Supplementary file 3: Proteome-wide GPCR identification in *Homo sapiens*

To provide a practical application of our predictor, we conducted a proteome-wide GPCR identification in *Homo sapiens*. The whole proteome of *Homo sapiens* was downloaded from the Ensembl database (ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/pep/Homo_sapiens.GRCh37.75.pep.abinitio.fa.gz) [1]. There are 104,763 protein sequences in the *Homo sapiens*. To quickly select the potential targets, we used TMHMM [2] to collect the proteins that have more than 5 TM segments and we obtained 4108 proteins. All the 4108 protein sequences were directly submitted to the GPCRserver.

The experimentally verified GPCRs should be known before benchmark. There are 998 human GPCRs complied by skolnick group (http://cssb2.biology.gatech.edu/gpcr2011/new_human_998.seq) [3]. Among the 998 proteins, 748 proteins can be mapped to Ensembl entries. For the remaining 250 proteins, we BLASTed them against the *Homo sapiens* proteins and the top ones of each BLAST search with e-value< 1e-7 are regarded as GPCR. Only one protein of Swiss-Prot (Swiss-Prot entry: Q8TDU0) was not used according to the fact that there is not hit to the *Homo sapiens* proteins searched by the BLAST for the protein. The mapped Ensembl proteins and the top ones of the 250 GPCRs BLASTed to *Homo sapiens* were considered as verified GPCRs in this manuscript.

The 4108 proteins, which selected by using TMHMM as potential targets, were fed into GPCRserver predictor. There are 681 proteins identified by our predictor at 1% false positive rate (i.e. 1% false positive rates by the Trans-GPCR, SSEA-GPCR and PPA-GPCR methods). And 486 out of the 681 proteins were verified GPCRs. Therefore, these 486 predicted GPCRs should be regarded as true positives with high confidence. Considering the highly imbalanced numbers of GPCRs and non-GPCRs in the proteome of *Homo sapiens* and the most targets selected by TMHMM were membrane proteins, it is not surprising that our predictor resulted in a certain number of false positives even at a false positive rate control of 1%. In order to reduce the false positives, we may resort to other bioinformatics tools. For the GPCRs that were not identified by our predictor, it may be ascribed to the fact that some of these GPCRs share dissimilar topologies and profiles with known GPCRs in our training set. Meanwhile, our method is not perfect, which is also an important factor. To maximize the performance of our predictor, a regularly-updated library which covers all sequence/structure space of known GPCRs is highly desired. Additionally, developing more effective scoring functions may also be an important direction to improve performance.

References

1. P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M.

McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino and S. M. Searle, *Nucleic acids research*, 2014, **42**, D749-755.

2. A. Krogh, B. Larsson, G. von Heijne and E. L. Sonnhammer, *Journal of molecular biology*, 2001, **305**, 567-580.
3. H. Zhou and J. Skolnick, *Molecular pharmaceutics*, 2012, **9**, 1775-1784.