Supplemental Materials 1

# Overview of predicates found in paths

Table 1 Overview of all 47 predicates that were found in the paths between the proteins in the EKP.

|  |  |
| --- | --- |
| affects | is a |
| augments | is associated with |
| binds with | is compared with |
| catalysis precedes | is functionally related to |
| coexists with | is higher than |
| consumption controlled by | is lower than |
| controls expression of | is manifestation of |
| controls phosphorylation of | is not higher than |
| controls state change of | is not lower than |
| controls transport of | is not part of |
| converts to | is parent of |
| disrupts | is part of |
| does not coexist with | is spatially related to |
| does not convert to | is the same as |
| does not inhibit | is tributary of |
| does not interact with | is variant of |
| does not produce | manages |
| does not stimulate | ortholog is associated with |
| forms protein complex with | performs |
| gene product is biomarker type | produces |
| gene product is encoded by | stimulates |
| gene product variant causes | targets |
| inhibits | uses |
| interacts with |  |

# Feature sets where all predicates are directed

In addition to the undirected and the mixed feature sets, we also experimented with feature sets where all predicates were considered to be directed, as indicated by the subject and object of the triple. Figure 1 shows example feature sets for both the metapaths and the split paths.



Figure 1 Examples of the directed feature sets for both the metapaths and the split paths. These features were derived from the same knowledge graph as shown in Figure 1 in the article. Because all predicates in this variation are directed, the “binds with” predicate is also directed in accordance with the subject and the object of the triple. Directionality of triples is only included for the direct and indirect scenarios, and not for the overlap scenario.

The classification performance for both reference sets, and their six possible feature sets is shown in Table 2. As can be seen, the directed feature sets achieved almost identical performance to the mixed feature sets.

Table 2 Classification performance results based on 10 repeats of a ten-fold cross validation experiment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Jensen reference set | | Van den Akker reference set | |
|  | Metapaths | Split paths | Metapaths | Split paths |
| Undirected | 83.3 (1.7) | 78.3 (1.7) | 72.5 (11.8) | 68.4 (13.0) |
| Mixed | 89.8 (0.9) | 82.8 (1.2) | 74.5 (10.5) | 70.3 (11.4) |
| Directed | 90.0 (0.8) | 83.1 (1.4) | 74.1 (10.3) | 70.8 (11.5) |

To determine whether the differences in AUC between the mixed and directed feature sets was statistically significant, we again performed paired T-tests by sampling identical folds between experiments. The differences between the mixed and directed feature sets for both reference sets were not significant. See Table 3 for a complete overview of all p-values.

Table 3 p-values of the paired t-tests of the AUCs of the different cross-validation experiments.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Jensen set | | Van den Akker set | |
|  | Metapaths | Split paths | Metapaths | Split paths |
| Undirected – Mixed | 2.2 × 10-16 | 2.2 × 10-16 | 0.02177 | 0.001282 |
| Mixed - Directed | 0.107 | 0.8825 | 0.8721 | 0.3176 |
| Undirected - Directed | 2.2 × 10-16 | 2.2 × 10-16 | 0.4884 | 0.0007527 |