Supplementary file 5: PSIPRED for protein GPCR secondary structure prediction

We used PSIPRED, which was developed by Jones [1] and not specifically designed for GPCRs, to predict secondary structures of GPCRs in this work. But since the physical environment of GPCRs is completely different from that of globular proteins, it is therefore very necessary to benchmark the performance of PSIPRED on GPCRs. We downloaded 108 structurally known GPCRs from PDB database.

The experimental secondary structures of these GPCRs were defined by using STRIDE [2] program. We assessed the performance of PSIPRED method utilizing Q_3 measure, which is the total number of correctly predicted residue states divided by the total number of residues. In addition, another three measures Q_H, Q_E and Q_C, which describe the fractions of correctly predicted residues out of the total numbers of residues in α-helix, β-strand and coil, were also used to evaluate the performance. The 108 GPCR structures were split into chains. Protein names are denoted by Protein Data Bank entries (first 4 characters) and followed by the chain identifiers. The proteins were filtered by CD-HIT [3] by removing redundant sequences at 95% identity cutoff and 55 chains were obtained. The 55 non-redundant chains were directly fed into PSIPRED. Although PSIPRED is not specifically designed for GPCRs, it is surprising that the prediction accuracies of PSIPRED for GPCRs are reasonably high. The Q_3 accuracy is 76.6%. The Q_H, Q_E and Q_C are 74.6%, 68.4% and 84.3%, respectively. As the accuracy of Q_E is relatively lower, this might be attributed to the fact that formation of β-strand is strongly influenced by long-range interactions [4]. Whether the performance of PSIPRED will persist for structurally unknown GPCRs is not clear, but at least the secondary structure prediction accuracy for structurally known GPCRs is at practical level in our benchmark. Here, we list the prediction for each protein of 55 non-redundant chains in Table S3. Symbol '-' stands for there is no corresponding secondary structure element in a protein.

Table S3. The performance of PSIPRED for GPCRs

Protein	Q_3	Q_H	Q_E	Q_C	Sequence Length
1F88B	0.836	0.869	0.583	0.790	305
1U19A	0.798	0.860	0.5	0.715	348
2LNLA	0.870	0.920	-	0.734	296
2R4RH	0.898	-	0.862	0.921	217
2R4RL	0.896	0.666	0.883	0.929	214
2RH1A	0.646	0.608	0	0.864	442
2YDVA	0.811	0.829	0	0.806	315
2ZIYA	0.872	0.890	0.833	0.833	370
3EMLA	0.647	0.634	0.1	0.780	448
3KJ6A	0.822	0.908	-	0.623	222
3ODUA	0.737	0.772	0.384	0.711	466
3OE6A	0.769	0.837	0.227	0.673	418

3P0GB	0.816	-	0.680	1	121
3PBLA	0.619	0.577	0	0.869	432
3PWHA	0.905	0.925	0.333	0.877	291
3RZEA	0.656	0.634	0	0.803	428
3SN6A	0.898	0.898	0.833	0.910	349
3SN6B	0.893	0.785	0.842	0.960	340
3SN6G	0.877	0.911	-	0.791	58
3SN6N	0.920	-	0.803	1	128
3UONA	0.666	0.630	0	0.862	438
3V2YA	0.700	0.658	0	0.864	455
3VG9B	0.913	0.666	0.879	0.978	212
3VG9C	0.879	-	0.790	0.98	224
3VW7A	0.778	0.788	0.636	0.765	442
3ZPQB	0.811	0.795	-	0.870	299
4AMIA	0.790	0.776	-	0.844	282
4BUOB	0.843	0.834	0.6	0.904	314
4BWBA	0.835	0.829	0.6	0.890	305
4DAJD	0.631	0.616	0	0.768	432
4DJHB	0.729	0.702	0.666	0.845	448
4DKLA	0.732	0.718	0.6	0.811	442
4EA3A	0.880	0.884	0.833	0.847	278
4EA3B	0.744	0.729	0.5	0.822	376
4EIYA	0.714	0.707	0.166	0.796	390
4EJ4A	0.700	0.673	0.727	0.782	442
4GRVA	0.746	0.722	0.545	0.836	454
4IARA	0.751	0.722	-	0.863	379
4IB4A	0.767	0.738	-	0.903	375
4K5YA	0.753	0.727	0.714	0.871	407
4K5YC	0.878	0.866	-	0.913	248
4L6RA	0.668	0.639	-	0.758	398
4LDEA	0.741	0.718	0.642	0.823	454
4LDEB	0.897	-	0.793	1	120
4MBSA	0.863	0.899	0.5	0.839	346
4MQSA	0.829	0.823	-	0.839	277
4MQSB	0.825	-	0.681	0.981	121
4N4WA	0.745	0.722	0.416	0.825	457
4N6HA	0.783	0.755	0.833	0.907	408
4NTJA	0.725	0.724	-	0.716	369
4O9RA	0.827	0.853	0.285	0.817	441
4OR2B	0.565	0.476	0.5	0.873	366

References

1. D. T. Jones, *Journal of molecular biology*, 1999, **292**, 195-202.
2. D. Frishman and P. Argos, *Proteins*, 1995, **23**, 566-579.

3. W. Li and A. Godzik, *Bioinformatics (Oxford, England)*, 2006, **22**, 1658-1659.
4. D. Kihara, *Protein Sci*, 2005, **14**, 1955-1963.