to express class A GPCRs by receptor protein assembly regulation, i.e., selective monomer/nonobligate dimer formation and is experimentally testable and will be used to confirm in vivo that a low S/N ratio interaction between $A_{2A}R$ and D2LR functions in dopamine neurotransmission in the striatum.

ACKNOWLEDGMENTS

We thank Drs. O. Saitoh (Nagahama Institute of Bio-Science and Technology, Nagahama) and K. Yoshioka (Kanazawa University, Kanazawa) for their respective contributions [12], Mr. M. Woolfenden for proofreading the English of a part of this manuscript and Dr. K. Fuxe (Karolinska Institutet, Stockholm) and Dr. H. Saya (Keio University, Tokyo) for encouragement. We appreciate all the services with that Keio University Medical Library provided us.

REFERENCES

01. Spano, P.F., M. Trabucchi and G. Di Chiara, 1997. Localization of nigral dopamine-sensitive adenylate cyclase on neurons originating from the corpus striatum. Science, 196: 1343-1345.

02. Kreitzer, A.C. and R.C. Malenka, 2008. Striatal plasticity and basal ganglia circuit function. Neuron, 60: 543-554.
03. Donahue, C.H. and A.C. Kreitzer, 2015. A direct path to action initiation. Neuron, 88: 240-241.
04. Glaser, T., A.R. Cappellari, M.M. Pillat, I.C. Iser, M.R. Wink, A.M.O. Battastini and H. Ulrich, 2012. Perspectives of purinergic signaling in stem cell differentiation and tissue regeneration. Purinergic Signalling, 8: 523-537.
05. Ralevic, V. and G. Burnstock, 1998. Receptors for purines and pyrimidines. Pharmacol. Rev., 50: 413-492.
06. Wall, N.R., M. De La Parra, E.M. Callaway and A.C. Kreitzer, 2013. Differential innervation of direct-and indirect-pathway striatal projection neurons. Neuron, 79: 347-360.
07. Liu, C., L. Kershberg, J. Wang, S. Schneeberger and P.S. Kaeser, 2018. Dopamine secretion is mediated by sparse active zone-like release sites. Cell, 172: 706-718.
08. Fuxe, K., A. Cintra, L.F. Agnati, A. Harfstrand and M. Goldstein, 1988. Studies on the relationship of tyrosine hydroxylase, dopamine and cyclic AMP-regulated phosphoprotein-32 immunoreactive neuronal structures and D1 receptor antagonist binding sites in various brain regions of the male rat-mismatches indicate a role of D1 receptors in volume transmission. Neurochem. Int., 13: 179-197.
09. Navarro, G., D.O. Borroto-Escuela, K. Fuxe and R. Franco, 2016. Purinergic signaling in Parkinson's disease. Relevance for treatment. Neuropharmacology, 104: 161-168.
10. Missale, C., S.R. Nash, S.W. Robinson, M. Jaber and M.G. Caron, 1998. Dopamine receptors: From structure to function. Physiol. Rev., 78: 189-225.
11. Kamiya, T., O. Saitoh and H. Nakata, 2005. Functional expression of single-chain heterodimeric g-protein-coupled receptors for adenosine and dopamine. Cell Struct. Funct., 29: 139-145.
12. Kamiya T., K. Yoshioka and H. Nakata, 2015. Analysis of various types of single-polypeptide-chain (sc) heterodimeric A_{2A}R/d2r complexes and their allosteric receptor–receptor interactions. Biochemical and Biophysical Research Communications 456: 573-579.
13. Kamiya, T., T. Masuko, D.O. Borroto-Escuela, H. Okado and H. Nakata, 2018. A transmembrane single-polypeptide-chain (sc) linker to connect the two g-protein–coupled receptors in tandem and the design for an *in vivo* analysis of their allosteric receptor- receptor interactions. In: Polypeptide-The New Insight into Drug Discovery and Development Tambunan, U.S.F., InTech, London, pp: 41-60.
14. Chernomordik, L.V. and M.M. Kozlov, 2008. Mechanics of membrane fusion. Nat. Struct. Mol. Biol., 15: 675-683.
15. Kurylowicz, M., H. Paulin, J. Mogyoros, M. Giuliani and J.R. Dutcher, 2014. The effect of nanoscale surface curvature on the oligomerization of surface-bound proteins. J. R. Soc. Interface, Vol. 11, No. 94. 10.1098/rsif.2013.0818
16. Soubias, O., W.E. Teague, K.G. Hines and K. Gawrisch, 2014. The role of membrane curvature elastic stress for function of rhodopsin-like G protein-coupled receptors. Biochimie, 107: 28-32.
17. Bryson-Richardson, R.J., D.W. Logan, P.D. Currie and I.J. Jackson, 2004. Large-scale analysis of gene structure in rhodopsin-like GPCRs: Evidence for widespread loss of an ancient intron. Gene, 338: 15-23.
18. Yee, D.C., M.A. Shlykov, A. Vastermark, V.S. Reddy, S. Arora, E.I. Sun and M.H. Saier Jr, 2013. The transporter-opsin-G protein coupled receptor (TOG) superfamily. FEBS J., 280: 5780-5800.
19. Krishnan, A. and H.B. Schioth, 2015. The role of G protein-coupled receptors in the early evolution of neurotransmission and the nervous system. J. Exp. Biol., 218: 562-571.
20. Tsirigos, K.D., S. Govindarajan, C. Bassot, A. Vastermark, J. Lamb, N. Shu and A. Elofsson, 2018. Topology of membrane proteins-predictions, limitations and variations. Curr. Opin. Struct. Biol., 50: 9-17.
21. Kall, L., A. Krogh and E.L. Sonnhammer, 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol., 338: 1027-1036.
22. Bausch-Fluck, D., U. Goldmann, S. Muller, M. van Oostrum, M. Muller, O.T. Schubert and B. Wollscheid, 2018. The in silico human surfaceome. Proc. National Acad. Sci., 115: E10988-E10997.
23. Lehman, C.W., J.D. Lee and C.F. Komives, 2005. Ubiquitously expressed GPCR membrane-trafficking orthologs. Genomics, 85: 386-391.
24. Gruber, M. and A.N. Lupas, 2003. Historical review: Another 50th anniversary-new periodicities in coiled coils. Trends Bioch Sci., 28: 679-685.
25. Mount, D.W., 2004. Bioinformatics: Sequence and Genome Analysis. 2nd Edn., Genetics and Genome Science Biotechnology, Laboratory Manuals/Handbooks, University of Arizona, Tucson.
26. Green, M.R. and J. Sambrook, 2012. Molecular Cloning: A Laboratory Manual. 4th Edn., Cold Spring Harbor Laboratory Press, New York, USA.,.
27. Dungan, A.M., L.M. Hartman, G. Tortorelli, R. Belderok and A.M. Lamb *et al*., 2020. *Exaiptasia diaphana* from the great barrier reef: A valuable resource for coral symbiosis research. Symbiosis, 80: 195-206.

28. Perfus-Barbeoch, L., A.M. Jones and S.M. Assmann, 2004. Plant heterotrimeric G protein function: Insights from *Arabidopsis* and rice mutants. Curr. Opin. Plant Boil., 7: 719-731.
29. Grigston, J.C., D. Osuna, W.R. Scheible, C. Liu, M. Stitt and A.M. Jones, 2008. D-glucose sensing by a plasma membrane regulator of G signaling protein, AtRGS1. FEBS Lett., 582: 3577-3584.
30. Saier Jr, M.H., 2016. Transport protein evolution deduced from analysis of sequence, topology and structure. Curr. Opin. Struct. Biol., 38: 9-17.
31. Liu, Y., M. Gerstein and D.M. Engelman, 2004. Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. Proc. National Acad. Sci., 101: 3495-3497.
32. Hennerdal, A., J. Falk, E. Lindahl and A. Elofsson, 2010. Internal duplications in α helical membrane protein topologies are common but the nonduplicated forms are rare. Protein Sci., 19: 2305-2318.
33. Lehnert, E.M., M.S. Burriesci and J.R. Pringle, 2012. Developing the anemone Aiptasia as a tractable model for cnidarian-dinoflagellate symbiosis: The transcriptome of aposymbiotic *A. pallid*. BMC. Genomics, Vol. 13, 10.1186/1471-2164-13-271
34. Dwyer, N.D., E.R. Troemel, P. Sengupta and C.I. Bargmann, 1998. Odorant receptor localization to olfactory cilia is mediated by ODR-4, a novel membrane-associated protein. Cell, 93: 455-466.
35. Thal, D.M., A. Glukhova, P.M. Sexton and A. Christopoulos, 2018. Structural insights into G-protein-coupled receptor allostery. Nature, 559: 45-53.
36. Jabs, F., M. Plum, N.S. Laursen, R.K. Jensen and B. Molgaard *et al*., 2018. Trapping IgE in a closed conformation by mimicking CD23 binding prevents and disrupts FcεRI interaction. Nature Commun., 9: 1-11.
37. Huang, M.C., J.M. Seyer and A.H. Kang, 1990. Comparison and accuracy of methodologies employed for analysis of hydropathy, flexibility and secondary structure of proteins. J. Immunol. Methods, 129: 77-88.
38. Brunette, T.J., F. Parmeggiani, P.S. Huang, G. Bhabha and D.C. Ekiert *et al*., 2015. Exploring the repeat protein universe through computational protein design. Nature, 528: 580-584.
39. Doyle, L., J. Hallinan, J. Bolduc, F. Parmeggiani, D. Baker, B.L. Stoddard and P. Bradley, 2015. Rational design of α-helical tandem repeat proteins with closed architectures. Nature, 528: 585-588.
40. Mondal, S., L. Adler-Abramovich, A. Lampel, Y. Bram, S. Lipstman and E. Gazit, 2015. Formation of functional super-helical assemblies by constrained single heptad repeat. Nat. Commun., 6: 1-8.
41. Bhardwaj, G., V.K. Mulligan, C.D. Bahl, J.M. Gilmore and P.J. Harvey *et al*., 2016. Accurate de novo design of hyperstable constrained peptides. Nature, 538: 329-335.
42. Marcos, E., B. Basanta, T.M. Chidyausiku, Y. Tang and G. Oberdorfer *et al*., 2017. Principles for designing proteins with cavities formed by curved β sheets. Science, 355: 201-206.
43. Overvelde, J.T., J.C. Weaver, C. Hoberman and K. Bertoldi, 2017. Rational design of reconfigurable prismatic architected materials. Nature, 541: 347-352.
44. Chen, X., J.L. Zaro and W.C. Shen, 2013. Fusion protein linkers: Property, design and functionality. Adv. Drug Delivery Rev., 65: 1357-1369.
45. Skorupka, K., S.K. Han, H.J. Nam, S. Kim and S. Faham, 2013. Protein design by fusion: Implications for protein structure prediction and evolution. Acta Crystallogr. Sect. D. Biol. Crystallogr., 69: 2451-2460.
46. Yu, K., C. Liu, B.G. Kim and D.Y. Lee, 2015. Synthetic fusion protein design and applications. Biotechnol. Adv., 33: 155-164.
47. Ferreira, K.N.T., M. Iverson, K. Maghlaoui, J. Barber and S. Iwata, 2004. Architecture of the photosynthetic Oxygen-evolving center. Science, 203: 1831-1838.
48. Guskov, A., J. Kern, A. Gabdulkhakov, M. Broser, A. Zouni and W. Saenger, 2009. Cyanobacterial photosystem II at 2.9-A resolution and the role of quinones, lipids, channels and chloride. Nat. Struct. Mol. Biol., 16: 334-342.
49. Sharpe, H.J., T.J. Stevens and S. Munro, 2010. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. Cell, 142: 158-169.
50. Petersen, T.N., S. Brunak, G. von Heijne and H. Nielsen, 2011. SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nat. Methods, 8: 785-786.
51. Henderson, I.R., F. Navarro-Garcia, M. Desvaux, R.C. Fernandez and D. Ala'Aldeen, 2004. Type V protein secretion pathway: The autotransporter story. Microbiol. Mol. Biol. Rev., 68: 692-744.
52. Bingle, L.E., C.M. Bailey and M.J. Pallen, 2008. Type VI secretion: A beginner's guide. Curr. Opin. Microbiol., 11: 3-8.
53. Leiman, P.G., M. Basler, U.A. Ramagopal, J.B. Bonanno and J.M. Sauder *et al*., 2009. Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. Proc. National Acad. Sci., 106: 4154-4159.

54. Faruque, O.M., D. Le-Nguyen, A.D. Lajoix, E. Vives, P. Petit, D. Bataille and E.H. Hani, 2009. Cell-permeable peptide-based disruption of endogenous PKA-AKAP complexes: A tool for studying the molecular roles of AKAP-mediated PKA subcellular anchoring. Am. J. Physiol. Cell Physiol., 296: C306-C316.
55. Frankel, A.D. and C.O. Pabo, 1988. Cellular uptake of the tat protein from human immunodeficiency virus. Cell, 55: 1189-1193.
56. Green, M. and P.M. Loewenstein, 1988. Autonomous functional domains of chemically synthesized human immunodeficiency virus tat trans-activator protein. Cell, 55: 1179-1188.
57. Kikutani, H., S. Inui, R. Sato, E.L. Barsumian and H. Owaki *et al*., 1986. Molecular structure of human lymphocyte receptor for immunoglobulin E. Cell, 47: 657-665.
58. Yokota, A., H. Kikutani, T. Tanaka, R. Sato, E.L. Barsumian, M. Suemura and T. Kishimoto, 1988. Two species of human Fcε receptor II (FcεRIICD23): Tissue-specific and IL-4-specific regulation of gene expression. Cell, 55: 611-618.
59. Richards, M.L., D.H. Katz and F.T. Liu, 1991. Complete genomic sequence of the murine low affinity Fc receptor for IgE. Demonstration of alternative transcripts and conserved sequence elements. J. Immunol., 147: 1067-1074.
60. Kamiya, T., O. Saitoh, K. Yoshioka and H. Nakata, 2003. Oligomerization of adenosine A_2A and dopamine D_2 receptors in living cells. Biochem. Biophys. Res. Commun., 306: 544-549.
61. Jaakola, V.P., M.T. Griffith, M.A. Hanson, V. Cherezov and E.Y. Chien *et al*., 2008. The 2.6 angstrom crystal structure of a human A_2A adenosine receptor bound to an antagonist. Science, 322: 1211-1217.
62. Li, J.B. and G.M. Church, 2013. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. Nature Neurosci., 16: 1518-1522.
63. Alivisatos, A.P., M. Chun, G.M. Church, K. Deisseroth and J.P. Donoghue *et al*., 2013. The brain activity map. Sci., 339: 1284-1285.
64. Ramaswami, G., R. Zhang, R. Piskol, L.P. Keegan, P. Deng, M.A. O'connell and J.B. Li, 2013. Identifying RNA editing sites using RNA sequencing data alone. Nature Methods, 10: 128-132.
65. Isberg, V., C. de Graaf, A. Bortolato, V. Cherezov and V. Katritch *et al*., 2015. Generic GPCR residue numbers-aligning topology maps while minding the gaps. Trends Pharmacol. Sci., 36: 22-31.
66. Luecke, H., B. Schobert, H.T. Richter, J.P. Cartailler and J.K. Lanyi, 1999. Structure of bacteriorhodopsin at 1.55 A resolution. J. Mol. Biol., 291: 899-911.
67. Xu, J., M. Gui, D. Wang and Y. Xiang, 2016. The bacteriophage φ29 tail possesses a pore-forming loop for cell membrane penetration. Nature, 534: 544-547.
68. Kuzniar, A., R.C. Van Ham, S. Pongor and J.A. Leunissen, 2008. The quest for orthologs: Finding the corresponding gene across genomes. Trends Genet., 24: 539-551.
69. Sugiura, M., S. Fuke, S. Suo, N. Sasagawa, H.H. Van Tol and S. Ishiura, 2005. Characterization of a novel D2 like dopamine receptor with a truncated splice variant and a D1 like dopamine receptor unique to invertebrates from *Caenorhabditis elegans*. J. Neurochem., 94: 1146-1157.
70. Brattelid, T., A.M. Kvingedal, K.A. Krobert, K.W. Andressen and T. Bach *et al*., 2004. Cloning, pharmacological characterisation and tissue distribution of a novel 5-HT_4 receptor splice variant, 5-HT_4(i). Naunyn-Schmiedeberg's Arch. Pharmacol., 369: 616-628.
71. Huang, P.S., S.E. Boyken and D. Baker, 2016. The coming of age of de novo protein design. Nature, 537: 320-327.
72. Van Regenmortel, M.H., 2009. What Is a B-Cell Epitope?. In: Epitope Mapping Protocols, Schutkowski, M. and U. Reineke (Eds.). Springer, Berlin, Germany, pp: 3-20.
73. McCorvy, J.D. and B.L. Roth, 2015. Structure and function of serotonin G protein-coupled receptors. Pharmacol. Ther., 150: 129-142.
74. Hunter, T., 2007. The age of crosstalk: Phosphorylation, ubiquitination and beyond. Mol. Cell, 28: 730-738.
75. Schwarz, F. and M. Aebi, 2011. Mechanisms and principles of N-linked protein glycosylation. Curr. Opin. Struct. Biol., 21: 576-582.
76. Korkuc, P. and D. Walther, 2017. Towards understanding the crosstalk between protein post translational modifications: Homo and heterotypic PTM pair distances on protein surfaces are not random. Proteins: Struct. Function Bioinf., 85: 78-92.
77. Rossmann, M.G. and P. Argos, 1981. Protein folding. Annu. Rev. Biochem., 50: 497-532.
78. Baker, E.G., G.J. Bartlett, M.P. Crump, R.B. Sessions, N. Linden, C.F. Faul and D.N. Woolfson, 2015. Local and macroscopic electrostatic interactions in single α-helices. Nature Chem. Biol., 11: 221-228.
79. Chen, Z., S.E. Boyken, M. Jia, F. Busch and D. Flores-Solis *et al*., 2019. Programmable design of orthogonal protein heterodimers. Nature, 565: 106-111.
80. Munter, L.M., P. Voigt, A. Harmeier, D. Kaden and K.E. Gottschalk *et al*., 2007. GxxxG motifs within the amyloid precursor protein transmembrane sequence are critical for the etiology of Aβ42. EMBO J., 26: 1702-1712.
81. Langosch, D. and I.T. Arkin, 2009. Interaction and conformational dynamics of membrane spanning protein helices. Protein Sci., 18: 1343-1358.