# In silico discoveries for biomedical sciences

Haagen, H. van

# Chapter 3

## Novel protein-protein interactions inferred from literature context

H.H.H.B.M. van Haagen[1], P.A.C. ’t Hoen[1], A. Botelho Bovo[2], A. de Morree[1], E.M. van Mulligen[1], C. Chichester[1], J.A. Kors[1], J.T. den Dunnen[1], G.J.B. van Ommen[1], S.M. van der Maarel[1], V. Medina Kern[2], B. Mons[1], M.J. Schuemie[1]

**1** Biosemantics Association, Department of Human Genetics, Leiden University Medical Center, Leiden, and Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
**2** Post-Graduate Program in Knowledge Engineering and Management (EGC), Federal University of Santa Catarina (UFSC), Florianópolis, Brazil

PLoS One. 2009 Nov 18;4(11):e7894.

## Abstract

We have developed a method that predicts Protein-Protein Interactions (PPIs) based on the similarity of the context in which proteins appear in literature. This method outperforms previously developed PPI prediction algorithms that rely on the conjunction of two protein names in MEDLINE abstracts. We show significant increases in coverage (76% versus 32%) and sensitivity (66% versus 41% at a specificity of 95%) for the prediction of PPIs currently archived in 6 PPI databases. A retrospective analysis shows that PPIs can efficiently be predicted before they enter PPI databases and before their interaction is explicitly described in the literature. The practical value of the method for discovery of novel PPIs is illustrated by the experimental confirmation of the inferred physical interaction between CAPN3 and PARVB, which was based on frequent co-occurrence of both proteins with concepts like Z-disc, dysferlin, and alpha-actinin. The relationships between proteins predicted by our method are broader than PPIs, and include proteins in the same complex or pathway. Dependent on the type of relationships deemed useful, the precision of our method can be as high as 90%. The full set of predicted interactions is available in a downloadable matrix and through the webtool Nermal, which lists the most likely interaction partners for a given protein. Our framework can be used for prioritizing potential interaction partners, hitherto undiscovered, for follow-up studies and to aid the generation of accurate protein interaction maps.

## Introduction

Protein-protein interactions (PPIs), which we define as proteins that physically interact, are crucial in most complex biological processes. Experimental high-throughput methods such as yeast two-hybrid screens have been used to make large inventories of PPIs and to create protein interaction maps[1-6]. However, it is well known that these methods merely show physical interaction under experimental condition and not necessarily indicate a common involvement in a biological process. Computational methods for the prediction of PPIs could theoretically aid the discovery of candidate biological interaction partners. There are many different sources of information that can be used in PPI prediction[7], including protein structures, phylogenetic distribution, interactions between homologous proteins in other organisms, genomic neighborhood, and gene fusions. In this article, we will focus on one source of information, which is arguably the most comprehensive, but also the least structured: biomedical literature itself. Until now text mining techniques are mainly used to rediscover PPIs explicitly described in literature. Often, the now 18 million freely available abstract records of MEDLINE are used for this purpose. PPIs extracted this way have been shown to improve the accuracy of predicted biological networks[8, 9]. Structured information on explicit PPIs extracted from MEDLINE  and other sources is freely available in the STRING database[10], or can be found by querying the iHOP website[11].

However, text mining can go one step further; by combining known associations, previously unknown PPIs can be inferred. Because most text mining research, including this study, limits itself to MEDLINE abstracts, these 'previously unknown' interactions also include interactions that are effectively known, but not explicit in MEDLINE as they are only mentioned in a full text article. Swanson[12, 13] *et al.* were the first to demonstrate that text mining can lead to the discovery of new knowledge (e.g. the treatment of Raynaud's disease by fish oil). Other studies in the biomedical domain verified the importance of implicit information for knowledge discovery[14-16]. Whereas Swanson used a word-based approach, linking entities by intermediate words that appeared frequently in the contexts of both entities, in our work we use a concept-based approach: different terms denoting the same concept (*i.e.* synonyms) are mapped to a single concept identifier, and ambiguous terms, e.g., identical terms used to indicate different concepts (*i.e.* homonyms) are resolved by a disambiguation algorithm. Such an approach is essential given the wide diversity and many ambiguities in gene and protein nomenclature[17, 18].

In order to predict PPIs, we summarize the typical context in which each protein appears into *concept profiles[15, 16, 19]*. We hypothesize that a high similarity between the concept profiles of two proteins is indicative for an actual biological interaction. For example, if two proteins are consistently mentioned together with a particular disease, the probability that these proteins interact is  higher than the a priori probability of two randomly selected proteins[20, 21]. This probability should increase further when they are also frequently co-mentioned with a particular pathway, a sub-cellular localization, or other proteins.

In this article, we first demonstrate the added value of a concept-based approach over a traditional term-based approach in detecting explicitly described relations. We proceed to show the added value of the concept profile-based approach over classical direct relation extraction, including the text-mining techniques used in the STRING database. Subsequently, we show the predictive power of our method by doing a retrospective study; we demonstrate that we can employ the literature available in 2005 to predict 52% of the PPIs newly described in Swiss-Prot in 2007 at a specificity level of 95%. We show that in addition, some of the PPIs that we predicted but are not yet recorded in any database represent indirect protein interactions and have biological relevance. Finally, we confirm one of the many predicted PPIs in three wet lab experiments, supporting our claim that the concept profiling method is capable of previously unknown PPI prediction from current literature.

These predictions will be useful for (i) the ranking of potential PPIs for more specific experimental analysis, and (ii) complementing other types of data such as co-expression and yeast two-hybrid data when using an integrative systems biology approach.

## Results

### Improved PPI detection using concept profiles

We compared the performance of different PPI prediction approaches in detecting known human PPIs in MEDLINE. The online human-curated databases Biogrid, DIP, HPRD, MINT, Reactome, and UniProt/Swiss-Prot were used to establish a set of 61,807 known human PPIs. A set of probable Non-Interacting Protein Pairs (NIPPs) was generated from all pairs of proteins that do not occur in the above databases nor in the IntAct[22] database, which includes, in addition to all PPIs recorded in UniProt/Swiss-Prot, many non-curated PPIs from high-throughput experiments. We compare four approaches:

- *Word-based direct relation*. This approach uses direct PubMed queries (words) to detect if proteins co-occur in the same abstract. This is the simplest approach and represents how biologists might use PubMed to search for information.
- *Concept-based direct relation*. This approach uses concept-recognition software to find PPIs, taking synonyms into account, and resolving homonyms. Here two concepts (in our case two proteins) are detected if they co-occur in the same abstract.
- *STRING[10]*. The STRING database contains a text mining score which is based on direct co-occurrences in literature.
- *Concept profile-based relation*. This approach uses the similarity in literature context. Here two proteins (concepts) can also be indirectly related via the concepts in their profiles. More detail on concept profiles and their construction can be found in the Methods section.

The word-based and concept-based direct relation methods could find at least one abstract containing both proteins for respectively 33% and 32% of the pairs in the PPI set. A text mining score from STRING could be obtained for 30% of the PPIs, in line with the co-occurrence based approach used to create STRING. Thus, a majority of the known PPIs cannot be found explicitly in MEDLINE. For the concept profile-based approach, we could create concept profiles and calculate a similarity score for 76% of the PPI set.

Similar to STRING, the other three approaches can also be used to calculate a continuous score that indicates the strength of the relation between two proteins. Figure S1 displays the distribution of the similarity scores of the concept profile-based method for the PPI and NIPP sets. This figure shows that the scores for the PPI set are higher although there is also overlap between the two distributions. The continuous scores can be used to rank protein pairs. After ranking the pairs in the PPI and in the NIPP set, we calculated the sensitivity at a specificity of 99% and 95%, and the Area under the Curve (AuC), which is often used in the evaluation of classifiers, and expresses the area under the Receiver Operator Characteristics (ROC) curve (see supplement S5). An AuC of 0.5 indicates a random classifier; an AuC of 1 indicates a perfect classifier. For this analysis, we limited ourselves to those pairs in the PPI and NIPP set for which all methods could make a prediction. We analyzed 44,920 pairs in the PPI set, and 58,388,409 pairs in the NIPP set.

The results show that, using concept profiles, we can detect 43% of the known PPIs, with a specificity of 99%, and 66% of all known PPIs with a specificity of only 95%. In contrast, the direct relations methods and STRING show much lower scores (Table S1).

**Table 1. Performance of different PPI prediction approaches on detecting known PPIs in MEDLINE. CDR stands for Concept-based Direct Relation method.**

|  | Word-based | CDR | Concept profiles | STRING |
|---|---|---|---|---|
| Sensitivity at spec = 99% | 28% | 37% | 43% | 39% |
| Sensitivity at spec = 95% | 33% | 41% | 66% | 41% |
| Area under Curve | 0.62 | 0.69 | 0.90 | 0.69 |

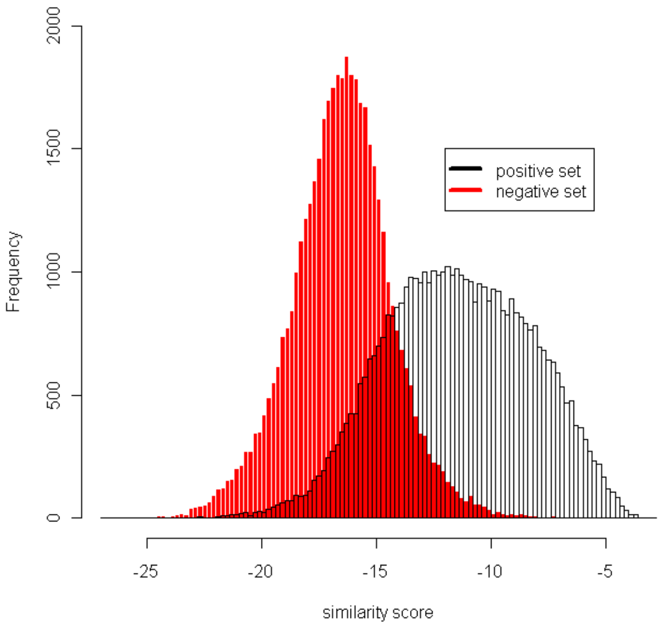


**Figure 1. Histogram of the distributions of similarity scores of the concept profile-based method for the PPI and NIPP sets. A log transformation is applied to the similarity scores for better visualization.**

**Proteins connected via one intermediate protein**

The results reported in the previous section indicate that not all proteins with high similarity scores are known to interact according to the combined protein databases. One possible explanation for this is that the proteins are related in another way, *e.g.* they could be involved in the same pathway or be part of the same protein complex, but do not physically interact. To determine whether this occurs, we also tested both concept-based approaches on the detection of known connections via one intermediate protein. For instance, if the protein pairs A-B and B-C are recorded as PPIs in databases, we form the additional protein pair A-C. In total we were able to create 1,028,265 of such pairs to serve as an independent test set. When the pairs are filtered on coverage by all methods the remaining set contains 790,245 pairs. At a specificity level of 99% and 95% the sensitivity level of the different methods was determined for those pairs. The results are given in Table S2 and indicate that the concept profile-based approach is indeed superior in predicting relationships between proteins potentially present in the same complex or pathway.

**Table 2. Performance on predicting proteins that are connected via an intermediate protein.**

| | Concept-based | CDR | STRING |
|---|---|---|---|
| Sensitivity at spec = 99% | 8% | 9% | 8% |
| Sensitivity at spec = 95% | 13% | 29% | 12% |
| Area under Curve | 0.54 | 0.78 | 0.53 |

### Average prediction performance per protein

Most researchers will not be interested in all PPIs, but only in those interactions involving a (set of) protein(s) of interest. Therefore, for each protein we created a top 10, top 100, and top 1,000 best matching proteins according to the concept-based direct relation, the concept profile method, and STRING. In these lists, we calculated the number of PPIs that are either (i) part of the PPI set, or (ii) described in the IntAct database, or else (iii) part of the pairs that are connected through intermediate proteins as described in the previous section. We limited our analyses to the 10,812 proteins that were detected in at least five MEDLINE abstracts (covered by the concept profiles method). The averages of these performance measures in terms of precision and recall are shown in Table S3. For comparison, the average total number of pairs per protein in each set is provided in the third column. For instance, on average each protein is involved in 8.73 interactions according to the PPI set, of which on average 6.34 are found in the top 1,000 of the concept profile method (precision and recall of 0.006 and 0.73 respectively) , and only 3.93 and 3.83 in the top 1,000 of the concept-based direct relation method and STRING respectively. The latter two methods show a slightly better performance for the top 10. Thus, it appears that co-occurrence-based methods can detect a smaller number of PPIs with a somewhat higher accuracy, but the concept profile method, by including indirect evidence, can predict more PPIs and is therefore likely to be more valuable for actual knowledge discovery.

**Table 3. Analysis of the top 10, 100, and 1,000 returned by the Concept Profile (CP) method, the Concept-based Direct Relation (CDR) method, and by STRING. The analysis shows the precision and recall of protein pairs that are in the PPI set, of additional pairs**

| | | | Top 10 | | Top 100 | | Top 1,000 | |
|---|---|---|---|---|---|---|---|---|
| | Method | Total | Precision | Recall | Precision | Recall | Precision | Recall |
| PPI | CP | 8.73 | 0.096 | 0.110 | 0.033 | 0.37 | 0.006 | 0.73 |
| | CDR | 8.73 | 0.108 | 0.124 | 0.026 | 0.30 | 0.004 | 0.45 |
| | STRING | 8.73 | 0.112 | 0.128 | 0.026 | 0.30 | 0.004 | 0.44 |
| IntAct | CP | 1.61 | 0.009 | 0.056 | 0.002 | 0.12 | 0.000 | 0.29 |
| | CDR | 1.61 | 0.009 | 0.056 | 0.002 | 0.11 | 0.000 | 0.24 |
| | STRING | 1.61 | 0.008 | 0.050 | 0.002 | 0.11 | 0.000 | 0.24 |
| Indirectly connected | CP | 190.21 | 0.105 | 0.006 | 0.080 | 0.042 | 0.048 | 0.25 |
| | CDR | 190.21 | 0.137 | 0.007 | 0.068 | 0.036 | 0.027 | 0.14 |
| | STRING | 190.21 | 0.100 | 0.005 | 0.062 | 0.033 | 0.026 | 0.14 |

## Retrospective prediction of currently known PPIs

Protein annotation databases are struggling to stay up-to-date with the literature, and there is often a substantial time lag between the first publication of a finding, and the time the PPI is entered in a database. It could therefore be postulated that many of the unknown PPIs predicted today are in fact correct, but may not be entered in a database for several years. We have performed a retrospective study to answer the question: how many of the PPIs that would have been predicted by the different methods in 2005 were confirmed in 2007?

Both direct relation and concept profile method-based PPI prediction scores were created using a MEDLINE corpus with publication dates up to February 2005. We ranked the PPIs according to the scores, and set a cut-off value at the 95% and 99% specificity levels based on PPIs present in Swiss-Prot 2005 (this is the only database for which historic versions are available). We subsequently evaluated how many of the 3,295 PPIs that were added to Swiss-Prot between 2005 and 2007 were above these cut-off values in 2005. These are the sensitivity values reported in Table S4. We also calculated the AuC based on Swiss-Prot 2007 alone.

The prediction performance is much better for concept profiles (52% versus 38% for a specificity level of 95%). This indicates that the majority of currently known PPIs were not yet explicitly described in MEDLINE at our testing point, but would have been predicted at a specificity rate of 95%. We postulate that this finding is indicative for the assumption that based on the full current literature a meaningful percentage of the 'unknowns' that pass the prediction threshold will be actual pairs worth studying in more detail.

**Table 4. Results of the retrospective prediction of PPIs added to Swiss-Prot between 2005 and 2007. PPIs are ranked based on MEDLINE up to 2005, and specificity levels are based on Swiss-Prot 2005.The sensitivity is determined on Swiss-Prot 2007**

| | Concept-based | Concept profiles |
|---|---|---|
| Sensitivity at spec = 99% | 27% | 33% |
| Sensitivity at spec = 95% | 38% | 52% |
| Area under Curve | 0.70 | 0.84 |

## Case Studies

The next logical step was therefore to investigate whether this method can only predict PPIs that are 'known' but not explicit in the literature corpus used, or whether it would also be able to effectively predict unknown, but real PPIs. We investigated this in two case studies. We generated predicted interactions for proteins with two proteins that are intensively investigated in our group: (i) Dystrophin (DMD), a structural protein causing Duchenne muscular dystrophy when defective, and (ii) Calpain 3 (CAPN3), a protease when mutated causing Limb-girdle muscular dystrophy (LGMD).

## DMD

We presented the list of predicted interacting proteins with DMD ordered by descending association scores, to two experts for evaluation. At a specificity of 99%, there are 196 proteins predicted to interact with DMD. This list was too long to manually evaluate and we therefore restricted the human curation analysis to the 99.8% specificity level (top 42 proteins, Table S5). The full list is presented as Table 7 in the supplementary file. The 42 proteins include 7 of the 19 dystrophin-interacting proteins that are known from curated databases (sensitivity of 37% at this very high specificity level). The remaining established interaction partners generally rank high in the list of literature-predicted targets (13/19 in the top 196, p-value from Kolmogorov-Smirnov test for comparison with overall ranking: 3.4 · $10^{-10}$). There are three proteins in the predicted set with at least indirect evidence in the literature for a physical interaction with DMD (CAV3, SPTB, ACTN2). One protein (SLMAP) may well interact given its distribution and localization but this needs experimental testing. Ten proteins in the list are found in the same protein complex as DMD but do not interact directly as far as known. Four proteins in the list were found wrongly associated with DMD due to homonym problems during literature indexing.

The remaining 17 proteins in the list are associated with DMD for other reasons (e.g. also involved in muscular dystrophy, or structural or functional homology) but are not likely to physically interact. If we only allow direct physical interaction pairs as true positives (11 proteins) the estimated precision is 26%. If predictions of protein pairs in a complex also are counted as true positives (21 proteins in total), the estimated precision would be 50%. Since also conceptually-related proteins that do not physically interact may be of interest to the biologist, the overall precision of our prediction method may be as high as 90%.

**Table 5. Top 42 ranked proteins with DMD. In total 10,812 proteins were matched against DMD. 7 proteins as known to interact with DMD. Only 4 proteins are real false positives due to homonyms problem resulting in a precision over 0.9.**

| **Rank** | **Protein symbol** | **Swiss-Prot id** | **Log similarity score** | **Direct relations** | **In PPI set** | **False positives (homonym)** |
|---|---|---|---|---|---|---|
| 1 | **UTRN** | P46939 | -5.14 | 214 | x | |
| 2 | SGCA | Q16586 | -6.13 | 119 | | |
| 3 | **DAG1** | Q14118 | -6.22 | 139 | x | |
| 4 | SGCB | Q16585 | -6.60 | 54 | | |
| 5 | SGCD | Q53XA5 | -6.95 | 46 | | |
| 6 | FCMD | O75072 | -7.05 | 29 | | |
| 7 | DYSF | O75923 | -7.19 | 43 | | |
| 8 | **DTNA** | Q9BS59 | -7.31 | 17 | x | |
| 9 | DRP2 | Q13474 | -7.34 | 9 | | |
| 10 | SSPN | Q0JV68 | -7.45 | 17 | | |
| 11 | LAMA2 | P24043 | -7.46 | 25 | | |
| 12 | GK1 | P32189 | -7.56 | 33 | | x |
| 13 | CAPN3 | P20807 | -7.93 | 28 | | |
| 14 | CAV3 | P56539 | -7.95 | 24 | | |
| 15 | **SNTA1** | Q13424 | -7.97 | 8 | x | |
| 16 | EIF3S12 | Q9UBQ5 | -8.05 | 91 | | x |
| 17 | BEST1 | O76090 | -8.13 | 26 | | x |
| 18 | SPTB | P11277 | -8.15 | 15 | | |
| 19 | FKRP | Q9H9S5 | -8.16 | 4 | | |
| 20 | MEB | 6988 | -8.17 | 7 | | |
| 21 | SLMAP | Q14BN4 | -8.20 | 4 | | |
| 22 | **SNTB1** | Q13884 | -8.20 | 6 | x | |
| 23 | NEB | P20929 | -8.33 | 16 | | |
| 24 | SGCE | O43556 | -8.35 | 10 | | |
| 25 | SGCG | Q13326 | -8.46 | 305 | | |
| 26 | ACTN2 | P35609 | -8.49 | 11 | | |
| 27 | POMT1 | Q5JT03 | -8.50 | 3 | | |
| 28 | LOC130074 | Q6NZ40 | -8.50 | 16 | | x |
| 29 | CMD1K | 14541 | -8.50 | 27 | | |
| 30 | FER1L3 | Q9NZM1 | -8.51 | 3 | | |
| 31 | NOS1 | P29475 | -8.53 | 42 | | |
| 32 | IKBKAP | O95163 | -8.63 | 10 | | |
| 33 | MACF1 | Q5T3B3 | -8.66 | 9 | | |
| 34 | AQP4 | P55087 | -8.67 | 13 | | |
| 35 | CKM | P06732 | -8.70 | 11 | | |
| 36 | FSHMD1A | 3966 | -8.74 | 8 | | |
| 37 | TCAP | O15273 | -8.75 | 7 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 38 | **DTNB** | O60941 | -8.76 | 9 | x | |
| 39 | LOC619409 | 619409 | -8.82 | 5 | | |
| 40 | VCL | P18206 | -8.87 | 36 | | |
| 41 | LGMD1A | 6574 | -8.88 | 3 | | |
| 42 | **SNTG1** | Q9NSN8 | -8.90 | 5 | x | |

## CAPN3

For CAPN3, an evaluation of the precision is more difficult since there is, compared to an intensively studied protein such as DMD, not enough established knowledge about its regulatory partners and substrates. Table S6 summarizes the currently known interaction partners for CAPN3: 13 interactions have been described in the literature (not necessarily in the abstracts that were used for our predictions, see column ‘direct relation’) and of those, six interactions have been entered in PPI databases. These known interaction partners generally rank high in the list of literature-predicted targets (Table S6, p-value from Kolmogorov-Smirnov test: $5.7 \cdot 10^{-5}$). Interestingly, the concept profiling method correctly predicted the interaction between myosin light chain 1 (MYL1) and CAPN3 on the basis of conceptual overlap in MEDLINE abstracts (specificity > 99%), although this interaction was only described in a full text paper[23] and not in any MEDLINE abstract used to generate the concept profiles.

Apart from literature based rediscovery of known interactions, we also set out to actually find new interaction partners for CAPN3. We selected predicted interaction partners that have not been entered in PPI databases so far and that do not have a direct co-occurrence in MEDLINE. The top ranked conceptual match is with Sarcoglycan-epsilon (SGCE), which is the smooth muscle counterpart of SGCA. Like for CAPN3, mutations in SGCA cause LGMD, but as far as we know, the protein is not expressed in skeletal muscle.

The second highest ranking protein was deemed to be an interesting candidate by the experts: Parvalbumin B (PARVB). The concept profiling method yielded a high association score because both proteins are described to have a physical interaction with dysferlin (DYSF)[24, 25], and with α-actinin (ACTN2)[26, 27], and they are both located at the Z-disc[28, 29]. For this predicted protein pair, we experimentally demonstrated a physical interaction, using three different set-ups.

First, it was shown that immobilized GST-fused PARVB could pull down recombinant T7-CAPN3 from bacterial lysates. Second, immobilized GST-PARVB could pull down endogenous CAPN3 from IM2 mouse myoblasts, and vice versa (Figure S2).

CAPN3 is hypothesized to act as a cytoskeleton remodeler and has been shown to interact with other focal adhesion proteins like Talin and Vinexin[30] (see Table S6). Ectopic CAPN3 over-expression results in cell rounding and cleavage and loss of co-expressed Talin and Vinexin[30]. This suggests that CAPN3 is a modulator of focal adhesions. Like CAPN3, PARVB is predominantly expressed in skeletal muscle, where it plays a role in cell spreading and localizes to focal adhesions[26] (for a review, see [31]). The predicted interaction is coherent with this hypothesis, and substantiates the evidence for a role for CAPN3 outside the sarcomere.

This showcase is just one example of a correct and meaningful PPI prediction using concept profiles. This exemplary case study can not be seen proof that many of the other high ranking predictions will also be true physical and biologically relevant interactions. However none of the other consulted applications (STRING, iHOP) predicted this pair of interacting proteins. As the predictions using concept profiling are based on conceptual relatedness rather than an explicit co-occurrence in MEDLINE, this case study is indicative of the power of concept profiles to discover new, implicitly related pairs of interacting proteins. The statistics presented in this paper support the conclusion that predicted PPIs using our method, especially the subset that remains after expert analysis of the top ranking list are likely to be very significantly enriched for proteins that are worthwhile studying in wet lab experiments.

**Table 6. List of proteins known to interact with Calpain-3. In total 10,812 proteins known to have a concept profile are matched against Calpain-3.**

| Name | Symbol | In PPI set | In literature (full text) | Direct relation (abstract) | Rank in literature-based prediction | Significant at specificity of 95 % |
|---|---|---|---|---|---|---|
| Dysferlin | DYSF | x | x | x | 2 | x |
| Titin | TTN | x | x | x | 4 | x |
| Filamin C | FLNC | x | x | x | 27 | x |
| Alpha-actinin | ACTN2 |  | x | x | 43 | x |
| Calpastatin | CAST |  | x | x | 55 | x |
| IkappaBalpha | NFKBIA | x | x | x | 126 | x |
| Myosin light chain 1 | MYL1 |  | x |  | 398 | x |
| Alpha-spectrin | SPTAN1 | x | x |  | 426 | x |
| Filamin A | FLNA |  | x |  | 853 |  |
| Ezrin | VIL2 |  | x |  | 2739 |  |
| Vinexin | SORBS3 |  | x |  | 3301 |  |
| Talin | TLN1 |  | x |  | 4725 |  |
| AHNAK | AHNAK |  | x | No (*) | 7371 |  |
| YWHAQ | YWHAQ | x |  |  | 7617 |  |

(*) paper describing this interaction in the abstract appeared in June 2008 and was not in the literature corpus used for the prediction

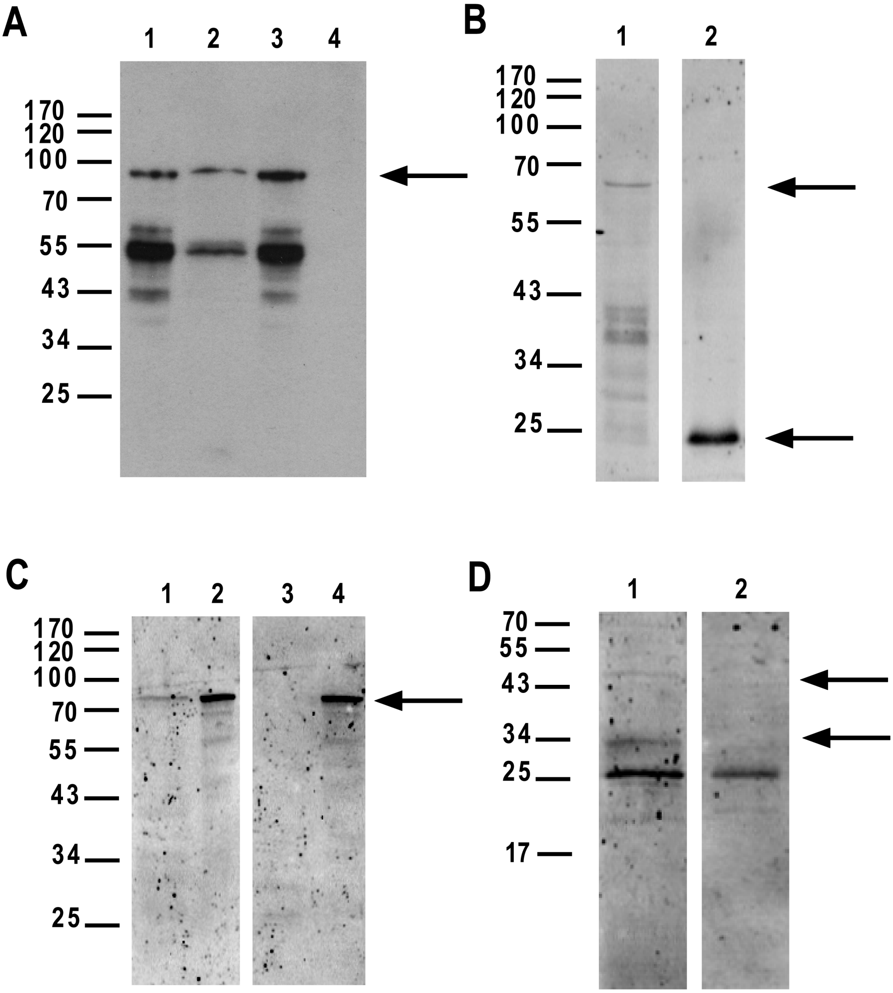


**Figure 2. CAPN3 and PARVB can directly interact. A: Immobilized GST-fused PARVB can pull down recombinant CAPN3 from a bacterial T7-tagged CAPN3 lysate (Lane 2 vs 1), where unfused GST cannot (Lane 4 vs 3). As CAPN3 is an unstable protein that outside skeletal muscle rapidly autolyses we used the active site mutant C129S48. All fractions were resolved on SDS-PAGE gel and analyzed by immunoblotting with anti-CAPN3. The lanes represent: GST-PARVB non-bound fraction (1), GST-PARVB bound fraction (2), GST non-bound fraction (3), GST bound fraction (4). B: Equal loading was confirmed with anti-GST (Lane 1 GST-PARVB, Lane 2 GST). C: GST-fused PARVB can pull down endogenous full-length CAPN3 from an IM2 lysate (Lane 1 vs 2), contrary to unfused GST (Lane 3 vs 4). Lane 1 GST-PARVB bound fraction, Lane 2 non-bound fraction, Lane 3 GST bound fraction, Lane 4 non bound fraction. D: Likewise, GST-CAPN3 can pull down endogenous PARVB (Lane 1), contrary to GST (Lane 2). Both PARVB translation products bind. Here we used the Δ6 variant of Capn3 that does not autolyse yet retains function30, 49, and is expressed in the proliferating IM2 myoblasts. The arrows indicate the detected proteins and in all panels a molecular marker is depicted on the left.**

## Discussion

Scientists in general and scientific annotators in particular derive their knowledge on PPIs not directly discovered by their own experiments from the literature. However, as we show here, only 32% of the known PPIs covered by curated PPI databases can be found in MEDLINE abstracts (Table S1), the resource that is most commonly used for concept searches in the biomedical domain. This is despite the use of a sophisticated synonym expansion and homonym disambiguation systems . It is likely that many of these interactions are only mentioned in the full text of articles, or that the interactions have never been explicitly described in literature but were directly submitted to a database. In either case, the applicability of the most commonly used approach for PPI detection - the direct relation method in publicly available literature - appears to be severely limited.

The specificity and sensitivity levels achieved by our novel prediction method appear to be very promising. However, when we predict interaction partners for a specific protein, the estimated precision levels (*i.e.* how many of the predicted proteins are true interaction partners) are still seemingly quite moderate. A first consideration is that we are intrinsically unable to determine an accurate 'true false positive rate' for the predicted PPIs, due to the fact that many PPIs have simply not been discovered and described yet. This unavoidable complication most certainly will lead to an underestimation of precision levels. The case study of CAPN3 and PARVB signifies this point; initially this pair would have been classified as a 'false positive'.

For a realistic estimation of the precision of our prediction method, effectively each predicted protein pair should be validated in a wet lab experiment, which is out of the realistic scope of this study. For this reason we developed Nermal. (http://biosemantics.org/nermal). In Nermal, researchers can enter the UniProt identifier of a protein of interest, and the tool will return a ranked list of proteins that are most likely to interact with the query protein, in combination with information on whether the PPI has already been described explicitly in MEDLINE and/or in one of the protein databases.

A second complicating factor is the size of the 'negative' set (>50 million) compared to the 'positive' set (44,920) . This aspect is illustrated by the average prediction performance for each protein in Table S3 and by the case study with DMD in Table S5, where the top 42 proteins yielded a precision of only 26%, whilst the specificity was 99.8%. We are currently working on a further improvement of the precision by including data sources other than the literature in the PPI prediction algorithms. A final consideration is that our predictions are yielding more conceptual connections than physically interacting proteins only. Conceptual overlap obviously can indicate a variety of other types of relations between proteins. For instance, we demonstrate that many proteins with high concept profile similarity do not interact directly, but are connected through intermediary proteins and are potentially part of the same complex or pathway. Therefore, the precision is to a certain extent dependent on the definition of a useful prediction. When other relationships than direct physical interactions are also deemed of interest, the precision of our method can become as high as 90%. The practical use of concept profiles will be in knowledge discovery in general, which is much broader than discovery of PPIs alone. In fact the hypothetical connection between any given pair of concepts can be calculated using our method.

To allow researchers to incorporate conceptual overlap data into their own analyses, we have made the concept profile similarity scores publicly available in two forms; first, a table containing similarity scores between all human proteins can be downloaded from our website; second, the previous mentioned web tool dubbed Nermal.

We conclude that concept profile similarity is a significantly better literature based predictor of PPIs than co-occurrence based methods. These improved predictions can be used to increase the biological interpretation and accuracy of interaction maps generated by high-throughput experiments, or can be used to prioritize proteins for further testing. In further studies, we will evaluate whether the use of concept profiles can also be applied in the prediction of other types of relations, for instance between drugs and diseases, and between genes and diseases.

## Methods

### Direct relation detection

Direct relations are typically extracted from literature based on co-occurrence[32]; if two proteins are mentioned in the same sentence or document more often than can be expected by chance, they are presumably related. We evaluated two alternatives for the detection of protein occurrences: a word-based approach and a concept-based approach. The word-based approach consists of combining the names of two proteins in an 'AND' query in the PubMed search engine. For the concept-based approach we have used the concept-recognition software Peregrine[33, 34], which includes synonyms and spelling variations[35] 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[36]. Even though a previous study has shown that Peregrine achieves state-of-the-art performance (75% precision and 76% recall on the BioCreactive II gene normalization testset[33, 34]), the concept recognition process is still error prone.

We used the likelihood ratio[19] to indicate the strength of the relation between two proteins. This ratio increases with the likelihood of there being a dependency between the occurrence of two proteins. Two hypotheses are used: (i) the occurrence of one protein is statistically dependent on the occurrence of the other protein; (ii) the occurrences are statistically independent. For each hypothesis a likelihood is calculated based on the observed data using the binomial distribution. The ratio of these likelihoods tells us how much more likely one hypothesis is over the other, or, in other words, how sure we are that there is a dependency. The following equations give the likelihood ratio $\lambda$ of concepts *i* and *j*.

$$\lambda(i,j)=\frac{L(n_{ij},n_i,p_j)L(n_j-n_{ij},N-n_i,p_j)}{L(n_{ij},n_i,p_1)L(n_j-n_{ij},N-n_i,p_2)}$$

where *N* is the total number of documents in the corpus, $n_i$ , $n_j$, and $n_{ij}$ are the number of documents containing *i, j,* and both *i* and *j*, respectively. $p=\frac{n_j}{N}$ , the probability *j* occurs in an abstract irrespective of *i,* $p_1=\frac{n_{ij}}{n_i}$ , the probability *j* occurs in an abstract containing *i*, $p_2=\frac{n_j-n_{ij}}{N-n_i}$ , the probability *j* occurs in a document not containing *i*, and $L(k,l,x)=x^k(1-x)^{l-k}$ , the likelihood function according to the binomial distribution.

**Concept profile-based relation detection**

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 contains all concepts that have a direct relation with a protein as found using the direct relation method described above. We evaluated two possible ways of applying this method: (i) using co-occurrences within a sentence, and (ii) using co-occurrences within an abstract. As shown in supplement S6, co-occurrence within an abstract yields a slightly higher AuC on predicting PPIs. We therefore used the abstract-based method in our study. The concepts in a profile include, in addition to proteins, all other concepts described in the Unified Medical Language System (UMLS) [37], such as diseases, symptoms, tissues, biological processes and many other types of concepts. We used the uncertainty coefficient[19] to calculate the weights of the concepts in the profiles. The uncertainty coefficient for the stochastic variables X and Y is given by

$$U(X|Y)=\frac{H(X)-H(X|Y)}{H(X)}$$

with *H(X)* is the entropy for *X* and *H(X|Y)* is the entropy for *X* given *Y*. X and Y can be any concept known in the ontology, e.g. drugs, proteins, diseases, disorders, chemicals, etc. The uncertainty coefficient is an information-theoretical measure that takes the a priori probability of direct relations into account. It gives extra weight to those concepts that are very specific for the set of documents belonging to the protein for which the concept profile is constructed. For a detailed description of concept profiles we refer to Jelier *et al.*[19].

The similarity score between two concept profiles A and B is taken as the inner product of the concept profile vectors, following Jelier *et al.*[38].

$$ip = \sum_{k=1}^{N} A_{uc(k)} B_{uc(k)}$$

with *uc(k)* the $k^{th}$ uncertainty coefficient in the profile and *N* the total number of concepts the two profiles have in common. The inner product increases with increasing overlap in concept profiles. If two proteins co-occur, the inner product of their concept profiles is in general high. This is shown in supplement S4.

**MEDLINE corpus**

We extracted the title and abstract of subsections of MEDLINE. The corpus used in our main study has a time span from 1980 up to July 2007 and contains 12,098,042 citations. The corpus used for the retrospective study has a time span from 1980 up to February 2005 and contains 10,363,027 citations. This is an increase in time of 9.8% whereas the increase in published articles over the last two years is 17%.

**Generation of the PPI and NIPP sets**

There are many protein databases that describe PPIs. Not all of these use protein identifiers that could be linked to our protein ontology and the databases also show a high degree of overlap (see supplement S2). In our analysis we use BioGRID[39], DIP[40], HPRD[41], IntAct[42], MINT[43], Reactome[44], and Swiss-Prot[45] and only consider human proteins. Except for IntAct, all these databases are curated, meaning that they only contain PPIs that were judged to be correct according to strict criteria. IntAct, on the other hand, also contains unchecked results from high-throughput experiments which could contain many false positives. For a comparison of the prediction performance of our method on the individual databases we refer to supplement S3. The release dates and dates of download can be found in supplement S1.

For the construction of our set of known PPIs, we only rely on the curated databases; if a PPI was mentioned in one of these databases, we assumed it to be a true PPI. The resulting positive set contains 61,807 PPIs. After removing pairs that are not covered by all four prediction methods, 44,920 PPIs remain. Unfortunately, there is no database of proteins that are known not to interact. We can therefore only create a set of proteins which are less likely to interact. For our NIPP set we took all pairs of human proteins that are not in the PPI set, and are not in the high-throughput part of the IntAct database. For computational reasons the calculation of the specificity and AuC was done on a random sample of 44,920 pairs of this set, setting both the positive and negative set size equal. Two randomly selected proteins form a pair and are checked if (i) they are not in the positive PPI set, (ii) not the same protein, e.g. proteins that interact with themselves are not taken into account, (iii) the protein pair is not already in the NIPP set, e.g. protein pairs can only occur once in a set. The random sample is actually quite small compared to the total NIPP set, however the ROC curve analysis is set size independent if the sample size is sufficiently large.

One last remark is that the positive set is incomplete. Therefore the creation of the NIPP set will introduce false negatives (PPIs that should have been in the positive set and recorded in a curated database). However the bias introduced by false negatives is negligible since the ratio of expected PPIs in human compared to the total set of formable protein pairs (~60 million) is very small[22].

**STRING database**

A copy of the STRING database, version 7.1, was downloaded from the STRING website. STRING is a pre-calculated database in PostgreSQL format. Only the text mining score table was used in our analysis.

**Sensitivity, Specificity, Precision**

In information retrieval terms like the sensitivity, specificity and precision are frequently used. The definitions are:

$$sensitivity=\frac{TP}{TP+FN}$$

$$specificity=\frac{TN}{TN+FP}$$

$$precision=\frac{TP}{TP+FP}$$

where TP are the number of true positives, FN number of false negatives, FP number of false positives, and TN number of true negatives. A perfect predictor has a specificity and sensitivity of 1.

When both set sizes are equal (#NIPP=#PPI) the precision equals the sensitivity. The specificity is sometimes confused with the precision. The distinction is critical when the classes are different sizes. A test with very high specificity can have a very low precision if there are far more true negatives than true positives, and vice versa.

**Online web tool Nermal**

Nermal is a web tool that prioritizes proteins that are most likely to be related with the protein you study. Given a query protein, the similarity scores are calculated between this protein and all other proteins in the ontology. The proteins are ranked on the similarity scores and presented in a table. Each row shows the similarity score between the two proteins, the databases in which the protein pair is known, and the sensitivity and (1-specificity) for that similarity score. These two rates should be interpreted as follows: given a similarity score between two proteins, (1-specificity) is the probability that a protein pair passing that score is a false positive. The sensitivity is the probability that you will miss a true PPI at that same score. Nermal can be found on http://biosemantics.org/nermal/. The full set of all protein pair match scores for human proteins can be downloaded at this link as well as the PPI and NIPP set used in the study.

**DNA cloning**

PARVB was amplified from proliferating IM2 myoblast cDNA with the following UTR primers: fw cgcactcgcttatgtcctc, rv ctccacatccttgtacttggtg. The ORF was amplified with a nested PCR introducing restriction sites for cloning into pET28aGST (modified pET28a vector with GST tag instead of T7 [46]). Primers were: fw aatatggatcctcctccgcgccaccacggt, rv atattctcgagctccacatccttgtacttgg. CAPN3 was similarly amplified with primers fw atgccaactgttattagtc, and rv ctaggcatacatggtaagc, and cloned into pET28aGST using fw tattacggatccatgccaactgttattagtc, and rv gtaatactcgagctaggcatacatggtaagc. The exon 6 deletion that does not autolyse was used for this experiment.

CAPN3c129s in pET28c was described previously[47]. All DNA constructs were verified by direct sequencing (LGTC, Leiden, The Netherlands), and subsequently transformed into BL21 (DE3)-RIL *E. coli* cells (Stratagene) for protein production.

**Protein production and preparation of lysates**

BL21 cells transformed with pET28aGST, pET28aGST-PARVB, pET28aGST-CAPN3 or pET28cCAPN3c129s were grown to log phase and stimulated with 1mM IPTG (Fermentas), and left to grow for 3 h at 37 °C. Next cells were spun down at 3,000 g and 4°C for 15 min. Pellets were dissolved in lysis buffer A (50 mM Tris-HCl pH 7.4, 1mM EDTA, 1.5 mg/ml lysozyme, 0.15 M NaCl, 1% Triton, Benozonase, 2x protease inhibitor cocktail tablet (Roche Molecular Biochemicals, Basel, Switzerland)), and sonicated on ice. Lysate was cleared by centrifugation at 13,000 g, and 4 °C for 30 min.

IM2 cells were grown at 33°C and 10% CO2 in DMEM 60196 (GIBCO-BRL, Grand-Island, NY) supplemented with 20% FCS, INFγ, glucose, pen/strep, glutamine and chick embryo extract. 15 cm plates (2x) were grown 75% confluent, washed 1x with PBS (37 °C) and lysed on ice with 1 ml lysis buffer B (50 mM

Tris-HCl pH 7.5, 150 mM NaCl, 0.2% Triton X-100, 2x protease inhibitor cocktail tablet). Lysate was spun down at 13,000 g and 4 °C for 30 min.

### Pull-down

GST sepharose beads (4B, Amersham, Uppsala, Sweden) were washed with PBS (2x) and pre-equilibrated with lysis buffer (2x), and added to the cleared GST fusion lysates. Lysates were incubated at 4 °C and tumbling for 2 h. Next the lysates were spun down at 500 g, 4 °C for 5 min, and washed 3x with lysis buffer A. Separately, IM2 lysates were treated with washed and pre-equilibrated GST sepharose beads (buffer B). An aliquot of the GST fusion proteins was loaded on SDS-PAGE gel and Coomassie stained to confirm equal loading.

IM2 lysate, or T7-CAPN3c129s lysate, was added to the bait, and incubated O/N at 4 °C and tumbling. GST sepharose beads were spun down and the sup was stored as non-bound fraction. The beads were washed 5x with ice cold lysisbuffer (A or B, 3x short, 2x five minutes tumbling). All remaining sup was removed with an insulin syringe and proteins were eluted with 2x Laemmli sample buffer and boiled 5 min. An aliquot of the non-bound fraction was similarly prepared.

### Western blot

Samples were loaded onto SDS-PAGE gels, separated and blotted to PVDF membrane. Blots were blocked in 4% skimmed milk PBS (Marvel) and incubated with primary antibody O/N at 4°C. Next morning blots were washed with 0.05% Tween in PBS, and incubated with secondary antibody for 1 h. Blots were washed again and scanned with an Odyssey scanner (Licor) or incubated with ECL plus (Amersham) and exposed to a Kodak XAR film. The following antibodies were used for Western detection: GaGST (1;10,000 Stratagene) MaCAPN3 (1;100, 12A2 Novocasta, Newcastle, UK), GaPARVB (1;200 Santa Cruz), GaMouseIRDye680 (1;5,000 Westburg, Leusden, NL), DaGIRDye800 (1;5,000 Westburg), RaMouseHRP (1;2,000 Dako Cytomation, Glostrup, Denmark), DaGoatHRP (1;10,000 Promega).

### References

1. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
2. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.* Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1143-7.

3. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.

4. Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.

5. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.

6. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.

7. Harrington, E.D., Jensen, L.J., and Bork, P., *Predicting biological networks from genomic data.* FEBS Lett, 2008. **582**(8): p. 1251-8.

8. Li, S., Wu, L., and Zhang, Z., *Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach.* Bioinformatics, 2006. **22**(17): p. 2143-50.

9. Kuffner, R., Fundel, K., and Zimmer, R., *Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts.* Bioinformatics, 2005. **21 Suppl 2**: p. ii259-67.

10. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., et al., *STRING 7--recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.

11. Hoffmann, R. and Valencia, A., *A Gene Network for Navigating the Literature.* Nature Genetics, 2004. **36**: p. 664.

12. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.

13. Swanson, D.R., *Medical literature as a potential source of new knowledge.* Bull Med Libr Assoc, 1990. **78**(1): p. 29-37.

14. Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., and Garner, H.R., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.

15. Schuemie, M.J., Chichester, C., Lisacek, F., Coute, Y., Roes, P.J., et al., *Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE.* Proteomics, 2007. **7**(6): p. 921-31.

16. Jelier, R., Jenster, G., Dorssers, L.C., Wouters, B.J., Hendriksen, P.J., et al., *Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation.* BMC Bioinformatics, 2007. **8**: p. 14.

17. Tuason, O., Chen, L., Liu, H., Blake, J.A., and Friedman, C., *Biological nomenclatures: a source of lexical knowledge and ambiguity.* Pac Symp Biocomput, 2004: p. 238-49.

18. Chen, L., Liu, H., and Friedman, C., *Gene name ambiguity of eukaryotic nomenclatures.* Bioinformatics, 2005. **21**(2): p. 248-56.

19. 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.

20. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., and Leunissen, J.A., *A text-mining analysis of the human phenome.* Eur J Hum Genet, 2006. **14**(5): p. 535-42.

21. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.

22. Ben-Hur, A. and Noble, W., *Choosing negative examples for the prediction of protein-protein interactions*. 2006. p. S2.

23. Cohen, N., Kudryashova, E., Kramerova, I., Anderson, L.V., Beckmann, J.S., et al., *Identification of putative in vivo substrates of calpain 3 by comparative proteomics of overexpressing transgenic and nontransgenic mice.* Proteomics, 2006. **6**(22): p. 6075-84.

24. Matsuda, C., Kameyama, K., Tagawa, K., Ogawa, M., Suzuki, A., et al., *Dysferlin interacts with affixin (beta-parvin) at the sarcolemma.* J Neuropathol Exp Neurol, 2005. **64**(4): p. 334-40.

25. Huang, M., Zhu, X., Hao, Y., Payan, D.G., Qu, K., et al., *Discovering patterns to extract protein-protein interactions from full texts.* Bioinformatics, 2004. **20**(18): p. 3604-12.

26. Yamaji, S., Suzuki, A., Kanamori, H., Mishima, W., Yoshimi, R., et al., *Affixin interacts with alpha-actinin and mediates integrin signaling for reorganization of F-actin induced by initial cell-substrate interaction.* J Cell Biol, 2004. **165**(4): p. 539-51.

27. Ojima, K., Ono, Y., Doi, N., Yoshioka, K., Kawabata, Y., et al., *Myogenic stage, sarcomere length, and protease activity modulate localization of muscle-specific calpain.* J Biol Chem, 2007. **282**(19): p. 14493-504.

28. Sorimachi, H., Kinbara, K., Kimura, S., Takahashi, M., Ishiura, S., et al., *Muscle-specific calpain, p94, responsible for limb girdle muscular dystrophy type 2A, associates with connectin through IS2, a p94-specific sequence.* J Biol Chem, 1995. **270**(52): p. 31158-62.

29. Bendig, G., Grimmler, M., Huttner, I.G., Wessels, G., Dahme, T., et al., *Integrin-linked kinase, a novel component of the cardiac mechanical stretch sensor, controls contractility in the zebrafish heart.* Genes Dev, 2006. **20**(17): p. 2361-72.

30. Taveau, M., Bourg, N., Sillon, G., Roudaut, C., Bartoli, M., et al., *Calpain 3 is activated through autolysis within the active site and lyses sarcomeric and sarcolemmal components.* Mol Cell Biol, 2003. **23**(24): p. 9127-35.

31. Sepulveda, J.L. and Wu, C., *The parvins.* Cell Mol Life Sci, 2006. **63**(1): p. 25-35.

32. Cohen, A.M. and Hersh, W.R., *A survey of current work in biomedical text mining.* Brief Bioinform, 2005. **6**(1): p. 57-71.

33. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.

34. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., et al., *Overview of BioCreative II gene normalization.* Genome Biol, 2008. **9 Suppl 2**: p. S3.

35. Schuemie, M.J., Mons, B., Weeber, M., and Kors, J.A., *Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification.* J Biomed Inform, 2007. **40**(3): p. 316-24.

36. Kors, J.A., Schuemie, M.J., Schijvenaars, B.J.A., Weeber, M., and Mons, B., *Combination of genetic databases for improving identification of genes and proteins in text.* BioLINK, 2005

37. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.

38. Jelier, R., Schuemie, M.J., Veldhoven, A., Dorssers, L.C., Jenster, G., et al., *Anni 2.0: a multipurpose text-mining tool for the life sciences.* Genome Biol, 2008. **9**(6): p. R96.

39. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Research, 2006. **34**(Database): p. 535-539.

40. 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.

41. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.

42. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.

43. 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.

44. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.

45. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A., *UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase.* Methods Mol Biol, 2007. **406**: p. 89-112.

46. Huang, Y., Laval, S.H., van Remoortere, A., Baudier, J., Benaud, C., et al., *AHNAK, a novel component of the dysferlin protein complex, redistributes to the cytoplasm with dysferlin during skeletal muscle regeneration.* Faseb J, 2007. **21**(3): p. 732-42.

47. Huang, Y., de Morree, A., van Remoortere, A., Bushby, K., Frants, R.R., et al., *Calpain 3 is a modulator of the dysferlin protein complex in skeletal muscle.* Hum Mol Genet, 2008. **17**(12): p. 1855-66.

**Supplementary information belonging to the article "Novel protein-protein interactions inferred from literature context"**

## S1 Downloaded protein database and release dates

In total seven protein databases are used in the study. The UniProt database consists of Swiss-Prot and TrEMBL.

| Protein database | Date of download |
|---|---|
| Biogrid | September 28, 2007 |
| DIP | September 20, 2007 |
| HPRD | August 22, 2007* |
| IntAct | January 26, 2008 |
| MINT | September 24, 2007* |
| Reactome | September 20, 2007 |
| UniProt | February 14, 2008* |

* For these databases it is possible to retrieve the original release dates. HPRD was released at January 9, 2007, MINT at June 28, 2007. Swiss-Prot and TrEMBL are combined in the database UniProt and have different release versions. UniProt release 12.0 contains Swiss-Prot release 54.0 and TrEMBL release 37.0. Both are dated from July 24, 2007.

## S2 PPI overlap between the seven databases

Many of the PPIs appear in several databases. The following table shows the distribution and overlap over the seven protein databases.

| | Biogrid | DIP | HPRD | IntAct | MINT | Reactome | Swiss-Prot |
|---|---|---|---|---|---|---|---|
| Biogrid | **16240** | 205 | 15476 | 3006 | 2637 | 909 | 827 |
| DIP | | **365** | 278 | 84 | 118 | 66 | 53 |
| HPRD | | | **34957** | 8031 | 7046 | 1401 | 1839 |
| IntAct | | | | **17456** | 5754 | 595 | 3839 |
| MINT | | | | | **10772** | 375 | 650 |
| Reactome | | | | | | **29672** | 290 |
| Swiss-Prot | | | | | | | **3841** |

## S3 Performance on individual databases

The positive set is a combination of six protein databases. The databases vary in size and also the level of curation of each PPI. The following table gives the Area under the ROC (AuC) curve for each database individually. The last row is the AuC for the complete positive set.

| Database | Concept profiles | Log likelihood | String |
|---|---|---|---|
| Biogrid | 0.95 | 0.82 | 0.82 |
| Dip | 0.99 | 0.96 | 0.94 |
| Hprd | 0.93 | 0.79 | 0.78 |
| Intact | 0.71 | 0.57 | 0.56 |
| Mint | 0.87 | 0.72 | 0.70 |
| Reactome | 0.90 | 0.60 | 0.60 |
| Swiss-Prot | 0.84 | 0.71 | 0.71 |
| Positive set | 0.90 | 0.69 | 0.69 |

## S4 Relationship between direct relation detection and concept profiles

The coverage in S3 shows that some PPIs have both overlap in concept profiles and a direct relation, while others have only concept profile overlap. The similarity score for proteins that share a direct relation is generally high. This is illustrated in figure 1.

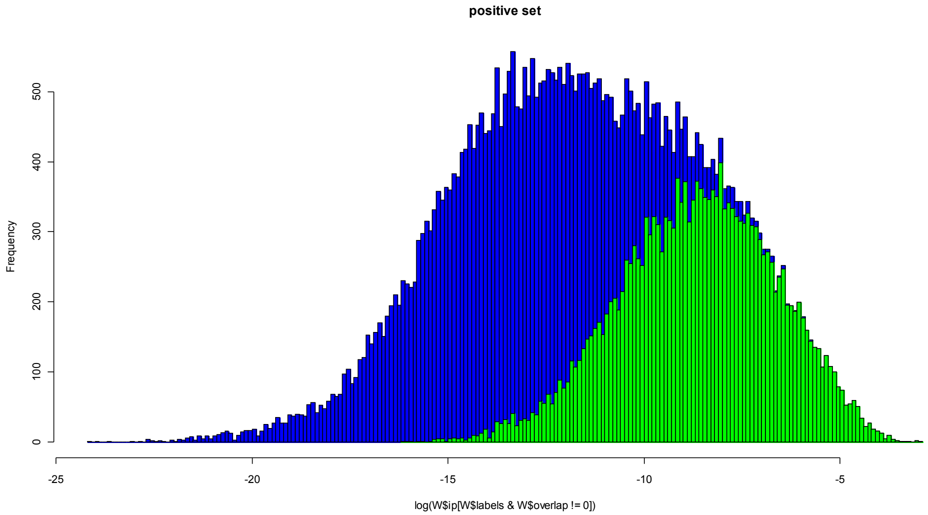
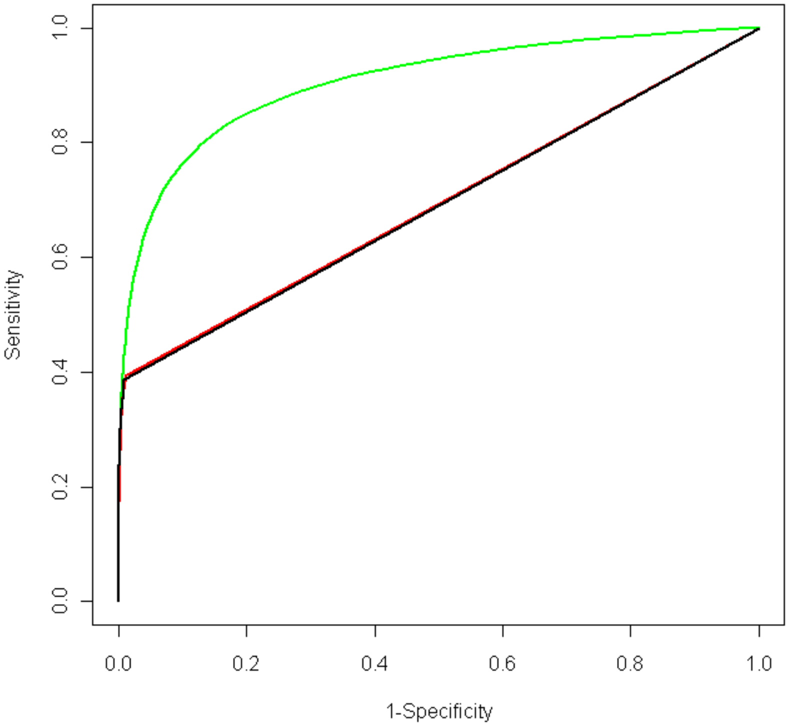


**Figure 1. Histogram of the distribution of the similarity scores of: (blue) PPIs with concept profile overlap and no direct relation, and (green) PPIs with both a concept profile overlap and a direct relation.**

## S5 ROC curve analysis

The next figure shows the ROC curves for the concept profile similarity score (green), and the likelihood ratio of the direct relation method (red). For the direct relation method we discern two special cases: (i) each protein individual occurs in Medline but they are never mentioned together, and (ii) one of the proteins does not occur in MedLine at all. In the first case the likelihood score is –infinity, in the second case the likelihood score is 0. These cases are quite frequent resulting in many duplicate values, and no natural ordering of the PPIs. We assume a perfect random ordering, resulting in the straight line at the end of the ROC curve in the figure (red for concept based method and black for the String database).



## S6 Relation detection at the abstract and sentence level

For the construction of concept profiles, we investigated two options: assume two concepts are related when they co-occur (i) in the same sentence, and (ii) in the same abstract. For each option we evaluated the performance on the prediction of PPIs.

| | Abstract level | Sentence level |
|---|---|---|
| *AuC** | 0.93 | 0.91 |

The difference in results are neglectable. There is a very small decrease in performance using sentence based detection of relations.

* this analysis was done using a MedLine corpus up to April 2007 and using an older ontology.

## S7 Ranked list of proteins predicted to interact with dystrophin (DMD)

The following table shows the proteins which similarity score with DMD have a specificity higher than 99%.

| Rank | Protein symbol | Swiss-Prot id | Log similarity score | Direct relations | FP rate | TP rate | Biogrid | Dip | Hprd | Intact | Mint | Reactome | Swiss-Prot |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | UTRN | P46939 | -5.14 | 214 | 0.003 | 0.856 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | SGCA | Q16586 | -6.13 | 119 | 0.013 | 4.047 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | DAG1 | Q14118 | -6.22 | 139 | 0.013 | 4.047 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | SGCB | Q16585 | -6.6 | 54 | 0.022 | 5.853 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | SGCD | Q53XA5 | -6.95 | 46 | 0.032 | 8.168 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | FCMD | O75072 | -7.05 | 29 | 0.034 | 8.62 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | DYSF | O75923 | -7.19 | 43 | 0.039 | 9.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | DTNA | Q9BS59 | -7.31 | 17 | 0.048 | 10.576 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 9 | DRP2 | Q13474 | -7.34 | 9 | 0.049 | 10.625 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | SSPN | Q0JV68 | -7.45 | 17 | 0.055 | 11.543 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | LAMA2 | P24043 | -7.46 | 25 | 0.055 | 11.543 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | GK1 | P32189 | -7.56 | 33 | 0.059 | 12.306 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | CAPN3 | P20807 | -7.93 | 28 | 0.08 | 15.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | CAV3 | P56539 | -7.95 | 24 | 0.08 | 15.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | SNTA1 | Q13424 | -7.97 | 8 | 0.081 | 15.274 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 16 | EIF3S12 | Q9UBQ5 | -8.05 | 91 | 0.091 | 16.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | BEST1 | O76090 | -8.13 | 26 | 0.096 | 16.703 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | SPTB | P11277 | -8.15 | 15 | 0.097 | 16.896 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | FKRP | Q9H9S5 | -8.16 | 4 | 0.098 | 17.046 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | MEB | 6988 | -8.17 | 7 | 0.099 | 17.106 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | SLMAP | Q14BN4 | -8.2 | 4 | 0.102 | 17.288 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | SNTB1 | Q13884 | -8.2 | 6 | 0.102 | 17.288 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 23 | NEB | P20929 | -8.33 | 16 | 0.117 | 18.497 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | SGCE | O43556 | -8.35 | 10 | 0.117 | 18.497 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | SGCG | Q13326 | -8.46 | 305 | 0.132 | 19.584 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | ACTN2 | P35609 | -8.49 | 11 | 0.137 | 19.754 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | POMT1 | Q5JT03 | -8.5 | 3 | 0.137 | 19.754 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | LOC130074 | Q6NZ40 | -8.5 | 16 | 0.138 | 19.925 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | CMD1K | 14541 | -8.5 | 27 | 0.138 | 19.925 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | FER1L3 | Q9NZM1 | -8.51 | 3 | 0.138 | 19.925 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | NOS1 | P29475 | -8.53 | 42 | 0.139 | 20.11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | IKBKAP | O95163 | -8.63 | 10 | 0.152 | 21.011 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | MACF1 | Q5T3B3 | -8.66 | 9 | 0.162 | 21.337 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 34 | AQP4 | P55087 | -8.67 | 13 | 0.162 | 21.337 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 35 | CKM | P06732 | -8.7 | 11 | 0.167 | 21.668 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 36 | FSHMD1A | 3966 | -8.74 | 8 | 0.172 | 21.859 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 37 | TCAP | O15273 | -8.75 | 7 | 0.173 | 22.153 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 38 | DTNB | O60941 | -8.76 | 9 | 0.173 | 22.153 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | LOC619409 | 619409 | -8.82 | 5 | 0.181 | 22.675 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 40 | VCL | P18206 | -8.87 | 36 | 0.189 | 23.173 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 41 | LGMD1A | 6574 | -8.88 | 3 | 0.192 | 23.273 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 42 | SNTG1 | Q9NSN8 | -8.9 | 5 | 0.194 | 23.459 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 43 | EMD | P50402 | -8.94 | 12 | 0.201 | 23.864 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 44 | GNE | Q6QNY6 | -9 | 7 | 0.205 | 24.407 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | MYOZ2 | Q9NPC6 | -9.03 | 7 | 0.209 | 24.632 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 46 | PGM5 | Q15124 | -9.04 | 3 | 0.212 | 24.733 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 47 | CASQ1 | P31415 | -9.05 | 5 | 0.213 | 24.892 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 48 | NR0B1 | P51843 | -9.06 | 18 | 0.218 | 25.047 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 49 | SYNC1 | Q9H7C4 | -9.08 | 4 | 0.219 | 25.066 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | TTN | Q8WZ42 | -9.08 | 7 | 0.22 | 25.157 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 51 | DENR | O43583 | -9.12 | 3 | 0.228 | 25.497 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 52 | POMGNT1 | Q8WZA1 | -9.15 | 7 | 0.233 | 25.802 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 53 | RAPSN | Q13702 | -9.19 | 8 | 0.239 | 26.192 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 54 | MYOT | Q9UBF9 | -9.27 | 5 | 0.253 | 27.025 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 55 | GDF8 | O14793 | -9.28 | 5 | 0.254 | 27.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 56 | AIED | 351 | -9.3 | 2 | 0.256 | 27.193 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 57 | TRIM32 | Q13049 | -9.31 | 3 | 0.256 | 27.193 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 58 | MYH7 | P13533 | -9.36 | 18 | 0.265 | 27.894 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 59 | LAMB1 | P07942 | -9.36 | 6 | 0.266 | 27.898 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 60 | RP23 | 10277 | -9.41 | 6 | 0.274 | 28.242 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 61 | SNTG2 | Q05AH5 | -9.42 | 2 | 0.275 | 28.462 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 62 | ACTN3 | Q08043 | -9.46 | 5 | 0.284 | 28.783 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 63 | LMNA | P02545 | -9.46 | 17 | 0.285 | 28.814 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 64 | SPTBN4 | Q9H254 | -9.51 | 1 | 0.289 | 29.342 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 65 | OTC | P00480 | -9.55 | 8 | 0.298 | 29.59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 66 | DTNBP1 | Q96EV8 | -9.56 | 5 | 0.299 | 29.616 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 67 | SNTB2 | Q13425 | -9.56 | 2 | 0.302 | 29.732 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 68 | LGMD1B | 6575 | -9.57 | 0 | 0.304 | 29.838 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 69 | SYNPO2 | Q9UMS6 | -9.57 | 3 | 0.307 | 29.862 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 70 | RPGR | Q4VX65 | -9.59 | 5 | 0.314 | 29.997 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 71 | SPTBN1 | Q01082 | -9.59 | 7 | 0.314 | 29.997 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 72 | GYPC | P04921 | -9.6 | 3 | 0.318 | 30.105 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 73 | TAZ | Q16635 | -9.63 | 8 | 0.329 | 30.387 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 74 | SNORD95 | 32757 | -9.63 | 3 | 0.329 | 30.387 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75 | DMN | O15061 | -9.64 | 3 | 0.33 | 30.483 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 76 | SEPN1 | Q9NZV5 | -9.74 | 2 | 0.364 | 31.413 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 77 | GATM | P50440 | -9.76 | 2 | 0.37 | 31.678 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 78 | MTM1 | Q13496 | -9.78 | 5 | 0.372 | 31.823 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 79 | PLEC1 | Q15149 | -9.82 | 1 | 0.384 | 32.306 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 80 | NRG4 | Q0P6N4 | -9.82 | 1 | 0.387 | 32.363 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 81 | AAVS1 | 22 | -9.83 | 4 | 0.389 | 32.414 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 82 | MYOD1 | O75321 | -9.84 | 9 | 0.389 | 32.414 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 83 | FLNC | Q14315 | -9.87 | 3 | 0.398 | 32.802 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 84 | VAULTRC3 | 12656 | -9.88 | 1 | 0.4 | 32.846 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 85 | CFC1 | Q9GZR3 | -9.89 | 16 | 0.401 | 32.965 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 86 | IL1RAPL1 | Q7Z2K4 | -9.9 | 4 | 0.403 | 33.116 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 87 | DYNLT3 | P51808 | -9.91 | 3 | 0.406 | 33.239 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 88 | DTL | Q9NZJ0 | -9.93 | 2 | 0.411 | 33.417 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 89 | DMPK | Q09013 | -9.93 | 5 | 0.411 | 33.417 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 90 | MYOG | P15173 | -9.94 | 8 | 0.414 | 33.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 91 | DGKZ | Q13574 | -9.95 | 2 | 0.417 | 33.614 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 92 | SRRM2 | O60382 | -9.96 | 2 | 0.418 | 33.686 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 93 | SMN1 | Q16637 | -10.04 | 3 | 0.441 | 34.576 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 94 | MYL2 | P10916 | -10.05 | 2 | 0.445 | 34.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 95 | MYLPF | Q6IB41 | -10.09 | 2 | 0.457 | 35.062 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 96 | PVALB | P02144 | -10.1 | 22 | 0.464 | 35.21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 97 | COL6A1 | P12109 | -10.14 | 2 | 0.473 | 35.566 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 98 | MYH7 | P12883 | -10.14 | 2 | 0.474 | 35.583 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 99 | CAPN8 | 1485 | -10.14 | 1 | 0.476 | 35.607 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | MEAX | 6987 | -10.15 | 2 | 0.477 | 35.638 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 101 | POMT2 | Q59GJ5 | -10.15 | 0 | 0.479 | 35.702 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 102 | AGRN | O00468 | -10.18 | 3 | 0.483 | 35.963 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 103 | DNPEP | Q9HAC6 | -10.18 | 2 | 0.484 | 35.967 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 104 | XIC | 12809 | -10.19 | 0 | 0.491 | 36.045 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 105 | PDLIM3 | Q53GG5 | -10.2 | 2 | 0.499 | 36.198 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 106 | COL6A2 | P12110 | -10.21 | 1 | 0.5 | 36.268 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 107 | GAA | P10253 | -10.21 | 7 | 0.501 | 36.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 108 | LAMA1 | P25391 | -10.26 | 0 | 0.52 | 36.811 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 109 | MYF6 | P23409 | -10.27 | 2 | 0.524 | 36.845 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 110 | CHRNG | P07510 | -10.29 | 1 | 0.531 | 37.012 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 111 | SPTA1 | O60686 | -10.3 | 2 | 0.535 | 37.144 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 112 | CSRP3 | P50461 | -10.3 | 3 | 0.542 | 37.216 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 113 | EPB41 | P11171 | -10.31 | 4 | 0.548 | 37.322 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 114 | PBDX | P55808 | -10.32 | 1 | 0.548 | 37.322 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 115 | LAMB2 | P55268 | -10.32 | 1 | 0.549 | 37.398 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 116 | WDM | 50988 | -10.34 | 1 | 0.561 | 37.627 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 117 | HHG | 4902 | -10.35 | 1 | 0.563 | 37.68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 118 | RPS4Y1 | P22090 | -10.35 | 2 | 0.563 | 37.68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 119 | ITGA7 | Q13683 | -10.35 | 1 | 0.563 | 37.68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 120 | TNNT2 | P45379 | -10.37 | 4 | 0.574 | 37.877 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 121 | CMD1B | 2102 | -10.37 | 2 | 0.574 | 37.877 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 122 | FOSL2 | P15408 | -10.38 | 1 | 0.578 | 38.028 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 123 | SFRS2 | Q01130 | -10.39 | 3 | 0.582 | 38.131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 124 | MIB2 | Q0JSM5 | -10.4 | 1 | 0.59 | 38.257 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 125 | MSRB2 | Q9Y3D2 | -10.4 | 1 | 0.59 | 38.257 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 126 | DNM1L | O00429 | -10.4 | 1 | 0.59 | 38.257 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 127 | XKR1 | P51811 | -10.41 | 2 | 0.593 | 38.307 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 128 | XIST | 12810 | -10.42 | 0 | 0.605 | 38.416 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 129 | TPM1 | O15513 | -10.42 | 5 | 0.606 | 38.439 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 130 | COL6A3 | P12111 | -10.42 | 1 | 0.61 | 38.522 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 131 | SUCLG1 | P53597 | -10.42 | 1 | 0.613 | 38.547 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 132 | NRG3 | P56975 | -10.43 | 1 | 0.614 | 38.587 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 133 | PPP1R10 | Q96QC0 | -10.43 | 1 | 0.614 | 38.587 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 134 | RNPS1 | Q15287 | -10.44 | 2 | 0.623 | 38.738 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 135 | MYEF2 | Q9P2K5 | -10.46 | 2 | 0.632 | 38.88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 136 | GAMT | Q14353 | -10.48 | 1 | 0.646 | 39.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 137 | TNNC1 | P63316 | -10.49 | 3 | 0.65 | 39.232 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 138 | RP2 | O75695 | -10.51 | 3 | 0.661 | 39.446 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 139 | MYL6 | P60660 | -10.51 | 1 | 0.663 | 39.469 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 140 | CTSH | P09668 | -10.51 | 2 | 0.664 | 39.48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 141 | CXADR | P78310 | -10.53 | 4 | 0.681 | 39.609 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 142 | ELOVL4 | Q9GZR5 | -10.53 | 1 | 0.682 | 39.635 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 143 | MYF5 | P13349 | -10.57 | 3 | 0.703 | 40.025 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 144 | FBXO32 | Q969P5 | -10.59 | 1 | 0.716 | 40.275 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 145 | PRX | Q9BXM0 | -10.6 | 2 | 0.716 | 40.275 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 146 | BLOC1S1 | P78537 | -10.6 | 1 | 0.718 | 40.311 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 147 | MUSK | O15146 | -10.6 | 1 | 0.718 | 40.311 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 148 | SMN2 | Q16637 | -10.62 | 1 | 0.731 | 40.491 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 149 | IS2 | 282552 | -10.62 | 0 | 0.734 | 40.516 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 150 | DM1 | 2923 | -10.62 | 2 | 0.735 | 40.521 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 151 | DYNLL1 | P63167 | -10.63 | 1 | 0.737 | 40.567 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 152 | PDAP1 | Q13442 | -10.63 | 1 | 0.737 | 40.567 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 153 | INVS | Q5JS85 | -10.65 | 3 | 0.748 | 40.76 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 154 | PABPN1 | Q86U42 | -10.66 | 1 | 0.76 | 40.877 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 155 | NOS1AP | O75052 | -10.67 | 1 | 0.762 | 40.915 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 156 | KCNJ10 | P78508 | -10.67 | 3 | 0.765 | 40.985 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 157 | TCTA | P57738 | -10.68 | 0 | 0.767 | 41.034 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 158 | ACTA1 | P68133 | -10.68 | 2 | 0.769 | 41.074 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 159 | CACNA1I | Q9P0X4 | -10.68 | 1 | 0.776 | 41.155 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 160 | MST4 | Q8NC04 | -10.71 | 1 | 0.786 | 41.424 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 161 | KFSD | 6313 | -10.72 | 2 | 0.786 | 41.424 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 162 | IGFBP5 | P24593 | -10.72 | 1 | 0.789 | 41.494 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 163 | DST | O94833 | -10.75 | 7 | 0.812 | 41.721 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 164 | FRG1 | Q14331 | -10.76 | 0 | 0.82 | 41.913 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 165 | CD5L | O43866 | -10.77 | 0 | 0.823 | 41.983 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 166 | ITPR1 | Q14643 | -10.79 | 1 | 0.83 | 42.123 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 167 | PARVB | Q9HBI1 | -10.8 | 0 | 0.838 | 42.202 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 168 | RIMS1 | Q5SZK2 | -10.8 | 1 | 0.838 | 42.202 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 169 | GAS2 | O43903 | -10.8 | 3 | 0.845 | 42.293 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 170 | WAS | P42768 | -10.8 | 238 | 0.846 | 42.308 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 171 | CDH15 | P55291 | -10.81 | 2 | 0.849 | 42.363 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 172 | ACTC1 | P68032 | -10.82 | 1 | 0.85 | 42.429 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 173 | MLS | 7145 | -10.82 | 1 | 0.85 | 42.429 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 174 | CACNA1S | Q13698 | -10.82 | 3 | 0.851 | 42.507 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 175 | ERF | P50548 | -10.83 | 1 | 0.856 | 42.558 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 176 | SFRS1 | Q07955 | -10.84 | 1 | 0.865 | 42.672 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 177 | DCTN3 | O75935 | -10.87 | 1 | 0.894 | 42.976 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 178 | DDX3Y | O15523 | -10.87 | 1 | 0.894 | 42.976 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 179 | SFRS5 | Q13243 | -10.87 | 1 | 0.896 | 43.018 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 180 | ALG3 | Q92685 | -10.87 | 109 | 0.896 | 43.018 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 181 | RYR1 | O75591 | -10.88 | 1 | 0.908 | 43.141 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 182 | GAS2L1 | Q99501 | -10.9 | 1 | 0.919 | 43.308 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 183 | COL4A5 | P29400 | -10.9 | 0 | 0.919 | 43.308 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 184 | PTBP2 | O95652 | -10.91 | 0 | 0.926 | 43.393 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 185 | MYH6 | P13533 | -10.91 | 4 | 0.928 | 43.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 186 | IGFBP4 | P22692 | -10.92 | 3 | 0.933 | 43.582 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 187 | SYNE1 | Q5JV23 | -10.93 | 1 | 0.934 | 43.62 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 188 | ZNF91 | Q05481 | -10.93 | 1 | 0.939 | 43.637 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 189 | SP1 | P08047 | -10.93 | 7 | 0.94 | 43.669 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 190 | PTPN22 | Q5TBC0 | -10.93 | 1 | 0.941 | 43.671 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 191 | LOC619511 | 619511 | -10.94 | 1 | 0.943 | 43.701 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 192 | EIF4EBP1 | Q13541 | -10.95 | 3 | 0.953 | 43.838 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 193 | MYOZ1 | Q9NP98 | -10.97 | 0 | 0.977 | 44.029 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 194 | BSN | Q2NLD3 | -10.98 | 0 | 0.988 | 44.167 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 195 | FBXO11 | Q86XK2 | -10.99 | 1 | 0.999 | 44.248 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 196 | ZBTB20 | Q9HC78 | -10.99 | 1 | 0.999 | 44.248 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |