SIGNOR 3.0 score

The SIGNOR 3.0 significance score ranges from 0.1 and 1, we set 0.1 as the minimum score as 0 stands for no evidence of interaction. The significance scores of interactions between entities of classes *protein*, *proteinfamily* and *complex* is calculated by using a principal component regression (PCR) approach ([reference link](https://www.jstatsoft.org/article/view/v018i02)). This approach produces a model that predicts whether the annotated functional relationships between an entity pair is part of a molecular pathway annotated in our resource. The dataset of functional interactions which are part of SIGNOR molecular pathways is taken as the golden standard of functional interactions

Briefly, we first created a matrix where each interacting pair is associated with the following features:

* Number of PubMed IDs (PMID) associated to the pair, as annotated in SIGNOR;
* A value that reflects whether the interaction has annotation direct yes/no. Specifically, yes=2, no=0, unknown= 1;
* STRING score ‘coexpression ‘, ranging from 0 to 1000;
* STRING score ‘database ‘, ranging from 0 to 1000;
* STRING score ‘experimental ‘, ranging from 0 to 1000;
* STRING score ‘textmining ‘, ranging from 0 to 1000.

A score was also associated with interactions between proteins and protein families (PF) or complexes (CPX) by remapping STRING scores to interactions involving these entities according to the following criteria:

1. To calculate the interaction score between a protein X and a PF we obtained from STRING the interaction scores between protein X and each PF -member and we associated the highest value (Fig.1).

Figure 1 protein-protein family interaction score

1. For interactions between a protein X and a CPX, for each feature, we retrieved from STRING the interaction scores between each complex subunit and protein X and we associated to the CPX/protein X interaction the mean of the subunit scores (Fig.2)



Figure 2 Figure 1 protein-complex interaction score

1. For PF-PF, CPX-CPX, PF-CPX and CPX-PF interactions, combinations of the aforementioned approaches were applied (Fig 1 and 2).

The PCR model was trained using the score matrix for protein-protein interactions having non-zero values for STRING scores.

We used the ‘pls’ package in R (‘pcr’ function, using cross validation as validation choice).

The regression coefficients to predict the dependent variable (probability of being part of an annotated pathway) were estimated on different numbers of principal components of the explanatory variables, ranging from 2 to 6. We eventually selected to use a model based on three principal components. The predicted score is then normalized between 0.1 and 1.

We evaluated the performance achieved by using the different numbers of principal components, by visually checking the score distribution obtained with each scoring model for the full interactome (in blue in Fig. 3) and for interactions annotated to pathways (in yellow in Fig. 3).

As shown in Fig. 3, by using three components (Fig.3B, framed in green) we obtained a model that maximizes the number of components used, still scoring low the interactions for which we have poor supporting information (framed in red), while scoring high interactions.

We, therefore, decided to use a model based on three principal components to calculate the score.



Figure 3 Score distribution of interactions involving protein, protein complexes and protein families by number of components. In yellow interactions annotated to a pathway, in blue the entire interactome

Scores for Interactions involving non protein entities, which are considered in the SIGNOR network (*smallmolecule, phenotypes, stimuli, chemicals, fusion proteins, miRNA and antibodies)*, could not be assigned using the PCR model as many of the features considered for building the model were not available. For these interactions that represent a minority In the SIGNOR network we assigned an ad hoc score that reflects our confidence in the overall reliability of the supporting evidence. For Interactions involving

* chemical and smallmolecule -> SIGNOR 3.0 significance score = 0.8
* Stimulus and phenotype -> SIGNOR 3.0 significance score = 0.7
* Antibody and mirna -> SIGNOR 3.0 significance score = 0.4
* Fusion proteins -> SIGNOR 3.0 significance score = 0.1