Title: Systematic prediction of ligand-receptor pair across the immunoglobulin superfamily using a novel sequence homology measure

Authors: Eng-Hui Yap, Tyler Rosche, Andras Fiser

Full Abstract

The immunoglobulin superfamily (IgSF) is a large group of cell surface and soluble proteins that play a key role in cell recognition, signaling and adhesion (Barclay 2003). Members of IgSF contain at least one immunoglobulin (Ig) domain comprising of 70-110 residues as a beta sandwich fold. The IgSF constitutes one of the largest group of protein families in the human genome, accounting for about 1600 of the known human proteins from Uniprot (Wu, Apweiler et al. 2006). Of particular interest to us are the 477 IGSF cell surface or secreted IGSF proteins that remained after excluding antibodies, T-cell receptors (TCRs) and multi-histocompatibility complex (MHC) proteins. Many of these IgSF are involved in critical cell-to-cell (trans) receptor-ligand interaction, often with other IgSF. Unfortunately, while some of the receptors have known binding partners, many remained orphan receptors without known ligands or function. Moreover, it is possible that a single receptor can have multiple binding partners, or vice versa.

We performed a large scale, IgSF-family wide prediction of new receptor-ligand relationships based on sequence homology of known IgSF receptor-ligand pairs. Our method differs from the usual sequence homology methods in the following way: (i) we assessed the sequence similarity between two proteins by comparing their respective profile hidden Markov models (HMMs) (Soding 2005). This amplifies the signal from the conserved (and often functionally important) portions of the sequences and downplays the less conserved segments; (ii) the scoring framework allows us to include empirical information available about the protein function. For instance, most IgSF proteins interact with their trans-binding partners using the N-terminal domain. We have included this information as a functional criterion in our scoring function, so IgSF pairs that share sequence homology starting at the N-terminus are considered more functionally similar than those that share homology over disparate segments of their sequences. (iii) We further increase the confidence of the prediction by hierarchically clustering the IgSF proteins using the above-mentioned sequence similarity measure with average linkage. This ensures that if two proteins A and B are considered similar, then a third protein C that wants to join the group must be similar to both A and B. (iv) Lastly, we built a training set of 17 IgSF pairs that shared the same experimentally verified ligands from the STRING database (Szklarczyk, Franceschini et al. 2011), and used it to guide our selection of the similarity score cutoff for predicting that two IgSFs have similar ligand.

We tested our predictions on a separate, independent validation set of 53 IgSF pairs that shared similar ligands. Of these, our method correctly predicted 40 pairs to share common ligands (true positives) and missed 13 (false negatives), giving a sensitivity of 75.47%. Given that some of these IgSF pairs have sequence identities lower than 25%, our HMM-based tree clustering method thus provides a highly sensitive approach to scan
for such hard-to-detect homology in binding. The method was then applied to all 477 IgSF proteins of interest in the human Ig superfamily, of which 380 IgSFs can be assigned to a functional family that binds the same ligand. The method can be readily adopted to handle other classes of proteins, and can be easily updated to include additional empirical information about the proteins’ binding modes.

References