Journal of
Proteomics & Bioinformatics

Kafkas et al., J Proteomics Bioinform 2011, 4:11
DOI: 10.4172/jpb.1000198

Research Article Open Access

Diversity in the Interactions of Isoforms Linked to Clustered Transcripts: A Systematic Literature Analysis

Şenay Kafkas[1,2*], Ekrem Varoğlu[1], Dietrich Rebholz-Schuhmann[2] and Bahar Taneri[3,4]

[1]*Department of Computer Engineering, Eastern Mediterranean University, Famagusta, North Cyprus*
[2]*European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, Hinxton, CB10 1SD, UK*
[3]*Department of Biological Sciences, Faculty of Arts and Sciences, Eastern Mediterranean University, Famagusta, North Cyprus*
[4]*Institute of Public Health Genomics, Department of Genetics and Cell Biology, Research Institutes CAPHRI and GROW, Faculty of Health, Medicine and Life Sciences, Maastricht University, 6202 AZ Maastricht, The Netherlands*

Abstract

Existing protein-protein interactions databases cover only a portion of the interactome and interaction information on protein isoforms is underrepresented. This leads to a lack of information on the functional similarity of protein isoforms and the effects of transcript diversity on the protein interaction networks. We present a comprehensive automated literature analysis that extracts interactions involving human protein isoforms linked to clusters of transcripts with high sequence similarity and deliver them in a database called TBIID for knowledge discovery.

We measure the interaction variability of the isoforms from the clustered transcripts by analysing the distribution of their interaction partners in TBIID. Almost all clusters analyzed (99%) contain isoforms with unique partners indicating that isoforms are specialized towards forming unique interactions and thus achieving functional diversity, which is similar to the results from public resources. TBIID is available at http://tbiid.emu.edu.tr containing most relevant candidates for future experiments focusing on understanding the isoform interaction networks and the resulting functional implications.

Keywords: Protein isoforms; Protein-Protein Interactions; Machine learning

Abbreviations: PPI: Protein-Protein Interaction; TBIID: Transcript Based Isoform Interaction Database; DT: Defined Transcript; CMT: Cluster with Multiple defined Transcripts; CST: Cluster with Single defined Transcript; CUT: Cluster with Undefined Transcript; HumanSDB3: Human Splicing DataBase version 3; SVM: Support Vector Machine; PPIE: Protein-Protein Interaction Extraction; IAS: Interaction Article Sub-Task; TF: Term Frequency; ID: Identifier

Introduction

Recent research in molecular biology has focussed on the identification of protein-protein interactions (PPIs) and the analysis of PPI networks to fully understand the organism's functionality. These efforts have produced collections of PPI data by using high-throughput methods such as yeast two hybrid (Y2H) and affinity purification [1], as well as literature mining methods [2]. High-throughput methods are experimental while the literature mining methods are computational approaches which rely on biomedical text mining to gather the PPIs from textual data. The collected PPI data is stored in structured databases, which are generally accessible through the World Wide Web. Several comprehensive PPI databases are the Database of Interacting Proteins (DIP) [3], the Molecular INTeraction Database (MINT) [4] and IntAct [5]. However, these databases still cover only a portion of the interactome [6,7] and show limitations regarding PPIs involving protein isoforms. For example, in the PINA database [8] only a small portion of the interaction pairs (772, i.e. 1.3% of all interactions in PINA) involve a protein that is a splicing variant according to Uniprot Knowledge Base [9].

High-throughput technologies such as large-scale sequencing enable scientists to perform genome-wide searches for regions with similar transcripts. Such transcripts form the origin of proteome diversity and are induced by alternative splicing events. Constitutive RNA splicing removes introns (non-coding regions) from the premature messenger RNA (pre-mRNA) and ligates exons (protein-coding regions) in the order as they appear in the genomic DNA to form the final mRNA. On the other hand, alternative splicing generates multiple different mRNAs with different exon-intron combinations from a single gene, by making use of alternate splice-sites within the pre-mRNA [10]. Such mRNAs lead to the production of protein isoforms from the same gene possibly with differences in their structures and in their functions generated as a result of their sequence variations [11]. Hence, alternative splicing highly increases the coding potential of the genome, which can lead to a diverse proteome [10,12].

In principle, such isoforms either share the same function, show minimal functional differences, or have entirely opposite functions. We would expect such functional differences to be reflected in other properties of the isoforms, such as the variability of a protein in its interactions and interaction partners. An example can be given from the ROBO proteins, where their interactions with Slit ligands play role in neurogenesis regulation. The Slit receptor Robo3 has two isoforms, namely Robo3.1 and Robo3.2, which differ in their carboxy terminal groups leading to opposite functions. Robo3.1 silences Slit repulsion while Robo3.2 favours Slit repulsion. This difference in function induces opposite results regarding the midline crossing events in the commissural axons [13]. Alternative splicing is a widespread cellular

***Corresponding author:** Şenay Kafkas, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, Hinxton, CB10 1SD, UK.Tel: + 44 (0) 1223 494 545; Fax: +44 (0) 1223 494 468; E-mail: kafkas@ebi.ac.uk

Received October 12, 2011; **Accepted** November 12, 2011; **Published** November 29, 2011

Citation: Kafkas Ş, Varoğlu E, Rebholz-Schuhmann D, Taneri B (2011) Diversity in the Interactions of Isoforms Linked to Clustered Transcripts: A Systematic Literature Analysis. J Proteomics Bioinform 4: 250-259. doi:10.4172/jpb.1000198

mechanism present across eukaryotic genomes [10]. Another example can be given from the *C. elegans* genome. The FGF receptor, EGL-15 has two alternative splicing variants (EGL-15(5A) and EGL-15(5B)) which differ in their extracellular domains leading to different functions. These isoforms play a role in the gonadal chemoattraction of the migrating sex myoblasts (SMs). Isoform 5A is required for attraction of the migrating SMs to the gonad, while isoform 5B is required for repulsion of the migrating SMs from the gonad [14].

High throughput methods have initiated the development of reference databases such as ASTD [15], ProSAS [16] and ECgene [17] that gather transcript diversity and alternative splicing events. Through sequence analysis across the reference databases it has been revealed that a large portion of genes exhibit alternative splicing events [15,18] and thus contribute to the transcript diversity to different degrees in various species: in *homo sapiens* 81-94%[15,18,19], in *mus musculus* 74-79% [15,18], in *rattus norvegicus* 39-61% [15,18,20] and in *arabidopsis thaliana* 42% [21]. Since alternative splicing and as a result also transcript diversity are both widespread within and across a number of genomes, it has been concluded that this process has been conserved evolutionarily [22]. The amount of experimentally identified transcript sequences is representative for proportion of alternative splicing detected in a given genome [23]. As the amount of sequence data increases, the relevance of transcript diversity will also increase in importance, which leads again to a higher detection rate for the functional variability of protein isoforms.

Here, we complement alternative splicing and transcript diversity studies with biomedical text mining in order to quantify the diversity of isoform interactions generated by these cellular mechanisms. Many studies have benefited from the automated analysis of the biomedical scientific literature [24-26]. However, until now only little effort has been spent on the identification of alternative splicing events or the analysis of isoform diversity from the literature [27,28]. This is despite the fact that both alternatively spliced forms and other kind of isoforms (i.e. isoforms having allelic origins and isoforms produced by gene duplication) contribute to the complexity of proteomes which can lead to significant variation in protein interactions. For example, Resh et al. has computationally shown that alternative splicing modifies biological structure of the isoforms, mainly by removing protein interaction domains which leads to redirection of protein interaction networks at key points [29]. In a more recent study reporting on the largest human testis protein phosphatase 1 (PP1) interactome, it has been experimentally shown that there is high diversity among the regulatory protein sets binding to PP1 isoforms in different tissues (77 proteins in testis and 7 proteins in sperm) [30]. Hence, it is important to better analyze the functional variability of isoforms at a large scale.

In this study, we analyzed the variability amongst the interactions of protein isoforms. For this purpose, we used the content of Human Splicing Database version 3 (HumanSDB3), which provides comprehensive genomic and transcriptomic data for human alternatively spliced variants and other kinds of isoforms but does not yet include protein interaction data for the isoforms [18].

Utilizing a comprehensive text mining pipeline, we systematically analyzed 4,083,094 Medline abstracts belonging to the clustered transcripts provided from HumanSDB3. We constructed an interaction database, which includes 7,161 proteins and 31,819 interactions, called the Transcript Based Isoform Interaction Database (TBIID). We used TBIID to quantify the variability in isoform interactions by analyzing the subset of interactions belonging to clusters having more than one distinct protein isoform. We quantified differences in the number of interaction partners for a total number of 1,226 proteins and a total number of 1,540 interactions and compare the results against reference PPI databases. This analysis demonstrates that almost all clusters analyzed (99%) contain isoforms exhibiting variation in their interactions.

To the best of our knowledge, this is the very first study which analyzes the effect of isoform diversity on the human interactome. TBIID is a novel database which supports further investigation on functional differences of isoforms based on this interaction variability.

Materials and Methods

HumanSDB3 development

For the analyses described in this work, we utilize HumanSDB3, an alternative splicing database for the human transcriptome, previously developed by Taneri et al. [18,23] as summarized here. HumanSDB3 consists of clusters, each one containing overlapping transcripts based on their sequences, mapping to the same genomic region. Transcripts are either full-length mRNAs or Expressed Sequence Tags (ESTs). During the development process, the transcripts in a cluster of HumanSDB3 were grouped according to the sequence alignment methods described in [18,23]. Briefly, around 4.5 million input transcripts were collected from UniGene human clusters and aligned to the genome (UCSC hg17). Only best aligned transcripts that show more than 75% sequence similarity to the genome having at least two exons where each of the exons matched the genomic DNA with 95% identity or have less than 5 mismatches were kept. The final database contains a total number of 1,459,966 transcripts from 20,707 different clusters each of which has 70.5 transcripts on the average [18,23]. (HumanSDB3 is accessible at http://emmy.ucsd.edu/sdb.php?db=HumanSDB3.)

HumanSDB3 clusters

As previously reported by Taneri et al. [18,23], HumanSDB3 contains variant (81.31%) and invariant (18.69%) clusters. Variant clusters are composed of transcripts exhibiting alternative splicing events, while invariant clusters represent genes for which alternative splicing was not revealed with the available input transcript data, at the time of database construction. Therefore, invariant clusters were excluded from our study (Clusters in HumanSDB3 are labelled to include database version number, chromosome number and cluster number. An example cluster id is Hs.3.chr15p.6725) [18,23].

For the purpose of the analyses presented here, we focus on the transcripts that have been annotated in the Entrez Gene Database [31] including an official symbol and name, termed here as Defined Transcripts (DTs) [32]. Furthermore, only the variant clusters that contain several different DTs (amongst additional undefined transcripts) are relevant for this study and are termed here as Clusters with Multiple defined Transcripts (CMTs). Clusters containing exactly one DT, i.e. Clusters with a Single defined Transcript (CSTs) and clusters containing none, i.e. Clusters with Undefined Transcripts (CUT) are not relevant. HumanSDB3 clusters were built via transcript alignments to the genome based on their sequence similarities. Therefore, a possibility remains that DTs from a given CMTs could also denote other kinds of isoforms, such as isoforms produced from allelic or duplicated genes, in addition to alternative splicing variants, but they have mapped to the same genomic locus based on very high sequence similarity. On the other hand, CSTs have a single DT and therefore are homogeneous. Table 1 provides two clusters as

Cluster Type	Cluster ID	Transcript ID	Gene ID	Official Symbol	Official Name
CST	Hs.3.chr15p.6725	X04665	7057	THBS1	thrombospondin 1
CMT	Hs.3.chr14p.5840	NM_000624	5104	SERPINA5	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5
CMT	Hs.3.chr14p.5840	CR601472	12	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3

CST: Cluster with Single defined Transcript, CMT: Cluster with Multiple Defined Transcripts

Table 1: Sample CST and CMT clusters.

examples of a CST (cluster ID Hs.3.chr15p.6725) and a CMT (cluster ID Hs.3.chr14p.5840). The CST contains a single DT, which encodes THBS1 (Entrez Gene ID:7057) protein. The CMT contains DTs denoting two different serpin isoforms, namely SERPINA3 (Entrez Gene ID:12) and SERPINA5 (Entrez Gene ID:5104). Although, the DTs map to the same locus in HumanSDB3, our literature based analysis on the cluster (based on the GenBank transcript IDs) shows that the DTs denote isoforms encoded by two structurally similar serpin genes located on human chromosome 14q32 [33]. Previous studies have shown that these serpin family genes are clustered together in serpin gene cluster indicating that they evolved through gene duplication [33,34].

Text-mining pipeline

The pipeline for the literature analysis is shown in Figure 1. In the first step, all names for all DTs of the variant clusters in HumanSDB3 were produced and used to retrieve all Medline abstracts linked to them. Every transcript was submitted to the Entrez Gene Database to retrieve its official symbol, name and additional term variants. In addition, all term variants from the SwissProt database [35] were added as well as synonyms of the retrieved symbols were produced to generate a rich term set for a comprehensive Medline search [32]. The search was limited to the human species only by using the Medical Subject Heading (MeSH) restrictions of PubMed [36].

In the next step, we identified all abstracts containing mentions of PPI. The protein mentions were tagged with the Genia tagger [37] and all abstracts were retained that contained two or more mentions of different protein. A Support Vector Machine (SVM) classifier was implemented using SVM^{Light} [38] and was trained on the BioCreative-II Interaction Article Subtask (IAS) dataset [39] to distinguish those abstracts that are likely to contain PPIs from the remaining ones (IAS SVM classifier). The effectiveness of SVM as a text classification tool has been demonstrated in various text classification problems [40,41]. The features used are: i) TF.χ^2 term weights [42], ii) number of distinct protein mentions in the abstract, and iii) document classification scores that represent likelihoods according to naive Bayesian calculation for a document to report on PPI [43]. ii) and iii) can be considered as domain specific features for PPI document classification.TF.χ^2 term weight is one of a large set of well known and frequently used term weighting schemes used in text classification. Distinct number of protein mentions has shown to be a good domain specific feature for selecting interaction abstracts, given that the probability of a randomly selected document being an interaction abstract increases with the number of distinct protein mentions in the document [44]. Document classification scores and the term weighting schemes lead to complementary precision/recall behaviours and their combination has shown to increase significantly the performance in document classification [45]. Our IAS SVM classifier was trained on the BioCreative-II IAS training dataset and has an F_1-measure of 81.31% on the IAS test set which is in agreement with state-of-the-art performances. It achieves 3.31% higher than the best performing system of the regarding challenge [46] and 1.06% and 0.41% better than the other state-of-the-art systems reported in [47] and [48] respectively.

All selected Medline abstracts were processed for Protein-Protein Interaction Extraction (PPIE). First, the protein mentions were translated into Entrez Gene Database IDs using gene normalization tool, GNAT [49]. Sentences with at least two different protein IDs were again classified for containing evidence of a PPI pair using an SVM with a tree kernel [50] (PPIE SVM classifier). The features were: i) all words between the two protein names in combination with three words prior to the first protein name and three words after the second one represented in a Bag of Words (BoW) representation. These features were used given that words surrounding the candidate entities potentially carry information regarding their relationships, ii) the features representing the relation between the two proteins identified by two different syntactic parsers used in the biological text mining domain: Ksdep [51] and Enju [52]. Significant contribution of such parsers to the accuracy in PPI extraction task has been demonstrated in several studies [53-55]. PPIE SVM classifier was trained on the AIMed corpus [56] which is one of the main gold standard PPI corpora in the biomedical domain. 10-fold cross validation experiments on the training data revealed a performance of 54.20% using F_1-measure.

Manual assessment of the text mining pipeline

For the assessment of our text mining pipeline, we selected 100 sentences at random and manually analysed a total number of 212 extracted protein pairs. A total of 91 pairs were true positives and a total of 80 pairs were false positives, while the remaining 41 pairs were identified as false negatives. Overall, the performance of the system was estimated at an F_1-measure of 60.07% with 68.94% recall and 53.22% precision. The F_1-score obtained manually here is at the state-of-the-art level obtained in many PPI tasks [39]. Our manual inspection revealed that the errors were mainly due to faulty protein name normalization. Protein names were normalized to their Entrez Gene Database IDs by using GNAT which has a relatively low recall (73.80%) [49] due to missing protein names, achieving only partial recognition or assigning wrong protein IDs. For example, the sentence "*Tudor domain missense mutations, including one found in an SMA patient, impair the interaction between SMN and fibrillarin (as well as the common snRNP protein SmB)*" (PubMed ID:11509571) [57] states that "SMN" and "SmB" do interact, but is not recognized (false negative) since GNAT does not recognize or resolve the symbol "SmB".

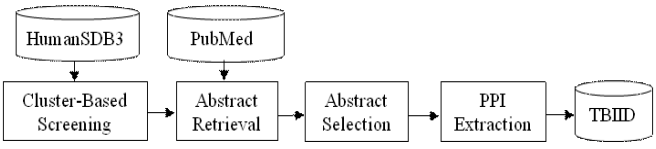


Figure 1: Text mining pipeline.

An example of a false positive PPI comes from the sentence "*CD26 mediates NH(2) terminus processing of CCL22, leading to the production of CCL22 (3-69) and CCL22 (5-69) that do not interact with CCR4*" (PubMed ID: 15067078) [58]. It contains a negation and a coordination and leads to the extraction of an interaction between "CCL22" and "CCR4".

Results and Discussions

PPI database for isoforms from the literature

HumanSDB3 contains 16,826 variant clusters (Table 2) and only a small portion represent CMTs (446 clusters, 2.65%). The majority of clusters are CSTs (12,192, 72.50%) and 3,568 (21.21%) clusters are CUTs. Furthermore, 620 (3.68%) clusters overlap with other clusters, since at least one DT from any of these clusters shares the description with a DT belonging to a different cluster. These clusters were discarded for the purposes of this study. A total of 13,174 DTs are contained in all CST and CMTs of HumanSDB3 (12,638 clusters in total) and all were used for abstract retrieval leading to a corpus of 4,083,094 abstracts (Table 3). In 2,465,692 abstracts, we found mentions of two different proteins. Of those abstracts 205,270 were classified as containing PPI information based on the IAS SVM classifier. From this subset of abstracts, we extracted 267,718 sentences containing two different protein names and finally, 33,158 distinct interaction pairs using the PPIE SVM classifier in comparison to over 1.2 million hypothetical interaction pairs from all pair-wise combinations in a sentence. Self-interacting proteins were excluded from our analysis since we focus on interactions between different protein isoforms.

We linked the extracted interaction pairs to DTs from HumanSDB3 clusters. For the majority of the interaction pairs (22,018, 66.40%) both protein partners were represented in HumanSDB3, whereas for 9,801 pairs (29.56%) one interaction partner was missing and for the remaining 1,339 (4.04%) interaction pairs none of the two interaction partners were contained in HumanSDB3. All interaction pairs with at least one interaction partner in HumanSDB3 (31,819 interaction pairs) have been imported into the new PPI database called TBIID.

Interaction variability in CMTs

We quantified the variability of interaction partners of isoforms linked to DTs in the CMTs to gain insight on whether different isoforms share interaction partners with other isoforms in the CMT (called *Shared Interactions*, Figure 2) or have unique interaction partners, i.e. the isoform is the sole isoform in the CMT to interact with the given partner (called *Unique Interactions*). We focused on those CMTs that contain references to multiple isoforms with known interaction partners and we compared our data from the literature analysis with the content from publicly available PPI databases. The variability in their interaction partners serve as an indicator for the functional variability of the isoforms, i.e. shared interactions indicate that the isoforms have kept their functional profile, whereas unique interactions indicate a higher level of functional diversity introduced to the interactome.

	Variant Clusters	CUT	Overlapping Clusters	CST	CMT	CST+CMT
Total	16,826	3,568	620	12,192	446	12,638
[%]	100	21.21	3.68	72.50	2.65	75.15

CUT: Cluster with Undefined Transcript, CST: Cluster with Single defined Transcript, CMT: Cluster with Multiple Defined Transcripts

Table 2: Overview of the distribution of HumanSDB3 clusters.

Phase		Total
Abstract Retrieval		4,083,094
Abstract Selection	Abstracts*	2,465,692
	Interaction abstracts	205,270
PPI Extraction	Sentences*	267,718
	Protein pairs generated	1,200,483
	Distinct interaction protein pairs	33,158

*Text containing at least two different protein mentions

Table 3: Literature analysis results for human alternatively spliced genes.

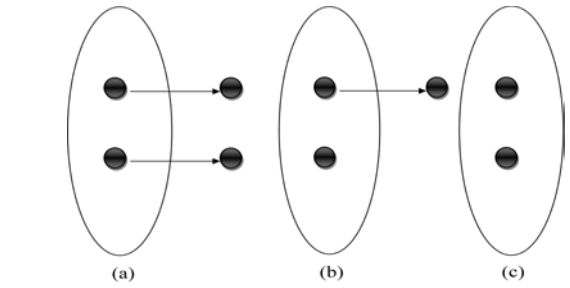


Figure 2: Illustration of CMTs containing distinct isoforms with possible interactions. (a) The CMT could contain isoforms with unique interaction partners only. (b) The CMT could contain isoforms with shared interaction partner only. (c) The CMT could contain isoforms having both; unique and shared interaction partners.

Nof Interacting Isoforms	Interaction Type	Nof CMTs
0	-	164
1	-	194
>1	Shared	1
>1	Unique	72
>1	Both	15

Nof:Number of

Table 4: Distribution of CMTs according to number and interaction types of isoforms based on literature analysis.

Table 4 gives an overview on our comprehensive Medline analysis. For 282 CMTs, at least one interacting isoform could be found while for 164 CMTs none was found. 194 out of 282 CMTs contain only one interacting isoform, while the remaining 88 CMTs contain multiple isoforms with known interactions. For the largest portion of CMTs (82%, 72 of the 88 CMTs), all the isoforms have unique interaction partners, whereas for 15 CMTs we found both shared and unique interactions of the isoforms, and only in 1 CMT, we found only shared interactions for all of its isoforms. As a significant finding of our study, we showed almost all CMTs (99%, 87 of 88 CMTs) exhibited unique interactions. Based on this finding we concluded that the isoforms of the CMTs have largely specialized towards having unique interaction partners to achieve functional diversity. Table 5 gives an overview of the distribution of shared and unique interactions across the 15 CMTs. The CMTs were categorized as having 2 or more than 2 isoforms and the average ratio of unique versus shared interactions in each category was found to be above 5.60.

It is noteworthy that CMTs with a single interacting isoform are

Iso/CMT	HumanSDB3 Cluster ID	Nof Iso	Nof S	Nof U	Nof S/Iso	Nof U/Iso	U/S	Avg U/S
2	Hs.3.chr6p.16643	2	22	38	11.00	19.00	1.73	
	Hs.3.chr17p.8013	2	20	50	10.00	25.00	2.50	
	Hs.3.chr11p.3558	2	12	24	6.00	12.00	2.00	
	Hs.3.chr6n.17144	2	10	46	5.00	23.00	4.60	
	Hs.3.chr1n.278	2	8	35	4.00	17.50	4.38	
	Hs.3.chr5n.15390	2	8	52	4.00	26.00	6.50	5.68
	Hs.3.chr14p.5840	2	4	25	2.00	12.50	6.25	
	Hs.3.chr12p.4823	2	2	40	1.00	20.00	20.00	
	Hs.3.chr17n.8529	2	2	6	1.00	3.00	3.00	
	Hs.3.chr19p.9432	2	2	19	1.00	9.50	9.50	
	Hs.3.chr22p.13094	2	2	4	1.00	2.00	2.00	
>2	Hs.3.chr6p.16595	3	14	38	4.67	12.67	2.71	
	Hs.3.chr3p.13906	3	2	11	0.67	3.67	5.50	5.62
	Hs.3.chr17n.8527	4	5	15	1.25	3.75	3.00	
	Hs.3.chr17n.8355	5	4	45	0.80	9.00	11.25	

Iso:Isoforms, Nof:Number of, Avg:Average, S:Shared interactions, U:Unique interactions

Table 5: Distribution of shared and unique interactions across CMTs based on literature analysis.

overrepresented possibly induced by the following reasons. First, some isoforms are more frequently represented in the literature since they have been studied more extensively in experiments. Second, some isoforms reported in the scientific literature could be missing from HumanSDB3 if mRNA or EST sequences were not available at the time of database generation or the available sequences did not meet the HumanSDB3 inclusion criteria. This would also depend on the transcript sequencing depth from given tissues, as some isoforms are known to be tissue-specific. Similarly, alternative splicing is known to be a developmental stage specific process, therefore presence of certain isoforms could depend on availability of sequencing from different developmental stages. In addition, the text mining pipeline employed could miss some interaction data.

Validation of the text mining results against PPI databases

In order to validate our results from the literature analysis, we compared the extracted PPIs for the isoforms linked to the selected CMTs to the content of the Protein Interaction Network Analysis Platform (PINA) [8]. PINA is a comprehensive and a state-of-the-art PPI dataset containing binary interactions from the six major PPI databases: DIP [3], MINT [4], IntAct [5], BioGRID [59], HPRD [60] and MIPS/MPact [61]. In contrast to many other resources, PINA suits to the purpose of this study taking into consideration that PINA excludes genetic interactions and complex formations.

PINA had to be pre-processed to remove all non-human interactions and all self-interactions of isoforms leading to 58,221 interactions between 11,856 different proteins. Then, Entrez Gene Database IDs of all proteins used in our study were mapped to their corresponding Uniprot accession number required for PINA using the Uniprot ID mapping system [62].

For all isoforms linked to the CMTs, the number of PPIs and their interaction type were identified in PINA (Table 6). For 345 CMTs, at least one interacting isoform could be identified within a PPI in PINA, whereas for 101 CMTs none could be found. A total of 158 CMTs have multiple interacting isoforms in contrast to 187 CMTs having a single interacting isoform only (Table 6). Amongst the 158 CMTs, we find 119 CMTs where the isoforms have only unique interactions whereas 9 have only shared interactions and 30 have both types of interactions (Table 7). Altogether, the majority of CMTs (94%, 149 of 158) do have multiple interacting isoforms exhibiting variability in their interactions. The average ratio of unique versus shared interactions for the clusters with two isoforms (8.82) was found to be slightly higher than the average values obtained for the clusters with more than two isoforms (7.53).

Altogether, the distribution of the interaction types in PINA was found to be in agreement with our results obtained through our comprehensive Medline analysis. Importantly, the average ratio of unique versus shared interaction values obtained in both categories by using PINA was slightly higher, indicating that a more fine-grained PPI dataset was obtained because of the interaction mapping process.

TBIID in comparison to PINA

We imported the results from our complete literature analysis into an interaction database called TBIID, which currently comprises 31,819 interactions for 7,161 unique proteins. A total of 5,615 of these proteins represent unique DTs belonging to either CSTs or CMTs and therefore can be linked to the corresponding gene/transcript sequence information in HumanSDB3. In particular, TBIID gives access to CMTs with multiple interacting isoforms exhibiting interaction variation, i.e. clusters with isoforms having either only unique or both unique and shared interactions. These clusters cover 1,540 interactions between 1,226 distinct proteins, where 994 of these proteins can be linked to the HumanSBD3.

When comparing TBIID against PINA, we found the following

Nof Interacting Isoforms	Interaction Type	Nof CMTs
0	-	101
1	-	187
>1	Shared	9
>1	Unique	119
>1	Both	30

Nof:Number of

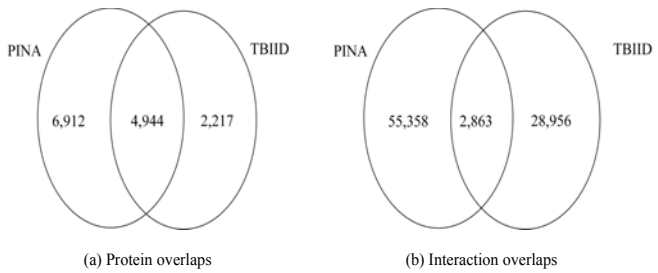
Table 6: Distribution of CMTs according to number and interaction type of isoforms based on the PINA PPI dataset.

Iso/CMT	Cluster ID	Nof Iso	Nof S	Nof U	Nof S/Iso	Nof U/Iso	U/S	Avg U/S
2	Hs.3.chr11p.3558	2	40	12	20	6	0.3	8.82
	Hs.3.chr17n.8529	2	22	54	11	27	2.45	
	Hs.3.chr5n.15390	2	18	71	9	35.5	3.94	
	Hs.3.chr6p.16643	2	10	13	5	6.5	1.3	
	Hs.3.chr12n.4463	2	8	6	4	3	0.75	
	Hs.3.chr19n.10450	2	6	25	3	12.5	4.17	
	Hs.3.chr6n.17144	2	6	26	3	13	4.33	
	Hs.3.chr17n.8585	2	4	4	2	2	1	
	Hs.3.chr1n.278	2	4	10	2	5	2.5	
	Hs.3.chr6n.17040	2	4	13	2	6.5	3.25	
	Hs.3.chr11n.3142	2	2	48	1	24	24	
	Hs.3.chr17p.8013	2	2	9	1	4.5	4.5	
	Hs.3.chr17p.8043	2	2	43	1	21.5	21.5	
	Hs.3.chr1n.361	2	2	134	1	67	67	
	Hs.3.chr2p.10772	2	2	4	1	2	2	
	Hs.3.chr4p.14617	2	2	11	1	5.5	5.5	
	Hs.3.chr4p.14694	2	2	3	1	1.5	1.5	
>2	Hs.3.chr16p.7233	3	27	2	9	0.67	0.07	7.53
	Hs.3.chr15p.6760	3	8	18	2.67	6	2.25	
	Hs.3.chr3p.13906	3	8	57	2.67	19	7.13	
	Hs.3.chr17n.8437	3	6	8	2	2.67	1.33	
	Hs.3.chr11n.3383	3	4	10	1.33	3.33	2.5	
	Hs.3.chr17n.8754	3	2	64	0.67	21.33	32	
	Hs.3.chr6p.16595	3	2	15	0.67	5	7.5	
	Hs.3.chr9n.19822	3	2	59	0.67	19.67	29.5	
	Hs.3.chr12n.4311	4	6	14	1.5	3.5	2.33	
	Hs.3.chr17n.8527	4	2	13	0.5	3.25	6.5	
	Hs.3.chr17n.8355	5	16	81	3.2	16.2	5.06	
	Hs.3.chr1p.1548	6	18	2	3	0.33	0.11	
	Hs.3.chr5p.15887	14	12	19	0.86	1.36	1.58	

Iso:Isoforms, Nof:Number of, Avg:Average, S:Shared interactions, U:Unique interactions

Table 7: Distribution of shared and unique interactions across CMTs based on the PINA PPI dataset.

results. On one hand, 4,944 (69.04%) proteins were shared between TBIID and PINA (this number was 927 (75.61%) proteins when only the clusters showing interaction variation were considered) (Figure 3). On the other hand, only 2,863 (9.00%) interactions in TBIID were also contained in PINA (this number was 141 (9.16%) when only the clusters exposing interaction variation were considered) (Figure 3). Altogether, TBIID provides access to a set of interactions that is rather complementary to PINA according to our analysis and since PINA integrates content from different primary data resources, we also conclude that TBIID is complementary to those primary data resources.



(a) Protein overlaps (b) Interaction overlaps

Figure 3: Venn diagrams showing overlaps between PINA and TBIID.

Dataset	Number of databases containing the interactions				
	1	2	3	4	5
PINA	33,725(57.92%)	16,086(27.63%)	5,182(8.90%)	3,102(5.33%)	126(0.22%)
TBIID*	1,198(41.84%)	1,210(42.26%)	338(11.81%)	97(3.39%)	20(0.70%)

*Overlapping interactions with PINA only

Table 8: Interaction pair distribution in PINA and overlapping datasets.

This result is not surprising, since the two PPI databases follow different standards and use different resources to identify relevant PPI information. The content analysis from PINA revealed that 57.92% of interaction entries were contained only in a single primary database, and only for 27.63% we found an interaction that was shared between any of the two databases constituting PINA (Table 8). In PINA, 8.90%, 5.33%, and 0.22% of interactions were shared amongst three, four, and five databases, respectively. This is also true for TBIID, i.e. 41.84% of the overlapping interactions in TBIID were reported in only one of the databases and 42.26% were reported in two databases constituting PINA. In TBIID, 11.81%, 3.39%, and 0.70% of the interactions were reported in three, four and five primary databases, respectively. It should be emphasized that, altogether 85.55% and 84.10% of interactions in PINA and TBIID respectively were reported in at most 2 primary databases.

These results show that the current PPI databases set a different focus in the selection of the PPIs. According to the previous studies, the heterogeneity of publicly available PPI databases is due to differences in the fact extraction methods, the curation methods and the utilized literature resources for the construction of the PPI database [6,7,63]. These reasons would explain the low rate of overlap (9.00%) between PINA and TBIID as well. We can reduce the emphasis on the selection of publication records, since 8,326 (42.98%) Medline abstracts in PINA are also contained in our retrieved set of abstracts (4,083,094 abstracts in total), and from this set 7,333 (37.85%) Medline abstracts are also included in our interaction abstract set (205,270 abstracts in total). Altogether, TBIID content was generated by automated text mining tools (recall: 68.94%, precision: 53.22%). In addition, TBIID relies only on freely available Medline abstracts, whereas many literature-curated databases use full text articles which would increase the rate of identified PPIs [38]. Another source of error is a 4% error rate in the assignment of Gene IDs from TBIID to their corresponding Uniprot accession numbers requiring that TBIID keeps the reference to the source text (Medline abstracts) to manually resolve questionable assignments. We conclude that the content in our interaction database has a specific focus but the distribution of entries is very similar to standard PPI databases.

TBIID Web-interface

A web interface was developed for the purposes of visualization of the TBIID content. Findings in TBIID are linked to the HumanSDB3 database, constructing a bridge between transcriptomic information of isoforms and their protein interactions. TBIID is publicly accessible at http://tbiid.emu.edu.tr.

A query system is embedded into the interface enabling users to search for the interactions of their protein isoforms of interest. Users can search for interactions either by using Entrez Gene Database IDs or official symbols of the protein isoforms. It is also possible to search interactions extracted from a given Medline abstract by submitting its PubMed ID to the query system.

We demonstrate the utility of TBIID and the usage of the web-interface by using CMT cluster Hs.3.chr1n.278 as an example. This particular cluster in HumanSDB3 contains two DTs for human IgG Fc Receptor III (FCGR3) coding two distinct 97% identical allelic isoforms, FCGR3A and FCGR3B [64].

Figure 4 illustrates a screenshoot from TBIID during the retrieval of the interactions involving low affinity immunoglobulin gamma Fc region receptor III-B from TBIID content by using its official symbol (FCGR3B). TBIID is unique in its interface since interactions of the

Transcipt Based Isoform Interaction Database

Home | Help | Contact |

Graph interactions of the whole cluster

Entrez gene ID	Entrez gene symbol	Transcript ID	Cluster type	HumanSDB3 cluster ID
2215	FCGR3B	X07934	CMT	Hs.3.chr1n.278
2214	FCGR3A	X52645	CMT	Hs.3.chr1n.278

FCGR3B interacts with 13 proteins:

Entrez gene Id	Entrez gene symbol	Transcript ID	Cluster type	HumanSDB3 cluster ID	PMIDs
920	CD4	S79267	CST	Hs.3.chr12p.5034	12918255, 8976730
7006	TEC	D29767	CST	Hs.3.chr4n.15163	15899983
22815	TDGF4		*		8757624
7124	TNF		*		14616797
325	APCS	BT006750	CST	Hs.3.chr1p.1318	11359830
925	CD8A	M12828	CST	Hs.3.chr2n.11681	8976730
6998	TDGF3		*		14971040, 8757624
6402	SELL		*		7507963
5657	PRTN3		*		16598772
3559	IL2RA	AF008556	CST	Hs.3.chr10n.2856	12918255
2214	FCGR3A	X52645	CMT	Hs.3.chr1n.278	7592758
5788	PTPRC	NM_002838	CST	Hs.3.chr1p.1185	9173906
2213	FCGR2B	AB050934	CST	Hs.3.chr1p.1293	15153797

Figure 4: Screen-shoot representing the retrieval of the FCGR3B interactions from TBIID content.

queried isoform are listed along with all other isoforms linked to the same HumanSDB3 cluster. Interactions of isoforms can also be visualized graphically. Hereby, we enable the end-user to simultaneously analyze the shared and unique interactions of all protein variants linked to the same cluster. Shared interactions of the queried isoform are highlighted with a different colour.

Proteins in TBIID are linked to the Entrez Gene Database providing access to additional information (e.g. functions - Gene Ontology (GO) terms, and metabolic pathways), which are crucial for good understanding of isoform interactions. Analyzing GO terms of FCGR3 isoforms by facilitating the web interface reveals that they share some molecular functions (Ig binding and receptor activity) and are involved in immune response processes. Given shared molecular functions, we could hypothesize that these isoforms have shared interactions.

The PINA database reports 12 interaction partners for FCGR3A (APCS, CD247, CD38, CD4, FCER1G, GP6, FCGR1A, IGHG1,LCK, PTPRC, SHC1, ZAP70) and only 4 partners for FCGR3B (APCS, IGHG1, M(2)21AB, Myb). Two of these partners, APRCS and IGHG1, are shared between the isoforms. However, utilizing TBIID, we could derive other potentially interesting interaction partners. For example, TBIID reports PTPRC (Entez Gene ID: 5788) as a shared interaction partner which is not reported by the primary PPI databases of PINA. This finding of TBIID is supported by experimental evidence reported from the literature (see PubMed IDs: 8157290 and 9173906). In addition, TBIID reports another unique interaction partner for FCGR3B isoform, TEC (see Entrez Gene ID:7006, PubMed ID: 15899983). As illustrated in the example discussed above, when compared to the other available analysis tools for the PPIs, TBIID provides differential interactions of isoforms. These functional features which are unique to TBIID are accessible through the database web-interface.

Conclusions

In this study, a new database, TBIID, which contains PPIs of human protein isoforms is presented. A comprehensive text mining pipeline is applied to the gene and transcript data contained in HumanSDB3 and a large scale analysis of PPIs is presented involving a significant portion of the proteome. State-of-the-art biomedical text mining tools are developed and utilized to automatically select abstracts that are likely to contain protein-protein interaction data and extract interaction annotations of protein isoforms from the interaction abstracts.

TBIID is screened for identifying and quantifying the variation in isoform interactions. The results based on our quantitative analysis reveal that an overwhelming majority of CMTs (99%) exhibit isoform interaction variability. Our findings have been validated against the literature-curated PPI data.

Up to now, neither a comprehensive PPI database for protein isoforms has been generated, nor has the variation in the isoform interactions been investigated on a large scale. TBIID brings both of these novel features to the PPI field. Undoubtedly, TBIID will help to initiate further studies on how alternative splicing and other transcript diversity mechanisms increase the complexity of proteomes and thus interactomes through potential differential interactions of protein isoforms. In this study, by investigating the data contained in TBIID, we for the first time provide quantitative evidence for the variability within the isoform interactions and thus functions. Presumably, the main source of this diversity is alternative splicing given that HumanSDB3 variant clusters contain mRNA and EST transcripts exhibiting alternative splicing events and thus are considered as splice variants. However, further detailed analysis on single CMTs is required to identify the exact transcript diversity mechanisms behind each isoform interaction. TBIID facilitates such further analysis on CMTs as well as representation of putative unique interactions of isoforms and thus within this context opens up the possibility for potential experimental exploration of different interactions of isoforms. Furthermore, the developed text mining tools used in the construction of TBIID are presented as efficient tools for abstract retrieval, protein interaction article selection and PPI extraction tasks on other platforms.

Our future research directions include extension of the study presented here to further investigate the functional variability of the protein isoforms. In order to assess the functional variation, we plan to analyze the distribution of functional annotations on the basis of Gene Ontology terms for all isoforms. Understanding the diversity in isoform functions and interactions is vital for successful drug discovery procedure, and not to mention drug docking. Interaction partners of isoforms exhibiting functional diversity are potentially good targets for pharmacological interventions [65]. Hence, the gathered data will be helpful in isoform-specific drug design. Isoform-specific drugs offer therapeutic advantages such as preventing disease progress over their non-specific types given different functions of isoforms. We also plan to gather disease-related information associated with CMTs. Such information would help to understand the mechanisms of transcript diversity, aberrant isoforms and their implications in abnormal protein functions as well as serving as an important information resource for molecular therapies.

Acknowledgements

Authors thank Terry Gaasterland of Scripps Genome Center, Scripps Institution of Oceanography, University of California San Diego, USA for access to the HumanSDB3 data, Christoph Grabmuller of Rebholz group, European Bioinformatics Institute for his useful suggestions on interaction variability validation as well as for his help on the web interface design, Vivian Lee of Rebholz group for her help on manual inspection of the data extracted from the literature and all other members of Rebholz Group for their critical feedback on this study.

This work is in part supported by the research grant MEKB-06-19 provided by the Ministry of Education and Culture of Northern Cyprus and in part by a European Commission grant to B.T. (grant no: 2010/249-026)

References

1. Cho S, Park SG, Lee DH, Park BC (2004) Protein-protein Interaction Networks: from Interactions to Networks. J Biochem Mol Biol 37:45-52.
2. Zhou D, He Y, Kwoh CK (2008) From biomedical literature to knowledge: mining protein-protein interactions. Computational intelligence in Biomedicine and Bioinformatics: Current Trends and Applications. Studies in Computational Intelligence (151), Springer, 397–421.
3. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, et al. (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28:289-291.
4. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett 513:135-140.
5. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, et al. (2004) IntAct: An open source molecular interaction database. Nucleic Acids Res 32:D452-D455.
6. Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, et al. (2006) An evaluation of human protein-protein interaction data in public domain. BCM Bioinformatics 7(Suppl. 5):S19.
7. Prieto C, Rivas J (2006) APID: Agile Protein Data Analyzer. Nucleic Acid Res 34:W298-W302.
8. Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP, et al. (2009) Integrated network analysis platform for protein-protein interactions. Nat Methods 6:75-77.

9. Uniprot Knowledge Base

10. Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457-463.

11. Modrek B, Lee C (2002) A genomic view of alternative splicing. Nat Genet 30:13-19.

12. Black DL (2000) Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. Cell 103:367–370.

13. Chen Z, Gore BB, Long H, Ma L, Tessier-Lavign M (2008) Alternative Splicing of the Robo3 Axon Guidance Receptor Governs the Midline Switch from Attraction to Repulsion. Neuron 58:325-332.

14. Lo TW, Branda CS, Huang P, Sasson IE, Goodman SJ, et al. (2008) Different isoforms of the C. elegans FGF receptor are required for attraction and repulsion of the migrating sex myoblasts. Dev Biol 318:268-275.

15. Koscielny G, Texier VL, Gopalakrishnan C, Kumanduri V, Riethoven JJ, et al. (2009). ASTD: The Alternative Splicing and Transcription Diversity database. Genomics 93:213-220.

16. Birzele F, Küffner R, Meier F, Oefinger F, Potthast C, et al. (2008) ProSAS: a database for analyzing in the context of protein structures. Nucleic Acids Res 36:D63-D68.

17. Kim P, Kim N, Lee Y, Kim B, Shin Y, et al. (2005) ECgene: genome annotation for alternative splicing. Nucleic Acids Res 33:D75-D79.

18. Taneri B, Snyder B, Novoradovsky A, Gaasterland T (2005) Databases for comparative analysis of human-mouse orthologous alternative splicing. Lec Notes in Computer Science 3388:123-131.

19. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.

20. Lee Y, Lee Y, Kim B, Shin Y, Nam S, et al. (2007) ECGene: an alternative splicing database update. Nucleic Acids Res 35:D99-D103.

21. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, et al. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res 20:45–58.

22. Sugnet CW, Kent WJ, Ares M Jr, Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput 66-77.

23. Taneri B (2005) Comparative Analysis of Alternative Splicing in Homo sapiens, Mus Musculus and Rattus norvegicus Transcriptomes. Ph.D Thesis, The Rockefeller University, USA.

24. Waagmeester A, Pezik P, Coort S, Tourniaire F, Evelo C et al (2009) Pathway enrichment based on text mining and its validation on carotenoid and vitamin A metabolism. OMICS 13:367-379.

25. Albert S, Gaudan S, Knigge H, Raetsch A, Delgado A, et al. (2003) Computer-assisted generation of a protein-interaction database for nuclear receptors. Mol Endocrinol 17:1555-1567.

26. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, et al. (2003) PreBIND and textomy-mining the biomedical literature for protein–protein interactions using a support vector machine. BMC Bioinformatics 4-11.

27. Shah PK, Jensen LJ, Boué S, Bork P (2005) Extraction of Transcript Diversity from Scientific Literature. PLoS Comput Biol 1:e10.

28. Cheng CY, Hsu FR, Tang CY (2008) Extracting Alternative Splicing Information from Captions and Abstracts Using Natural Language Processing. Proceedings of IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing SUTC2008 Taichung.

29. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, et al. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res 3:76-83.

30. Fardilha M, Esteves SL, Korrodi-Gregório L, Vintém AP, Domingues SC, et al. (2011) Identification of the human testis protein phosphatise 1 intreactome. Biochem Pharmacol 82:1403-1415.

31. Entrez Gene Database

32. Kafkas S, Varoglu E, Taneri B (2008) Methods for Abstract Retrieval from PubMed Database for Alternatively Spliced Genes. Proceedings of the Third International Symposium on Health Informatics and Bioinformatics HIBIT 2008.

33. Billingsley GD, Walter MA, Hammond GL, Cox DW (1993) Physical mapping of four serpin genes: alpha 1-antitrypsin, alpha 1-antichymotrypsin, corticosteroid-binding globulin, and protein C inhibitor within a 280-kb region on chromosome 14q32.1. Am J Hum Genet 52:343-353.

34. Pelissier P, Delourme D, Germot A, Blanchet X, Becila S, et al. (2008) An original SERPINA3 gene cluster: elucidation of genomic organization and gene expression in the Bos taurus 21q24 region. BMC Genomics 9:151.

35. Swissprot Database

36. NCBI, PubMed Database

37. Genia Tagger

38. SVM^{Light}

39. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A (2008) Overview of the protein-protein interaction extraction task of BioCreative II. Genome Biol 9:S4.

40. Leopold E, Kindermann J (2002) Text categorization with support vector machines: How to represent texts in input space. Mach Learn 46:423-444.

41. Burges CJB (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121-167.

42. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. Proceedings of the fourteenth International Conference on Machine Learning (ICML-97).

43. Marcotte EM, Xenarios I, Eisenberg D (2001) Mining literature for protein-protein interactions. Bioinformatics 17:359-363.

44. Abi-Haidar A, Kaur J, Maguitman A, Radivojac P, Retchsteiner A, et al. (2007) Uncovering Protein-Protein Interactions in the Bibliome. Proceedings of the Second BioCreative Workshop.

45. Kafkas S, Varoglu E, Taneri B (2009) Improving the Performance of Protein-Protein Interaction Article Selection Using Domain Specific Features. Proceedings of 2009 International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics BCBGC.

46. Lan M, Tan CL, Su J (2007) A Term Investigation and Majority Voting for Protein Interaction Article Sub-task 1 (IAS). Proceedings of the Second BioCreative Challenge Workshop.

47. Lan M, Tan CL, Su J (2009) Feature generation and representations for protein–protein interaction classification. J Biomed Inform 42:866-872.

48. Tsai RT, Hung HC, Dai HJ, Lin YW, Hsu WL (2008) Exploiting likely-positive and unlabeled data to improve the identification of protein-protein interaction articles. BMC Bioinformatics 9: S3.

49. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G (2008) Inter-species normalization of gene mentions with GNAT. Bioinformatics 24:i126-132.

50. Moschitti A (2006) Making tree kernels practical for natural language learning. Proceedings of the Eleventh International Conference on European Association for Computational Linguistics EACL.

51. Sagae K, Tsujii J (2007) Dependency parsing and domain adaptation with LR models and parser ensembles. Proceedings of the CoNLL 2007 Shared Task. Joint Conferences on Empirical Methods in Natural Language Processing and Computational Natural Language Learning EMNLP-CoNLL.

52. Miyao Y, Tsujii J (2008) Feature forest models for probabilistic HPSG parsing. Computational Linguistics 34:35-80.

53. Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, et al. (2008) All-path graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics 9: S2.

54. Miyao Y, Saetre R, Sagae K, Matsuzaki T, Tsujii J (2008) Task-oriented evaluation of syntactic parsers and their representation. Proceedings of the

Forty Sixth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies ACL:HLT.

55. Miwa M, Saetre R, Miyao Y, Tsujii J (2009) Protein-protein interaction extraction by leveraging multiple kernels and parsers. Int J Med Inform 78:e39-e46.

56. Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, et al. (2005) Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 33:139-155.

57. Jones KW, Gorzynski K, Hales CM, Fischer U, Badbanchi F, et al. (2001) Direct interaction of the spinal muscular atrophy disease protein SMN with the small nucleolar RNA-associated protein fibrillarin. J Biol Chem 276:38645-38651.

58. Bonecchi R, Locati M, Galliera E, Vulcano M, Sironi M, et al. (2004) Differential recognition and scavenging of native and truncated macrophage-derived chemokine (macrophage-derived chemokine/CC chemokine ligand 22) by the D6 decoy receptor. J Immunol 172:4972-4976.

59. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34:D535-D539.

60. Keshava PTS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database - 2009 Update. Nucleic Acids Res 37:D767-772.

61. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, et al. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21:832-834.

62. Uniprot ID Mapping System

63. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, et al. (2009) Literature-curated protein interaction datasets. Nat Methods 6:39-46.

64. Rogers KA, Scinicariello F, Attanasio R (2006) IgG Fc Receptor III Homologues in Nonhuman Primate Species: Genetic Characterization and Ligand Interactions. J Immunol 177:3848-3856.

65. da Cruz e Silva OA, Fardilha M, Henriques AG, Rebelo S, Vieira S, et al. (2004). Signal transduction therapeutics: relevance for Alzheimer's disease. J Mol Neurosci 23:123-142.