Supplementary file 1: Removing redundancies of datasets

The redundant sequences in the datasets are needed to be removed. For brevity, the datasets of 558 GPCRs, 2222 GPCRs and 721 non-GPCRs membrane proteins were named GPCR_TEST558, GPCR_TRAIN2222 and MEM_721, respectively. Here, CD-HIT [1] (http://www.bioinformatics.org/cd-hit/) was employed, which is a widely used program for clustering proteins and removing redundant sequences. Remaining of sequences in the datasets using CD-HIT with various identity cutoffs are given in Table S1.

Table S1. The remaining sequences in the datasets with various identities.

Datasets	70%	80%	90%	95%	99%
GPCR_TEST558 (558)[a]	377	413	464	492	544
GPCR_TRAIN2222 (2222)	781	1012	1374	1697	2081
MEM_721 (721)	325	393	454	504	555

[a] The value inside the parentheses denotes the number of proteins in the corresponding dataset.

As given in Table S1, there exist 377, 781 and 325 sequences in the GPCR_TEST558, GPCR_TRAIN2222 and MEM_721 datasets when similar sequences were removed at the cutoff of 70% sequence identity. We hope to collect as many non-redundant sequences as possible. In this work, 95% identity was employed to remove the highly similar sequences, and 492, 1697 and 504 sequences were obtained in the three datasets. These datasets were used and were named GPCR_TEST492, GPCR_TRAIN1697 and MEM_504, respectively.

References

1. W. Li and A. Godzik, *Bioinformatics (Oxford, England)*, 2006, **22**, 1658-1659.