Description of features used in Salmonella-human PPI prediction

Meghana Kshirsagar
LTI, Carnegie Mellon University, Pittsburgh

July 27, 2012

We use the 62 *Salmonella*-human protein interactions reported in [12] as the "positive class" in our gold standard data. Call this, set P. Let 'U' be the set of protein pairs obtained by pairing up all *Salmonella* proteins with all human proteins. Our dataset had 3592 unique *Salmonella*[1] and 24431 unique human proteins, resulting in $|U| \approx$90 million pairs. We sample some random pairs from the set $U \backslash P$, to derive the set N that will be the negative class. The number of random pairs chosen to be the negative class is decided by what we expect the interaction ratio to be. We chose a ratio of 1:100 meaning that we expect 1 in every 100 random pairs of proteins to interact with each other, based on prior work in [15, 5]. Our training data $D = P \cup N$ thus had 62 positives and 6200 negatives. To generate novel potential interactions given the gold-standard data, we apply the classifier model built on D to the rest of the protein pairs : $(U \backslash D)$ and report the positive pairs predicted by the classifier.

Feature set

The *Salmonella*-human PPI prediction problem is cast as a two-class classification problem: each protein pair $p =< p_s, p_h >$ is an instance belonging to either the positive, 'interaction' class or the negative, 'non-interaction' class. For each pair, we derived features which can belong to one of the three types: (a) feature derived on the pair p, (b) feature derived using only the human protein p_h and (c) feature derived using only the *Salmonella* protein p_s. Based on the source of information, we can also categorize our feature set into the following groups: (1) GO similarity, (2) graph based features using the human interactome, (3) gene expression, (4) RNAi expression, (5) sequence based features, (6) PFam overlap, (7) features from protein family and protein domain interactions, (8) interolog based and (9) *Salmonella* gene properties.

1. **GO similarity features**: These features model the similarity between the functional properties of two proteins. Gene Ontology [1] provides GO-term annotations for three important protein properties: molecular function (F), cellular component (C) and biological process (P). We derive 6 types of features using these properties. For each of 'F', 'C' and 'P', two types of GO similarity features were defined: (a) pair-level similarity and (b) similarity with human protein's binding partners. The similarity between two individual GO terms was computed using the G-Sesame algorithm [4]. This feature is a matrix of all the GO term combinations found in a given protein pair: $< p_s, p_h >$, the rows of the matrix represent GO terms from protein p_s and the columns represent GO terms from p_h. Analogously, the second feature type - (b) computes the similarity between the GO term sets of the *Salmonella* protein and the human protein's binding partners in the human interactome. We used HPRD to get the human interactome.

2. **Graph based features using the human interactome**: These features are derived using only the human protein 'p_h' from the pair. Pathogens generally target host proteins that are important in several host processes; these host proteins interact with many other host proteins to carry out their tasks. This insight is captured in the form of three graph properties: *degree*, *between-ness centrality* and *clustering coefficient* of the human protein "node" in the human interactome graph. The human interactome was downloaded from HPRD [10]. The degree of a node is the number of its neighbouring nodes in the graph. The clustering coefficient of a node 'n' is defined as: the ratio of the number of edges present amongst n's neighbors to the number of all possible edges that could be present between the neighbours. Betweenness centrality for a node 'n', is defined as the sum over all pairs of nodes (u, v), the fraction of shortest paths from u to v, that pass through n. Mathematically, it is: $\sum_{u,v \in V \backslash n} \frac{\text{shortest\_paths}_n(u,v)}{\text{shortest\_paths}(u,v)}$. Intuitively, nodes that occur on many

[1] We consider strains Typhimurium, Typhi and Enteritidis

shortest paths between other vertices have higher betweenness than those that do not.

3. **Gene expression features**: The intuition behind this feature is that genes that are significantly differentially regulated upon being subject to *Salmonella*, are more likely to be involved in the infection process, and thereby in interactions with bacterial proteins. These features are derived using the gene of the human protein 'p_h' from the pair. We selected 3 transcriptomic datasets GDS77, GDS78, GDS80 from the GEO database [2], which give the differential gene expression of human genes infected by *Salmonella*, under 7 different control conditions. The 3 datasets give us a total of 7 features: the dataset GDS77 has two samples representing two conditions and gave 2 features; datasets GDS78 and GDS80 had time series gene expression with 3 and 2 control conditions respectively – the time series in each condition was averaged resulting in 3 and 2 features, respectively. All datasets reported log-ratios and did not require further normalization.

4. **RNAi expression**: Genome-scale RNA interference (RNAi) screening is used to identify host cell factors that promote or inhibit infection when the host gene is silenced. The set of host genes that are found to be important for infection are called "hits" and such hits have been published for *Salmonella* infection in humans [9]. Similar to the work in [14], we define two types of features using the hits from the RNAi experiment: "pathway-enrichment feature" and "complex-enrichment feature". The first feature is "on" for all human proteins that belong to a pathway that is statistically enriched by the RNAi hit genes. For computing enrichment we used the hypergeometric distribution and a p-value cut-off of 0.01. The second feature is similarly "on" for only human proteins which belong to statistically enriched protein complexes. The pathway membership information for each protein was downloaded from Pathway Commons database [3] and the human protein complex membership was obtained from CORUM database [11].

5. **Sequence based features**: Protein sequence similarity indicates the presence of similar structures and domains between the two proteins. We define two pair-level features that use this information: (1) sequence similarity between p_h and p_s and (2) sequence similarity between p_s and the human protein - p_h's binding partners in the human interactome. Sequence similarity of the two proteins was computed using PSI-BLAST. The feature takes a log of the reported alignment's e-value.

6. **PFam overlap**: This is a pair-level feature that computes the overlap between the protein family (PFam) of the proteins in the pair. This feature is the size of the intersection between the PFam assignment of p_h with that of the p_s. We downloaded the PFam information for each protein from the PFam database [7].

7. **Features from PFam and protein domain interactions**: Two pair-level features were computed using protein family interactions from the iPFam database [6] and protein domain interactions from 3DID database [13]. For a pair, the first feature counts how many of all the possible interactions between the PFam families of the two proteins are present in iPFam. The second feature counts how many of the interactions between the domain sets of a protein pair are present as domain-domain interactions in 3DID.

8. **Interolog based features**: This feature uses known interactions between proteins from other organisms to infer new interactions. It was derived using the interologs information from the BIANA database [8]. For a given pair 'p', if 'p_{hom}' a homologous protein pair involving any other organisms, BIANA uses the databases: BIND, DIP, IntAct to check if p_{hom} is an interacting pair. If yes, then p is an inferred interacting pair. For every pair 'p', this feature counts the number of homologous protein pairs p_{hom} that are interacting as per BIANA.

9. ***Salmonella* gene property: Is_effector feature**: This feature is a binary feature and simply indicates whether the *Salmonella* protein p_s is a known effector protein.

References

[1] Gene ontology. `http://www.geneontology.org/`.

Table 1: Summary of the various categories of features. *Feature category* indicates the data-source used to generate the feature. *Feature level* indicates whether the feature was (a) pair-level, (b) human gene level or (c) *Salmonella* gene-level.

Feature category	1	2	3	4	5	6	7	8	9
Feature level	(a)	(b)	(b)	(b)	(a)	(a)	(a)	(a)	(c)

[2] T Barrett, Troup DB, Wilhite SE, and Ledoux P et al. Ncbi geo: archive for functional genomics data sets 10 years on. *Nucleic Acids Res.*, 39(Database issue):D1005–10, 2011.

[3] E. G. Cerami, B. E. Gross, E. Demir, and I. Rodchenkov et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39(Database issue):D685–90, 2010.

[4] Z. Du, L. Li, Chin-Fu Chen, P. S. Yu, and J. Z. Wang. G-sesame: web tools for go term based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, 37(Web Server issue):W345–9, 2009.

[5] M.D. Dyer et al. The human-bacterial pathogen protein interaction networks of bacillus anthracis, francisella tularensis, and yersinia pestis. *PLOS One*, 5(8), 2010.

[6] R.D. Finn, M. Marshall, and A. Bateman. ipfam: visualization of protein–protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–2, 2005.

[7] R.D. Finn, J. Mistry, J. Tate, and P. Coggill et al. The pfam protein families database. *Nucleic Acids Res.*, 38(Database issue):D211–22, 2010.

[8] J. Garcia-Garcia, E. Guney, R. Aragues, J. Planas-Iglesias, and B. Oliva. Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, 11(56), 2010.

[9] B. Misselwitz, S. Dilling, P. Vonaesch, and R. Sacher et al. Rnai screen of salmonella invasion shows role of copi in membrane targeting of cholesterol and cdc42. *Mol Syst Biol.*, 7(474), 2011.

[10] T. S. K. Prasad, Goel R, Kandasamy K, and Keerthikumar S et al. Human protein reference database - 2009 update. *Nucleic Acids Res.*, 3(Database issue):D767–72, 2009.

[11] A. Ruepp, Waegele B, Lechner M, and Brauner B et al. Corum: the comprehensive resource of mammalian protein complexes - 2009. *Nucleic Acids Res.*, 38(Database issue):D497–501, 2009.

[12] S. Schleker, J. Sun, B. Raghavan, and M. et al. Srnec. The current salmonella-host interactome. *Proteomics. In press.*, 2012.

[13] A. Stein, A. Ceol, and P. Aloy. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, 39(Database issue):D718–23, 2011.

[14] O. Tastan. Prediction of host-virus protein-protein interactions. *PhD. Thesis*, 2011.

[15] O. Tastan, Y. Qi, and et al. J. G. Carbonell. Prediction of interactions between hiv-1 and human proteins by information integration. *Pac. Symp. Biocomput.*, 2009.