Function Prediction for G Protein-Coupled Receptors through Text Mining and Induction Matrix Completion

Jiansheng Wu, Qin Yin, Chengxin Zhang, Jingjing Geng, Hongjie Wu, Haifeng Hu, Xiaoyan Ke, Yang Zhang

Supporting Information

Table S1. Performance dependence on keywords for querying literature in PubMed.

GO term aspect	Keyword	Recall				
		r=20	r=40	r=60	r=80	r=100
Molecular Function	Receptor	.1795	.2490	.2978	.3403	.3774
	G Protein-Coupled Receptors	.1802	.2501	.2896	.3433	.3751
	GPCRs	.1797	.2512	.2899	.3333	.3685
Biological Process	Receptor	.2125	.3101	.3783	.4330	.4790
	G Protein-Coupled Receptors	.2201	.3087	.3745	.4296	.4756
	GPCRs	.2101	.3135	.3813	.4321	.4668

Supplementary material 1. Code Usage

We have share the source code and dataset of this study at http://zhanglab.ccmb.med.umich.edu/TM-IMC. The code for TM-IMC was written in matlab2014, which can easily be implemented across multiple platforms, including Windows 10 and Linux. The Github repository includes two demo programs:

(1) demo_new: This provides a general learning framework integrating text mining and Inductive matrix completion for users to develop their own tools on the basis of our codes. Input: text information of samples, text information of labels, and the sample-label association matrix. Output: model performance. The steps are as follows: first, vector representation of sample and label text information is implemented by the Word2Vec tool; second, the transformation of multiple instances into single vector is performed by the miFV algorithm; finally, Inductive matrix completion models are constructed to get the predictive performance (Recall and Relative Error values).

(2) demo_predict: This provides the models for predicting the GO molecular functions and biological processes of GPCRs, and the predictive scores can be obtained for each GPCR protein on each GO term. Input: the GPCR feature spaces, the GO feature spaces, and the GPCR-GO association matrix. Output: the predictive scores of GPCRs on each GO term. The steps are as follows. Firstly, the GPCR-GO associations are randomly divided into three equal parts where two of them are used to train the model and the remaining one part is to test the model. The process is repeated three times to ensure that each association is tested exactly once. Finally, all the predictive scores are saved in the predict_scores.txt file.