"main" — 2018/10/26 — 15:04 — page 1 — #1

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Data and text mining

PPaxe: easy extraction of protein occurrence and interactions from the scientific literature

S. Castillo-Lara [1], J.F. Abril [1,*]

[1]Computational Genomics Lab; Genetics, Microbiology & Statistics Dept.; Universitat de Barcelona; Institut de Biomedicina (IBUB) Barcelona, Catalonia, Spain.

[*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein-protein interactions (PPIs) are very important to build models for understanding many biological processes. Although several databases hold many of these interactions, exploring them, selecting those relevant for a given subject, and contextualizing them can be a difficult task for researchers. Extracting PPIs directly from the scientific literature can be very helpful for providing such context, as the sentences describing these interactions may give insights to researchers in helpful ways.
Results: We have developed `PPaxe`, a python module and a web application that allows users to extract PPIs and protein occurrence from a given set of PubMed and PubMedCentral articles; It presents the results of the analysis in different ways to help researchers export, filter and analyze the results easily.
Availability: `PPaxe` web demo is freely available at https://compgen.bio.ub.edu/PPaxe. All the software can be downloaded from https://compgen.bio.ub.edu/PPaxe/download, including a command-line version and docker containers for an easy installation.
Contact: jabril@ub.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-protein interactions (PPIs) play a major role in many biological processes, such as cancer, regeneration, and the development of many diseases (Ian *et al.*, 2012); which makes the identification of these interactions a key point for the understanding of these processes. Databases such as iHOP (Hoffmann *et al.*, 2005) have helped researchers to navigate the scientific literature using genes and proteins as drivers. Further tools exist in order to automatically retrieve interactions described in the scientific literature (Raja *et al.*, 2013; Quan *et al.*, 2014; Zhao *et al.*, 2016); however, the accessibility to these tools for researchers is rather difficult, either because they can't be easily downloaded or because they lack a user-interface. Here we present `PPaxe`, a python module and a web application, superseding already developed code used for a previous analysis of the protein network of retinitis pigmentosa; that code was referred to as *Sparser* in the methods section of Boloc *et al.* 2015.

Implementation

`PPaxe` reads a text file with a list of PubMed identifiers and downloads all the necessary articles. `PPaxe` downloads either abstracts from Medline or full-text articles from PubMed Central, depending on the option provided by the user. On the web application, the user can also provide a well-formatted PubMed query, so that the tool can retrieve the list of publication identifiers directly from NCBI PubMed; or a plain-text file with the text to be analyzed.

`PPaxe` uses Stanford CoreNLP (Manning *et al.*, 2014) for name entity recognition (NER) of proteins and genes. Three datasets were used in order to train the The Stanford Named Entity Recognizer: AImed (Bunescu *et al.*, 2005), MedTag (Smith *et al.*, 2005) and BioInfer (Pyysalo *et al.*, 2007). First, each sentence was tokenized by Stanford CoreNLP; then, a Conditional Random Field (CRF) classifier was trained. Performance of the NER tagger was assessed by 2-fold cross-validation. Once the NER is trained, `PPaxe` extracts all the co-occurring proteins in each sentence and for each pair of them, computes several features, as described on Suppl. Mat. Table S1, which will be used in order to identify pairs as interacting

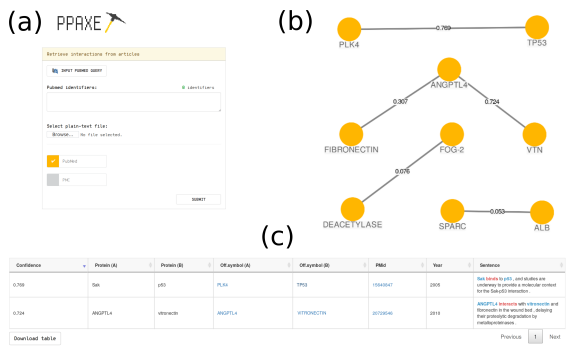

Fig. 1. PPaxe web interface and output examples. **a)** Web application input form; **b)** Graph visualization from the HTML report. **c)** Sentences containing PPIs retrieved from the analysis of four PubMed abstracts (PMIDs: 15640847, 20729546, 25196150, and 25211495), presented in the HTML output as a searchable table. The Confidence value (shown in the table and on the graph edges) corresponds to the normalized percentage of votes of the predictor (ranging from 0 to 1). Enlarged version available as Suppl. Mat. Figure S1.

or not. The prediction is based on a Random Forest Classifier from scikit-learn (Pedregosa *et al.*, 2011), trained over the annotated sentences of AImed, LLL-challenge (Nédellec, 2005) and BioInfer. A votes cut-off of 0.55 was selected to detect interacting proteins from sentences, based on the performance of the classifier on our evaluation (estimated using 10-fold cross-validation).

Finally, `PPaxe` produces several possible outputs: an HTML page with all the PPIs, the sentences in which they were found, a table with all the proteins found in the specified articles, a graph visualization made with the JavaScript library `cytoscape.js` (Franz *et al.*, 2015), and a PDF with a summary of the analyses. `PPaxe` is also distributed as two docker images, a command-line application and a web application, respectively; which can be retrieved from the downloads page. The web application docker image runs a `localhost` server on a customizable port. `PPaxe` python modules are available too.

Results and Discussion

Protein and gene name tagging by the Stanford CoreNLP was evaluated by 2-fold cross-validation on the datasets of AImed, MedTag and Bioinfer; which resulted in a precision of 74.5%, a recall of 70.0%, and an F1 of 72.1%. As `PPaxe` does not use any dictionary table in order to identify the proteins, it is able to tag protein symbols of any species, or even newly described proteins and genes. However, it cross-checks the identifiers against the aliases provided by the HUGO Gene Nomenclature Committee as a prior normalization step (Yates *et al.*, 2017).

`PPaxe` is able to retrieve protein-protein and genetic interactions without needing to define specific patterns or rules, which makes the application's use broader and more general than other previously mentioned approaches, such as PPInterFinder (Raja *et al.*, 2013). `PPaxe` considers only a narrow selection of features, namely POS composition, token distance, and keywords (see Suppl. Mat. Table S3), freeing `PPaxe` of syntactical parsing and the posterior processing of dependency trees. On an Intel Core i7 machine, `PPaxe` took ~2″ to download and analyze 10 articles, 16″ for 100, 2′36″ for 1,000, and 28′54″ for 10,000. An assessment of the interaction extraction performance and a validation over BioGRID (Stark *et al.*, 2006) are described in Suppl. Mat. evaluation section and is shown on Table S1 and Table S2. `PPaxe` facilitates the visualization of the results, without requiring any other additional software, thanks to the inclusion of an HTML report output. The output includes several summary tables, a dynamic visualization of the retrieved interactions, along with some statistics for the retrieved references (Fig.1).

In conclusion, yet some similar tools have already been developed, `PPaxe` allows researchers both to perform large-scale text-mining, but also to analyze small and focused sets. The newly implemented range of output options and reports should make `PPaxe` a valuable tool for researchers to retrieve and curate novel PPIs described in the scientific literature.

Acknowledgements

The authors are grateful to our beta tester users that played with the initial versions of the tool, especially to Rodrigo Arenas-Galnares, and to the referees for their useful comments.

Funding

This work was supported by research grants from Spanish Ministry of Economy (BFU2014-56055P), and from Generalitat de Catalunya (2014SGR687, 2017SGR1455). SC-L is fellow of Catalan Government "AGAUR" (FI-FDR: 2017FI_B_00191).

References

Boloc, D. *et al.* (2015). Distilling a Visual Network of Retinitis Pigmentosa Gene-Protein Interactions to Uncover New Disease Candidates. *PLoS ONE*, **10**(8), e0135307.

Bunescu, R. *et al.* (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.*, **33**(2), 139–155.

Franz, M. *et al.* (2015). Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, **32**(2), 309–311.

Hoffmann, R. *et al.* (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**, 252–258.

Ian, W. *et al.* (2012). Protein interaction networks in medicine and disease. *Proteomics*, **12**(10), 1706–1716.

Manning, C. *et al.* (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nédellec, C. (2005). Learning language in logic - genic interaction extraction challenge. In *Proc. of the Learning Language in Logic 2005 Workshop at the Int.Conf. on Machine Learning*.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *J.Mach.Learn.Res.*, **12**, 2825–2830.

Pyysalo, S. *et al.* (2007). Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, **8**(1), 50.

Quan, C. *et al.* (2014). An unsupervised text mining method for relation extraction from biomedical literature. *PLoS ONE*, **9**, e102039.

Raja, K. *et al.* (2013). PPInterFinder - A mining tool for extracting causal relations on human proteins from literature. *Database*, **2013**, bas052 (1–11).

Smith, L. H. *et al.* (2005). MedTag: A Collection of Biomedical Annotations. In *Proc. of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, pages 32–37.

Stark, C. *et al.* (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Yates, B. *et al.* (2017). Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**(D1), D619–D625.

Zhao, Z. *et al.* (2016). A protein-protein interaction extraction approach based on deep neural network. *Int. J. Data Mining and Bioinformatics*, **15**(2), 145–164.

Supplementary materials

PPaxe: easy extraction of protein occurrence and interactions from the scientific literature

S. Castillo-Lara and J.F. Abril

1 PPaxe features

`PPaxe` uses three types of features: token distance measures between the entities in the sentence (e.g. from each candidate protein to each verb between them), POS tag composition (both between the candidate proteins and in their vicinity), and finally keyword occurrence in the sentences. This approach makes `PPaxe` capable of retrieving interactions that don't follow any particular structure or pattern, such as "*Protein-A binds Protein-B*", "*Protein-A and Protein-B interact.*", and so on. All features are listed on Supplementary Materials Table S3.

2 PPaxe evaluation

Performance of the `PPaxe` estimated over the three datasets using 10-fold cross-validation, compared to previously published tools, is shown on Supplementary Materials Table S1. *Quan et. al* I and II correspond to the unsupervised and the semi-supervised methods described in that article respectively. The performance of `PPaxe` when it comes to retrieving interactions varies from dataset to dataset. Overall, `PPaxe` performed better than a previously described method on the same datasets (Zhao *et al.*, 2016), slightly worse than a rule-based approach (Raja *et al.*, 2013), and better and similarly to an unsupervised and a semi-supervised method respectively (Quan *et al.*, 2014).

2.1 Performance metrics

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

$$Precision = \frac{tp}{tp+fp}$$

$$Recall = \frac{tp}{tp+fn}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$$

where:

tp is true positives.

tn is true negatives.

fp is false positives.

fn is false negatives.

2.2 BioGRID test case

PPaxe interaction extraction was compared against the interactions annotated in the BioGRID database. In order to do so, the following query was performed on PubMed:

```
protein protein interactions
    AND (hasabstract[text]
    AND "last 5 years"[PDat]
    AND Humans[Mesh])
```

That PubMed query returned 41,286 entries on October 17^{th}, 2018. We picked up the first 2,000 PubMed identifiers for posterior analyses, after sorting them by decreasing publication date, to ensure retrieving the most recent papers that may contain novel interactions.

PPaxe retrieved 104 interactions (92 unique). From those interactions, 31 contained interactions already annotated in BioGRID, defined as "BioGRID positive" (valid sentence defining interactions already supported by BioGRID) and "BioGRID negative" (wrong sentences defining an interaction contained in BioGRID). We manually annotated the remaining 73 interactions as "PPaxe positive" (those sentences with a correct protein-protein or genetic interaction, according to the sentence alone) or "PPaxe negative" (interactions not supported by the sentence). The results of this annotation can be seen in Table S2.

Even though the comparison against BioGRID showed that most interactions retrieved by PPaxe were not annotated in the database, the manual annotation revealed that most of them were in fact described in the sentence: 40 out of the 73 sentences were annotated as "PPaxe positive". By considering both "BioGRID positive" and "PPaxe positive" together, 65.38% of the interactions retrieved by PPaxe can be considered correctly retrieved, which is slightly smaller than the estimated precision of interaction retrieval computed by 10-fold cross validation shown in Table S1.

We can't consider this test as a full assessment of PPaxe performance due to several limitations of this approach. For instance, we don't have a gold standard neither for sentences, nor interactions; we can't consider BioGRID as a full validated gold standard, and the annotation of the sentences was performed after their selection. Moreover, the impact of the protein/gene tagger can't be considered separately from the interaction extraction procedure as it has been done in the cross-validation tests. In summary, that implies we can't estimate some of the performance metrics such as the Recall to compare those results with the cross-validation assessment. Finally, the large number of correctly retrieved interactions by PPaxe that were not annotated in BioGRID could be explained by either differences in gene symbol normalization (although both sets of protein symbols were normalized using the HGNC alias dictionary), by the use of non-official gene names in the articles, or by the fact that the interactions are yet to be included in BioGRID.

References

Quan, C. *et al.* (2014). An unsupervised text mining method for relation extraction from biomedical literature. *PLoS ONE*, **9**, e102039.

Raja, K. *et al.* (2013). PPInterFinder - A mining tool for extracting causal relations on human proteins from literature. *Database*, **2013**, bas052 (1–11).

Zhao, Z. *et al.* (2016). A protein - protein interaction extraction approach based on deep neural network. *Int. J. Data Mining and Bioinformatics*, **15**(2), 145–164.

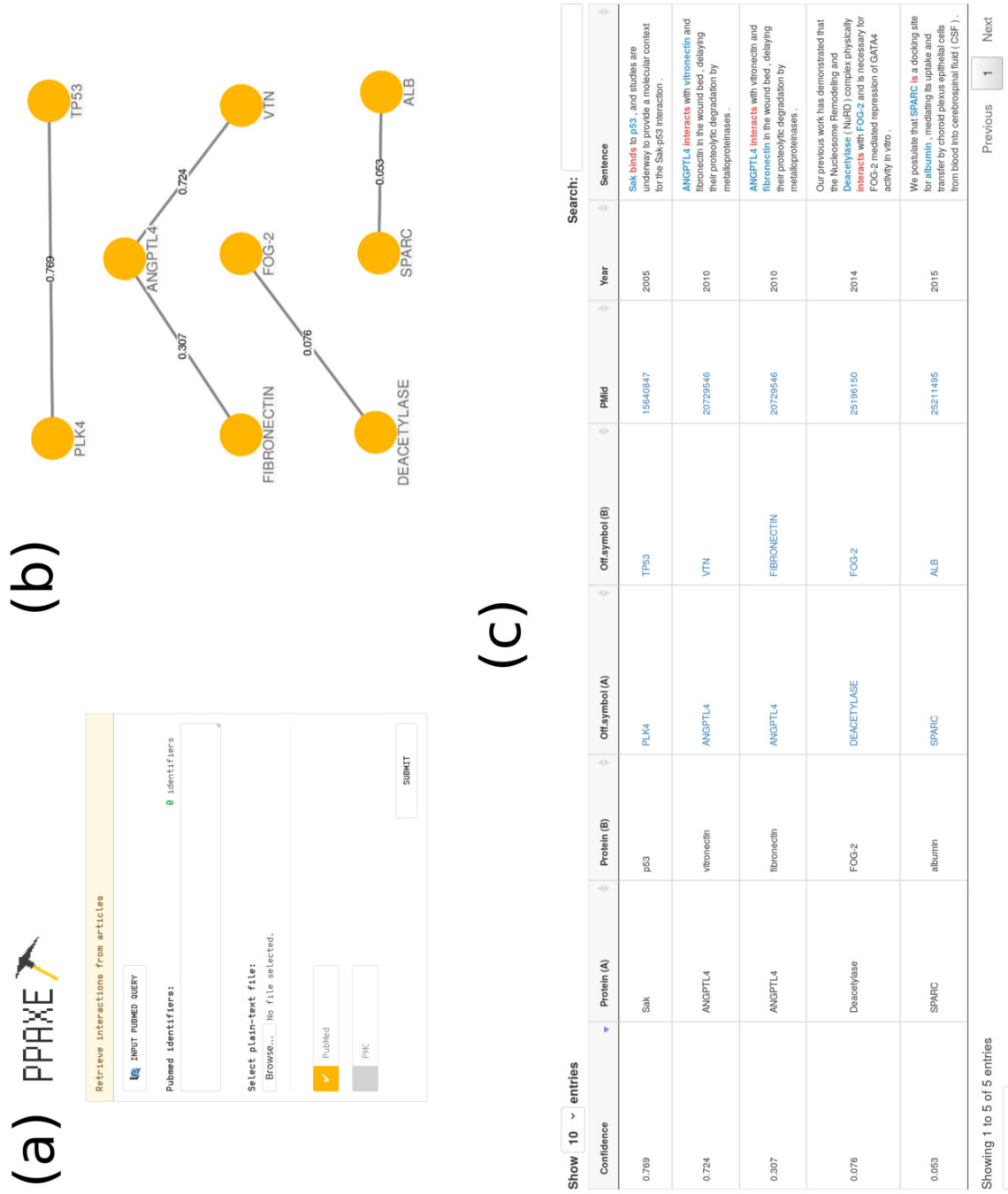

Figure S1: **A larger view of manuscript Figure 1**. **a)** `PPaxe` web application form, which allows users to both input PubMed identifiers directly or to write a PubMed query to retrieve the requested articles. **b)** Graph visualization made using Cytoscape.js from the HTML report of `PPaxe`. **c)** Sentences containing PPIs retrieved from the analysis of four PubMed abstracts (PMIDs: 15640847, 20729546, 25196150, and 25211495), presented in the HTML output as a searchable table. The Confidence value (shown in the table and on the graph edges) corresponds to the normalized percentage of votes of the predictor (ranging from 0 to 1).

Table S1: **Assessment of PPaxe interaction extraction performance**. PPaxe performance metrics were computed with a 10-fold cross-validation for each dataset. The same procedure was not applied to the other tools, due to the inability to get the corresponding software to run under the same conditions. Therefore, the metrics displayed on this table were retrieved from their respective articles. "All sets" correspond to the validation run on a merged dataset built from the other three (AImed + BioInfer + LLL).

	Tool	Accuracy	Precision	Recall	F1
AImed	**PPaxe**	**80.07**	**70.54**	**50.93**	**59.15**
	Zhao *et. al*	—	50.51	63.38	56.12
	Quan *et. al* I	—	44.80	71.40	55.10
	Quan *et. al* II	—	56.6	66.80	60.7
	Raja *et. al*	—	80.25	56.12	66.05
BioInfer	**PPaxe**	**88.42**	**81.27**	**65.55**	**72.57**
	Zhao *et. al*	—	53.89	72.9	61.63
LLL	**PPaxe**	**72.96**	**76.98**	**87.39**	**81.86**
	Zhao *et. al*	—	75.84	91.84	82.00
All sets	**PPaxe**	**84.48**	**76.50**	**58.68**	**66.41**

Table S2: **PPaxe extracted interactions when intersecting them with those in BioGRID.** The retrieved interactions were compared against the BioGRID database, and those not annotated in the database were manually curated. In blue, 28 interactions described in BioGRID ("BioGRID positive"); in yellow, 3 interactions described in BioGRID but annotated as incorrect after manual inspection ("BioGRID negative"); in green, 40 interactions not described in BioGRID but annotated as correct after manual curation ("PPaxe positive"); in red, 33 interactions not described in BioGRID annotated as incorrect ("PPaxe negative"). The "Conf." value corresponds to the normalized percentage of votes of the random forest classifier used by `PPaxe`.

PMID	INT A	INT B	Conf.	Sentence after tokenization
29899090	TRIM41 (TRIM41)	NP (ZNF384)	0.796	Here , we report that TRIM41 interacts with NP through its SPRY domain
29444082	S100B (S100B)	S100A1 (S100A1)	0.676	We found that S100B could interact with S100A1 via NMR 1H-15N HSQC titrations
29178343	GRP78 (HSPA5)	PRDM14 (PRDM14)	0.644	These results suggest that HSP90 and GRP78 interact with PRDM14 and participate in cancer regulation
29220652	PARP1 (TIPARP)	ALC1 (MYL4)	0.627	Its engagement with PARylated PARP1 activates ALC1 at sites of DNA damage , but the underlying mechanism remains unclear
29358401	DNAAF2 (DNAAF2)	SPAG1 (SPAG1)	0.564	FRET analysis of HEAT domain deletions and human mutations showed that HEATR2 interacted with itself and SPAG1 at multiple HEAT domains , while DNAAF2 interacted with SPAG1
30111544	REV7 (MAD2L2)	REV3 (REV3L)	0.556	Rev7 interacts with Rev3 by a mechanism conserved among HORMA proteins , whereby an open-to-closed transition locks the ligand underneath the " safety belt " loop
29328377	MYC (MYC)	CDK2 (CDK2)	0.518	The cyclindependent kinase inhibitor 1A (CDKN1A) , E2F transcription factor 1 (E2F1) , and MYC interacted with CDK2
29374759	VDR (VDR)	P53 (TP53)	0.480	VDR binding to p53 was confirmed by western blot analysis
30021902	CDC25A (CDC25A)	TBK1 (TBK1)	0.458	Further analysis indicated that Cdc25A can interact with TBK1 and reduce the phosphorylation of TBK1 at S172 , which in turn decreases the phosphorylation of its downstream substrate IRF3
29281729	IFITM3 (IFITM3)	LSD1 (KDM1A)	0.458	Our data suggest that the demethylation of IFITM3 by LSD1 is beneficial for the host to fight against RNA virus infection
29684085	KAT5 (KAT5)	CD4 (CD4)	0.449	The pro-latency effect of KAT5 is confirmed in a primary CD4 + T cell latency model as well as cells from ART-treated patients
29295922	PAK4 (PAK4)	CDC42 (CDC42)	0.436	Using solution scattering we find that the full-length PAK4 heterodimer with CDC42 adopts primarily a compact organization
29154191	CD11B (ITGAM)	RHO (RHOD)	0.413	Mechanistically , -synuclein bound to CD11b and subsequently activated Rho signaling pathway
29295922	CDC42 (CDC42)	PAK4 (PAK4)	0.387	These additional interactions modulate kinase activity and increase the binding affinity of CDC42 for full-length PAK4 compared with the CRIB domain alone
30021902	CDC25A (CDC25A)	IRF3 (IRF3)	0.382	Consistently , knockdown of Cdc25A upregulates the phosphorylation of both TBK1-S172 and IRF3 in Sendai virus-infected or TBK1-transfected 293T cells
29925658	STAT1 (STAT1)	NSP2 (RTN2)	0.378	Chemically blocking CRM1-mediated nuclear export in the presence of nsP2 additionally showed that nuclear translocation of STAT1 is not affected by nsP2
29218693	NCK2 (NCK2)	ITGB1 (ITGB1)	0.373	Co-IP showed that NCK2 can directly bind ITGB1 , but not VEGFA
29334217	P53 (TP53)	HSP90 (HSP90AA1)	0.360	The DNA-binding domain (DBD) of p53 is known to interact with the chaperone Hsp90 , but the role of other members of the chaperone network , including co-chaperones such as p23 , is unknown
29295922	CDC42 (CDC42)	PAK4 (PAK4)	0.351	We therefore show that the interaction of CDC42 with PAK4 can influence kinase activity in a previously unappreciated manner
29358401	HEATR2 (DNAAF5)	SPAG1 (SPAG1)	0.338	FRET analysis of HEAT domain deletions and human mutations showed that HEATR2 interacted with itself and SPAG1 at multiple HEAT domains , while DNAAF2 interacted with SPAG1

Continued on next page

PMID	INT A	INT B	Conf.	Sentence after tokenization
29232376	GCN5 (KAT2B)	CHE1 (BCHE)	0.329	In this study , we aimed to identify whether besides ADA3 , other components of the HAT modules of SAGA and ATAC complexes , human ADA2 and GCN5 also interact with Che-1 / AATF
29379028	NS1 (PTPN11)	TBK1 (TBK1)	0.324	This mutation enables NS1 binding to TBK1 and reduces TBK1 phosphorylation
29184850	CELLUGYRIN (SYNGR2)	CDTB (CDB2)	0.324	Furthermore , we demonstrate that cellugyrin is an intracellular binding partner for CdtB as demonstrated by immunoprecipitation
29512721	OSTEOCALCIN (BGLAP)	FOXO1 (FOXO1)	0.311	The HIF1induced expression of Runx2 and ALP may be completely dependent on the expression levels of Foxo1 , and in turn , osteocalcin may be partially dependent on Foxo1 expression
29950413	SOX2 (SOX2)	STAT3 (STAT3)	0.307	IE1 mediates SOX2 depletion by targeting STAT3 , a critical upstream regulator of SOX2 expression
29281729	LSD1 (KDM1A)	IFITM3 (IFITM3)	0.302	We have found that LSD1 is recruited to demethylate IFITM3 at position K88 under IFN treatment
29212519	VEGF (VEGFA)	IGF1 (IGF1)	0.298	We found that a functional cooperation between HIF-1 and GPER is essential for the transcriptional activation of VEGF induced by IGF1
29178989	NETS (SPINK5)	FSAP (HABP2)	0.298	Taken together , NETs bind to FSAP , but do not activate pro-FSAP unless histones are released from NETs by DNAse
29743362	FUBP1 (FUBP1)	P53 (TP53)	0.284	Here we report that human adenovirus 5 coopts the cellular protein FUBP1 to prevent the activation of the p53 stress response pathway that would block viral replication
29152905	L13A (RPL13A)	EIF4G (EIF4G1)	0.280	EPRS binds the GAIT element in target mRNAs , NSAP1 negatively regulates mRNA binding , L13a binds eIF4G to block ribosome recruitment , and GAPDH shields L13a from proteasomal degradation
29899107	MDA5 (IFIH1)	IFN (IFNA1)	0.271	Although RIG-I has been recognized as the leading cytoplasmic sensor against HCV for a long time , recent findings that MDA5 regulates the IFN response to HCV have emerged
29665350	PXR (NR1I2)	CYP3A4 (CYP3A4)	0.262	In conclusion , PXR activation and PXR-mediated induction of CYP3A4 expression by PAs seem to be structure-dependent
29669839	GP (RNF130)	TETHERIN (BST2)	0.258	To our knowledge , these findings demonstrate for the first time that GP can antagonize tetherin in infected cells and provide a tool to study the impact of GP-dependent tetherin counteraction on EBOV spread
29232376	GCN5 (KAT2B)	AATF (AATF)	0.258	In this study , we aimed to identify whether besides ADA3 , other components of the HAT modules of SAGA and ATAC complexes , human ADA2 and GCN5 also interact with Che-1 / AATF
29660231	S100A4 (S100A4)	P53 (TP53)	0.253	Wnt / -catenin targets , c-MYC and S100A4 were upregulated in p53 cells and were downregulated when plakoglobin was coexpressed
29690653	VEGFA (VEGFA)	VEGFR2 (KDR)	0.218	Although VEGF-A ligands bind to both VEGFR1 and VEGFR2 , they primarily signal via VEGFR2 leading to endothelial cell proliferation , survival , migration and vascular permeability
29925821	RHOA (RHOA)	RAC1 (RNASE1)	0.213	All genetic evidences indicate that in these disorders the RhoA pathway is hyperactive while the Rac1 and cdc42 pathways are consistently hypoactive
29690653	VEGFA (VEGFA)	VEGFR1 (FLT1)	0.213	Although VEGF-A ligands bind to both VEGFR1 and VEGFR2 , they primarily signal via VEGFR2 leading to endothelial cell proliferation , survival , migration and vascular permeability
29212519	GPER (GPER1)	VEGF (VEGFA)	0.204	We found that a functional cooperation between HIF-1 and GPER is essential for the transcriptional activation of VEGF induced by IGF1
29444113	KCNQ1 (KCNQ1)	KCNE1 (KCNE1)	0.200	The results reveal that interactions between KCNQ1 with KCNE1 causes a pore constriction in the former , which in-turn forms small energetic barriers in the ion-permeation pathway
30111544	REV7 (MAD2L2)	REV1 (REV1)	0.196	We demonstrate that Rev7 uses the conventional HORMA dimerization interface both to form a homodimer when tethered by the two RBMs in Rev3 and to heterodimerize with other HORMA domains , Mad2 and p31 Structurally , the Rev7 dimer can bind only one copy of Rev1 , revealing an unexpected Rev1/Pol architecture
29677136	NOP53 (NOP53)	RIGI (DDX58)	0.182	Cytoplasmic NOP53 interacts with the retinoic acid-inducible gene I (RIG-I) to remove its K63-linked ubiquitination , leading to attenuation of type I interferon IFN-

Continued on next page

PMID	INT A	INT B	Conf.	Sentence after tokenization
29269420	L1 (IGKV116)	CENPA (CENPA)	0.182	We show that human CENP-N confers binding specificity through interactions with the L1 loop of CENP-A , stabilized by electrostatic interactions with the nucleosomal DNA
29690653	VEGFA (VEGFA)	VEGFR2 (KDR)	0.178	This review explores the molecular pharmacology of VEGF-A isoforms at VEGFR2 in respect to ligand binding and downstream signalling
29167311	AND1 (WDHD1)	CTF4 (WDHD1)	0.178	AND-1 has maintained the trimeric structure of yeast Ctf4 , driven by its conserved SepB domain
29925658	NSP2 (RTN2)	IFN (IFNA1)	0.173	The research described here specifies where in the JAK/STAT signaling cascade the IFN response is inhibited and which protein domain of nsP2 is responsible for IFN inhibition
29207260	E2F1 (E2F1)	TEAD1 (TEAD1)	0.173	Further , we found that human E2F1 competes with YAP for TEAD1 binding , affecting YAP activity , indicating that this mode of cross-regulation is conserved
29444113	KCNE1 (KCNE1)	KCNQ1 (KCNQ1)	0.169	These findings correlate with the previous experimental reports that interactions of KCNE1 dramatically slows the activation of KCNQ1
29297316	CCR5 (CCR5)	CD4 (CD4)	0.169	It was found that the subnetworks formed by CCR5 and IFNAR1 and their neighbors were a fragments of two key pathways functioning during the course of tick-borne encephalitis : (1) the attenuation of interferon-I signaling pathway by the TBEV NS5 protein that targeted peptidase D ; (2) proinflammation and tissue damage pathway triggered by chemokine receptor CCR5 interacting with CD4 , CCL3 , CCL4 , CCL2
29669839	GP (RNF130)	TETHERIN (BST2)	0.160	Moreover , they provide the first evidence that GP can antagonize tetherin in the context of an infectious EBOV surrogate
29321315	RIGI (DDX58)	TRIM25 (TRIM25)	0.151	We report interactions between the Nipah virus V protein and both RIG-I regulatory protein TRIM25 and RIG-I
29471045	PRDM14 (PRDM14)	HOXA1 (HOXA1)	0.147	Here , we confirm PRDM14 is an interactor of HOXA1 and we identify the homeodomain of HOXA1 as well as the PR domain and Zinc fingers of PRDM14 to be required for the interaction
29367244	ATG12 (ATG12)	ATG5 (ATG5)	0.147	A close inspection of the HBV/autophagy cross talk revealed that the virus depended on Atg12 covalently conjugated to Atg5
29760086	STIM1 (STIM1)	ORAI1 (ORAI1)	0.138	Store-operated Orai1 channels are activated through a unique inside-out mechanism involving binding of the endoplasmic reticulum Ca sensor STIM1 to cytoplasmic sites on Orai1
29549180	P53 (TP53)	MDM2 (MDM2)	0.133	In unstressed cells , p53 is normally held in check by MDM2 that targets p53 for transcriptional repression , proteasomal degradation , and cytoplasmic localization
29614078	GDNF (GDNF)	AP1 (JUND)	0.124	GDNF stimulates MAP kinase , activating the transcription factors SRF and AP-1
29512721	ALP (SLPI)	FOXO1 (FOXO1)	0.116	The HIF1induced expression of Runx2 and ALP may be completely dependent on the expression levels of Foxo1 , and in turn , osteocalcin may be partially dependent on Foxo1 expression
29734338	RIT1 (RIT1)	RAC1 (RNASE1)	0.111	We found RIT1 also to directly interact with the RHO GTPases CDC42 and RAC1 , both of which are crucial regulators of actin dynamics upstream of PAK1
29232376	ADA2 (TADA2A)	AATF (AATF)	0.111	In this study , we aimed to identify whether besides ADA3 , other components of the HAT modules of SAGA and ATAC complexes , human ADA2 and GCN5 also interact with Che-1 / AATF
30021902	CDC25A (CDC25A)	TBK1 (TBK1)	0.107	These results demonstrate that Cdc25A inhibits the antiviral immune response by reducing the active form of TBK1
29339503	BTN3A2 (BTN3A2)	BTN3A1 (BTN3A1)	0.107	Addressing this paradox , we show that BTN3A2 regulates the subcellular localization of BTN3A1 , including functionally important associations with the endoplasmic reticulum (ER) , and is specifically required for optimal BTN3A1-mediated activation of V9V2 T cells
29328377	E2F1 (E2F1)	CDK2 (CDK2)	0.107	The cyclindependent kinase inhibitor 1A (CDKN1A) , E2F transcription factor 1 (E2F1) , and MYC interacted with CDK2
30181274	RAB5 (RAB5A)	RABGAP5 (SGSM3)	0.102	Thus , binding of Etf-2 to RAB5-GTP appears to delay RAB5 inactivation by impeding RABGAP5 localization to endosomes
29899144	LPL (LPL)	GPIHBP1 (GPIHBP1)	0.102	Third , we show that LPL accumulates near capillary endothelial cells even in the absence of GPIHBP1
29949917	DDX6 (DDX6)	RIGI (DDX58)	0.098	These findings imply a novel function for DDX6 as an RNA cosensor and signaling enhancer for RIG-I

Continued on next page

PMID	INT A	INT B	Conf.	Sentence after tokenization
29949917	DDX6 (DDX6)	RIGI (DDX58)	0.098	Notably , DDX6 was found to bind viral RNA capable to stimulate RIG-I
29712868	SOMATOSTATIN28 (SST)	INSULIN (INS)	0.098	To demonstrate the usefulness of this approach in revealing regulatory pathways , we identify from among the contacted sites the previously uncharacterized gene and show that it plays an important role in controlling the effect of somatostatin-28 on insulin secretion
29269420	L1 (IGKV116)	CENPA (CENPA)	0.098	Mutational analyses demonstrate analogous interactions in , which are further supported by residue-swapping experiments involving the L1 loop of CENP-A
29450990	RBM10 (RBM10)	RBM5 (RBM5)	0.089	Splicing factor RBM10 and its close homologues RBM5 and RBM6 govern the splicing of oncogenes such as Fas , NUMB , and Bcl-X
29321329	SAMHD1 (SAMHD1)	CYCLIN A2 (CCNA2)	0.089	Human SAMHD1 interacts with cell cycle regulatory proteins cyclin A2 , cyclin-dependent kinase 1 (CDK1) , and CDK2
29205500	HSAMIR1270 (MIR1270)	TAP2 (TAP2)	0.089	Our findings support the hypothesis that hsa-miR-1270 suppresses the production of TAP2 by binding to this SNP in the 3 ' - UTR of this gene
29281729	IFITM3 (IFITM3)	LSD1 (KDM1A)	0.087	However , infection by either Vesicular Stomatitis Virus (VSV) or Influenza A Virus (IAV) triggers methylation of IFITM3 by promoting its disassociation from LSD1
29297316	CCR5 (CCR5)	CCL4 (CCL4)	0.084	It was found that the subnetworks formed by CCR5 and IFNAR1 and their neighbors were a fragments of two key pathways functioning during the course of tick-borne encephalitis : (1) the attenuation of interferon-I signaling pathway by the TBEV NS5 protein that targeted peptidase D ; (2) proinflammation and tissue damage pathway triggered by chemokine receptor CCR5 interacting with CD4 , CCL3 , CCL4 , CCL2
29734338	RIT1 (RIT1)	CDC42 (CDC42)	0.080	This effect was prevented by co-expression of RIT1 with dominant-negative CDC42 or RAC1 and kinase-dead PAK1
29505800	RAB7A (RAB7A)	ER (EREG)	0.076	Collectively , these findings reveal a new role of Rab7a in ER homeostasis , and indicate that genetic and pharmacological ER stress manipulation may restore ER morphology in Rab7a silenced cells
29382845	TAX (CNTN2)	UPF1 (UPF1)	0.076	Tax interacts with the central helicase core domain of UPF1 and might plug the RNA channel of UPF1 , reducing its affinity for nucleic acids
29297316	CCR5 (CCR5)	CCL3 (CCL3)	0.076	It was found that the subnetworks formed by CCR5 and IFNAR1 and their neighbors were a fragments of two key pathways functioning during the course of tick-borne encephalitis : (1) the attenuation of interferon-I signaling pathway by the TBEV NS5 protein that targeted peptidase D ; (2) proinflammation and tissue damage pathway triggered by chemokine receptor CCR5 interacting with CD4 , CCL3 , CCL4 , CCL2
29925821	RHOA (RHOA)	CDC42 (CDC42)	0.071	All genetic evidences indicate that in these disorders the RhoA pathway is hyperactive while the Rac1 and cdc42 pathways are consistently hypoactive
29669839	TETHERIN (BST2)	GP (RNF130)	0.071	However , tetherin antagonism by GP has so far been demonstrated only with virus-like particles , and it is unknown whether GP can block tetherin in infected cells
30021902	CDC25A (CDC25A)	TBK1 (TBK1)	0.062	We demonstrated that Cdc25A reduces TBK1 activity and consequently restrains the activation of IFN- transcription as well as the antiviral status of nearby cells
29950413	STAT3 (STAT3)	SOX2 (SOX2)	0.062	IE1 mediates SOX2 depletion by targeting STAT3 , a critical upstream regulator of SOX2 expression
29501414	PP1C (PPP1CC)	MYPT1 (PPP1R12A)	0.062	Br-BASG and TFM-BASG suppressed partially binding of PP1c to MYPT1 in surface plasmon resonance based binding experiments
29669839	GP (RNF130)	TETHERIN (BST2)	0.058	However , tetherin antagonism by GP has so far been demonstrated only with virus-like particles , and it is unknown whether GP can block tetherin in infected cells
29614078	GDNF (GDNF)	SRF (SRF)	0.058	GDNF stimulates MAP kinase , activating the transcription factors SRF and AP-1
29571014	EIF4A3 (EIF4A3)	CASC2 (CASC2)	0.058	Knockdown of EIF4A3 reversed the effects of sanguinarine plus CASC2 silencing

Continued on next page

PMID	INT A	INT B	Conf.	Sentence after tokenization
29297316	CCR5 (CCR5)	CCL2 (CCL2)	0.053	It was found that the subnetworks formed by CCR5 and IFNAR1 and their neighbors were a fragments of two key pathways functioning during the course of tick-borne encephalitis : (1) the attenuation of interferon-I signaling pathway by the TBEV NS5 protein that targeted peptidase D ; (2) proinflammation and tissue damage pathway triggered by chemokine receptor CCR5 interacting with CD4 , CCL3 , CCL4 , CCL2
29444113	KCNQ1 (KCNQ1)	KCNE1 (KCNE1)	0.049	The voltage-gated KCNQ1 potassium ion channel interacts with the type I transmembrane protein minK (KCNE1) to generate the slow delayed rectifier (IKs) current in the heart
29212935	MTOR (MTOR)	CD63 (CD63)	0.049	Increases in mTOR activation following CD63 knockout are coincident with the development of serum-dependent autophagic vacuoles that are acidified in the presence of high LMP1 levels
29684085	BRD4 (BRD4)	TAT (TAT)	0.044	This unusual acetylation profile attracts Brd4 to suppress the interaction of Tat with the host super elongation complex (SEC) that is essential for productive HIV transcription and latency reversal
29450990	RBM10 (RBM10)	RBM6 (RBM6)	0.044	Splicing factor RBM10 and its close homologues RBM5 and RBM6 govern the splicing of oncogenes such as Fas , NUMB , and Bcl-X
29494137	CAM (KRIT1)	CSH2 (CSH2)	0.040	Because of the flexible nature of both CaM and cSH2 , multiple binding modes of the interactions are possible
29605296	ATF3 (ATF3)	JNK (MAPK8)	0.036	Mechanistically , TNF--induced Atf3 expression was significantly suppressed by the inhibition of the c-Jun N-terminal kinase (JNK) pathway
29925658	IFN (IFNA1)	STAT1 (STAT1)	0.031	This demonstrates that the C-terminal domain of nuclear nsP2 specifically inhibits the IFN response by promoting the nuclear export of STAT1
29705079	TLR2 (TLR2)	SAA1 (SAA1)	0.031	As TLR2 is known to be a functional receptor of SAA1 , a co-immunoprecipitation assay was performed
29375208	IFN (IFNA1)	HUH7 (MIR73HG)	0.031	HCV-1b core protein increased miR-93-5p expression and induced inactivation of the IFN signaling pathway in Huh7 cells
29269396	SP1 (SP1)	OSM (OSM)	0.031	Knockdown of Sp1 abrogated the expression and functionality of OSM
29614078	SRF (SRF)	EGR1 (EGR1)	0.022	SRF initiates an immediate transcriptional response , activating EGR1 and suppressing ER
29495654	NSP3 (SH2D3C)	PIN (DYNLL1)	0.018	The function of nsP3 has been more difficult to pin down and it has long been referred to as the more enigmatic of the nsPs
29328377	E2F TRANSCRIPTION FACTOR 1 (E2F1)	CDK2 (CDK2)	0.018	The cyclindependent kinase inhibitor 1A (CDKN1A) , E2F transcription factor 1 (E2F1) , and MYC interacted with CDK2
29184850	CELLUGYRIN (SYNGR2)	CDTB (CDB2)	0.018	We propose that cellugyrin plays a critical role in the internalization and translocation of CdtB to critical intracellular target sites
29203283	PD1 (SPATA2)	CDR2 (CDR2)	0.013	The residue-residue contact analysis further shows that PD-1 interacts with PD-L1 mainly by F and G strands while monoclonal antibodies like avelumab and BMS-936559 mainly interact with PD-L1 by CDR2 and CDR3 loops of the heavy chain
30181274	RAB5 (RAB5A)	RAB5 (RAB5A)	0.009	Ectopically expressed Etf-2-GFP also localized to inclusions and membranes of early endosomes marked with RAB5 and interacted with GTP-bound RAB5 but not with a GDP-bound RAB5
29895728	E2 (SNORA62)	E1 (SNORA73A)	0.004	These findings support a role for E2 beyond E1 recruitment in viral DNA replication , demonstrate a novel functional interaction in PV DNA replication , and further implicate cellular pol in PV DNA replication
29382759	MALT1 (MALT1)	BCL10 (BCL10)	0.001	Cooperative MALT1 interaction with BCL10 filaments observed under EM suggests immediate dimerization of MALT1 in the BCL10 filamentous scaffold

Table S3: **PPaxe features**.

Feature name	Description
TOK_DIST	Token distance between ProtA and ProtB.
TOTAL_TOK	Number of Tokens in sentence.
BETWEEN_VB_COUNT	Number of Tokens tagged as VB between ProtA and ProtB.
BETWEEN_VBD_COUNT	Number of Tokens tagged as VBD between ProtA and ProtB.
BETWEEN_VBG_COUNT	Number of Tokens tagged as VBG between ProtA and ProtB.
BETWEEN_VBN_COUNT	Number of Tokens tagged as VBN between ProtA and ProtB.
BETWEEN_VBP_COUNT	Number of Tokens tagged as VBP between ProtA and ProtB.
BETWEEN_VBZ_COUNT	Number of Tokens tagged as VBZ between ProtA and ProtB.
BETWEEN_VERB_MAXSCORE	Higher score for verbs between ProtA and ProtB.
BETWEEN_VERB_TOTALSCORE	Sum of verb scores between ProtA and ProtB.
BETWEEN_VERB_CLOSEST_DIST_A	Token distance to closer verb to ProtA located between ProtA and ProtB.
BETWEEN_VERB_FARTHEST_DIST_A	Token distance to farthest verb to ProtA located between ProtA and ProtB.
BETWEEN_VERB_CLOSEST_DIST_B	Token distance to closer verb to ProtB located between ProtA and ProtB.
BETWEEN_VERB_FARTHEST_DIST_B	Token distance to farthest verb to ProtB located between ProtA and ProtB.
ALL_VB_COUNT	Total number of Tokens tagged as VB in sentence.
ALL_VBD_COUNT	Total number of Tokens tagged as VBD in sentence.
ALL_VBG_COUNT	Total number of Tokens tagged as VBG in sentence.
ALL_VBN_COUNT	Total number of Tokens tagged as VBN in sentence.
ALL_VBP_COUNT	Total number of Tokens tagged as VBP in sentence.
ALL_VBZ_COUNT	Total number of Tokens tagged as VBZ in sentence.
ALL_VERB_MAXSCORE	Higher scoring verb in sentence.
ALL_VERB_TOTALSCORE	Sum of verb scores in sentence.
ALL_VERB_CLOSEST_DIST_A	Token distance to closer verb to ProtA in sentence.
ALL_VERB_FARTHEST_DIST_A	Token distance to farthest verb to ProtA in sentence.
ALL_VERB_CLOSEST_DIST_B	Token distance to closer verb to ProtB in sentence.
ALL_VERB_FARTHEST_DIST_B	Token distance to farthest verb to ProtB in sentence.
BETWEEN_POS_2AP	Number of Tokens tagged as 2AP between ProtA and ProtB.
BETWEEN_POS_1AP	Number of Tokens tagged as 1AP between ProtA and ProtB.
BETWEEN_POS_COMMA	Number of Tokens tagged as COMMA between ProtA and ProtB.
BETWEEN_POS_LRB	Number of Tokens tagged as LRB between ProtA and ProtB.
BETWEEN_POS_RRB	Number of Tokens tagged as RRB between ProtA and ProtB.
BETWEEN_POS_DOT	Number of Tokens tagged as DOT between ProtA and ProtB.
BETWEEN_POS_COLON	Number of Tokens tagged as COLON between ProtA and ProtB.
BETWEEN_POS_CC	Number of Tokens tagged as CC between ProtA and ProtB.
BETWEEN_POS_CD	Number of Tokens tagged as CD between ProtA and ProtB.
BETWEEN_POS_DT	Number of Tokens tagged as DT between ProtA and ProtB.
BETWEEN_POS_EX	Number of Tokens tagged as EX between ProtA and ProtB.
BETWEEN_POS_FW	Number of Tokens tagged as FW between ProtA and ProtB.
BETWEEN_POS_IN	Number of Tokens tagged as IN between ProtA and ProtB.
BETWEEN_POS_JJ	Number of Tokens tagged as JJ between ProtA and ProtB.
BETWEEN_POS_JJR	Number of Tokens tagged as JJR between ProtA and ProtB.
BETWEEN_POS_JJS	Number of Tokens tagged as JJS between ProtA and ProtB.
BETWEEN_POS_LS	Number of Tokens tagged as LS between ProtA and ProtB.
BETWEEN_POS_MD	Number of Tokens tagged as MD between ProtA and ProtB.

Continued on next page

Feature name	Description
BETWEEN_POS_NN	Number of Tokens tagged as NN between ProtA and ProtB.
BETWEEN_POS_NNP	Number of Tokens tagged as NNP between ProtA and ProtB.
BETWEEN_POS_NNPS	Number of Tokens tagged as NNPS between ProtA and ProtB.
BETWEEN_POS_NNS	Number of Tokens tagged as NNS between ProtA and ProtB.
BETWEEN_POS_PDT	Number of Tokens tagged as PDT between ProtA and ProtB.
BETWEEN_POS_POS	Number of Tokens tagged as POS between ProtA and ProtB.
BETWEEN_POS_PRP	Number of Tokens tagged as PRP between ProtA and ProtB.
BETWEEN_POS_PRP_DOLAR	Number of Tokens tagged as PRP_DOLAR between ProtA and ProtB.
BETWEEN_POS_RB	Number of Tokens tagged as RB between ProtA and ProtB.
BETWEEN_POS_RBR	Number of Tokens tagged as RBR between ProtA and ProtB.
BETWEEN_POS_RBS	Number of Tokens tagged as RBS between ProtA and ProtB.
BETWEEN_POS_RP	Number of Tokens tagged as RP between ProtA and ProtB.
BETWEEN_POS_SYM	Number of Tokens tagged as SYM between ProtA and ProtB.
BETWEEN_POS_TO	Number of Tokens tagged as TO between ProtA and ProtB.
BETWEEN_POS_UH	Number of Tokens tagged as UH between ProtA and ProtB.
BETWEEN_POS_VB	Number of Tokens tagged as VB between ProtA and ProtB.
BETWEEN_POS_VBD	Number of Tokens tagged as VBD between ProtA and ProtB.
BETWEEN_POS_VBG	Number of Tokens tagged as VBG between ProtA and ProtB.
BETWEEN_POS_VBN	Number of Tokens tagged as VBN between ProtA and ProtB.
BETWEEN_POS_VBP	Number of Tokens tagged as VBP between ProtA and ProtB.
BETWEEN_POS_VBZ	Number of Tokens tagged as VBZ between ProtA and ProtB.
BETWEEN_POS_WDT	Number of Tokens tagged as WDT between ProtA and ProtB.
BETWEEN_POS_WP	Number of Tokens tagged as WP between ProtA and ProtB.
BETWEEN_POS_WP_DOLAR	Number of Tokens tagged as WP_DOLAR between ProtA and ProtB.
BETWEEN_POS_WRB	Number of Tokens tagged as WRB between ProtA and ProtB.
ALL_POS_2AP	Number of Tokens tagged as 2AP in sentence.
ALL_POS_1AP	Number of Tokens tagged as 1AP in sentence.
ALL_POS_COMMA	Number of Tokens tagged as COMMA in sentence.
ALL_POS_LRB	Number of Tokens tagged as LRB in sentence.
ALL_POS_RRB	Number of Tokens tagged as RRB in sentence.
ALL_POS_DOT	Number of Tokens tagged as DOT in sentence.
ALL_POS_COLON	Number of Tokens tagged as COLON in sentence.
ALL_POS_CC	Number of Tokens tagged as CC in sentence.
ALL_POS_CD	Number of Tokens tagged as CD in sentence.
ALL_POS_DT	Number of Tokens tagged as DT in sentence.
ALL_POS_EX	Number of Tokens tagged as EX in sentence.
ALL_POS_FW	Number of Tokens tagged as FW in sentence.
ALL_POS_IN	Number of Tokens tagged as IN in sentence.
ALL_POS_JJ	Number of Tokens tagged as JJ in sentence.
ALL_POS_JJR	Number of Tokens tagged as JJR in sentence.
ALL_POS_JJS	Number of Tokens tagged as JJS in sentence.
ALL_POS_LS	Number of Tokens tagged as LS in sentence.
ALL_POS_MD	Number of Tokens tagged as MD in sentence.
ALL_POS_NN	Number of Tokens tagged as NN in sentence.
ALL_POS_NNP	Number of Tokens tagged as NNP in sentence.
ALL_POS_NNPS	Number of Tokens tagged as NNPS in sentence.

Continued on next page

Feature name	Description
ALL_POS_NNS	Number of Tokens tagged as NNS in sentence.
ALL_POS_PDT	Number of Tokens tagged as PDT in sentence.
ALL_POS_POS	Number of Tokens tagged as POS in sentence.
ALL_POS_PRP	Number of Tokens tagged as PRP in sentence.
ALL_POS_PRP_DOLAR	Number of Tokens tagged as PRP_DOLAR in sentence.
ALL_POS_RB	Number of Tokens tagged as RB in sentence.
ALL_POS_RBR	Number of Tokens tagged as RBR in sentence.
ALL_POS_RBS	Number of Tokens tagged as RBS in sentence.
ALL_POS_RP	Number of Tokens tagged as RP in sentence.
ALL_POS_SYM	Number of Tokens tagged as SYM in sentence.
ALL_POS_TO	Number of Tokens tagged as TO in sentence.
ALL_POS_UH	Number of Tokens tagged as UH in sentence.
ALL_POS_VB	Number of Tokens tagged as VB in sentence.
ALL_POS_VBD	Number of Tokens tagged as VBD in sentence.
ALL_POS_VBG	Number of Tokens tagged as VBG in sentence.
ALL_POS_VBN	Number of Tokens tagged as VBN in sentence.
ALL_POS_VBP	Number of Tokens tagged as VBP in sentence.
ALL_POS_VBZ	Number of Tokens tagged as VBZ in sentence.
ALL_POS_WDT	Number of Tokens tagged as WDT in sentence.
ALL_POS_WP	Number of Tokens tagged as WP in sentence.
ALL_POS_WP_DOLAR	Number of Tokens tagged as WP_DOLAR in sentence.
ALL_POS_WRB	Number of Tokens tagged as WRB in sentence.
BETWEEN_PROTA_COUNT	Number of times ProtA appears between ProtA and ProtB.
BETWEEN_PROTB_COUNT	Number of times ProtB appears between ProtA and ProtB.
ALL_PROTA_COUNT	Number of times ProtA appears in sentence.
ALL_PROTB_COUNT	Number of times ProtB appears in sentence.
KEYWORD_COUNT_acetylate	Number of times 'acetylate' appears between ProtA and ProtB.
KEYWORD_COUNT_activate	Number of times 'activate' appears between ProtA and ProtB.
KEYWORD_COUNT_acylate	Number of times 'acylate' appears between ProtA and ProtB.
KEYWORD_COUNT_amidate	Number of times 'amidate' appears between ProtA and ProtB.
KEYWORD_COUNT_assemble	Number of times 'assemble' appears between ProtA and ProtB.
KEYWORD_COUNT_attach	Number of times 'attach' appears between ProtA and ProtB.
KEYWORD_COUNT_bind	Number of times 'bind' appears between ProtA and ProtB.
KEYWORD_COUNT_biotinylate	Number of times 'biotinylate' appears between ProtA and ProtB.
KEYWORD_COUNT_block	Number of times 'block' appears between ProtA and ProtB.
KEYWORD_COUNT_brominate	Number of times 'brominate' appears between ProtA and ProtB.
KEYWORD_COUNT_carboxylate	Number of times 'carboxylate' appears between ProtA and ProtB.
KEYWORD_COUNT_catalyze	Number of times 'catalyze' appears between ProtA and ProtB.
KEYWORD_COUNT_cleave	Number of times 'cleave' appears between ProtA and ProtB.
KEYWORD_COUNT_complex	Number of times 'complex' appears between ProtA and ProtB.
KEYWORD_COUNT_conjugate	Number of times 'conjugate' appears between ProtA and ProtB.
KEYWORD_COUNT_contact	Number of times 'contact' appears between ProtA and ProtB.
KEYWORD_COUNT_couple	Number of times 'couple' appears between ProtA and ProtB.
KEYWORD_COUNT_cysteinylate	Number of times 'cysteinylate' appears between ProtA and ProtB.
KEYWORD_COUNT_demethylate	Number of times 'demethylate' appears between ProtA and ProtB.
KEYWORD_COUNT_dephosphorylate	Number of times 'dephosphorylate' appears between ProtA and ProtB.

Continued on next page

Feature name	Description
KEYWORD_COUNT_dimerise	Number of times 'dimerise' appears between ProtA and ProtB.
KEYWORD_COUNT_dimerize	Number of times 'dimerize' appears between ProtA and ProtB.
KEYWORD_COUNT_disassemble	Number of times 'disassemble' appears between ProtA and ProtB.
KEYWORD_COUNT_discharge	Number of times 'discharge' appears between ProtA and ProtB.
KEYWORD_COUNT_dissociate	Number of times 'dissociate' appears between ProtA and ProtB.
KEYWORD_COUNT_down-regulate	Number of times 'down-regulate' appears between ProtA and ProtB.
KEYWORD_COUNT_downregulate	Number of times 'downregulate' appears between ProtA and ProtB.
KEYWORD_COUNT_farnesylate	Number of times 'farnesylate' appears between ProtA and ProtB.
KEYWORD_COUNT_formylate	Number of times 'formylate' appears between ProtA and ProtB.
KEYWORD_COUNT_hydroxilate	Number of times 'hydroxilate' appears between ProtA and ProtB.
KEYWORD_COUNT_hydroxylate	Number of times 'hydroxylate' appears between ProtA and ProtB.
KEYWORD_COUNT_inactivate	Number of times 'inactivate' appears between ProtA and ProtB.
KEYWORD_COUNT_induce	Number of times 'induce' appears between ProtA and ProtB.
KEYWORD_COUNT_inhibit	Number of times 'inhibit' appears between ProtA and ProtB.
KEYWORD_COUNT_interact	Number of times 'interact' appears between ProtA and ProtB.
KEYWORD_COUNT_mediate	Number of times 'mediate' appears between ProtA and ProtB.
KEYWORD_COUNT_methylate	Number of times 'methylate' appears between ProtA and ProtB.
KEYWORD_COUNT_modify	Number of times 'modify' appears between ProtA and ProtB.
KEYWORD_COUNT_modulate	Number of times 'modulate' appears between ProtA and ProtB.
KEYWORD_COUNT_multimerise	Number of times 'multimerise' appears between ProtA and ProtB.
KEYWORD_COUNT_multimerize	Number of times 'multimerize' appears between ProtA and ProtB.
KEYWORD_COUNT_myristoylate	Number of times 'myristoylate' appears between ProtA and ProtB.
KEYWORD_COUNT_myristylate	Number of times 'myristylate' appears between ProtA and ProtB.
KEYWORD_COUNT_nitrosylate	Number of times 'nitrosylate' appears between ProtA and ProtB.
KEYWORD_COUNT_overexpress	Number of times 'overexpress' appears between ProtA and ProtB.
KEYWORD_COUNT_palmitoylate	Number of times 'palmitoylate' appears between ProtA and ProtB.
KEYWORD_COUNT_palmitylate	Number of times 'palmitylate' appears between ProtA and ProtB.
KEYWORD_COUNT_phosphorylate	Number of times 'phosphorylate' appears between ProtA and ProtB.
KEYWORD_COUNT_precipitate	Number of times 'precipitate' appears between ProtA and ProtB.
KEYWORD_COUNT_promote	Number of times 'promote' appears between ProtA and ProtB.
KEYWORD_COUNT_pyruvate	Number of times 'pyruvate' appears between ProtA and ProtB.
KEYWORD_COUNT_regulate	Number of times 'regulate' appears between ProtA and ProtB.
KEYWORD_COUNT_repress	Number of times 'repress' appears between ProtA and ProtB.
KEYWORD_COUNT_stimulate	Number of times 'stimulate' appears between ProtA and ProtB.
KEYWORD_COUNT_substitute	Number of times 'substitute' appears between ProtA and ProtB.
KEYWORD_COUNT_sumoylate	Number of times 'sumoylate' appears between ProtA and ProtB.
KEYWORD_COUNT_suppress	Number of times 'suppress' appears between ProtA and ProtB.
KEYWORD_COUNT_transactivate	Number of times 'transactivate' appears between ProtA and ProtB.
KEYWORD_COUNT_ubiquitinate	Number of times 'ubiquitinate' appears between ProtA and ProtB.
KEYWORD_COUNT_ubiquitinylate	Number of times 'ubiquitinylate' appears between ProtA and ProtB.
KEYWORD_COUNT_up-regulate	Number of times 'up-regulate' appears between ProtA and ProtB.
KEYWORD_COUNT_upregulate	Number of times 'upregulate' appears between ProtA and ProtB.