

## **UCL methods used for providing PPI predictions to the FuncNet meta-server (CODA, hiPPI and GECO).**

### **Introduction.**

We have implemented three methods to predict new protein-protein interactions (PPIs) which exploit different genetic and genomic signals, such as: homology to known PPIs (hiPPI), using domain co-occurrence (CODA) and gene expression similarity measures (GECO). All of them are based on standard, well established methodologies, extensively used in many other protein interaction prediction pipelines such as, for example, STRING [1]. The difference with other implementations is not due to the signal exploited but in the application of different scoring functions which give different performances in their prediction reliability. We have also exploited domains information rather than protein information based on the substantial experience of our group in determining and annotating protein domains (CATH and Gene3D [2,3]; [www.cathdb.info](http://www.cathdb.info)). We describe below each of the three individual methods and we demonstrate their performance in providing reliable PPI predictions in the whole human genome by validating them with a broad representation of gold standard PPI datasets. We also assess the performance of the Fisher integration of all the individual prediction methods:

### **Description of the Methods:**

**GECO (Gene Expression Comparison) Method:** Microarrays provide a high throughput approach for identifying functionally related proteins. A clear signal between microarray data and interacting proteins has previously been observed with the yeast dataset and the MIPS (MPACT) PPI dataset [4]. For human we use the E-TABM-185 compendium dataset of 6000 gcrma normalised HGU133-A affymetrix microarrays assembled by array-express [5]. A maximum of 5 values were allowed to be missing from a given genes expression profile, using the C-clustering libraries masking function. For the human hgu133a affymetrix chips 14,500 genes are well characterised giving a very large set of similarity scores.

**CODA (Co-Occurrence of Domain Analysis) Method:** CODA is based on domain fusion analysis. The aim of gene fusion methods is to infer protein-protein interactions or more generally functional associations between pairs of separate protein chains in a genome of interest whose orthologues have become fused in another species. Enright et al. (1999) and Marcotte et al. (1999) [6,7] were the first groups to introduce this approach. CODA uses a Multi-Domain Architecture (MDA) representation of proteins in complete genomes (target genomes) provided by Gene3D Multi-Domain Architecture datasets [3]. The Gene3D database contains protein sequences for all complete genomes with predictions for CATH [2] and Pfam [8] domains as well as functional annotations including GO. MDA CATH and PFAM datasets were created from 527 complete genomes (50 eukaryotes, 438 eubacteria and 39 archaea), CODA predictions were performed on these two (CATH and PFAM) generating the CODAcath and CODApfam datasets.

CODA scoring method: Here we consider how the method is implemented for a particular pair of proteins  $i = (p, q)$  in a query genome  $g$ .  $P$  is the set of domains in protein  $p$ .  $a \in P$  denotes that protein  $p$  contains a domain of superfamily  $a$ .  $J$  is the set of domain pairs  $j = (a, b)$  where  $a \in P$ ,  $b \in Q$ . In other words  $J$  consists of all the distinct pairs of domains between proteins  $p$  and  $q$ . It is also required that  $P \cap Q = \{\}$ , as the two proteins must not share any domains of the same superfamily.

To determine a fusion event we require that a target genome (one other than the query genome) contains a protein  $s$  where  $a \in S$  and  $b \in S$  i.e. domains which are separated in the query genome are found fused in the target genome. The set  $T$  comprises those genomes other than  $g$  which contain such proteins  $s$ . For a domain pair  $j$  in genome  $g$ , the fusion score  $C_j$  is taken as a maximum over all genomes in  $T$ :

$$C_j = \max_{t=1}^{|T|} \left( \frac{1}{n_{g_A} + n_{t_A}} + \frac{1}{n_{g_B} + n_{t_B}} \right) \quad (1)$$

Where  $|T|$  is the number of elements of set  $T$  (i.e. the number of target genomes),  $n_{g_A}$  and  $n_{g_B}$  are the frequencies of domain  $A$  and domain  $B$  respectively in genome  $g$  and  $n_{t_A}$  and  $n_{t_B}$  are the frequencies of domains  $A$  and  $B$  respectively in genome  $t$ . For a particular protein pair  $i$ , in query genome  $g$ , the maximum  $C_j$  is taken over all possible domain pairs  $j$ .

$$C_i = \max_{j=1}^{|J|} (C_j) \quad (2)$$

Where  $|J|$  is the number of elements in set  $J$  (i.e. distinct domain pairs). Thus  $C_i$  is the CODA score for proteins  $p, q$  (pair  $i$ ); the best (highest) score over all domain pairs between the proteins and over potential fusion proteins in all genomes (other than the query genome). The important novel aspect of this score is that it takes the maximum score over all the genomes whereas other methods do not consider target genomes individually. The score was chosen to reflect the uncertainty that fused domains and their unfused relatives are orthologues. The highest (best) possible score is 1, which is returned when there is only one example of each domain family in the query genome and one fused protein in a target genome, with no other domain homologues. In this case it is highly likely that the query protein domains are orthologous to the target protein.

**hiPPI (homology inherited Protein-Protein Interaction) Method:** The hiPPI method takes advantage of the Gene3D families of structurally conserved proteins as well as multiple sources of protein-protein physical interaction (PPI) data to reliably infer ('inherit') novel protein-protein interactions from homologues. hiPPI exploits the Gene3D protein families (G3D\_families) datasets [3]. These are families of proteins with similar multi-domain architectures generated using an automated, but conservative, clustering procedure. Interactions are only inherited between proteins belonging to the same Gene3D family, even though there may be recognisable sequence similarity with a

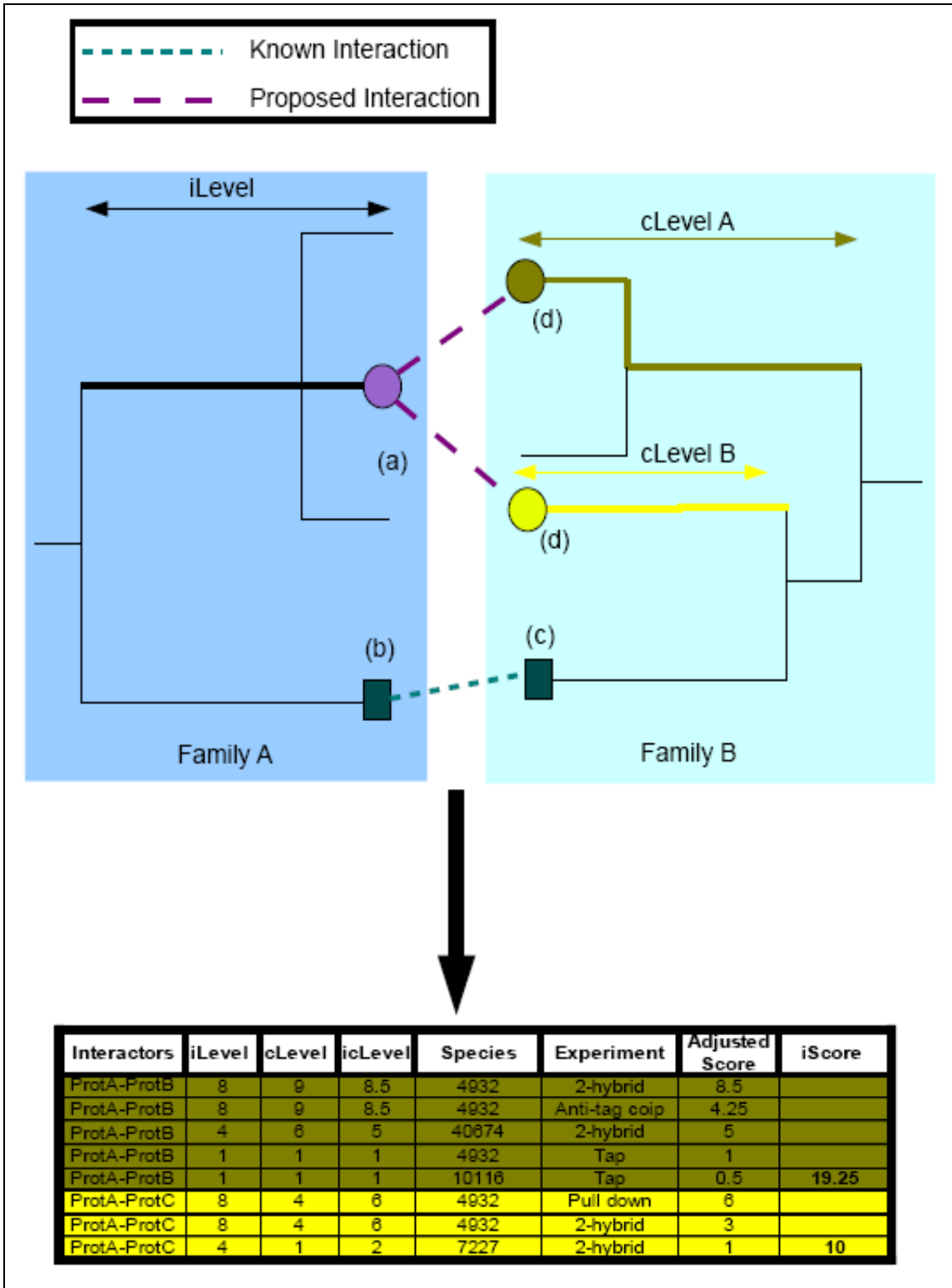
protein in another cluster. This step helps to reduce the amount of noise produced by attempting to inherit from overly-distantly related sequences.

The interaction dataset is formed from a merger of the PPI resources from MIPS, IntAct, HPRD and MINT protein-protein interaction datasets, obtained from the Gene3D database [3]. From each dataset the interacting proteins, their family, species, and experimental method was retrieved. Gene3D family codes consist of 11 elements. The first is the root family code, then the family is further sub clustered at 10 levels of sequence identity, termed S-levels (S30, S35, S40, S50, S60, S70, S80, S90, S95, S100).

hiPPI score (for a graphical description see Figure 1): For every test protein ('the inheritees'), all the relatives ('the inheritors') with interactions with proteins ('the complementors') with relatives in the inheritee's species ('the complementees') are identified. The inherited interaction ('inheritance') is that between the inheritee and the potential complementees. Known direct interactions were discarded. For each inheritance the similarity between the inheritee and the inheritor is measured by what Slevel they belong to, on a scale of 1 – 11 (1 is the family code and 11 is 100% identical); this is termed the 'iLevel'. Identically, the 'cLevel' is calculated for the complement. The two values are then averaged to create the 'icLevel'.

At this stage two alternative steps can be taken, and both are useful in different situations. The first assumes that if a protein interacts with one member of family then it is likely to at least show some affinity for another member in the same species. This can be considered biologically realistic, as the effect is seen in many genetic experiments (i.e. complementation tests). In this case all inheritances are counted. The second disregards this assumption in order to identify the probable biologically most important interaction. In this case, those inheritances with either a cLevel or an iLevel of 10 are disregarded. For the current study the former approach was used.

Since each protein-protein interaction can be inherited from more than one species, experimental method or iLevel, a summed score is created (the 'iScore') for each distinct pair of icLevels (NB iLevel = 10, cLevel = 8 is not the same as iLevel = 8, cLevel = 10). The first entry (non-redundant experiment) at that level contributes the full score of the icLevel. For subsequent entries at that icLevel if the experiment type or species is not new the score is halved; if neither is new but is a recombination of previously observed ones then the score is quartered. The final iScore is the sum of all the intermediate icLevel scores.



**Figure 1. The hiPPI approach.** (A) An example of identifying two potential interaction partners (labelled (d)) for the query protein (a). The interaction is inferred from the known interaction between (b) and (c), which are homologous to (a) and (d) respectively. Small example trees are shown for each family; each branch in the trees occurs at a particular family S-level (i.e. 80% sequence identity). The S-level in common between (a) and (b) is the iLevel, while the S-levels in common with (c) and the (d) s are the cLevels.

## Running the methods on the human proteomes.

The GECO, hiPPI, CODAcath and CODA pfam methods were run against all sequences in the human (*Homo sapiens*) proteome. Proteome file was downloaded from the Integr8 database [9]. GECO retrieved 26,292,126 protein pairs of predictions, hiPPI yielded 86,099 protein pairs of predictions; CODAcath yielded 32,259,881 and CODApfam generated 24,984,943.

**Integrating the prediction data:** The p-values from each method were calculated and integrated using two methods: Simple and Fisher [10] integration. The simple integration method was done by selecting the most significant prediction (lowest p-value) from all the prediction methods. While Fisher score was calculated with the formula:

$$-2 \sum_{i=1..n} \ln(p_i)$$

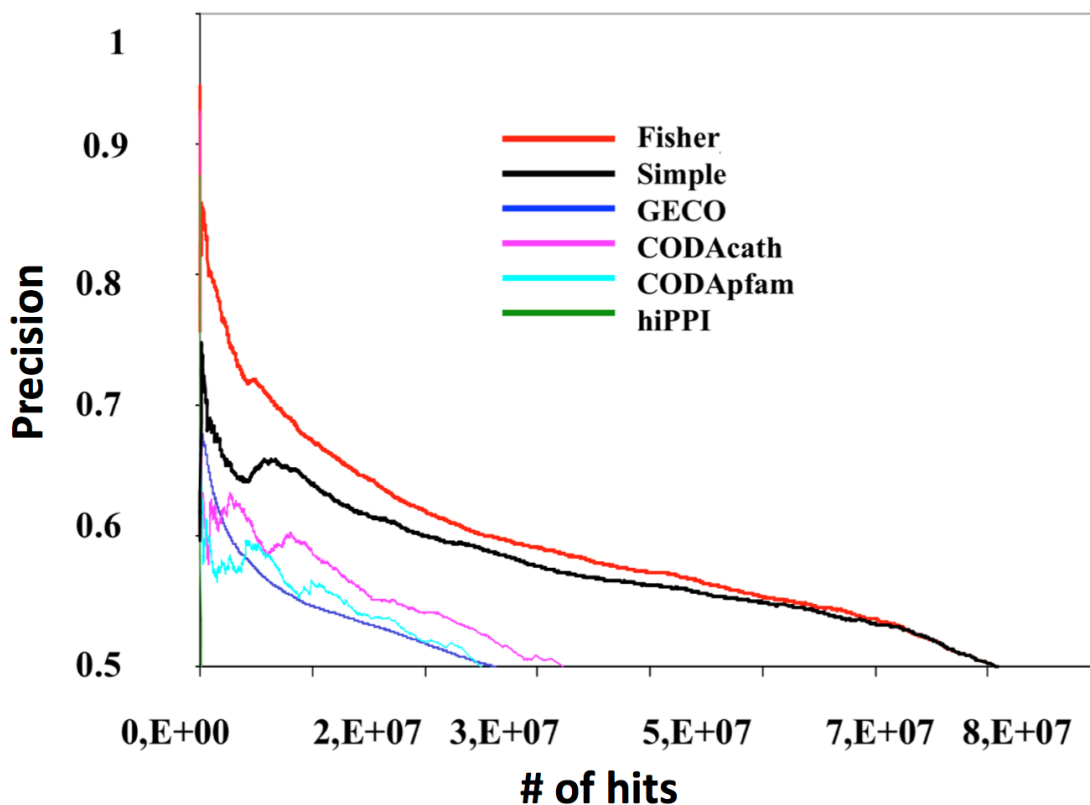
Normalised p-values ( $p_i$ ) were calculated based on the score distributions of the integrated methods (Simple and Fisher datasets). Simple and Fisher integration yielded both a total of 70,908,243 predicted pairs.

## Validation of the methods.

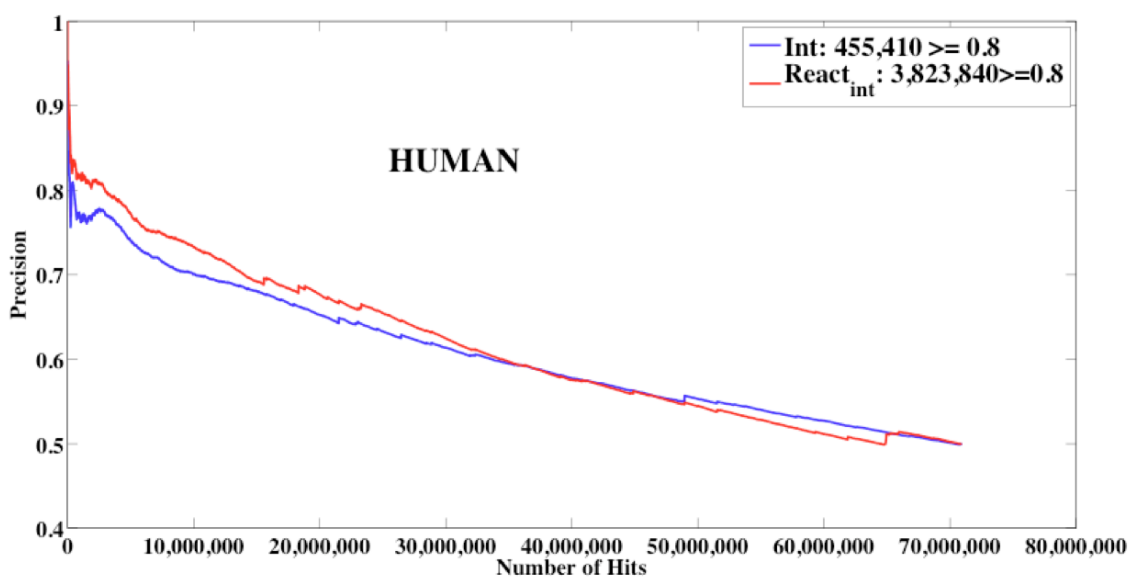
**We benchmarked our predictions using the following goldstandard datasets:**

- 1) **The GO Semantic Similarity refined dataset (Gossr):** We benchmarked our predictions using the highest quality annotations of the human proteomes in the Gene Ontology (GO) database [11]. The GO terms' Semantic Similarity (Goss) scores were calculated for all versus all protein pairs in human and yeast proteomes as described by Lord et al. 2003 [12], using the GO relational graph implicit in the GO ontology file (GO ontology files; OBO v1.0 format 30th-October-2008; <http://www.geneontology.org/>). Sets of protein pairs with significant Goss score (Goss  $\geq 5.0$ ; [13]) in the refined sets of GO annotations were selected as validating datasets
- 2) **The Int dataset:** Int dataset combines the interaction data from the HRPD, MINT, and Intact databases.
- 3) **The Reactome\_int dataset:** a dataset which contains the physical interactions annotated in the Reactome database.

**Precision and Recall calculation:** Precision was calculated as the ratio of accumulative TP/TP+FP at different prediction p-values, where TP (True Positives) is the number (#) of hits predicted within the validation dataset of true protein binary associations (e.g. GOSSr or Reactome\_int, see above), and FP (False positive) is the average # of hits predicted from 1000 random models of the same validation dataset. Recall is calculated as the accumulative number of predicted hits by a given method at different p-value levels.



**Figure 1.** Results of the benchmark studies for the individual prediction methods and the integrated Simple and Fisher methods using the GOSSr goldstandard dataset. Plot shows precision (y-axis) versus recall (# of hits, x-axis).



**Figure 2. Results of the benchmark for Fisher integration method using the Int and Reactome\_int goldstandard datasets.** Plot shows precision (y-axis) versus recall (# of hits, x-axis) obtained in the validation of human predictions using the Int (blue line) and Reactome\_int (red line) gold standard datasets. The boxes highlight the number of predictions retrieved with precision  $\geq 80\%$ .

We found that all single and integrated methods p-values correlate inversely with the precision scores in all the validations performed, as expected if functional relationship (GOSSr; see Fig. 1) and physical interaction (Int and Reactome\_int; see Fig.2) information is linked to the prediction p-value scores.

Fisher PPI predictions with p-values  $\leq 0,014$  reaches precision  $\geq 80\%$  in the GOSSr validation, precision  $\geq 76\%$  in the Int validation and precision  $\geq 82\%$  in the Reactome\_int validation.

Although precision calculations of the same Fisher integrated dataset vary depending on the gold standard used in the validation, we can clearly conclude from all the validation analyses we have performed that: 1) Fisher scores are linked to genuine physical and functional protein-protein relationships performing better than any single or Simple combination methods; 2) the Fisher PPI predictions with p-values  $\leq 0,014$  show consistent reliability in all the validations performed with precisions around 80%. 3) p-values correlate with precision allowing us to select Fisher predictions at different levels of reliability.

## Comparison against STRING prediction datasets.

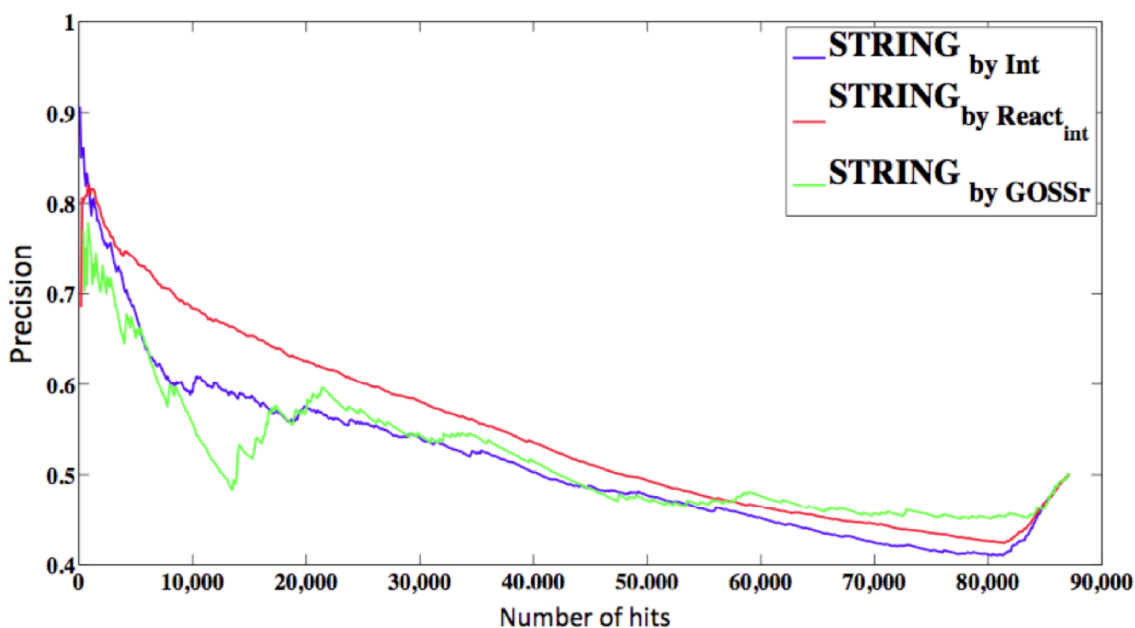
We have compared our Fisher predictions against the Fisher integration of similar methods from STRING [1]. STRING comprises, amongst other methods, four *ab-initio* prediction methods similar to our methods. STRING is a continuously updated and well-known resource for predicting protein interactions, making it a suitable gold standard to compare our integrated prediction pipeline against. STRING integrates the results from several methods including the following *ab-initio* predictions based on genome and gene expression comparison: gene neighbourhood, gene fusion (comparable to CODA), gene co-occurrence, and gene co-expression (comparable to GECCO). STRING does not include a prediction algorithm similar to hiPPI, but instead STRING implements a phylogenetic profiling method (gene co-occurrence) and a gene genome co-localization method (gene neighbourhood). In any case, hiPPI predictions represent a small percentage (around 0,1%) of the total predictions integrated by Fisher in our work, and therefore with little influence on the overall increase in prediction power observed.

The four individual *ab-initio* PPIs prediction datasets (neighbourhood, fusion co-occurrence, and co-expression) for the human complete proteomes were downloaded from the STRING server and integrated by Fisher method following the same protocol used in our work. The total number of predictions from STRING is significantly smaller than the total number of UCL\_FuncNET predictions indicating a much lower coverage. (Table I).

	Human	
Method	STRING	UCL_FuncNET
Total <i>ab-initio</i> predictions	87,102	70,908,243
Neighborhood	18,391	-
Fusion	3,943	32,259,881 (CODAcath) 24,984,943 (CODApfam)
co-occurrence	20,348	-
coexpression	49,382	26,292,126 (GECO)
hiPPI	-	86,099

**Table I. Comparison of STRING and UCL\_FuncNet predictions in human.** First column: methods. Following columns: number of predictions retrieved by each *ab-initio* method independently for the STRING and UCL\_FuncNet datasets.

Since a given method can predict large numbers of predictions without any functional information necessarily being associated with them, the total number of predictions is not a good indicator of the methods' performance in itself, but a useful measure is the number of accurate predictions provided above a significant precision level (e.g. 80% precision). Therefore, STRING predictions were validated with the same GOSSr, Int and Reactome\_int gold standard datasets used to validate the Fisher UCL\_FuncNet predictions in our work. The scores from the STRING Fisher predictions show correlation with precision (see Fig. 3). However, the Fisher integration's prediction power for STRING predictions is shown to be much lower than Fisher UCL\_FuncNet predictions. In the human validation, STRING *ab-initio* methods show poor performance. STRING human validation shows 100 predictions with precision  $\geq 80\%$  in GOSSr validation, 550 and 650 predictions in the Int and Reactome\_int validation, while the comparable figures in the Fisher UCL\_FuncNet human validations are 1,052,579 , 455,450, and 3,823,840 in the GOSSr, Int and Reactome\_int respectively with precision  $\geq 80\%$  (see and compare Figs. 3 vs. Figs. 1 and 2). Fisher UCL\_FuncNet datasets clearly outperforms the integrated STRING datasets.



**Figure 3. Validation of the Fisher integration of the STRING ab-initio prediction datasets in yeast and human.** *STRING\_Fisher* predictions were validated with *GOSSr*, *Int* and *Reactome\_int*. Plot shows Precision (y-axis) versus number of predictions – Recall- (x-axis).

## References:

1. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37: D412-D416.
2. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35:D291-297.
3. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, et al. (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res* 36:D414-418.
4. Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* 29:3513-3519.
5. Parkinson H, , Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35:D747-750.
6. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86-90.
7. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751-753.
8. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36: D281-288.
9. Pruess M, Kersey P, Apweiler R (2005) The Integr8 project; a resource for genomic and proteomic data. *In Silico Biol* 5:179-185.
10. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, et al. (2005) A data integration methodology for systems biology. *Proc Natl Acad Sci USA* 102:17296-17301.
11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25-29.
12. Lord PW, Stevens RD, Brass A, Goble CA (2003) Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocompu* :601-612.
13. Ranea JA, Yeats C, Grant A, Orengo CA (2007) Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS Comput Biol* 3:e237.