frontiers in Genetics

OPINION
published: 31 March 2016
doi: 10.3389/fgene.2016.00046

Biomolecular Relationships Discovered from Biological Labyrinth and Lost in Ocean of Literature: Community Efforts Can Rescue Until Automated Artificial Intelligence Takes Over

Rajinder Gupta and Shrikant S. Mantri *

Computational Biology Lab, National Agri-Food Biotechnology Institute, Mohali, India

Keywords: literature mining, named entity recognition, information retrieval, information extraction, community efforts, relationship database, interactions, relationships

OPEN ACCESS

Edited by:
Alessandro Laganà, Icahn School of Medicine at Mount Sinai, USA

Reviewed by:
Yongqun "Oliver" He, University of Michigan, USA

***Correspondence:**
Shrikant S. Mantri
shrikant@nabi.res.in;
bioinfoman3@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *20 January 2016*
Accepted: *15 March 2016*
Published: *31 March 2016*

Citation:
Gupta R and Mantri SS (2016) Biomolecular Relationships Discovered from Biological Labyrinth and Lost in Ocean of Literature: Community Efforts Can Rescue Until Automated Artificial Intelligence Takes Over. Front. Genet. 7:46. doi: 10.3389/fgene.2016.00046

Many brilliant minds are at work to decipher the biological labyrinth and as a result immense amount of information about biological entities and their relationships is getting accumulated in the form of published literature (Hunter and Cohen, 2006). To cater the needs of a researcher, many tools are designed to perform tasks of Named Entity Recognition (NER), Information Retrieval (IR), and Information Extraction (IE) viz. A Combined Clinical Concept Annotator (Kang et al., 2012), BANNER (Leaman and Gonzalez, 2008), Biblio-MetReS (Usie et al., 2014), BioTextQuest+ (Papanikolaou et al., 2014), BIOSMILE Web Search (Dai et al., 2008),E3Miner (Lee et al., 2008), EBIMed (Rebholz-Schuhmann et al., 2007), eFIP (Arighi et al., 2011), FACTA+ (Tsuruoka et al., 2008), GNSuite[1] iHOP (Hoffmann and Valencia, 2004), MyMiner (Salgado et al., 2012), RLIMS-P(Hu et al., 2005), Anni (Jelier et al., 2008), CoPub (Frijters et al., 2008), MedScan (Novichkova et al., 2003), PPInterFinder (Raja et al., 2012), pGenN (Ding et al., 2015), SciMiner (Hur et al., 2009), BIGNER (Li et al., 2009), hybrid named entity tagger (Raja et al., 2014), and more such tools can be obtained from BIONLP resource[2] and in detail analysis of many NLP tools is given by Krallinger et al. (2008) and Fleuren and Alkema (2015). **Table 1** gives an informational and statistical insight into some of these literature mining tools, shedding light on their efficiency translated by statistical parameters viz. F-score, recall, and precision. Many tools are domain specific like kinase family specific but still calls for human intervention for exactitude and thus limit their usage. Moreover, the data output formats are sometimes too vague as name highlighting; to be put to use for bigger literature searches.

The naming ambiguity in scientific literature is one of the major concerns for NER and sentence structure for IR and IE. Presently, NER tools need to maintain a comprehensive dictionary of all names, aliases and web-repository specific IDs or have their AI (Artificial Intelligence) defined algorithms trained on many test data sets. Many such dictionaries are available but the list is ever-increasing and so is the training data set. This results into investing more money, time and effort in obtaining a comprehensive list of names, aliases and IDs. A very comprehensive work on NLP can be found on BioNLP.[3] The availability of manpower or intellect is huge but there is acute scarcity

[1] https://www.idi.ntnu.no/~satre/biocreative/IAT/.
[2] http://zope.bioinfo.cnio.es/bionlp_tools/.
[3] http://bionlp.org/.

of funds (Bourne et al., 2015), so we have to device optimized approaches to take care of the issues discussed in subsequent section.

ISSUES IN LITERATURE TEXT MINING

Let's have a deeper look into major concerns in biological literature mining:

(A) Non-standard naming conventions:
The absence of any standard naming convention(s) for biological entities results in ambiguity and chaos. Presence of eponyms (Vedantam and Viswanathan, 2012) e.g., Bence Jones' protein, Wolfgram protein, Pokemon, Pikachurin etc., naming based on localization of proteins e.g., B-cell receptor-associated protein, naming based on function e.g. "101 kDa heat shock protein," naming based on function and/or sequence similarity e.g., Epidermal growth factor-like protein 7 etc.; have all added to the complexity. A lot of research has been done to systematically name proteins and genes but no universal standards have been approved so far.

(B) Too many names:
Owing to bad conventions followed to name biological entities many aliases (common name, acronym, descriptive name etc) for biological entities have come into existence (Iragne et al., 2004) e.g., 14-3-3 protein beta/alpha, Protein 1054, Protein kinase C inhibitor protein 1, KCIP-1 for one protein. Too many web-repositories have also resulted in many IDs for one entity e.g.,: P62258, P42655, CAB016200, CAB021109, CAB047350, HPA008445 for Uniprot Id P62258. And lastly non-uniform names e.g.,: AAD14 protein, AAD-14 protein, AAD 14 protein for Uniprot Id Q99415 adds to the problem.

(C) Defining relationships:
English is the prime language of published research and it is evolving, and because of it NLP (Natural Language Processing) algorithms will never be 100% precise. Moreover, different people have different ways of putting up information and expressing their thoughts resulting in varied sentence corpuses making NER and IR tasks more difficult (Nadkarni et al., 2011). For defining relationships there are absolutely no conventions followed making it harder for the NLP tools.

(D) Scarcity of funds:
The biological research demands too much of funds (Bourne et al., 2015) and more for its IT (Information Technology) support for the enormous amount of data that is generated. To provide a computational facility, that includes storage, data management, and making it available to the community through GUI (Graphical User Interface), it is an expensive affair. In addition, looking at amount of resources invested in devising NLP is too big to ignore.

(E) Unavailability of full text articles:
Many high reputed journals provide their content for a price and only abstracts are available for free, making it harder for the researchers working in the domain to get hands on the missing information (Mower and Youngkin, 2008; Singh et al., 2011). There are ∼3.7 million PMC full text articles and ∼14 million Pubmed abstracts[4] , conveying we are only having ∼25% of research at hand to go forward and this will increase further in days to come. Furthermore, the online unavailability of the supplementary material is a great setback for information extraction process (Evangelou et al., 2005).

MORE DATA LESS INFORMATION

NCBI[4] houses 14,096,969 publications and a total of 64,815,068 genes and proteins; Uniprot[5] houses 53,333,247 proteins collected from 1,007,941 publications. The data from Biogrid (Stark et al., 2006), one of the most extensive PPI repository has ∼760 K interactions (I) for ∼80 K proteins (Pr) and covers ∼55 K publications (P) of total ∼14 million present at Pubmed[4]. Some other PPI databases IntAct (Hermjakob et al., 2004; P = 13,892; Pr = 89,430; I = 564,831), DIP (Xenarios et al., 2000; P = 7,817; Pr = 28,215; I = 80,286), MINT (Zanzoni et al., 2002; P = 132,733; Pr = 35,553; I = 241,458), UniHI (Chaurasia et al., 2007; E = 22,300; I = 250,000), APID (Prieto and De Las Rivas, 2006; P = 416,124; Pr = 56,460; I = 322,579) also reflect the gap between the published literature and curated literature. No clear predictions can be made about how many interactions or relations we might be missing with such great amount of literature not being curated but surely a lot is missed. The gap will increase more and will become impassable if steps are not taken in time to bridge it.

The research also shows that so far we have been protein biased and all the relationship studies and repositories are dedicated to proteins (and on occasions, protein coding genes). We have totally missed the point that we are studying a system that comprises of rRNAs, ncRNAs, microsatellites, chemical components, drugs etc., and there is a crying need to bring them to the relationship databases too.

CURRENT PROGRESS

Many changes have been suggested and some have been implemented to take care of expanding biological literature and to make the information available as knowledge to the researchers in accessible formats or for computer programs to make sense of the text. Pubmed describes its own xml structure[6] to store and provide the literature data. Such a structure having dedicated headers for the sections of the article are well suited for storing and retrieving of data but provide no assistance in making inferences form the text. Such an xml structure is limited in its usefulness to the NLP tools in just defining the sections such as title, abstract etc. that needs to be parsed.

More prominent work on making the structure of the format in which the literature is submitted has been carried out by Seringhaus and Gerstein (2007); suggesting to have a

[4]ftp://ftp.ncbi.nlm.nih.gov/pubmed/.
[5]http://www.uniprot.org/downloads.
[6] https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html.

Structured Digital Abstract (SDA) and reporting of findings to appropriate databases, but community participation in populating databases/knowledgebases is very limited (Mazumder et al., 2010). SDA should be of great advantage to NLP and other computer programs to access the data (Superti-Furga et al., 2008) as it precisely defines the attributes such as species, gene, protein, mutation, interaction, experiment etc. in a well-organized and framed manner.

Winnenburg et al. (2008) proposed to have authors make annotations of their work and submit them according to some standard guidelines in addition to the original research paper. Shotton et al. (2009) also proposed many changes of which providing links to data from external sources; highlighting disease, organism, protein etc.; a document summary etc. are few important to take notice of. They also pressed for use of standard ontologies in biology literature. Clark et al. (2014) put forward an innovative approach to tackle the perishing literature issue by introduction of micropublications. They propose to have statement based models.

Ontologies also play a very important role in standardizing biological data such as classes of entities, relationships etc. (He and Xiang, 2013). Robinson and Bauer (2011) in their book have explained in depth about various aspects of bio-ontologies; data organization, integration, searching, computer reasoning etc. are few of them. The use of ontologies and their significance is well studied by Hur et al. (2011, 2015) in their work on gene-gene interactions and vaccines. Many more recommendations to improve the scientific literature's human and computer accessibility are available (Stevens et al., 2002; Leitner and Valencia, 2008; Sainani, 2008; Attwood et al., 2009; Fink et al., 2010) talking of liquid publications[7] etc. are discussed in greater depths.

THE WAYS TO PASS THE IMPASSABLE

The scientific community has already spent jillions of money to uncover various biological phenomena, now to spend more to extract it from literature seems like a trivial task. Points enlisted below can help in addressing the concerns:

(i) Universal biomolecular entity and relationship database:
A universal biomolecular relationships' database and an appropriate intuitive GUI needs to be designed and developed where researcher should submit their biomolecular relationship findings through an interactive data submission form. Every journal should encourage the authors to submit the data at this GUI in addition to submitting it to their journals and after the acceptance of the article the reported findings should go live. The database should house relationship data for all species and for all type of biomolecules in biological systems. All the entities of the database will be linked to external data sources to enhance the information of the entity, process etc. Inclusion of standard ontologies will further enrich the resource.

(ii) New section to the e-version of publications/articles:
A new section which defines the biomolecular entities and relationships in some standardized format should be added to the e-version of the publications/articles as described by many pioneers (Seringhaus and Gerstein, 2007; Clark et al., 2014). This way it should be easier for algorithm designers and developers to extract precise information from the published literature. The section can be in an XML (Extensible Markup Language) or OWL (Web Ontology Language) format (highly accepted across domains) that could be used by various tools and thus makes it easier to populate the relationship database. Journal editors need to take the big step and make it compulsory for the authors to add that new section.

(iii) Data from existing relationship databases:
Many relationship databases have manually curated relationship data (Xenarios et al., 2000; Hermjakob et al., 2004; Stark et al., 2006), that all can be added to the new repository and thus eliminating the need to redo the curation of the literature that has been done once or more. Using crawlers and APIs (Application Program Interface) that data should be integrated into the universal relationship database.

(iv) Data backlog:
Too much of literature is still lying in the dumps of data repositories viz. scientific journals that also need to be taken care of. We can start off with best of the tools (NER, IE, and IR) to handle them and over time let the community work on it to resolve clashes and normalize the relationships.

All the options should be used to eliminate the time gap between data availability i.e., publication of literature and its recognition in relevant databases for e.g., interaction databases, sequence database etc. The journals should provide programmatic access to their literature and supplementary data, allowing for speedy curation and fleeting integration in conformant databases. The authors from the journals open to such programmatic access will feel more to be a part of the knowledge evolution.

Currently the community efforts like Biocreative[8] to solve the literature mining labyrinth have brought to life many new tools and approaches. Huang and Lu (2015) give in depth insights into the community programs and efforts. Similar initiatives need to be taken to accomplish this task also. The task is big but is a needed one, so we appeal the community to participate in designing the structure of form, new section in e-version etc.; developing standards for data submission, xml structure (as discussed by; Seringhaus and Gerstein, 2007), or some more ways like micropublications (Clark et al., 2014) etc.; and populating the databases with their research data and curating the data back log. Scientific journals need to make collaborative efforts to make it obligatory to submit literature in accordance with community established standards. Tools like PALM-IST (Mandloi and Chakrabarti, 2015) that use readily available relationship data from relationship databases to construct the

[7] Casati, F.; Giunchiglia, F.; Marchese, M. Liquid Publications, Scientific Publications Meet the Web; Technical Rep. DIT-07-073, Informatica e Telecomunicazioni; University of Trento: Trento, Italy, 2007. http://eprints.biblio.unitn.it/1313/1/073.pdf.

[8] http://www.biocreative.org/.

TABLE 1 | Informational (viz. data used, parameters for evaluation and working platform) and statistical (viz. *f*-value, recall and precision) insights for a few literature mining tools with their brief description and links to the tools' home page.

Tool	Event/Data used	Parameters	Platform	*F*-value (%)	Recall (%)	Precision (%)	Link	Description
A Combined Clinical Concept Annotator (Kang et al., 2012)	i2b2 challenge	Concept exact match task	Web**	82.1	81.2	83.3	http://www.biosemantics.org/ACCCA_WEB	Concept annotation system for clinical records
Banner (Leaman and Gonzalez, 2008)	BioCreative 2 GM task	NER	Desktop	81.96	79.06	85.09	http://banner.sourceforge.net	Named entity recognition system, primarily intended for biomedical text
Biblio-MetReS (Usie et al., 2014)	Literature Databases and Journals	Biological entities and relationships	Desktop	37	27*	58	http://metres.udl.cat/	To reconstruct networks from an always up to date set of scientific documents
BIOSMILE Web Search (Dai et al., 2008)	BioCreAtIvE II GM tagging task and IAS task	NER and PPI article classifier	Web**	85.76	89.12	82.59	http://bws.iis.sinica.edu.tw/BWS/	Analyze articles for selected biomedical verbs and lists abstracts along with snippets by order of relevancy to protein–protein interaction
E3Miner (Lee et al., 2008)	100 random abstracts	E3 related data	Web	8*	74	97	http://e3miner.biopathway.org/e3miner.html	Extracts novel E3 discoveries and important findings related to specific E3s from the literature
RLIMS-P (Jelier et al., 2008)	BioCreative IV (BioCreative IAT)	Kinase, substrate and site	Web	92	96*	88	http://research.bioinformatics.udel.edu/rlimsp/	Rule-based text-mining program designed to extract protein phosphorylation information on protein kinase, substrate and phosphorylation sites from biomedical literature
Anni 2.0 (Frijters et al., 2008)	Micro-array data and multiple publications	Associations between biological entities	Web**	75.5*	76	75	http://biosemantics.org/anni/	Ontology-based interface to MEDLINE and retrieves documents and associations for several classes of biomedical concepts, including genes, drugs and diseases
PPInterFinder (Ding et al., 2015)	BioCreative workshop 2012	NER, IR	Web**	78.07	70.58	87.33	http://www.biominingbu.org/ppinterfinder/	Extracts human PPIs from biomedical literature using relation keyword co-occurrences with protein names to extract information on PPIs from MEDLINE abstracts
pGenN (Hur et al., 2009)	104 plant relevant abstracts	NER	Web	88.9	87.2	90.9	http://biotm.cis.udel.edu/gn/	A gene normalization tool for plant genes and proteins in scientific literature
SciMiner (Li et al., 2009)	BioCreAtIvE II	NER, IR	Desktop/Web	75.8	87.1	71.3	http://141.214.81.219/SciMiner/	Identifies genes and proteins using a context specific analysis of MEDLINE abstracts and full texts
BIGNER (Raja et al., 2014)	BioCreative 2 GM	NER	Web**	89.05	87.63	90.52	http://202.118.75.18:8080/bioner	To locate gene/protein names in biomedical literature

**These values were self calculated from the given values.*

***Out of order web-interfaces.*

i2b2, Informatics for Integrating Biology and the Bedside; GM, Gene Mention; IAS, Interaction Article Sub-task; E3, ubiquitin-protein ligase; IAT, Interactive Task.

biological interaction maps will be able to make good use of such relationship databases. Moreover, the precise relationship information will in turn provide diverse data sets for training our algorithms and should allow us to cover all literature that is not published with set norms. Semantics Scholar[9], Plato[10], and Aristo[11] are artificial intelligence based natural language processing, visual knowledge extraction and reasoning systems, respectively, developed for searching relevant relationships from text, inferring information from images and answering questions from varied sources of information. Predictive potential of such novel tools in the field of biology should improve drastically by utilizing community cumulated biomolecular relationship knowledge.

[9]http://allenai.org/semantic-scholar.html.

[10]http://allenai.org/plato.html.

[11]http://allenai.org/aristo.html.

Automating the literature mining process using NER, IE and IR has proved to be a costly affair with slow progress as compared to the speed of new research getting published. More robust approaches need to be thought of to accommodate the gap between the published literature and manually curated literature. One way to achieve this is by having a universal biomolecular relationship database and data submission GUI where all biological relationship information is shared by the authors themselves. Extensive community efforts will be required to achieve such an enormous task.

AUTHOR CONTRIBUTIONS

SM: Conceived and designed the study. RG and SM: Performed literature review, wrote the manuscript and approve this final draft.

FUNDING

This work was financially supported by the core grant of National Agri-Food Biotechnology Institute (NABI), Mohali, India and the Department of Biotechnology, Government of India.

REFERENCES

Arighi, C. N., Siu, A. Y., Tudor, C. O., Nchoutmboube, J. A., Wu, C. H., and Shanker, V. K. (2011). "eFIP: a tool for mining functional impact of phosphorylation from literature," in *Bioinformatics for Comparative Proteomics*, eds C. H. Wu and C. Chen (Newark, NJ: Humana Press), 63–75. doi: 10.1007/978-1-60761-977-2_5

Attwood, T., Kell, D., McDermott, P., Marsh, J., Pettifer, S., and Thorne, D. (2009). Calling International Rescue: knowledge lost in literature and data landslide!. *Biochem. J.* 424, 317–333. doi: 10.1042/BJ20091474

Bourne, P. E., Lorsch, J. R., and Green, E. D. (2015). Perspective: sustaining the big-data ecosystem. *Nature* 527, S16–S17. doi: 10.1038/527S16a

Chaurasia, G., Iqbal, Y., Hänig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.* 35, D590–D594. doi: 10.1093/nar/gkl817

Clark, T., Ciccarese, P., and Goble, C. (2014). Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J. Biomed. Semantics* 5:28. doi: 10.1186/2041-1480-5-28

Dai, H. J., Huang, C. H., Lin, R. T., Tsai, R. T. H., and Hsu, W. L. (2008). BIOSMILE web search: a web application for annotating biomedical entities and relations. *Nucleic Acids Res.* 36, W390–W398. doi: 10.1093/nar/gkn319

Ding, R., Arighi, C. N., Lee, J. Y., Wu, C. H., and Vijay-Shanker, K. (2015). pGenN, a gene normalization tool for plant genes and proteins in scientific literature. *PLoS ONE* 10:e0135305. doi: 10.1371/journal.pone.0135305

Evangelou, E., Trikalinos, T. A., and Ioannidis, J. P. (2005). Unavailability of online supplementary scientific information from articles published in major journals. *FASEB J.* 19, 1943–1944. doi: 10.1096/fj.05-4784lsf

Fink, J. L., Fernicola, P., Chandran, R., Parastatidis, S., Wade, A., Naim, O., et al. (2010). Word add-in for ontology recognition: semantic enrichment of scientific literature. *BMC Bioinformatics* 11:103. doi: 10.1186/1471-2105-11-103

Fleuren, W. W., and Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods* 74, 97–106. doi: 10.1016/j.ymeth.2015.01.015

Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., et al. (2008). CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.* 36, W406–W410. doi: 10.1093/nar/gkn215

He, Y., and Xiang, Z. (2013). HINO: a BFO-aligned ontology representing human molecular interactions and pathways. arXiv preprint arXiv:1311.3355.

Hermjakob, H., Montecchi−Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–D455. doi: 10.1093/nar/gkh052

Hoffmann, R., and Valencia, A. (2004). A gene network for navigating the literature. *Nat. Genet.* 36, 664–664. doi: 10.1038/ng0704-664

Hu, Z. Z., Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K., and Wu, C. H. (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* 21, 2759–2765. doi: 10.1093/bioinformatics/bti390

Huang, C. C., and Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.* 17, 132–144. doi: 10.1093/bib/bbv024

Hunter, L., and Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Mol. Cell* 21, 589–594. doi: 10.1016/j.molcel.2006.02.012

Hur, J., Özgür, A., Xiang, Z., and He, Y. (2015). Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions. *J. Biomed. Semantics* 6, 1. doi: 10.1186/2041-1480-6-2

Hur, J., Schuyler, A. D., and Feldman, E. L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics* 25, 838–840. doi: 10.1093/bioinformatics/btp049

Hur, J., Xiang, Z., Feldman, E. L., and He, Y. (2011). Ontology-based Brucella vaccine literature indexing and systematic analysis of gene-vaccine association network. *BMC Immunol.*12:49. doi: 10.1186/1471-2172-12-49

Iragne, F., Barré, A., Goffard, N., and De Daruvar, A. (2004). AliasServer: a web server to handle multiple aliases used to refer to proteins. *Bioinformatics* 20, 2331–2332. doi:10.1093/bioinformatics/bth241

Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G., and Kors, J. A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.* 9:R96. doi: 10.1186/gb-2008-9-6-r96

Kang, N., Afzal, Z., Singh, B., Van Mulligen, E. M., and Kors, J. A. (2012). Using an ensemble system to improve concept extraction from clinical records. *J. Biomed. Inform.* 45, 423–428. doi: 10.1016/j.jbi.2011.12.009

Krallinger, M., Valencia, A., and Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 9:S8. doi: 10.1186/gb-2008-9-s2-s8

Leaman, R., and Gonzalez, G. (2008). "BANNER: an executable survey of advances in biomedical named entity recognition," in *Pacific Symposium on Biocomputing, Vol. 13* (Kona, HI), 652–663.

Lee, H., Yi, G. S., and Park, J. C. (2008). E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucleic Acids Res.* 36, W416–W422. doi: 10.1093/nar/gkn286

Leitner, F., and Valencia, A. (2008). A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Lett.* 582, 1178–1181. doi: 10.1016/j.febslet.2008.02.072

Li, Y., Lin, H., and Yang, Z. (2009). Incorporating rich background knowledge for gene named entity classification and recognition. *BMC Bioinformatics* 10:223. doi: 10.1186/1471-2105-10-223

Mandloi, S., and Chakrabarti, S. (2015). PALM-IST: pathway assembly from literature mining-an information search tool. *Sci. Rep.* 5:10021. doi: 10.1038/srep10021

Mazumder, R., Natale, D. A., Julio, J. A., Yeh, L. S., and Wu, C. H. (2010). Community annotation in biology. *Biol. Direct*. 5:12. doi: 10.1186/1745-6150-5-12

Mower, A., and Youngkin, M. E. (2008). Expanding access to published research: open access and self-archiving. *J. Neuro Ophthalmol.* 28, 69–71. doi: 10.1097/WNO.0b013e318167730b

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.* 18, 544–551. doi: 10.1136/amiajnl-2011-000464

Novichkova, S., Egorov, S., and Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* 19, 1699–1706. doi: 10.1093/bioinformatics/btg207

Papanikolaou, N., Pavlopoulos, G. A., Pafilis, E., Theodosiou, T., Schneider, R., Satagopam, V. P.,et al. (2014). BioTextQuest+: a knowledge integration platform for literature mining and concept discovery. *Bioinformatics* 30, 3249–3256. doi: 10.1093/bioinformatics/btu524

Prieto, C., and De Las Rivas, J. (2006). APID: agile protein interaction dataanalyzer. *Nucleic Acids Res.* 34, W298–W302. doi: 10.1093/nar/gkl128

Raja, K., Subramani, S., and Natarajan, J. (2012). "PPInterFinder–a web server for mining human protein-protein interactions," in *2012 BioCreative Workshop* (Washington, DC), 151–163.

Raja, K., Subramani, S., and Natarajan, J. (2014). A hybrid named entity tagger for tagging human proteins/genes. *Int. J. Data Min. Bioinform.* 10, 315–328. doi: 10.1504/IJDMB.2014.064545

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., and Stoehr, P. (2007). EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics* 23, e237–e244. doi: 10.1093/bioinformatics/btl302

Robinson, P. N., and Bauer, S. (2011). *Introduction to Bio-Ontologies*. Boca Raton, FL: CRC Press.

Sainani, K. (2008). Mining biomedical literature: using computers to extract knowledge nuggets. *Biomed. Comput. Rev.* 4, 16–27.

Salgado, D., Krallinger, M., Depaule, M., Drula, E., Tendulkar, A. V., Leitner, F., et al. (2012). MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinformatics* 28, 2285–2287. doi: 10.1093/bioinformatics/bts435

Seringhaus, M. R., and Gerstein, M. B. (2007). Publishing perishing*?* Towards tomorrow's information architecture. *BMC Bioinformatics* 8:17. doi: 10.1186/1471-2105-8-17

Shotton, D., Portwin, K., Klyne, G., and Miles, A. (2009). Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput. Biol.* 5:e1000361. doi: 10.1371/journal.pcbi.1000361

Singh, A., Singh, M., Singh, A. K., Singh, D., Singh, P., and Sharma, A. (2011). "Free Full Text Articles": where to Search for Them?. *Int. J. Trichol.* 3, 75. doi: 10.4103/0974-7753.90803

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539. doi: 10.1093/nar/gkj109

Stevens, R., Goble, C., Horrocks, I., and Bechhofer, S. (2002). OILing the way to machine understandable bioinformatics resources. *IEEE Trans. Inf. Technol. Biomed.* 6, 129–134. doi: 10.1109/TITB.2002.1006300

Superti-Furga, G., Wieland, F., and Cesareni, G. (2008). Finally: The digital, democratic age of scientific abstracts. *FEBS Lett.* 582, 1169. doi: 10.1016/j.febslet.2008.02.070

Tsuruoka, Y., Tsujii, J. I., and Ananiadou, S. (2008). FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24, 2559–2560. doi: 10.1093/bioinformatics/btn469

Usie, A., Karathia, H., Teixidó, I., Alves, R., and Solsona, F. (2014). BiblioMetReS for user-friendly mining of genes and biological processes in scientific documents. *Peer J.* 2:e276. doi: 10.7717/peerj.276

Vedantam, G., and Viswanathan, V. K. (2012). Naming names: eponyms and biological history. *Gut Microbes* 3, 173–175. doi: 10.4161/gmic.20454

Winnenburg, R., Wächter, T., Plake, C., Doms, A., and Schroeder, M. (2008). Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?. *Brief Bioinformatics* 9, 466–478. doi: 10.1093/bib/bbn043

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291. doi: 10.1093/nar/28.1.289

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). MINT: a Molecular INTeraction database. *FEBS Lett.* 513, 135–140. doi: 10.1093/nar/gkp983

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.