In silico discoveries for biomedical sciences

Haagen, H. van

Chapter 4

In silico discovery and experimental validation of new protein-protein interactions

Herman H.H.B.M. van Haagen, Peter A.C. 't Hoen, Antoine de Morrée, Willeke M.C. van Roon-Mom, Dorien J. M. Peters, Marco Roos, Barend Mons, Gert-Jan van Ommen, Martijn J. Schuemie

Manuscript accepted to Proteomics at November 25, 2010

Abstract
We introduce a framework for predicting novel protein-protein interactions (PPIs), based on Fisher's method for combining probabilities of predictions that are based on different data sources, such as the biomedical literature, protein domain and mRNA expression information. Our method compares favorably to our previous method based on text-mining alone and other methods such as STRING. We evaluated our algorithms through the prediction of experimentally found protein interactions underlying Muscular Dystrophy, Huntington's Disease, and Polycystic Kidney Disease, which had not yet been recorded in protein-protein interaction databases. We found a 1.74 fold increase in mean average prediction precision for dysferlin and a 3.09 fold for huntingtin when compared to STRING. The top 10 of predicted interaction partners of huntingtin were analysed in depth. Five were identified previously, and the other five were new potential interaction partners. The full matrix of human protein pairs and their prediction scores is available for download. Our framework can be extended to predict other types of relationships such as proteins in a complex, pathway or related disease mechanisms.

Introduction
The biomedical literature and domain-specific databases contain a wealth of background information, which should aid biomedical researchers in the design and interpretation of their experiments. Many databases compile information from several resources for use as reference and lookup. Databases such as KEGG, STRING, and IntNetDB are examples that are useful for studying protein-protein relations. These resources represent existing knowledge well. However, of particular interest is the potential to reveal genuinely novel relations by data mining algorithms [1]. In previous work we showed the ability to predict protein-protein interactions using information contained in literature alone [2]. With the so called concept profile technology, we found novel protein interaction pairs that could not have been found by a simple MEDLINE query. This was illustrated by the prediction of the physical interaction between calpain 3, which causes a form of muscular dystrophy, and parvalbumin B, which is found mainly in skeletal muscle. However, this method does not exploit the full potential of information available for data mining. Combining data sets beyond literature may increase coverage and the reliability of our predictions.
Combining data from different sources for extracting relevant knowledge is a general objective in bioinformatics. Here, we distinguish data concatenation and evidence score combination. Data concatenation merely summarizes the results of queries to a number of individual databases (*e.g.* www.genecards.org [3]). The summaries are provided to an investigator for interpretation. When investigating PPIs, a summary may contain information on the presence of certain PPIs in a curated PPI database, Gene Ontology terms that are shared, and co-expression of

genes in certain tissues or cellular compartments. No additional algorithms are provided to predict and highlight putatively novel relations.
Evidence score combination provides a score to order the information from a combination of data sets. For each set, the score reflects the contribution of the set to the overall result of a query across a number of data sets. The individual scores are combined using one of several combination techniques. Evidence score combination can be used to predict new relationships between biological concepts, including protein-protein interactions (PPIs).
Several web tools are available that provide some form of data integration and evidence score combination for the extraction of PPIs [4-6]. (i) STRING[6], which is maintained by EMBL, contains functional associations for over 600 species. STRING uses information on genomic content, high throughput experiments, co-expression, and co-mentioning in PubMed abstracts and recorded in public curated databases like KEGG or Reactome. STRING uses a combination technique based on the product of p-values to provide a confidence score for predicted PPIs. (ii) FunCoup[5] provides a predicted protein-protein network for eight eukaryotes. It uses information on PPIs, mRNA expression, sub-cellular co-localization, phylogenetic profiles, miRNA-mRNA targets, transcription factor regulation, protein expression, and protein domain interactions. The network is optimized using a Bayesian approach. (iii) IntNetDB v1.0.[4] is restricted to a few species and mainly focuses on human data. IntNetDB uses physical interactions, phenotype similarity, genetic interactions, shared GO annotation, domain-domain interactions, co-expression, and gene context in PubMed articles. IntNetDB uses a Naive Bayes classifier as combination technique. As stated previously, these web tools perform well on reproducing existing PPIs. STRING for instance aggregates known interactions from several databases and predictions made by several predicting methods. Their evidence score reflects how well supported an interaction or association is by these sources. Our aim is to develop a true interaction predictor, and a score that reflects the likelihood that the prediction is true. In contrast to STRING, our method will predict known as well as unknown interactions that can have equally high scores. The correspondence with known protein-protein interactions validates our approach.
The remainder of this article is structured as follows. We give a brief introduction of our framework which is based on Fisher's method for combining p-values based on different data sources. Next our framework was validated by evaluating three show cases. The first case is on dysferlin (*DYSF* encoded protein), its deficiency causing progressive Limb Girdle Muscular Dystrophy type 2B. We aimed for the discovery of dysferlin interaction partners by immunoprecipitation experiments and show how well we could predict these new interactions. The second case relates to the huntingtin protein which is associated with Huntington's disease. We took PPIs from the article by Kaltenbach et. al. [7], which had not been stored (yet) in PPI

databases and which had not been described in MEDLINE abstracts. They serve as a good test set where we simulate that our framework is able to predict proteins from these lists. The last show case is on Polycystic Kidney Disease caused by the mutated *PKD1* gene. This case illustrated how to solve homonym problems encountered in text by including additional expression data. We end with summarizing the results for each show case and conclude that we significantly improve the discovery of novel PPIs over previous methods.

Materials and methods

Performance measurement

For measuring performance we used receiver operating characteristics (ROC) curves and the area under the curve (AuC). Second, we used the mean average precision (MAP). Both are measurements often used in information retrieval. In the case studies the test sets used for dysferlin and huntingtin are labeled as positive instances. The rest of the proteins in our ontology are labeled as negative instances. The AuC values have a range between 0.5 and 1. 0. A value of 0.5 means that the systems is no different than a random ordering of the samples, *i.e.* the positive instances are equally distributed over the ordered list (ordered by match score) of all proteins. An AuC of 1 mean the system is a perfect predictor, *i.e.* al positive instances first rank at the top followed by all negative instances.

The mean average precision is a measurement more sensitive to samples size of both the positive and negative set. The MAP is calculated by averaging all precisions where each precision is calculated at the occurrence of a positive instance in an ordered list (ordered by match score).

Match scores for each individual database.

Concept profiles

To calculate the similarity of the contexts in which proteins appear in literature, we summarize the context of each protein in a concept profile. This profile for a protein contains all concepts that are co-mentioned with the protein as found in MEDLINE abstracts. To find concepts in text we have used the concept-recognition software Peregrine [8], which includes synonyms and spelling variations of concepts and uses simple heuristics to resolve homonyms. For this, Peregrine uses a protein ontology that was constructed by combining several gene and protein databases. Proteins from different species are fused together and we do not distinguish between a gene and a protein.

Each concept in the profile is assigned a weight. The weight reflects the strength of the association between the concept and the protein. The concepts that appear in both protein profiles are used to calculate a match score. The match score is the

inner product calculated over the weights for the shared concepts. For a detailed description of concept profiles and weight calculation we refer to [9].

Gene Ontology

Match scores defined for the Gene Ontology were investigated by Mistry *et al.* [10]. They compared the term overlap with other well known similarity measures adapted from the work of Resnik[11], Lin[12], and Jiang[13]. We did a ROC curve analyses on all four similarity measures and obtained the highest AuC value for the method by Resnik. The score we use is inferred from Resnik. Resnik originally defines the score to find the similarity between two GO terms, whereas we want to find the similarity between two proteins. First the information content for a GO term t_i is defined

$$IC(t_i) = -\log(p(t_i))$$

where $p(t_i)$ is the probability of a gene being annotated to that term. $p(t_i)$ can be calculated as follows

$$p(t_i) = \frac{\#genes\_annot(t_i)}{\#genes\_annot(rootnode)}$$

In words, the number of genes annotated to GO term t_i divided by all the genes under consideration. All the genes are annotated in the root node of the GO graph. The information content of the root node therefore is 0 as would be expected. Resnik's similarity measure is then calculated by taking the IC of the lowest common ancestor (LCA) shared between two proteins.

$$sim_{Resnik}(p1, p2) = IC(LCA)$$

With *p1* and *p2* the two proteins that form a pair (either random or a PPI).

Microarray data

Microarray co-expression values are pre-calculated for COXPRESdb [14] and can be used directly after download. For Gene Atlas [15] the human GNF1H chip is used. First, the log was taken from the MAS5.0 normalized expression values for each tissue (78 in total), and probes with the same EntrezGene IDs were averaged. Subsequently, a Pearson correlation was calculated for the gene expression values for all pairs of genes.

Tissue specificity

TiGER[16] contains expressed sequence tags that are defined for 30 tissues. For TiGER we evaluated a number of vector similarity measures namely, Pearson's correlation coefficient, inner product, cosine, euclidean distance, and the Tanimoto coefficient. The latter one showed the best prediction performance. The Tanimoto coefficient between two vectors A and B is defined as follows:

$$T(A,B)=\frac{A\cdot B}{\|A\|^2+\|B\|^2-A\cdot B}$$

Tanimoto coefficient values > 0.85 are generally considered similar to each other.

Domain-domain interactions

We used InterPro[17] to annotate each protein in our ontology with domains. Subsequently we used DOMINE[18] to determine which domains (one of protein A and the other of protein B) interact. The final score is simply the number of interacting domains.

Probable non interacting protein pairs

A null hypothesis was generated by choosing random protein pairs[19]. This null hypothesis is used to calculate a single sided p-value for Fisher's Method. The only constraint that we applied is that the protein pair should not be in a curated database nor in the high-throughput database IntAct [20]. The curated databases used are listed in the supplementary files. The complete random protein pair set consisted of over 500 millions proteins pairs (all possible combinations of two proteins). For computational reasons our analysis was limited to a random subset of 100,000.

Combined match score: Fisher's method

Fisher's method combines one sided p-values from different databases into one test statistics which follows a χ^2 distribution with $2*L$ degrees of freedom using the formula

$$\chi^2=-2\sum_{i=1}^{L}\log(p_i)$$

When the p-values tend to be small, the test statistic χ^2 will be large. The p-values are obtained from the random protein pairs distribution described earlier.

In the first version of Fisher, missing values for this combiner are also completely ignored. This is done by setting the p-value to 1. The log becomes 0 and the missing value does not contribute to the score. The degrees of freedom are fixed and are the same for each sample (a protein pair). The second version takes into account the degrees of freedom (dof). The dof is only taken for databases that have a match score. The last two variations are where the individual database scores are weighted. They are weighted with the AuC and MAP values and then the previous formula is applied. Fisher's method can be sensitive to databases if p-values become 0. Then the combined score is dominated by one database only. This could result in false positives. We added a small offset to each p-value of 10^{-4} to filter for this side effect when p-values are too small (or 0).

STRING

We benchmarked our system against the STRING database. We downloaded STRING version 8.1 that was last updated on October 18, 2009. A current version of STRING can be found online: http://string.embl.de. The databases used by STRING are:

- Neighborhood in the genome (nscore)
- Gene fusion (fscore)
- Co-occurrence across genomes (homology; pscore and hscore)
- (Co-expression (ascore)
- Experimental/biochemical data (escore),
- Association in curated databases (dscore)
- Co-mentioned in PubMed abstracts (tscore; text-mining based on direct co-occurrences)

String uses a combiner based on the product of probabilities using the following formula

$$S = 1 - \prod_{i}^{N} (1 - S_{i)}$$

With *Si* the probability score for database *i*, *S* the combined score, and *N* the total number of databases to be combined.

Dataset

The raw scores for each database, and the combined Fisher Method score, are merged together in a tab delimited text file which can be downloaded from our website http://www.biosemantics.org/ppi-prediction

Results and Discussion

Our previous approach used only text-mining for the prediction of PPIs [2]. We postulated that a combination of information indicative of protein-protein associations, such as co-expression and functional and structural similarities, increases the overall probability of a genuine PPI. Therefore, we included information from these five additional databases:

- Gene Ontology: manual functional annotation
- COXPRESdb[14]: mRNA co-expression over a wide range of conditions
- Gene Atlas[15]: mRNA co-expression in 78 tissues
- Tiger[16]: expressed sequence tags (EST) counts in 30 tissues
- InterPro/DOMINE[17, 18]: domain annotation and domain-domain interactions

Our hypothesis was that these data sources are valuable for the prediction or protein interactions since two interacting proteins should be expressed in the same tissue and cell, are likely to be co-regulated at the transcriptional level, and interact via a specific combination of protein domains. The selected databases are publically available and have suitable data formats for processing (xml, tab delimited files, Entrez Gene or Uniprot accession numbers, etc). For each database a score was defined that reflects a degree of association between two proteins. Individual scores were then combined to obtain a final score for a protein pair.
For combining the score we used a method developed by Fisher (see Materials and Method for detailed explanation of this method). This method is based on combining p-values taken from different predictions. Briefly, the match score for every database is converted into a single sided p-value. Then, the p-values are log transformed and summed resulting in a Fisher score with 2*N degrees of freedom (N the number of p-values to be summated). We made two variations of this Fisher method. In the first one, the degrees of freedom are fixed (missing values taken into account). In the second one, each p-value is weighted with AuC or MAP values, giving more weight to the data sources that are most important for the prediction. The AuC stands for Area under the ROC curve and MAP for Mean Average Precision. Both measures are well known in the field of information retrieval and data-mining. An AuC of 0.5 reflects a prediction with random behavior (like flipping a coin). An AuC of 1 correlates to a perfect prediction. MAP values range from 0 to 1 (perfect prediction). All performance measures (AuC and MAP values) are given in the supplementary files.
We choose Fisher's method after evaluating three other methods for combining databases (see supplementary files for the other methods and the evaluation). Fisher's method showed the best overall results both in AuC and MAP.
In the analysis we will benchmark our system against STRING. STRING is a web tool that has been intensively optimized and updated since 2000. It enables downloading of previous releases. STRING has the same approach for predicting PPIs, *e.g.* it defines evidence scores for several databases and combines them into a single score.

Example 1: Predicting proteins interacting with Dysferlin (*DYSF*, MIM: 603009)
Dysferlin is a 230 kDa C2-domain containing transmembrane protein. Dysferlin is highly expressed in skeletal muscle, but is also found in other tissues such as kidney, heart and monocytes. Mutations in dysferlin cause progressive muscular dystrophies like Limb Girdle Muscular Dystrophy type 2B (MIM: 253601), Miyoshi Myopathy (MIM: 254130) and Distal Anterior Compartment Myopathy (MIM: 606768), collectively referred to as dysferlinopathies [21]. From cellular studies it is known that dysferlin participates in membrane repair. Cultured

myotubes show a calcium-dependent accumulation of dysferlin at sites of membrane damage upon laser-inflicated membrane wounding [22]. In absence of calcium or dysferlin the muscle fiber cannot repair the damage, and undergoes necrosis [22].
We performed a high-throughput screen for proteins interacting with dysferlin and evaluated whether our PPI prediction algorithm could predict dysferlin's experimentally identified interaction partners. To date, nine physical interaction partners were described in literature, and all are believed to aid dysferlin in its membrane repair function. However, it is not completely understood how dysferlin functions, and possibly it does more than membrane repair alone.
We have developed a specific, robust and reproducible immunoprecipitation (IP) method to isolate dysferlin protein complexes from biological sources ([23], de Morrée et al in preparation). We *in vitro* differentiated mouse myoblasts to spontaneously contracting myotubes, and immunoprecipitated dysferlin protein complexes. Mass spectrometry analysis yielded a list of 352 putative interaction partners (manuscript in preparation), including the previously described *ANXA2, AHNAK, CAPN3, TRIM72* encoded proteins, underlining the validity of the method. The proteins already known to interact with dysferlin (recorded in a database) were omitted from this IP list. We created a prioritized list of 25,036 proteins with our Fisher combiner, by matching dysferlin against all other proteins known in our ontology, and compared the IP dataset with this list. Figure 1a shows that text-mining yields a high AuC of 0.78, indicating that implicit information contained in the literature is able to correctly predict interaction partners for dysferlin. As shown in figure 1a the other nine databases yield AuC's between 0.6 and 0.7, and as a result the Fisher combiner AuC does not differ much from text-mining alone. Thus, most predictive value is contained in text and to a lesser extent in gene expression and Gene Ontology. STRING gives an AuC of 0.63, close to random behavior, confirming that our system performs better than STRING. The MAP reflects how many IP partners are present in lists of predicted proteins, a useful measure for those interested in validation of candidates. In figure 1b the MAP are plotted for the IP partners. The MAP achieved by the AuC weighted Fisher combiner was 1.74 fold better than STRING's. Again, literature had the highest predictive value, and the addition of other databases to the prediction led to only small improvement in precision. Finally, we evaluated how many dysferlin interaction partners from the IP list were found in the top 50 of predicted interaction partners. As shown in table 4, the Fisher combiner yields 9 hits, whereas STRING finds only 6. The top 50 of predicted proteins are given in Supplementary table 6.

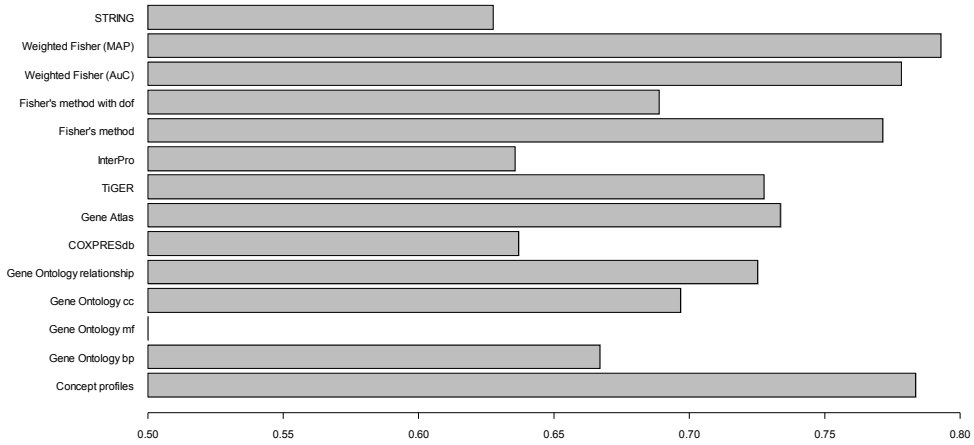

Figure 1a. AuC values (ranging from 0.5 till 1) for the individual databases, the Fisher methods, and STRING, for the dysferlin case study.

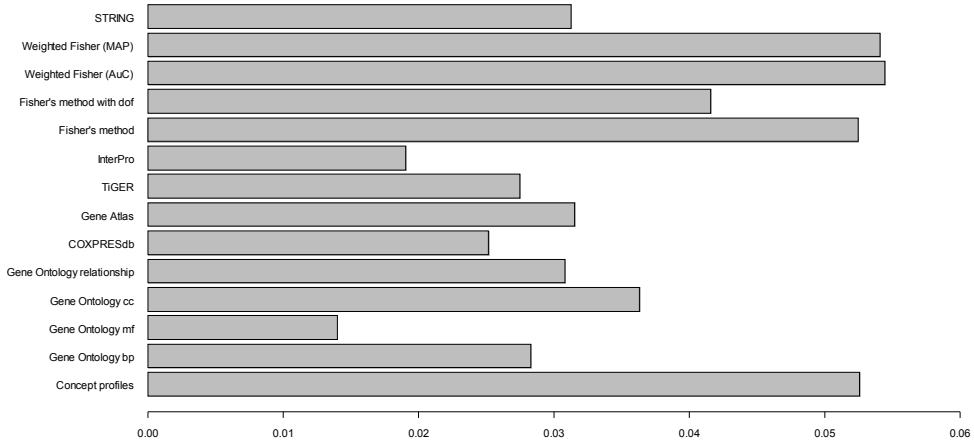

Figure 1b. MAP values for the individual date sources, the Fisher methods, and STRING, for the dysferlin case study.

Example 2: Predicting proteins interacting with Huntingtin (*HTT*, MIM: 613004)

Huntington's disease (HD, MIM: 143100) is a progressive autosomal dominant neurodegenerative disorder that is caused by a CAG repeat expansion in the *HTT*

gene, which results in an expansion of polyglutamines at the N-terminal end of the huntingtin protein, and the accumulation of cytoplasmic and nuclear aggregates in neurons. The polyglutamine expansion in the protein plays a central role in the disease and the size of this expansion has a direct link to the aggregation-proneness as well as the severity of pathology and clinical features [24]. When the mutation for HD was found, huntingtin was a protein of unknown function but extensive research over the past decade has revealed numerous functions for huntingtin and many cellular processes are affected in HD, such as transcriptional de-regulation, mitochondrial dysfunction, and vesicle transport dysfunction [25]. Although the precise underlying disease mechanism of HD is still unknown there is evidence to support a role for aberrant protein-protein interactions in HD pathogenesis [26].

A recent study by Kaltenbach *et al*. [7] identified a comprehensive set of huntingtin-interacting proteins. (1) With yeast two-hybrid screening (Y2H) 104 interacting proteins were identified and (2) affinity pull down followed by mass spectrometry identified 130 proteins. Subsequently, Kaltenbach *et al*. tested if the interacting proteins they had identified could influence mutant huntingtin toxicity. (3) An arbitrary sample of 60, out of the 234, proteins were tested in either over-expressing or partial loss of function *Drosophila* strains expressing the first 336 amino acids of the huntingtin protein containing an expanded 128 glutamines.

For the current study, the already known interacting proteins were omitted from these three datasets to serve as a test panel to examine if our framework can predict proteins from these lists, leaving 92 proteins from the Y2H experiment, 120 from the pull down experiments, and 42 from the Drosophila huntingtin-induced neurodegeneration. With our Fisher method (figure 2b), we obtained a MAP of 0.025 for the Drosophila interaction partners. This is a 3.09 fold increase compared to STRING. The Y2H, and IP experiments showed a 1.48, and 2.56 fold increase over the STRING method respectively. The top 50 of predicted proteins out of 25,036 proteins, are shown in table 3.

From the top 50 proteins identified by our system, 3 proteins namely syntaxin 1A (*STX1A* encoded protein), catenin beta 1 (*CTNNB1* encoded protein) and adaptor-related protein complex 2 (*AP2A1* encoded protein) were in the group of 60 that we-re tested in the *Drosophila* model (compared to 0 by STRING, table 1), and all three were confirmed to modify phenotype, validating that these PPIs are functional.

The interaction between huntingtin and syntaxin 1A has been proposed previously (PMID: 16162412) but the direct interaction between catenin beta 1 and huntingtin was a novel prediction in the Kaltenbach paper that was also high in our list (rank 16) of potential interacting proteins. This protein shows no co-occurrences in MEDLINE abstracts with the huntingtin protein (also not in STRING), but it has been reported in some papers that beta catenin overexpression protects cells from poly(Q) toxicity (PMID: 12097329). AP2A1 is part of the adaptor protein 2 (AP-2)

complex found in clathrin coated vesicles . Although AP2A1 has never been associated with HD previously, the AP-2 complex is involved in the clathrin mediated endocytosis of GABA(A) receptors (PMID: 17690529) and GABA(A) receptors are present on the class of striatal GABAergic neurons that are affected in Huntington's disease[27].

There are 5 proteins out of the top 10 predicted interacting partners for huntingtin that are new potential huntingtin-interacting proteins:

(1) **Platelet-activating factor acetylhydrolase 1b**, regulatory subunit 1 (*PAFAH1B1* encoded protein) inactivates Platelet-Activating Factor (PAF) by removing the acetyl group at the sn-2 position. It is required for induction of nuclear movement and control of microtubule organization[28]. *PAFAH1B1* is also known as *LIS1* [29]and deletions in *LIS1* cause Lissencephaly, a disorder of neuronal migration[30]. A possible link to HD might lie in the fact that PAF induces Clathrin-Mediated Endocytosis [31], which is a common pathway used by G protein-linked receptors to transduce extracellular signals. Both huntingtin interacting protein 1 (*HIP1* encoded protein) and huntingtin interacting protein 1 related (*HIP1R* encoded protein) have been implicated in this process (see below).

(2) **Adenomatous polyposis coli protein** (Protein APC or FPC, ranks position 7 in table 3) is a tumor suppressor protein that acts as an antagonist of the Wnt signaling pathway and has a role in regulating microtubules and actin in polarized epithelia [32]. The *APC* gene is highly expressed in the embryonic and postnatal developing brain. In addition, APC is present in astrocytes, although its role in astrocytes is, as yet, unknown [33].

(3) **Metabotropic glutamate receptor 3** (*GRM3* encoded protein) is an interesting protein because it has been implicated in Huntington's Disease (contributes 22.21% to the concept profile score, PMID: 9600992) while there was no evidence found in STRING (http://string.embl.de/). There is convincing evidence showing that glutamate-mediated excitotoxicity plays a role in HD pathology [34, 35] but there have been no reports to our knowledge directly implicating mGluR3 in HD.

(4) **Vesicle-associated membrane protein-associated protein B** (*VAPB* encoded protein) is a protein that plays an important role in protein folding [36]. To function efficiently, the endoplasmic reticulum relies on a system that detects a buildup of unfolded or misfolded proteins. The cell's response to prevent or correct this buildup is called the unfolded protein response. VAPB is implicated in the autosomal dominant adult-onset form of Amyotrophic Lateral Sclerosis 8 (*ALS8* encoded protein) and in this disease cytosolic aggregates were present in all cell types examined, including mouse and human nonneuronal cells[37]. Protein aggregates can

impair the ability of cells to function normally and huntingtin aggregates are a hallmark of HD[38].

(5) **The GABA(A) receptor-associated protein** (*GABARAP* encoded protein) protein clusters neurotransmitter receptors by mediating interaction with the cytoskeleton[39]. Although there were no co-occurrences for GABARAP and STRING did not find any functional links between GABARAP and huntingtin, it is highly likely that this protein is involved in HD, since GABA(A) receptors are present on the class of striatal GABAergic neurons that are affected in Huntington's disease[27].

Of the top 10 predicted interacting partners for huntingtin, there are 5 proteins that have been identified previously:

(1) **Syntaxin1A** (*STX1A* encoded protein) was identified by Kaltenbach *et al.* and when tested in an HD fruitfly model, STX1A influenced the phenotype [7]. Previous studies have shown that huntingtin enhances calcium influx by blocking STX1A inhibition of N-type calcium channels[40, 41].

(2) **Solute carrier family 1** (glial high affinity glutamate transporter) member 2 (*SLC1A2* encoded protein) was also identified by Kaltenbach *et al.* SLC1A2 is also called glutamate transporter 1 (GLT1). It is a membrane-bound protein that is the principal transporter clearing the excitatory neurotransmitter glutamate from the extracellular space at synapses in the central nervous system and was found to be increased in HD[42, 43].

(3) **Microtubule-associated protein tau** (*MAPT* encoded protein) promotes microtubule assembly and stability, might be involved in the establishment and maintenance of neuronal polarity. Tau is involved in several neurodegenerative disorders such as Alzheimer's disease (AD) and although AD and HD are both protein aggregation disorders, Tau has never been documented to interact with mutant huntingtin. However, it was recently suggested that the level of tau phosphorylation could limit the severity and/or progression of HD[44]. The tau protein in most cases could not be detected by our text-mining algorithm or by STRING resulting in no co-occurring hits with huntingtin. However this problem is solved by intermediate concepts that relate huntingtin with tau (Neurodegenerative Disorders, and Nerve Degeneration).

(4) **Dopamine receptor D2** (*DRD2* encoded protein) is a G-protein-coupled receptor that inhibits adenylyl cyclase activity. In HD there is a major loss of DRD2 binding in the caudate nucleus, putamen and globus pallidus externus[45].

(5) **Huntingtin interacting protein 1 related** (*HIP1R* encoded protein) has a role in clathrin-mediated endocytosis (CME)[46]. It binds to huntingtin interacting protein 1 (*HIP1* encoded protein) and links actin to clathrin[47].

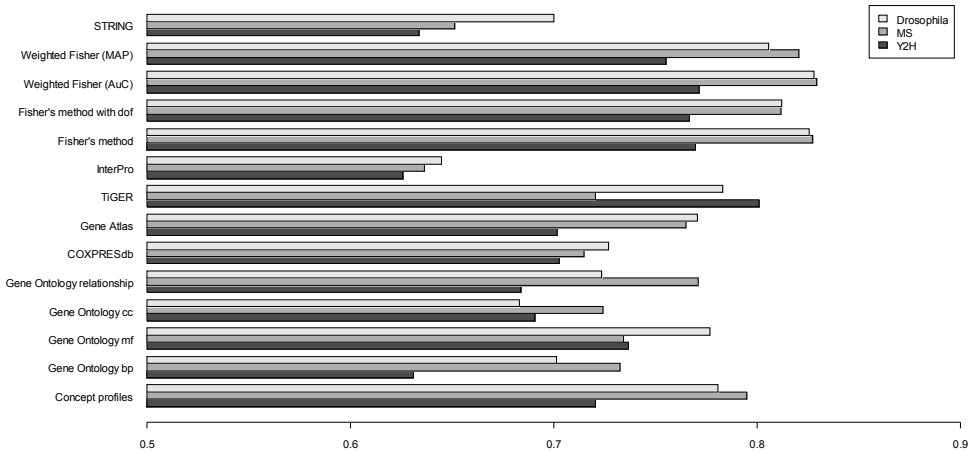

Figure 2a. AuC results for the huntingtin case study.

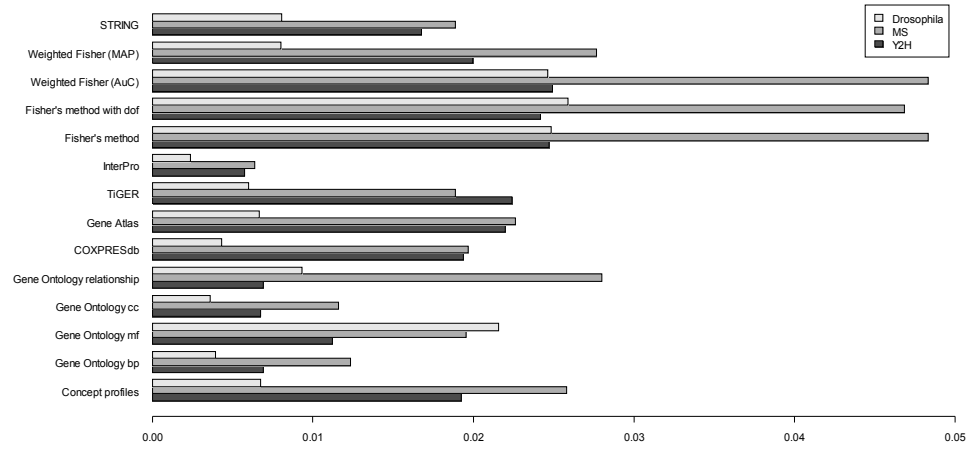

Figure 2b. MAP results for the huntingtin case study.

Showcase 3: polycystic kidney disease 1 (*PKD1*, MIM: 601313). Filtering by feature selection and solving homonyms

In specific cases, certain databases may add noise instead of valuable information. We evaluated a ranked list for the *PKD1* gene that causes polycystic kidney disease 1. The extracellular part of *PKD1* encoded protein contains many domains important for physical interactions with other proteins. The protein domain

information therefore dominated the prediction of PKD's interactions partners. This effect was undesired and therefore the InterPro/DOMINE score was left out.
The prediction of interaction partners for PKD1 by literature analysis alone was also not ideal. Although the literature remains the biggest information source, it is also the information source which requires the most preprocessing. Text-mining on its own is a challenging field of research with involves many steps such as extracting public articles, defining an ontology containing concepts and their synonyms, and disambiguating words in text using concept recognition software. Disambiguation is the process of mapping a word in text to a unique concept and labels it with a unique identifier. A term is considered to be ambiguous if it has multiple meanings. We investigated this homonym problem for PKD1. The first homonym problem is that the name 'polycystic kidney disease 1' itself can refer to the gene or the disease. When only concept profiles were used the top of most associated proteins with PKD1 showed six proteins that ranked high due to homonyms. Two proteins, protein kinase D1 (*PRKD1* encoded protein) and ectonucleotide pyrophosphatase/ phosphodiesterase 1 (*ENPP1* encoded protein), were caused by direct homonym problems. In literature, PRKD1 is also referred to as PKD1. ENPP1 has a synonym PC1 that is also used as a synonym for PKD1. The other four proteins had synonym problems in the overlapping concepts of their concept profiles. These can be seen as indirect homonym problems. In literature protein kinase D2(*PRKD2* encoded protein) is referred to as polycystic kidney disease 2 (*PKD2* encoded protein) which has a close relationship with PKD1. Protein kinase D3 (*PRKD3* encoded protein) is referred to as protein kinase C and has many relationships with PRKD1 in literature. The same holds for protein kinase C substrate 80K-H (*PRKCSH* encoded protein) which is referred to as protein kinase C substrate. Phosphoglycolate phosphatase (*PGP* encoded protein) is referred to as PRKD1 which on itself causes homonym problems. When the concept profiles are used in combination with expression data these homonyms can be suppressed.
For PKD1 we generated a ranked list while omitting the InterPro domain information from the prediction. We calculated the match score based on Fisher's method and checked if mentioned homonyms were suppressed since these proteins are not likely to be co-expressed with PKD1. The last column in table 2 shows that mentioned proteins with homonym problems indeed had much lower rankings than in the prediction based on literature only. Further manual curation by an expert showed that Fisher's method gives better associations with PKD1 in the top predictions when concept profiles are used in combination with microarray expression data and eliminating the InterPro domain information. In practice an expert should be able to choose which databases are being combined for the best prediction.

Table 1. Ranks of the homonyms associated with PKD1. The first rank is based on concept profiles, showing that the homonyms rank high. Fisher's methods suppressed these homonyms and the rank becomes lower

Gene symbol	Gene name	Rank Concept profiles	Rank Fisher's method
PRKD1	Serine/threonine-protein kinase D1	2	46
PRKD2	Serine/threonine-protein kinase D2	4	164
PRKD3	Serine/threonine-protein kinase D3	7	328
ENPP1	Ectonucleotide pyrophosphatase/phosphodiesterase family member 1	15	258
PRKCSH	Glucosidase 2 subunit beta	30	283
PGP	Phosphoglycolate phosphatase	36	1983

Concluding remarks

In this study we have shown that combining information from the biomedical literature and from different databases using Fisher's method significantly improves the prediction of novel protein interactions compared to previously applied methods. We evaluated three case studies on dysferlin, huntingtin, and polycystin-1 and predicted proteins previously not recorded in any protein interaction database. For huntingtin, besides the literature, other databases like Gene Atlas and The Gene Ontology contributed to the matchscore. An evaluation of the top 10 predicted huntingtin interacting proteins showed 5 proteins known to be associated with huntingtin. The other 5 were novel ones that have been curated and are potential interaction partners with huntingtin. From these top 10 proteins 5 could not be detected with a MEDLINE query, indicating that implicit knowledge extraction is possible.

For dysferlin we showed that the literature remains the biggest information source and that the other databases to a lesser extent contribute to the match score. Although for dysferlin the contribution of other databases to the literature alone seems low, the aid of other databases has been shown to be useful in solving homonym problems. This was shown in the PKD1 study. PKD1 showed 6 proteins that were caused by homonyms and these were suppressed when the concept profiles were combined with other databases. Thus the combination of literature and non-textual information makes our algorithm more robust.

Fisher's Method is a simple and robust method to combine several databases. In addition its interpretation is very intuitive. For every database you first define a p-value for a sample that needs to be evaluated. Fisher's methods then tells if the

combination of individual p-values (taken from different databases) for that sample is significant.
We made a list available of Fisher match scores between every two proteins in our ontology. The list can be downloaded from www.biosemantics.org/ppi-prediction.

Acknowledgements
This project was supported by the Biorange project SP 3.5.1 of the Netherlands Bioinformatics Center and the Center for Medical Systems Biology, both financed by the Netherlands Genome Initiative, and by the Dutch Prinses Beatrix Fonds.

Table 2. Prediction in the top 50

Hungtintin	Fisher fixed dof	Fisher variable dof	Weighted Fisher AuC	Weighted Fisher MAP	STRING
Y2H	2	3	2	0	0
IP	3	3	3	1	0
Drosophila	3	3	3	2	0
Dysferlin					
IP	9	9	9	6	6

Table 3. Top 50 for huntingtin predicted interacting partners

rank	name	Y2H	MS	Drosophila	PPI	cooccurrences
1	HTT	x	x			1131
2	STX1A		x			2
3	SLC1A2		x			2
4	PAFAH1B1					0
5	GABARAP					0
6	MAPT					0
7	FPC					0
8	DRD2					9
9	GRM3					0
10	VAPB					0
11	HIP1R				x	4
12	HIP1	x			x	23
13	KIF5B					0
14	MAPRE1					0
15	GSK3B					2
16	CTNNB1	x		x		0
17	ATN1					19
18	STX6					0
19	CLASP1	x				0
20	BID					0
21	TMED10					0
22	KIF1B					0
23	CDK5					5
24	NTRK2					0
25	HIPK2					0
26	MAP1S					0
27	AP2A1			x		0
28	CLASP2					0
29	RAE1					0
30	BBS4					0
31	GIPC1					0
32	PACSIN1	x		x	x	3
33	AKT1				x	2
34	KLC1					0
35	SYT1		x			0
36	NRCAM					0
37	ATXN1					4
38	BCL2L11					2
39	RAB3A					1
40	CDK5R1					1
41	ULK1					0
42	HIF1A					0
43	DIAPH1					0
44	SNCA					18
45	SOD1					5
46	YKT6					0
47	BDNF					38
48	AP2A2	x		x	x	4
49	TPPP					0
50	DYNC1I1					0

References

1. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.
2. van Haagen, H.H.H.B.M., t Hoen, P.A.C., Botelho Bovo, A., de MorrÃ©e, A., van Mulligen, E.M., et al., *Novel Protein-Protein Interactions Inferred from Literature Context.* PLoS ONE, 2009. **4**(11): p. e7894.

3. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D., *GeneCards: integrating information about genes, proteins and diseases.* Trends Genet, 1997. **13**(4): p. 163.
4. Xia, K., Dong, D., and Han, J.D., *IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model.* BMC Bioinformatics, 2006. **7**: p. 508.
5. Alexeyenko, A. and Sonnhammer, E.L., *Global networks of functional coupling in eukaryotes from comprehensive data integration.* Genome Res, 2009. **19**(6): p. 1107-16.
6. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., et al., *STRING 8--a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Res, 2009. **37**(Database issue): p. D412-6.
7. Kaltenbach, L.S., Romero, E., Becklin, R.R., Chettier, R., Bell, R., et al., *Huntingtin interacting proteins are genetic modifiers of neurodegeneration.* PLoS Genet, 2007. **3**(5): p. e82.
8. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.
9. Jelier, R., Schuemie, M.J., Roes, P.J., van Mulligen, E.M., and Kors, J.A., *Literature-based concept profiles for gene annotation: the issue of weighting.* Int J Med Inform, 2008. **77**(5): p. 354-62.
10. Mistry, M. and Pavlidis, P., *Gene Ontology term overlap as a measure of gene functional similarity.* BMC Bioinformatics, 2008. **9**: p. 327.
11. Philip, R., *Using information content to evaluate semantic similarity in a taxonomy*, in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. 1995, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada.
12. Dekang, L., *An Information-Theoretic Definition of Similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998, Morgan Kaufmann Publishers Inc.
13. Jiang, J.J. and Conrath, D.W. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. in *International Conference Research on Computational Linguistics (ROCLING X)*. 1997.
14. Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H., et al., *COXPRESdb: a database of coexpressed gene networks in mammals.* Nucleic Acids Res, 2008. **36**(Database issue): p. D77-82.
15. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.

16. Liu, X., Yu, X., Zack, D.J., Zhu, H., and Qian, J., *TiGER: a database for tissue-specific gene expression and regulation.* BMC Bioinformatics, 2008. **9**: p. 271.
17. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., et al., *InterPro: an integrated documentation resource for protein families, domains and functional sites.* Brief Bioinform, 2002. **3**(3): p. 225-35.
18. Raghavachari, B., Tasneem, A., Przytycka, T.M., and Jothi, R., *DOMINE: a database of protein domain interactions.* Nucleic Acids Res, 2008. **36**(Database issue): p. D656-61.
19. Ben-Hur, A. and Noble, W.S., *Choosing negative examples for the prediction of protein-protein interactions.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S2.
20. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., et al., *The IntAct molecular interaction database in 2010.* Nucleic Acids Res, 2009.
21. Laval, S.H. and Bushby, K.M., *Limb-girdle muscular dystrophies--from genetics to molecular pathology.* Neuropathol Appl Neurobiol, 2004. **30**(2): p. 91-105.
22. Bansal, D., Miyake, K., Vogel, S.S., Groh, S., Chen, C.C., et al., *Defective membrane repair in dysferlin-deficient muscular dystrophy.* Nature, 2003. **423**(6936): p. 168-72.
23. Huang, Y., Verheesen, P., Roussis, A., Frankhuizen, W., Ginjaar, I., et al., *Protein studies in dysferlinopathy patients using llama-derived antibody fragments selected by phage display.* Eur J Hum Genet, 2005. **13**(6): p. 721-30.
24. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., et al., *The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease.* Nat Genet, 1993. **4**(4): p. 398-403.
25. Landles, C. and Bates, G.P., *Huntingtin and the molecular pathogenesis of Huntington's disease. Fourth in molecular medicine review series.* EMBO Rep, 2004. **5**(10): p. 958-63.
26. Harjes, P. and Wanker, E.E., *The hunt for huntingtin function: interaction partners tell many different stories.* Trends Biochem Sci, 2003. **28**(8): p. 425-33.
27. Waldvogel, H.J., Kubota, Y., Fritschy, J., Mohler, H., and Faull, R.L., *Regional and cellular localisation of GABA(A) receptor subunits in the human basal ganglia: An autoradiographic and immunohistochemical study.* J Comp Neurol, 1999. **415**(3): p. 313-340.
28. Arai, H., Koizumi, H., Aoki, J., and Inoue, K., *Platelet-activating factor acetylhydrolase (PAF-AH).* J Biochem, 2002. **131**(5): p. 635-640.

29. Hattori, M., Adachi, H., Tsujimoto, M., Arai, H., and Inoue, K., *Miller-Dieker lissencephaly gene encodes a subunit of brain platelet-activating factor acetylhydrolase [corrected].* Nature, 1994. **370**(6486): p. 216-218.
30. Reiner, O., Bar-Am, I., Sapir, T., Shmueli, O., Carrozzo, R., et al., *LIS2, gene and pseudogene, homologous to LIS1 (lissencephaly 1), located on the short and long arms of chromosome 2.* Genomics, 1995. **30**(2): p. 251-256.
31. McLaughlin, N.J.D., Banerjee, A., Kelher, M.R., Gamboni-Robertson, F., Hamiel, C., et al., *Platelet-Activating Factor-Induced Clathrin-Mediated Endocytosis Requires beta-Arrestin-1 Recruitment and Activation of the p38 MAPK Signalosome at the Plasma Membrane for Actin Bundle Formation.* The Journal of Immunology, 2006. **176**(11): p. 7039-7050.
32. Caldwell, C.M. and Kaplan, K.B., *The role of APC in mitosis and in chromosome instability.* Adv.Exp.Med Biol, 2009. **656**: p. 51-64.
33. Senda, T., Shimomura, A., and Iizuka-Kogo, A., *Adenomatous polyposis coli (Apc) tumor suppressor gene as a multifunctional gene.* Anat.Sci Int., 2005. **80**(3): p. 121-131.
34. Grunewald, T. and Beal, M.F., *Bioenergetics in Huntington's disease.* Ann.N.Y.Acad.Sci, 1999. **893**: p. 203-213.
35. Cicchetti, F., Prensa, L., Wu, Y., and Parent, A., *Chemical anatomy of striatal interneurons in normal individuals and in patients with Huntington's disease.* Brain Research Reviews, 2000. **34**(1-2): p. 80-101.
36. Kanekura, K., Suzuki, H., Aiso, S., and Matsuoka, M., *ER stress and unfolded protein response in amyotrophic lateral sclerosis.* Mol Neurobiol., 2009. **39**(2): p. 81-89.
37. Teuling, E., Ahmed, S., Haasdijk, E., Demmers, J., Steinmetz, M.O., et al., *Motor Neuron Disease-Associated Mutant Vesicle-Associated Membrane Protein-Associated Protein (VAP) B Recruits Wild-Type VAPs into Endoplasmic Reticulum-Derived Tubular Aggregates.* Journal of Neuroscience, 2007. **27**(36): p. 9801-9815.
38. Stefani, M. and Dobson, C.M., *Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution.* J Mol Med, 2003.
39. Wang, H., Bedford, F.K., Brandon, N.J., Moss, S.J., and Olsen, R.W., *GABAA-receptor-associated protein links GABAA receptors and the cytoskeleton.* Nature, 1999. **397**(6714): p. 69-72.
40. Romero, E., Cha, G.H., Verstreken, P., Ly, C.V., Hughes, R.E., et al., *Suppression of Neurodegeneration and Increased Neurotransmission Caused by Expanded Full-Length Huntingtin Accumulating in the Cytoplasm.* Neuron, 2008. **57**(1): p. 27-40.

41. Swayne, L.A., Chen, L., Hameed, S., Barr, W., Charlesworth, E., et al., *Crosstalk between huntingtin and syntaxin 1A regulates N-type calcium channels.* Molecular and Cellular Neuroscience, 2005. **30**(3): p. 339-351.
42. Miller, B.R., Dorner, J.L., Shou, M., Sari, Y., Barton, S.J., et al., *Up-regulation of GLT1 expression increases glutamate uptake and attenuates the Huntington's disease phenotype in the R6/2 mouse.* Neuroscience, 2008. **153**(1): p. 329-337.
43. Arzberger, T., Krampfl, K., Leimgruber, S., and Weindl, A., *Changes of NMDA receptor subunit (NR1, NR2B) and glutamate transporter (GLT1) mRNA expression in Huntington's disease--an in situ hybridization study.* J Neuropathol.Exp.Neurol, 1997. **56**(4): p. 440-454.
44. Caparros-Lefebvre, D., Kerdraon, O., Devos, D., Dhaenens, C.M., Blum, D., et al., *Association of corticobasal degeneration and Huntington's disease: Can Tau aggregates protect Huntingtin toxicity?* Movement Disorders, 2009. **24**(7): p. 1089-1090.
45. Glass, M., Dragunow, M., and Faull, R.L., *The pattern of neurodegeneration in Huntington's disease: a comparative study of cannabinoid, dopamine, adenosine and GABA(A) receptor alterations in the human basal ganglia in Huntington's disease.* Neuroscience, 2000. **97**(3): p. 505-519.
46. Gottfried, I., Ehrlich, M., and Ashery, U., *The Sla2p/HIP1/HIP1R family: similar structure, similar function in endocytosis?* Biochem Soc Trans. **38**(Pt 1): p. 187-191.
47. Engqvist-Goldstein, A.E., Kessels, M.M., Chopra, V.S., Hayden, M.R., and Drubin, D.G., *An actin-binding protein of the Sla2/Huntingtin interacting protein 1 family is a novel component of clathrin-coated pits and vesicles.* J Cell Biol, 1999. **147**(7): p. 1503-1518.

Supplementary information

S1 Databases for information extraction

Table 1 shows the databases that are used in our analysis and the date of download.

Table 1. Databases that are combined and their date of download.

Database	Date of download
Concept profiles	May 2009
Gene ontology	July 23, 2009
Gene Atlas	April, 2004
Coxpresdb	April 17, 2008
TiGER	February 19, 2009
InterPro	July 22, 2009
DOMINE	April 16, 2007
STRING version 8.1	October 18, 2009

S2 Curated Protein-Protein interaction databases

For training, testing and optimizing our system we constructed a set of known human protein-protein interactions (PPIs) taken from public, curated databases. We called this set of known PPIs the positive set. The databases used were Biogrid[1], DIP[2], HPRD[3], Mint[4], Reactome[5], and Uniprot/Swiss-Prot[6]. Table 2 shows the date of download for these databases. If a PPI was mentioned in one of these databases, we assumed it to be a true PPI. There is a level of redundancy between these databases meaning that some protein-protein interaction pairs occur in multiple databases, which is a good indication that it is a true PPI. These protein pairs count only once. We restricted our analysis to human proteins only. The resulting positive set contains 83,930 PPIs.
A negative set was constructed as described in the materials and method section of the paper. The negative set is the same as the null distribution used for the Fisher Method and has a size of 100.000 samples.

Table 2. Protein databases used for the positive set and their date of download.

Protein database	Data of download
BioGrid	July 1, 2009
DIP	October 15, 2008
HPRD	July 6, 2009
IntAct	July 11, 2009
MINT	July 23, 2009
Reactome	June 11, 2009
UniProt	June 17, 2009

S3 Cross validation

All performance measures (AuC and MAP) were calculated in a 5-fold cross validation loop. First the data consisting of positive and negative instances (*e.g* PPIs and random protein pairs) were splitted in five equally sized parts. Then at each iteration four parts were used for straining classifiers and combination methods and the remaining fifth part was used for testing. This was repeated until each part was used once for testing.

S4: Coverage and prediction accuracy of individual databases

In the analyses that follow we first defined a positive set that consists of protein-protein interactions recorded in six curated databases (see supplement S2), and a negative set of probable non-interacting random protein pairs. We evaluated how well each database covers samples from the positive and negative set. Table 3 shows the coverage for each individual database, the combination of databases and STRING. A protein pair is covered if at least one on the individual databases has a match score for that protein pair. Our combined databases cover almost the complete positive set. The coverage is similar to STRING.

To evaluate prediction performance for PPIs, we used the AuC and MAP criteria. A third measure is used to reflect the predictions made in the top of a ranked list. It counts the number of predicted true positives when the number of predicted false positives is fixed to 50. We refer to this measure as ROC50.

For each database the AuC and MAP were calculated and the results are given in table 4. The AuC and MAP were calculated in a 5-fold cross validation loop (see S3). Figure 1 shows the ROC50. Here it is shown that the concept profiles (cp) has the highest number of true positives (over 6,000). The STRING curated database score (dscore) also gives a performance of over 6,000 predicted true positives. This result is expected since this score is based on curated protein databases, some of which were also used to create our evaluation set. The Gene Ontology gives an overall best performance with a high AuC and high coverage.

Table 3. Coverage for each database and the databases combined. As a benchmark the coverage of STRING is given.

Database	Positive set (%)	Negative set (%)
Concept profiles	84	24
Gene Ontology biological process	94	34
Gene Ontology molecular function	95	39
Gene Ontology cellular component	95	43
Gene Ontology relationships	99	52
COXPRESdb	95	51
Gene Atlas	69	12
TiGER	72	25
InterPro/DOMINE	95	49
Combined system	99.97	67
All STRING databases	99.13	62

Table 4. AuC and MAP for the individual databases for a 5-fold cross validation. The standard errors are not shown because they were negligible small. The Gene Ontology is separated into the three main categories and the relationships.

Database	AuC	MAP
Concept profiles	0.88	0.90
Gene Ontology biological process	0.91	0.90
Gene Ontology molecular function	0.88	0.85
Gene Ontology cellular component	0.89	0.88
Gene Ontology relationships	0.91	0.88
COXPRESdb	0.82	0.79
Gene atlas	0.80	0.81
TiGER	0.78	0.75
InterPro/DOMINE	0.80	0.77
STRING DATABASES		
Neighborhood in the genome	0.69	0.59
Gene fusion	0.69	0.58
Cooccurrences across genomes	0.69	0.59
Coexpression	0.69	0.59
Experimental/biochemical data	0.81	0.82
Association in curated databases	0.82	0.82
Co-mentioned in PubMed abstracts	0.83	0.84

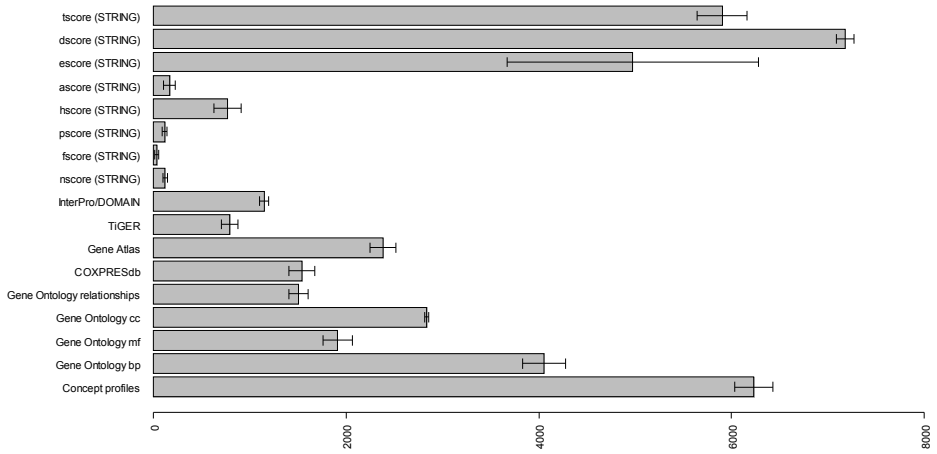

Figure 1. Number of true positives that are retrieved at 50 false positives for each individual database. The concept profiles (cp) retrieves the highest amount of true positives, reflecting that the literature is still the most important source of information. The assocation in curated database score (dscore) for STRING shows the best result as expected. The errorbars are the standard deviation around the mean calculated over 5-fold cross validation.

S5 Different combining techniques

Before we came to our final approach based on Fisher's Method we evaluated four other combining methods described below.

(1) Combining rules by Kuncheva

The first combiner is the one defined by Kuncheva [7]. In total there are five combining rules namely the product, sum, maximum, minimum, and majority vote. The combiners defined by Kuncheva are applied to the output of each classifier trained on a single database; hence this step requires training data. In our case we used a simple logistic regression classifier [8]. Each raw match score defined for a database is converted to a probability value between 0 and 1. The concept profiles score was first log transformed to produce more normal distributed classes. Since we evaluate a two class problem (the class of protein-protein interactions (PPI) and the class of non interacting protein pairs (NIPP)) the probability of the second class can be calculated once the probability of the first one is known. If *p1* is the probability for a sample in class one then *p2=1-p1* is the probability for that

sample in class two. On the output of each classifier the combining rule was applied. The product rule for a sample *x* is defined as follows

$$\mu_j = \prod_{i=1}^{L} p_{i,j}(x)$$

With μ_j the combined probability for class *j* and $p_{i,j}(x)$ the probability of x belonging to class j according to database i, and *L* the total number of databases to combine (our case *L=6*). After the rule is applied, the combined probabilities can be normalized to add up to 1. In the same way the sum, maximum and minimum rule can be defined. Missing values are completely ignored. If one database has a missing value the rule is applied to the remaining databases. If a sample has missing values for all the databases the probabilities are set to 0 and 1 for class one (PPIs) and two respectively.
The advantage is that these combiners do not require training data. The disadvantage is that the classifiers trained on each database in the first step make assumptions about your data, for instance that the classes follow a normal distribution. This could result in false predictions if the assumptions are not true.

(2) Rank combiners
Calculating a rank combiner is similar to the Kuncheva combiners. The same rules such as, product and sum, can be applied to ranks. For instance the rank product is a non-parametric statistic that is often used for gene expression profiling [9]. Here the formula for the rank product is given.

$$RP(x) = \left(\prod_{i=1}^{L} r_{x,i} \right)^{1/L}$$

With $r_{x,i}$ the rank obtained for database *i* for a sample *x*. In the same way the other combiners based on ranks can be derived. *L* are the number of databases with no missing value for that sample. The rank for samples where all values are missing is set to positive infinite. The advantage of this combiner is that it also does not require any training data. Furthermore, it does not put any constraints on the data. The disadvantage is that it is highly sensitive to the presence of poorly performing databases.

(3) Maximum AuC linear classifier (MALC)
Marrocco et. al. [10] describes a method where a linear classifier is trained such that the resulting trained classifier maximizes the AuC. Mainstream classifiers are designed to minimize the classification error, *e.g.* taken into account the false negatives, whereas the MALC is designed to minimize the false positives, *e.g.* maximizing the AuC. We implemented a different version of their algorithm which is both fast and robust. The features (match scores defined for databases) are combined in an iterative manner and at each iteration step two features are

combined and result in a new feature. Step one is to normalize the date between 0 and 1. The inner product score between concept profiles were first log transformed before normalization. Step two is to calculate a Pearson correlation matrix (all pair wise correlations between any two features). The two features with the lowest correlation are combined first. Step three is to apply the linear classifier to the features *h* and *k* which is given as

$$x_{lc} = \alpha x_h + (1-\alpha) x_k$$

where x_{lc} is the weighted sum of the two features, and alpha the weight parameter that needs to be optimized. Step four is to vary the alpha level between 0 and 1 in steps of 0.01 (or any other step size) and calculate the AuC for each alpha. Then choose the alpha level that corresponds with the highest AuC value. Step five is to replace the two features with xlc features and repeat steps two till five until all features are combined to a single feature.

4) Fisher's method

The Fisher method was described in the article. The advantage of this method is that it is robust, simple, and no information is needed on the positive set (PPIs) since it only uses the null distribution (negative set of probable non interacting protein pairs).

S6 Choosing the best combining method

The four different combining methods (and each with a number of variations) are compared which each other using the AuC and MAP as performance criteria in a 5-fold cross validation loop. The results for all combiners are given in table 3. Fisher's method and the MALC show the best results in both MAP and AuC. To further evaluate the accuracy of these combiners we looked at the ROC50 results. That is the number of true positives predicted when the number of false positives was fixed to 50. The results are given in figure 2. Here the Fisher method shows slightly better result than the MALC. Figure 3 shows the ROC curve for the Fisher method with fixed degrees of freedom. The histogram for the positive and negative set is given in figure 4.

Table 5. AuC and MAP for the different combining techniques. The standard errors are not shown.

Kuncheva combining rule	**AuC**	**MAP**
Product	0.94	0.92
Sum	0.91	0.86
Maximum	0.84	0.73
Minimum	0.90	0.86
Majority vote	0.90	0.83
Rank combiners		
Mean	0.94	0.83
Max	0.83	0.46
Min	0.93	0.93
Product	0.95	0.90
Fisher's method		
Fixed number of dof (=9)	0.97	0.97
Fisher with variable dof	0.97	0.96
Weighted Fisher with AuC	0.97	0.97
Weighted Fisher with MAP	0.97	0.97
Fisher +4 features from String	0.97	0.97
Maximize AuC		
MALC	0.97	0.96

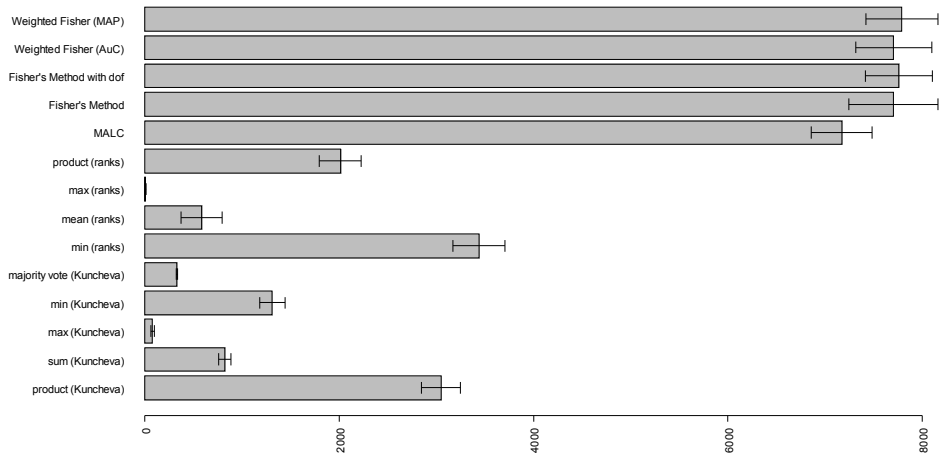

Figure 2. Number of true positives at 50 false positives for the different combination techniques

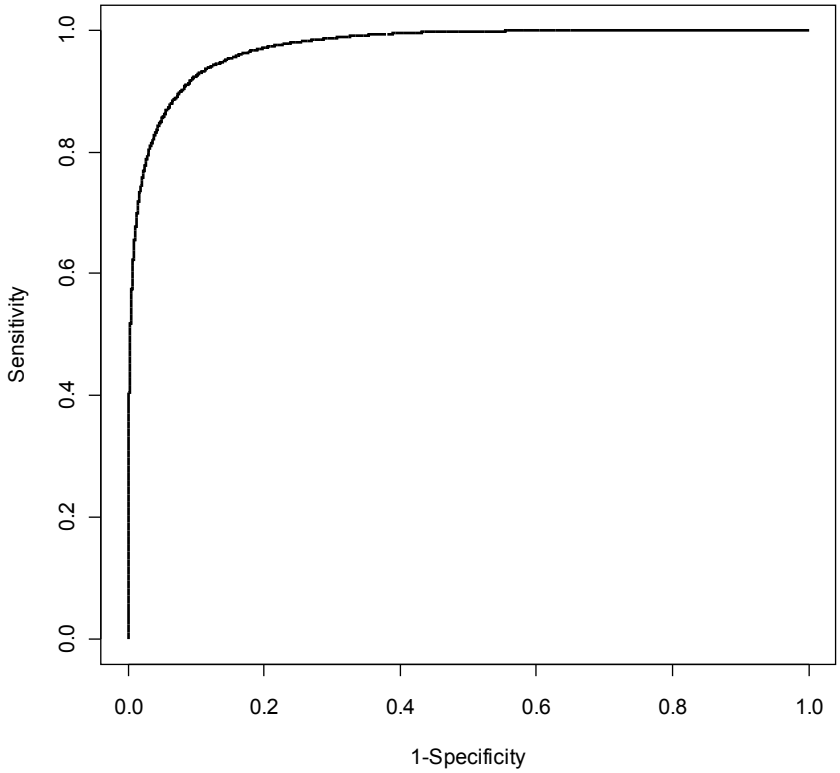

Figure 3. ROC plot of the Fisher method combiner.

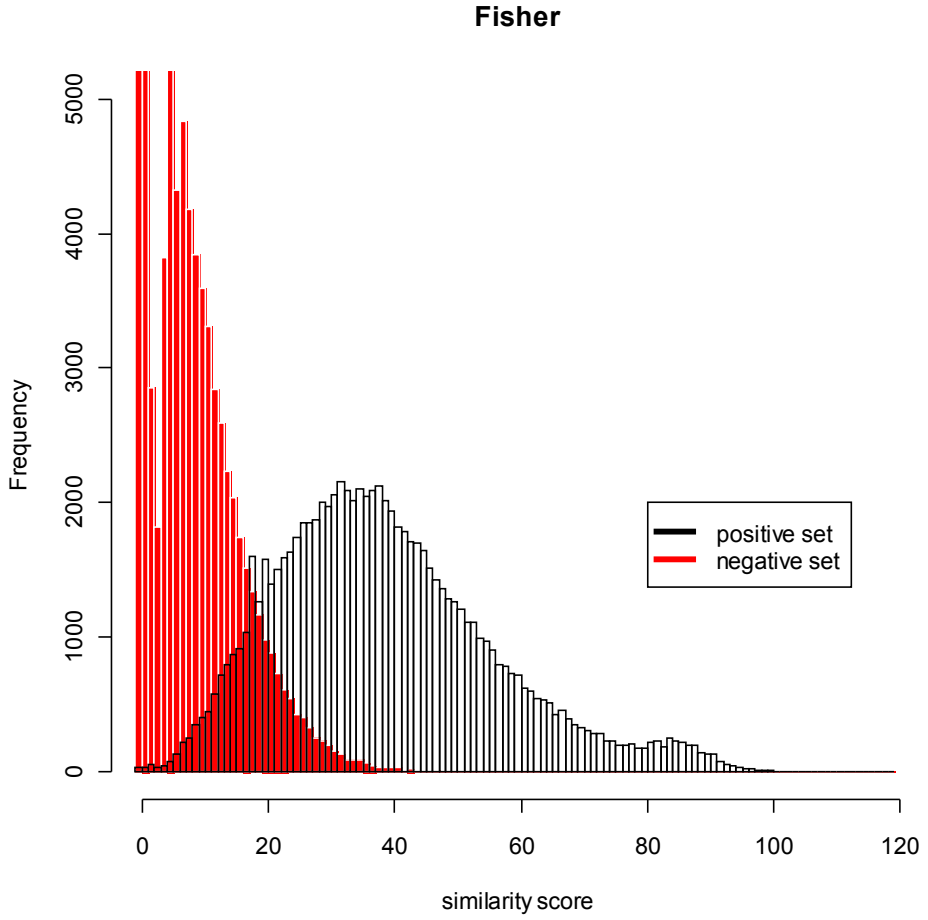

Figure 4. Histogram plot of the positive PPI set and the negative random set for the Fisher combiner.

Table 6. Top 50 of most associated proteins with Dysferlin.

rank	name	Co-occurrences	PPI
1	DYSF	246	
2	MYOF	13	
3	TGFB1	0	
4	RYR2	0	
5	TTN	2	
6	MYH7	1	
7	KCNQ1	0	
8	MYL3	0	
9	TNNC2	0	

10	SNTA1	2	
11	TCAP	13	
12	ADRBK1	0	
13	SGCA	18	
14	TPM1	0	
15	TNNC1	0	
16	MYH4	0	
17	IL1B	0	
18	MYOT	10	
19	C5AR1	0	
20	OTOF	6	
21	SSPN	1	
22	FKBP1B	0	
23	MYH2	0	
24	Cf5	0	
25	ACTN2	1	
26	UTRN	3	
27	TNNT1	0	
28	TNNT3	0	
29	CSF3R	0	
30	CACNA1H	0	
31	HMOX1	0	
32	MYBPC3	0	
33	DES	0	
34	GAA	0	
35	TPP1	0	
36	SNTB1	0	
37	KCNE1	0	
38	ACTA1	0	
39	HCK	0	
40	CAV3	35	X
41	CAPN3	44	X
42	FPR1	0	
43	RYR1	1	
44	KCNMA1	0	
45	MYLK2	0	
46	MYH6	0	
47	TNNI3	0	
48	NCF2	0	
49	NOS3	0	
50	CAV1	0	

1. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.

2. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
3. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
4. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
5. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., et al., *Reactome knowledgebase of human biological pathways and processes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D619-22.
6. *The Universal Protein Resource (UniProt) 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D169-74.
7. Kuncheva, L.I., *Combining pattern classifiers*. 1 ed. 2004: Wiley-Interscience. 376.
8. Mitchell, T.M., *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*, in *Machine Learning*. 2005. p. 1-17.
9. Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P., *Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.* FEBS Lett, 2004. **573**(1-3): p. 83-92.
10. Marrocco, C., Duin, R.P.W., and Tortorella, F., *Maximizing the area under the ROC curve by pairwise feature combination.* Pattern Recognition, 2008. **41**(6): p. 1961-1974.