In: Internet Policies and Issues, Volume 7
Editor: B.G. Kutais
ISBN: 978-1-61668-745-8

Chapter 11

BIOMEDICAL LITERATURE ANALYSIS: CURRENT STATE AND CHALLENGES

Maurice HT Ling[1, 2], Christophe Lefevre[3] and Kevin R. Nicholas[2, 3]
[1]School of Chemical and Life Sciences, Singapore Polytechnic, Singapore
[2]Department of Zoology, The University of Melbourne, Australia
[3]Institute for Technology Research and Innovation, Deakin University, Australia

ABSTRACT

Advances in molecular biology tools and techniques from the end of the last century had shifted the focus of biomedical research from the study of individual proteins and genes to the interactions within an entire biological systems. At the same time, advanced tools generates large sets of experimental data which required collaborations of groups of biologists to decipher. This resulted in a need to have a diverse research knowledge. However, the amount of published research information in the form of published articles is increasing exponentially, making it difficult to maintain a productive edge. Biomedical literature analysis is seen as a means to manage the increased amount of information – to gather relevant articles and extract relevant information from these articles. We review the central (information retrieval, information extraction and text mining) and allied (corpus collection, databases and system evaluation methods) domains of computational biomedical literature analysis to present the current state of biomedical literature analysis for protein-protein and protein-gene interactions and the challenges ahead.

ACKNOWLEDGEMENT

We wish to thank Professor Thomas Rindflesch, National Institute of Health, USA; Professor Jonathan Wren, Associate Editor for Bioinformatics, for his comments on improving the initial drafts.

REFERENCES

[1] Hunter, L. & Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular Cell*, *21*, 589-594.

[2] Cohen, A. M. & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*, 57-71.

[3] He, M., Wang, Y. & Li, W. (2009). PPI finder: a mining tool for human protein-protein interactions. *PLoS ONE, 4*, e4554.

[4] Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*, 7-18.

[5] Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, *31*, 526-557.

[6] Prange, J. D. (1996). *Evaluation driven research: the foundation of the TIPSTER text program.* Tipster Text Program Phase II, May 6-8, 1996.

[7] Hersh, W., Bhupatiraju, R. T. & Corley, S. (2004). Enhancing access to the Bibliome: the TREC Genomics Track. *Medinfo*, *11*, 773-777.

[8] Hirschman L. (1998). The evolution of evaluation: lessons from the Message Understanding Conferences. *Information Processing and Management, 37*, 383-402.

[9] Leek TR. Information extraction using Hidden Markov Model. *Department of Computer Science*. University of California, San Diego (1997)..

[10] Fukuda K., Tsunoda T., Tamura A., Takagi T. (1998). Toward information extraction: identifying protein names from biological papers. *Proceedings of the Pacific Symposium on Biocomputing* (PSB'98): 705 - 716.

[11] Craven M., Kumlien J. (1999). Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol,* 77-86.

[12] Blaschke C., Andrade, M. A., Ouzounis, C. & Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, 60-67.

[13] Shatkay H. & Wilbur, W. J. (2000). *Finding themes in Medline documents: probabilistic similarity search.* IEEE Conference on Advances in Digital Libraries., pp. 183-192.

[14] Friedman C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*., *17*, S74-S82.

[15] Leser, U. & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings on Bioinformatics.*, *6*, 357-369.

[16] Natarajan, J., Berrar, D., Hack, C. J. & Dubitzky, W. (2005). Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology*, *25*, 31-52.

[17] Tsai, R. T., Wu, S. H., Chou, W. C., Lin, Y. C., He, D. & Hsiang, J. et al. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, *7*, 92.

[18] Han, B., Obradovic, Z., Hu, Z. Z., Wu, C. H. & Vucetic S. (2006). Substring selection for biomedical document classification. *Bioinformatics*.

[19] Chen, D., Muller, H. M. & Sternberg, P. W. (2006). Automatic document classification of biological literature. BMC *Bioinformatics*, *7*, 370.

[20] Gerard, S., Edward, A. F. & Harry, W. (1983). Extended Boolean information retrieval. Communications of the *ACM*, *26*, 1022-1036.

[21] Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G. & Nelson, S. J. et al. (2000). The NLM Indexing Initiative. *Proc AMIA Symp,* 17-21.

[22] Wilbur, W. J. & Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med*, *26*, 209-222.

[23] Wilbur, W. J. (2002). A thematic analysis of the AIDS literature. *Pacific Symposium on Biocomputing*., *7*, 386-397.

[24] Muller, H. M., Kenny, E. E. & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*., *2*, e309.

[25] Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L. & Weinstein, J. N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*., *27*, 1210-1214, 1216-1217.

[26] Shatkay, H., Pan, F., Rzhetsky, A. & Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, *24*, 2086-2093.

[27] Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M. & Andrade-Navarro, M. A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res*., *37*, W141-146.

[28] Doms, A. & Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, *33*, W783-786.

[29] Chiang, J. H., Shin, J. W., Liu, H. H. & Chin, C. L. (2006). GeneLibrarian: an effective gene-information summarization and visualization system. *BMC Bioinformatics*, *7*, 392.

[30] Simon, M. L., Patrick, M., Kimberly, F. J. & Jennifer, S. (2004). MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics*, *20*, 3659.

[31] Yoo, I., Hu, X. & Song, I. Y. (2007). Biomedical ontology improves biomedical literature clustering performance: a comparison study. International Journal of *Bioinformatics Research and Applications*., *3*, 414-428.

[32] Ding, J., Viswanathan, K., Berleant, D., Hughes, L., Wurtele, E. S. & Ashlock, D. et al. (2005). Using the biological taxonomy to access biological literature with PathBinderH. *Bioinformatics*., *21*, 2560-2562.

[33] Baker, C. J., Kanagasabai, R., Ang, W. T., Veeramani, A., Low, H. S. & Wenk, M. R. (2008). Towards ontology-driven navigation of the lipid bibliosphere. *BMC Bioinformatics*, *9,* Suppl *1*, S5.

[34] Aronson, A. R. & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. Proc AMIA *Annu Fall Symp*, 485-489.

[35] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21.

[36] Hersh, W., Price, S. & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proc AMIA Symp*, 344-348.

[37] Witte, R., Kappler, T. & Baker, C. J. (2007). Enhanced semantic access to the protein engineering literature using ontologies populated by text mining. International Journal of Bioinformatics *Research and Applications*., *3*, 389-413.

[38] Jensen, L. J., Saric, J. & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Review Genetics*., *7*, 119-129.

[39]Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*., *21*, 543-565.

[40]Garten, Y. & Altman, R. B. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*., *10*, Suppl *2*, S6.

[41]Smith, L., Rindflesch, T. & Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*., *20*, 2320-2321.

[42]Kudo, T., Matsumoto, Y. (2000). Use of support vector learning for chunk identification. *4th Conference on CoNLL-2000 and LLL*-142-144.

[43]Liu, H. & Friedman, C. (2003). Mining terminological knowledge in large biomedical corpora. *Pacific Symposium on Biocomputing*., *8*, 415-426.

[44]Chang, J. T. (2003). *Using machine learning to extract drug and gene relationships from text.* pp. 183. Stanford University.

[45]Chang, J. T., Schutze, H. & Altman, R. B. (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*., *9*, 612-620.

[46]Yu, H., Hripcsak, G. & Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*., *9*, 262-272.

[47]Schwartz, A. S. & Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing*., *8*, 451-462.

[48]Adar, E. (2004). SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics*., *20*, 527-533.

[49]Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M. & Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo*., *10*, 371-375.

[50]Wren, J. D. & Garner, H. R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine*., *41*, 426-434.

[51]Wren, J. D ., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R. & Altman, R. B. (2005). Biomedical term mapping databases. *Nucleic Acids Research*., *33*, D289-293.

[52]Okazaki, N. & Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*., *22*, 3089-3095.

[53]Sohn, S., Comeau, D., Kim, W. & Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*., *9*, 402.

[54]Xu, Y., Wang, Z., Lei, Y., Zhao, Y. & Xue, Y. (2009). MBA: a literature mining system for extracting biomedical abbreviations. *BMC Bioinformatics*., *10*, 14.

[55]Kuo, C. J., Ling, M. H. T., Lin, K. T. & Hsu, C. N. (2009). *BIOADI: A machine learning approach to identifying abbreviations and definitions in biological literature.* 9th International Conference on Bioinformatics., Singapore.

[56]Torii, M., Hu, Z. Z., Song, M., Wu, C. H. & Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*., *8* Suppl *9*, S5.

[57]Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P. & Coster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*., *67*, 49-61.

[58] Hanisch, D., Fluck, J., Mevissen, H. T. & Zimmer, R. (2003). Playing biology's name game: identifying protein names in scientific text. *Pacific Symposium on Biocomputing*., 403-414.

[59] Krauthammer, M. & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Medical Bioinformatics*., *37*, 512-526.

[60] Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V. & Jacq, B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome informatics Workshop on Genome Informatics*., *9*, 72-80.

[61] Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R. & Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*., 9 Suppl *3*, S3.

[62] Nenadic, G., Spasic, I. & Ananiadou, S. (2004). Mining biomedical abstracts: What is in a term? In: K.Y. Su, (ed), *Natural Language Processing - IJCNLP 2004*. Springer., Berlin, 797-806.

[63] Jacquemn, C. (2001). *Spotting and discovering terms through natural language processing,* MIT Press, Cambridge, MA.

[64] Liu, H., Hu, Z. Z., Torii, M., Wu, C. & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. *J Am Med Inform Assoc., 13*, 497-507.

[65] Tsuruoka, Y. & Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform*., *37*, 461-470.

[66] Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene*., *259*, 245-252.

[67] Tanabe, L. & Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*., *18*, 1124-1132.

[68] Chang, J. T., Schutze, H. & Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*., *20*, 216-225.

[69] Egorov, S., Yuryev, A. & Daraselia, N. (2004). A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc*., *11*, 174-178.

[70] Hanisch, D., Fundel, K., Mevissen, H. T ., Zimmer, R. & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*., *6*, Suppl *1*, S14.

[71] Ryan, T. M., Winters, R. S., Mark, M., Yang, J., Peter, S. W. & Fernando, P. (2004). An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*., *20*, 3249.

[72] McDonald, R. & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*., 6 Suppl *1*, S6.

[73] Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*., *21*, 3191-3192.

[74] Zhou, G., Shen, D., Zhang, J., Su, J. & Tan, S. (2005). Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*., *6*, Suppl *1*, S7.

[75] Hatzivassiloglou, V., Duboue, P. A. & Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*., *17*, Suppl *1*, S97-106.

[76] Hou, W. J. & Chen, H. H. (2004). Enhancing performance of protein and gene name recognizers with filtering and integration strategies. *Journal of Biomedical Informatics*., *37*, 448-460.

[77]Majoros, W., Subramanian, G. & Yandell, M. (2003). Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics*., *19*, 402-407.

[78]Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B. & Grover, C. (2005). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, *6*, Suppl *1*, S5.

[79]Li, L., Zhou, R. & Huang, D. (2009). Two-phase biomedical named entity recognition using CRFs. *Comput Biol Chem*., *33*, 334-338.

[80]Kou, Z., Cohen, W. W. & Murphy, R. F. (2005). High-recall protein entity recognition using a dictionary. Bioinformatics., 21 Suppl *1*, i266-273.

[81]Fundel, K., Guttler, D., Zimmer, R. & Apostolakis, J. (2005). A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*., 6 Suppl *1*, S15.

[82]de Bruijn B., Martin J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. *International Journal of Medical Informatics*., *67*, 7-18.

[83]Gaizauskas, R., Demetriou, G., Artymiuk, P. J. & Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*., *19*, 135-143.

[84]Daniel, M. M., Hsinchun, C., Hua, S., Byron, B. M. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*., *20*, 3370.

[85]Crim, J., McDonald, R. & Pereira, F. (2005). Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*., *6*, Suppl *1*, S13.

[86]Hakenberg, J., Plake C., Royer L., Strobelt H., Leser U., Schroeder M. (2008). Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*., *9*, Suppl *2*, S14.

[87]Huang, M., Ding, S., Wang, H. & Zhu, X. (2008). Mining physical protein-protein interactions from the literature. *Genome Biology*., *9*, Suppl *2*, S12.

[88]Krallinger, M., Valencia, A. & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biolgy*., *9*, Suppl *2*, S8.

[89]Zhou, W., Torvik, V. I. & Smalheiser, N. R. (2006). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*., *22*, 2813-2818.

[90]Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C. & Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE.*, pe21.

[91]Skusa, A., Ruegg, A. & Kohler, J. (2005). Extraction of biological interaction networks from scientific literature. *Briefings in Bioinformatics*., *6*, 263-276.

[92]Cohen, K. B. & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology*., *4*, e20.

[93]Stapley, B. J. & Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pacific Symposium on Biocomputing*., *5*, 526-537.

[94]Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V. & Tuekam, B. et al. (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*., *4*, 11.

[95]Hoffmann, R. & Valencia, A. (2004). A gene network for navigating the literature. *Nature Genetics*., *36*, 664.

[96] Stephens, M., aplakal, M., Mukhopadhyay, S., Raje, R. & Mostafa, J. (2001). Detecting gene relations from MEDLINE abstracts. *Pacific Symposium on Biocomputing*., *6*, 483-496.

[97] Cooper, J. W. & Kershenbaum, A. (2005). Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics*., *6*, 143.

[98] Ray, S. & Craven, M. (2005). Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*., *6*, Suppl *1*, S18.

[99] Jelier, R., Jenster, G., Dorssers, L. C., van der Eijk, C. C., van Mulligen, E. M. & Mons, B. et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*., *21*, 2049-2058.

[100] Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*., *28*, 21-28.

[101] Alako, B. T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S. & Rullmann, T. et al. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. BMC *Bioinformatics*., *6*, 51.

[102] Becker, K. G., Hosack, D. A., Dennis, G., Jr.., Lempicki, R. A., Bright, T. J. & Cheadle, C., et al. (2003). PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*., *4*, 61.

[103] Fang, Y. C., Huang, H. C. & Juan, H. F. (2008). MeInfoText: associated gene methylation and cancer information from text mining. *BMC Bioinformatics*, *9*, 22.

[104] Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques.,* Morgan Kaufmann.

[105] Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A. & Wilbur, W. J. (2002). *Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. AMIA Symp,* 2002., 919-923.

[106] Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K. & Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*., *20*, 3604-3612.

[107] Yu, H. & Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*., *19*, Suppl *1*, i340-349.

[108] Florence, H., Anthony, L. L. & Fred, E. C. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*., *20*, 557.

[109] Divoli, A. & Attwood, T. K. (2005). BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics*., *21*, 2138-2139.

[110] Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*., *19*, 313-330.

[111] Dernatas, E. & Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Computational Linguistics., 21*, 137-163.

[112] Kupiec, J. (1992). Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language*., *6*.

[113] Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. (2005). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*.

[114] Temkin, J. M. & Gilder, M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*., *19*, 2046-2053.

[115] Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P. & Morris, M. et al. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*., *37*, 43-53.

[116] Li, Q. & Wu, Y. F. (2006). Identifying important concepts from medical documents. *J Biomed Inform*., *39*, 668-679.

[117] Crystal, D. (1997). *The Cambridge Encyclopedia of Languages (2nd ed.)*, Cambridge University Press, Cambridge.

[118] Santos, C., Eggle, D. & States, D. J. (2005). Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics*., *21*, 1653-1658.

[119] Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. & Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*., *43*, 516-533.

[120] Rindflesch, T. C., Rajan, J. & Lawrence Hunter, L. (2000). *Extracting molecular binding relationships from biomedical text.* 6th Applied Natural Language Processing Conference, 188-195.

[121] Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C. & Konstandi, O. et al. (2007). Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine*., *39*, 127-136.

[122] Friedman, C. (2000). A Broad Coverage Natural Language Processing System. *American Medical Informatics Association Symposium*, 270 - 274.

[123] Novichkova, S., Egorov, S. & Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*., *19*, 1699-1706.

[124] Allen, J. (1994). *Natural language understanding,* Benjamin-Cummings Publishing Company, New York.

[125] Sells, P. (1984). *Lectures on contemporary syntactic theories,* C S L I Publications.

[126] Chiang, J. H. & Yu, H. C. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*., *19*, 1417-1422.

[127] Kim, J. D., Ohta, T., Tateisi, Y. & Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*, i180-i182.

[128] David, P. A. C., Bernard, F. B., William, B. L. & David, T. J. (2004). BioRAT: extracting biological information from full-length papers. *Bioinformatics*, *20*, 3206.

[129] Cunningham, H. (2000). *Software Architecture for Language Engineering.* Department of Computer Science. pp. 244. University of Sheffield.

[130] Chen, H. & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*., *5*, 147.

[131] Brants, T. (2000). *TnT - a statistical part-of-speech tagger.* 6th Applied Natural Language Processing Conference.

[132] Chiang, J. H., Yu, H. C., Hsu, H. J. (2004). GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*., *20*, 120.

[133] Nasukawa, T. & Nagono, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal*., *40*, 967-984.

[134] Karopka, T., Scheel, T., Bansemer, S. & Glass, A. (2004). Automatic construction of gene relation networks using text mining and gene expression data. Medical Informatics and the Internet in *Medicine*., *29*, 169-183.

[135] Neff, M. S., Byrd, R. J. & Boguraev, B. K. (2004). The Talent system: TEXTRACT architecture and data model. *Natural Language Engineering*., *10*, 307-326.

[136] Cooper, J. & Byrd R. (1998). *Lexical navigation: visually prompted query refinement.* ACM Digital Libraries Conference.

[137] Jang, H., Lim, J., Lim, J. H., Park, S. J., Lee, K. C. & Park, S. H. (2006). Finding the evidence for protein-protein interactions from PubMed abstracts. *Bioinformatics*., *22*, e220-226.

[138] Malik, R., Franke, L. & Siebes, A. (2006). Combination of text-mining algorithms increases the performance. *Bioinformatics*., *22*, 2151-2157.

[139] Cussens, J. & Nédellec, C. (eds) (2005). P*roceedings of the 4th Learning Language in Logic Workshop,* (LLL05), Bonn.

[140] Mika, S. & Rost, B. (2004). NLProt: extracting protein names and sequences from papers. *Nucleic Acids Research*., *32*, W634-637.

[141] Horn, F., Lau, A. L. & Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*., *20*, 557-568.

[142] Schneider, G., Rinaldi, F. & Dowdall, J. (2004). *Fast., deep-linguistic statistical dependency parsing.* 20th International Conference on Computational Linguistics. Association of Computational Linguistics, University of Geneva, Switzerland.

[143] Reynar, J., Ratnaparkhi, A. (1997). *A maximum entrophy approach to identifying sentence boundaries.* Fifth Conference on Applied Natural Language Processing, Washington, DC: University of Pennsylvania.

[144] Ratnaparkhi, A. (1996). *A Maximum Entropy Model for Part-of-Speech Tagging.* Conference on Empirical Methods in Natural Language Processing., 133-142.

[145] Minnen, G., Carroll, J. & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering*., *7*, 207-223.

[146] Mikheev, A. (1997). Automatics rule induction for unknown word guessing. *Computational Linguistics*., *23*, 405-423.

[147] Feng, C., Yamashita, F. & Hashida, M. (2007). Automated extraction of information from the literature on chemical-CYP3A4 interactions. *Journal of Chemical Information and Modeling*., *47*, 2449-2455.

[148] Ling, M. H., Lefevre, C., Nicholas, K. R. & Lin, F. (2007). *Re-construction of Protein-Protein Interaction Pathways by Mining Subject-Verb-Objects Intermediates.,* Second IAPR Workshop on Pattern Recognition in Bioinformatics (PRIB 2007). Springer-Verlag, Singapore.

[149] Liu, H. & Lieberman, H. (2005). *Metafor: visualizing stories as code.* 10th International Conference on Intelligent User Interfaces.

[150] Ling, M. H. (2006). An Anthological Review of Research Utilizing MontyLingua, a Python-Based End-to-End Text Processor. *The Python Papers*, *1*, 5-12.

[151] Chen, L. (2006). *Automatic construction of domain-specific concept structures.* Technischen Universitat Darmstadt.

[152] van Eck, N. J. & van den Berg, J. (2005). *A novel algorithm for visualizing concept associations.* 16th International Workshop on Database and Expert System Applications, (DEXA'05).

[153] Lee, H., Yi, G. S. & Park, J. C. (2008). E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucleic Acids Research*., *36*, W416-422.

[154] Kim, S., Shin, S. Y., Lee, I. H., Kim, S. J., Sriram, R. & Zhang, B. T. (2008). PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Research, 36*, W411-415.

[155] Plake, C., Hakenberg, J. & Leser, U. Optimizing syntax patterns for discovering protein-protein interactions. *ACM Symposium on Applied Computing*., 187-192. ACM Press (2005).

[156] Barnickel, T., Weston, J., Collobert, R., Mewes, H. W. & Stumpflen, V. (2009) Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE*., *4*, e6393.

[157] Jiao, D. & Wild, D. J. (2009). Extraction of CYP chemical interactions from biomedical literature using natural language processing methods. *J Chem Inf Model*, *49*, 263-269.

[158] National Library of Medicine. (2003). UMLS Knowledge Sources (14th ed.).

[159] Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*., *1*, 161-174.

[160] Sleator, D. & Temperley, D. (1991). *Parsing English with a Link Grammar.* Third International Workshop on Parsing Technologies.

[161] Hao, Y., Zhu, X., Huang, M. & Li, M. (2005). Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*., *21*, 3294-3300.

[162] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M. & Foglierini, M. et al. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*., *33*, D433-437.

[163] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. & Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics., *20*, 604-611.

[164] Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. *Pacific Symposium on Biocomputing*., *6*, 408-419.

[165] Chen, E. S., Hripcsak, G., Xu, H., Markatou, M. & Friedman, C. (2008). Automated Acquisition of Disease Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *Journal of the American Medical Informatics Association*., *15*, 87-98.

[166] Osborne, J. D., Lin, S., Zhu, L. & Kibbe, W. A. (2007). Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS). *Methods in Molecular Biology*., *408*, 153-169.

[167] Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B. & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*., *33*, 1544-1552.

[168] Aerts, S., Haeussler, M., van Vooren, S., Griffith, O. L., Hulpiau, P. & Jones, S. J. et al. (2008). Text-mining assisted regulatory annotation. *Genome Biology*., *9*, R31.

[169] Hu, Z. Z., Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K. & Wu, C. H. (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*., *21*, 2759-2765.

[170] Shah, P. K., Jensen, L. J., Boue, S. & Bork, P. (2005). Extraction of transcript diversity from scientific literature. *PLoS Computational Biology*., *1*, e10.

[171] Yang, H., Nenadic, G. & Keane, J. A. (2008). Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics*., 9 Suppl *3*, S11.

[172] Xu, H., Anderson, K., Grann, V. & Friedman, C. (2004). Facilitating cancer research using natural languge procesiong of pathological reports. 11th World Congress on *Medical Informatics*., 565-569.

[173] Yuryev, A., Mulyukov, Z., Kotelnikova, E., Maslov, S., Egorov, S. & Nikitin, A. et al. (2006). Automatic pathway building in biological association networks. *BMC Bioinformatics*., *7*, 171.

[174] Matsuzawa, H. & Fukuda, T. (2000). *Mining structured association patterns from database*. 4th Pacific and Asia International Conference on Knowledge Discovery and Data Mining (PAKDD-2000)., 233-244.

[175] Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T. & Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics., 25*, 394-400.

[176] Rodriguez-Esteban, R., Iossifov, I. & Rzhetsky, A. (2006). Imitating Manual Curation of Text-Mined Facts in Biomedicine. *PLoS Comput Biol., 2*.

[177] Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E. & Matthews, M. et al. (2008). Assisted curation: does text mining really help? *Pac Symp Biocomput*, 556-567.

[178] Swanson, D. R. (1990). Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*., *33*, 157-186.

[179] Swanson, D. R. (1990). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association*., *78*, 29-37.

[180] Weeber, M., Kors, J. A. & Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*., *6*, 277-286.

[181] Srinivasan, P., Libbus, B. & Sehgal, A. K. (2004). Mining MEDLINE: postulating a beneficial role for Curcumin Longa in retinal diseases. BioLink Linking Biological Literature, *Ontologies, and Databases*., 33-40.

[182] Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, *3*, 2.

[183] Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G. & Kors, J. A. (2008). Anni 2.*0*, a multipurpose text-mining tool for the life sciences. *Genome Biology*., *9*, R96.

[184] Smalheiser, N. R., Torvik, V. I. & Zhou, W. (2009). Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine*., *94*, 190-197.

[185] Sarkar, I. N. & Agrawal, A. (2006). Literature based discovery of gene clusters using phylogenetic methods. *AMIA Annu Symp Proc*: 689-693.

[186] Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*., *74*, 289-298.

[187] Yetisgen-Yildiz, M. & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform*., *39*, 600-611.

[188] Colosimo, M. E., Morgan, A. A., Yeh, A. S., Colombe, J. B. & Hirschman, L. (2005). Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*., *6*, Suppl *1*, S12.

[189] Wilbur, W. J., Rzhetsky, A. & Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*., *7*, 356.

[190] Tanabe, L., Xie, N., Thom, L. H., Matten, W. & Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*., *6*, Suppl *1*, S3.

[191] Pustejovsky, J., Castaño, J., Saurí, R., Rumshisky, A., Zhang, J. & Luo, W. (2002). *Medstract: Creating Large-scale Information Servers for Biomedical Libraries.* ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain.

[192] Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C. & Ohta, T. et al. (1999). *The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers.* Ninth Conference of the European Chapter of the Association for Computational Linguistics.

[193] Ohta, T., Tateisi, Y., Mima, H. & Tsujii, J. (2002). *The GENIA corpus: an annotated research abstract corpus in molecular biology domain.* Human Language Technology Conference.

[194] Cussens, J. & Dzeroski, S. (eds) (2000). *Learning Languages in Logic.* Springer, Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

[195] Vincze, V., Szarvas, G., Farkas, R., Mora, G. & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*, Suppl *11*, S9.

[196] Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F. & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, *9*, Suppl *3*, S6.

[197] Zhang, Z., Tang, S. & Ng, S. K. (2005). Towards discovering disease-specific gene networks from online literature. *Advances in Bioinformatics and Computational Biology.* 3rd Asia-Pacific Bioinformatics Conference, 161-169.

[198] Chatr-aryamontri, A., Kerrien, S., Khadake, J., Orchard, S., Ceol, A. & Licata, L. et al. (2008). MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biology., 9* Suppl *2*, S5.

[199] Lee, L. C., Horn, F. & Cohen, F. E. (2007). Automatic Extraction of Protein Point Mutations Using a Graph Bigram Association. *PLoS Comput Biol*, *3*, e16.

[200] Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C. E. & Rocha, L. M. (2006). Large-scale testing of bibliome informatics using Pfam protein families. *Pac Symp Biocomput*: 76-87.

[201] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, *32*, D449-451.

[202] Miotto, O., Tan, T. W. & Brusic, V. (2005). Supporting the curation of biological databases with reusable text mining. *Genome Inform*, *16*, 32-44.

[203] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M. & Bantoft, K. et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*., *33*, D418-424.

[204] Zhou, D. & He, Y. (2008). Extracting interactions between proteins from the literature. *Journal of Biomedical Informatics*., *41*, 393-407.

[205] Scott, M., Lu, G., Hallett, M. & Thomas, D. Y. (2004). The Hera database and its use in the characterization of endoplasmic reticulum proteins. *Bioinformatics*., *20*, 937-944.

[206] Basu, S., Bremer, E., Zhou, C. & Bogenhagen, D. F. (2006). MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation. *Bioinformatics*., *22*, 485-492.

[207] Martinez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L. & Tobes, R. (2004). BacTregulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics*., *20*, 2787-2791.

[208] Percudani, R. & Peracchi, A. (2009). The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics*., *10*, 273.

[209] Blaineau, S. V. & Aouacheria, A. (2009). BCL2DB: moving 'helix-bundled' BCL-2 family members to their database. *Apoptosis*., *14*, 923-925.

[210] Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle, N. M., Murray-Rust, P. & Thornton, J. M. et al. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*., *21*, 4315-4316.

[211] Mao, C., Qiu, J., Wang, C., Charles, T. C. & Sobral, B. W. (2005). NodMutDB: a database for genes and mutants involved in symbiosis. *Bioinformatics., 21*, 2927-2929.

[212] Wood, D. L., Miljenovic, T., Cai, S., Raven, R. J., Kaas, Q. & Escoubas, P. et al. (2009). ArachnoServer: a database of protein toxins from spiders. *BMC Genomics*., *10*, 375.

[213] Testa, O. D., Moutevelis, E. & Woolfson, D. N. (2009). CC+: a relational database of coiled-coil structures. *Nucleic Acids Res*., *37*, D315-322.

[214] Li, Y. & Chen, Z. (2008). RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol Lett*., *289*, 126-129.

[215] Jacobs, G. H., Chen, A., Stevens, S. G., Stockwell, P. A., Black, M. A. & Tate, W. P. et al. (2009). Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res*., *37*, D72-76.

[216] Jayakanthan, M., Muthukumaran, J., Chandrasekar, S., Chawla, K., Punetha, A. & Sundar, D. (2009). ZifBASE: a database of zinc finger proteins and associated resources. *BMC Genomics*., *10*, 421.

[217] Kim, C., Kwon, S., Lee, G., Lee, H., Choi, J. & Kim, Y. et al. (2009). A database for allergenic proteins and tools for allergenicity prediction. *Bioinformation*., *3*, 344-345.

[218] Gao, J., Agrawal, G. K., Thelen, J. J. & Xu, D. (2009). P3DB: a plant protein phosphorylation database. *Nucleic Acids Res*., *37*, D960-962.

[219] Encinar, J. A., Fernandez-Ballester, G., Sanchez, I. E., Hurtado-Gomez, E., Stricher, F. & Beltrao, P. et al. (2009). ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*., *25*, 2418-2424.

[220] Magkrioti, C. K., Spyropoulos, I. C., Iconomidou, V. A., Willis, J. H. & Hamodrakas, S. J. (2004). cuticleDB: a relational database of Arthropod cuticular proteins. *BMC Bioinformatics*., *5*, 138.

[221] George, R. A., Spriggs, R. V., Thornton, J. M., Al-Lazikani, B. & Swindells, M. B. (2004). SCOPEC: a database of protein catalytic domains. *Bioinformatics., 20,* Suppl *1*, I130-I136.

[222] Li, B. & Gallin, W. J. (2004). VKCDB: voltage-gated potassium channel database. *BMC Bioinformatics*., *5*, 3.

[223] Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., et al. (2005). DisProt: a database of protein disorder. *Bioinformatics*., *21*, 137-140.

[224] Guo, A., He, K., Liu, D., Bai, S., Gu, X. & Wei, L. et al. (2005). DATF: a database of Arabidopsis transcription factors. *Bioinformatics*., *21*, 2568-2569.

[225] Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W. & Zheng, W. et al. (2006). DRTF: a database of rice transcription factors. *Bioinformatics*., *22*, 1286-1287.

[226] Tung, M. & Gallagher, D. T. (2009). The Biomolecular Crystallization Database Version *4*, expanded content and new features. *Acta Crystallogr D Biol Crystallogr*., *65*, 18-23.

[227] Sun, Q., Zybailov, B., Majeran, W., Friso, G., Olinares, P. D. & van Wijk, K. J. (2009). PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res*., *37*, D969-974.

[228] Chandra, N. R., Kumar, N., Jeyakani, J., Singh, D. D., Gowda, S. B. , &Prathima, M. N. (2006). Lectindb: a plant lectin database. *Glycobiology*., *16*, 938-946.

[229] Nogales-Cadenas, R., Abascal, F., Diez-Perez, J., Carazo, J. M. & Pascual-Montano, A. (2009). CentrosomeDB: a human centrosomal proteins database. *Nucleic Acids Res*., *37*, D175-180.

[230] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M. & Hannay, T. et al. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res*., *37*, D674-679.

[231] McDowall, M. D., Scott, M. S. & Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res*., *37*, D651-656.

[232] Zhao, X. M., Zhang, X. W., Tang, W. H. & Chen, L. (2009). FPPI: Fusarium graminearum Protein-Protein Interaction Database. *J Proteome Res.*

[233] Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S. & Sacco, F. et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res., 37*, D669-673.

[234] Chen, J. Y., Mamidipalli, S. & Huan, T. (2009). HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics*., 10 Suppl *1*, S16.

[235] Andres Leon, E., Ezkurdia, I., Garcia, B., Valencia, A. & Juan, D. (2009). EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res.*, *37*, D629-635.

[236] Lin, C. Y., Chen, C. L., Cho, C. S., Wang, L. M., Chang, C. M. & Chen, P. Y. et al. (2005). hp-DPI: Helicobacter pylori database of protein interactomes--embracing experimental and inferred interactions. *Bioinformatics*., *21*, 1288-1290.

[237] Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T. & Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics*., *24*, 1743-1744.

[238] Pawlicki, S., Le Bechec, A. & Delamarche, C. (2008). AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics*., *9*, 273.

[239] Theodoropoulou, M. C., Bagos, P. G., Spyropoulos, I. C., Hamodrakas, S. J. (2008). gpDB: a database of GPCRs., G-proteins, effectors and their interactions. *Bioinformatics*., *24*, 1471-1472.

[240] Chuan Tong, J., Meng Song, C., Thiam Joo Tan, P. & Chee Ren, E. A AS. (2008). BEID: Database for sequence-structure-function information on antigen-antibody interactions. *Bioinformation*., *3*, 58-60.

[241] Yimeng, D., Pierre-FranÁois, B., Gianluca, P., Yann, P., James, N. & Pierre, B. (2004). ICBS: a database of interactions between protein chains mediated by ≤-sheet formation. *Bioinformatics*., *20*, 2767.

[242] Beuming, T., Skrabanek, L., Niv, M. Y., Mukherjee, P. & Weinstein, H. (2005). PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics*., *21*, 827-828.

[243] Dou, Y., Baisnee, P. F., Pollastri, G., Pecout, Y., Nowick, J. & Baldi, P. (2004). ICBS: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics*., *20*, 2767-2777.

[244] Yang, C. Y., Chang, C. H., Yu, Y. L., Lin, T. C., Lee, S. A. & Yen, C. C., et al. (2008). PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*., *24*, i14-20.

[245] Song, S., Huang, Y., Wang, X., Wei, G., Qu, H., Wang, W. & et al. (2009). HRGD: a database for mining potential heterosis-related genes in plants. *Plant Mol Biol., 69*, 255-260.

[246] Richardson, C. J., Gao, Q., Mitsopoulous, C., Zvelebil, M., Pearl, L. H. & Pearl, F. M. (2009). MoKCa database--mutations of kinases in cancer. *Nucleic Acids Res*., *37*, D824-831.

[247] Sagar, S., Kaur, M., Dawe, A., Seshadri, S. V., Christoffels, A. & Schaefer, U. et al. (2008). DDESC: Dragon database for exploration of sodium channels in human. *BMC Genomics*., *9*, 622.

[248] Miranda-Saavedra, D., De, S., Trotter, M. W., Teichmann, S. A. & Gottgens, B. (2009). BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res*., *37*, D873-879.

[249] Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. (2009). DOOR: a database for prokaryotic operons. *Nucleic Acids Res*., *37*, D459-463.

[250] Liu, B. & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res*., *37*, D443-447.

[251] Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M. & Mattick, J. S. (2009). NRED: a database of long noncoding RNA expression. *Nucleic Acids Res., 37*, D122-126.

[252] Essack, M., Radovanovic, A., Schaefer, U., Schmeier, S., Seshadri, S. V. & Christoffels, A. et al. (2009). DDEC: Dragon database of genes implicated in esophageal cancer. *BMC Cancer*., *9*, 219.

[253] Kim, C. K., Kim, J. S., Lee, G. S., Park, B. S. & Hahn, J. H. (2008). PlantGM: a database for genetic markers in rice (Oryza sativa) and Chinese cabbage (Brassica rapa). *Bioinformation*., *3*, 61-62.

[254] Ding, G., Lorenz, P., Kreutzer, M., Li, Y. & Thiesen, H. J. (2009). SysZNF: the C2H2 zinc finger gene database. *Nucleic Acids Res*., *37*, D267-273.

[255] Boby, T., Patch, A. M. & Aves, S. J. (2005). TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*., *21*, 811-816.

[256] Sakharkar, M. K. & Kangueane, P. (2004). Genome SEGE: a database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics*., *5*, 67.

[257] Brockman, J. M., Singh, P., Liu, D., Quinlan, S., Salisbury, J. & Graber, J. H. (2005). PACdb: PolyA Cleavage Site and 3'-UTR Database. *Bioinformatics*., *21*, 3691-3693.

[258] Shimada, M. K., Matsumoto, R., Hayakawa, Y., Sanbonmatsu, R., Gough, C. & Yamaguchi-Kabata, Y. et al. (2009). VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res*., *37*, D810-815.

[259] Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K. & Go, M. (2009). AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res*., *37*, D305-309.

[260] Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J. J. & Nardone, F. et al. (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics., 93*, 213-220.

[261] Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. (2008). FstSNP-HapMap*3*, a database of SNPs with high population differentiation for HapMap3. *Bioinformation*., *3*, 139-141.

[262] Ackermann, A. A., Carmona, S. J. & Aguero, F. (2009). TcSNP: a database of genetic variation in Trypanosoma cruzi. *Nucleic Acids Res*., *37*, D544-549.

[263] Palaniswamy, S. K., Jin, V. X., Sun, H. & Davuluri, R. V. (2005). OMGProm: a database of orthologous mammalian gene promoters. *Bioinformatics*., *21*, 835-836.

[264] Kim, J., Seo, J., Lee, Y. S. & Kim, S. (2005). TFExplorer: integrated analysis database for predicted transcription regulatory elements. *Bioinformatics*., *21*, 548-550.

[265] Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. (2006). REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics*., *22*, 381-383.

[266] Morris, R. T., O'Connor, T. R. & Wyrick, J. J. (2008). Osiris: an integrated promoter database for Oryza sativa L. *Bioinformatics., 24*, 2915-2917.

[267] Rushton, P. J., Bokowiec, M. T., Laudeman, T. W., Brannock, J. F., Chen, X., Timko, M. P. (2008). TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics*., *9*, 53.

[268] Kaas, Q., Westermann, J. C., Halai, R., Wang, C. K., Craik, D. J. (2008). ConoServer, a database for conopeptide sequences and structures. *Bioinformatics*., *24*, 445-446.

[269] O'Brien, E. A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B. F. et al. (2009). GOBASE: an organelle genome database. *Nucleic Acids Res*., *37*, D946-950.

[270] Maselli, V., Di Bernardo, D. & Banfi, S. (2008). CoGemiR: a comparative genomics microRNA database. *BMC Genomics*., *9*, 457.

[271] Lu, T., Huang, X., Zhu, C., Huang, T., Zhao, Q. & Xie, K. et al. (2008). RICD: a rice indica cDNA database resource for rice functional genomics. *BMC Plant Biol*., *8*, 118.

[272] Lim, D., Cho, Y.M ., Lee, K. T., Kang, Y., Sung, S. & Nam, J. et al. (2009). The Pig Genome Database (PiGenome): an integrated database for pig genome research. *Mamm Genome*., *20*, 60-66.

[273] Gauthier, J. P., Legeai, F., Zasadzinski, A., Rispe, C. & Tagu, D. (2007). AphidBase: a database for aphid genomic resources. *Bioinformatics*., *23*, 783-784.

[274] Cameron, R. A., Samanta, M., Yuan, A., He, D. & Davidson, E. (2009). SpBase: the sea urchin genome database and web site. *Nucleic Acids Res*., *37*, D750-754.

[275] Nystrom, J., Fierlbeck, W., Granqvist, A., Kulak, S. C. & Ballermann, B. J. (2009). A human glomerular SAGE transcriptome database. *BMC Nephrol*., *10*, 13.

[276] Lee, B. & Shin, G. (2009). CleanEST: a database of cleansed EST libraries. *Nucleic Acids Res., 37*, D686-689.

[277] Beldade, P., Rudd, S., Gruber, J. D. & Long, A. D. (2006). A wing expressed sequence tag resource for Bicyclus anynana butterflies., an evo-devo model. *BMC Genomics*., *7*, 130.

[278] Schlamp, K., Weinmann, A., Krupp, M., Maass, T., Galle, P. & Teufel, A. (2008). BlotBase: a northern blot database. *Gene.*, *427*, 47-50.

[279] Sherlock, G., Hernanadez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C. & Dwight, S. S. et al. (2001). The Stanford microarray database. *Nucleic Acid Research*., *29*, 152-155.

[280] Markus, R., Srinivas, V., Samuel, A., Johan, S. & Jari, H. k. (2004). ACID: a database for microarray clone information. *Bioinformatics*., *20*, 2305.

[281] Singh, M. K., Srivastava, S., Raghava, G. P. & Varshney, G. C. (2006). HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. *Bioinformatics., 22*, 253-255.

[282] Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*., *22*, 623-625.

[283] Zhang, S., Xia, X., Shen, J., Zhou, Y. & Sun, Z. (2008). DBMLoc: a Database of proteins with multiple subcellular localizations. *BMC Bioinformatics*., *9*, 127.

[284] Zheng, C. J., Zhou, H., Xie, B., Han, L. Y., Yap, C. W. & Chen, Y. Z. (2004). TRMP: a database of therapeutically relevant multiple pathways. *Bioinformatics., 20*, 2236-2241.

[285] Barrett, T. & Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*., *411*, 352-369.

[286] Lakshmanan, L. V. S., Sadri, F. & Subramanian, S. N. (2001). SchemaSQL - An extention to SQL for multidatabase interoperability. *ACM Transactions on Database Systems, 26*, 476-519.

[287] Wyss, C. M. & Robertson, E. L. (2005). Relational languages for metadata integration. *ACM Transactions on Database Systems*, *30*, 624-660.

[288] Fristensky, B. (2007). BIRCH: a user-oriented, locally-customizable, bioinformatics system. *BMC Bioinformatics*, *8*, 54.

[289] Garcia Castro, A., Chen, Y. P. & Ragan, M. A. (2005). Information integration in molecular bioscience. *Appl Bioinformatics*, *4*, 157-173.

[290] Zhou, W., Smalheiser, N. R. & Yu, C. (2006). A tutorial on information retrieval: basic terms and concepts. *J Biomed Discov Collab*., *1*, 2.

[291] Hirschman, L., Park, J. C., Tsujii, J., Wong, L. & Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*., *18*, 1553-1561.

[292] Biagini, R. E., Krieg, E. F., Pinkerton, L. E. & Hamilton, R. G. (2001). Receiver operating characteristics analyses of Food and Drug Administration-cleared serological assays for natural rubber latex-specific immunoglobulin E antibody. *Clinical and Diagnostic Laboratory Immunology.*, *8*, 1145-1149.

[293] Gjengsto, P., Paus, E., Halvorsen, O. J., Eide, J., Akslen, L. A. & Wentzel-Larsen, T. et al. (2005). Predictors of prostate cancer evaluated by receiver operating characteristics partial area index: a prospective institutional study. *Journal of Urology*., *173*, 425-428.

[294] Margolis, D. J., Bilker, W., Boston, R., Localio, R., Berlin, J. A. (2002). Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology*., *55*, 518-524.

[295] Rosman, A. S. & Korsten, M. A. (2007). Application of summary receiver operating characteristics (sROC) analysis to diagnostic clinical testing. *Advances in Medical Science., 52*, 76-82.

[296] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*., *240*, 1285-1293.

[297] Hersh, W. (2005). Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Briefings in Bioinformatics*., *6*, 344-356.

[298] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K. & Chetvernin, V. et al. (2006). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*., *34*, D173-180.

[299] Stolovitzky, G. A., Kundaje, A., Held, G. A., Duggar, K. H., Haudenschild, C. D. & Zhou, D. et al. (2005). Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression. *Proceedings of the National Academy of Science*, U S A., *102*, 1402-1407.

[300] Shi, L. & Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics*., *6*, 88.

[301] Rzhetsky, A., Zheng, T. & Weinreb, C. (2006). Self-correcting maps of molecular pathways. *PLoS ONE*., *1*, e61.

[302] Wren, J. D. & Garner, H. R. (2004). Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*., *20*, 191-198.

[303] Ling, M. H. T., Leferve, C. & Nicholas, K. R. (2008). Filtering microarray correlations by statistical literature analysis yields potential hypotheses for lactation research. *The Python Papers*., *3*, 4.

[304] Grover, C., Klein, E., Lascarides, A. & Lapata, M. (2002). *XML-based NLP Tools for Analysing and Annotating Medical Language.* Second International Workshop on NLP and XML (NLPXML-2002).

[305] Grover, C., Lapata, M. & Lascarides, A. (2003). A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering*., *1*, 1-38.

[306] Ling, M. H. T., Leferve, C. & Nicholas, K. R. (2008). A Case Study where Parts-of-Speech Tagging Error Does Not Adversely Affect Extraction of Protein-Protein Interactions from Text. *The Python Papers*., *3*, 65-80.

[307] Voorhees, E. & Buckland, L. P. (eds) (2005). *The Fourteen Text REtrieval Conference Proceedings. National Institute of Standards and Technology (NIST).,* the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA), Gaithersburg, Maryland.

[308] Cano, C., Monaghan, T., Blanco, A., Wall, D. P. & Peshkin, L. (2009). Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *Journal of Biomedical Informatics*.

[309] Newman, D., Hettich, S., Blake, C., Merz, C. (1998). *UCI Repository of machine learning databases.* University of California, Department of Information and Computer Science, Irvine., CA.

[310] Kano, Y., Nguyen, N., Saetre, R., Yoshida, K., Miyao, Y. & Tsuruoka, Y. et al. (2008). Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. *Pacific Symposium on Biocomputing*: 616-627.

[311] Lourenco, A., Carreira, R., Carneiro, S., Maia, P., Glez-Pena, D. & Fdez-Riverola, F. et al. (2009). @Note: a workbench for biomedical text mining. *J Biomed Inform.*, *42*, 710-720.

[312] Altman, R. B., Bergman, C. M., Blake, J., Blaschke, C., Cohen, A. & Gannon, F. et al. (2008). Text mining for biology--the way forward: opinions from leading scientists. *Genome Biology.*, *9*, Suppl *2*, S7.

[313] Muller, M., Marko, K., Daumke, P., Paetzold, J., Roesner, A. & Klar, R. (2007). Biomedical data mining in clinical routine: expanding the impact of hospital information systems. *Medinfo.*, *12*, 340-344.

[314] Caporaso, J. G., Deshpande, N., Fink, J. L., Bourne, P. E., Cohen, K. B. & Hunter, L. (2008). Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pacific Symposium on Biocomputing*, 640-651.

[315] Roberts, P. M. & Hayes, W. S. (2008). Information needs and the role of text mining in drug development. *Pacific Symposium on Biocomputing,* 592-603.

[316] Kabiljo, R., Clegg, A. B. & Shepherd, A. J. (2009). A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics.*, *10*, 233.

[317] Roberts, P. M. (2006). Mining literature for systems biology. *Briefings in Bioinformatics.*, *7*, 399-406.

[318] Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E. & Apweiler, R. et al. (2006). GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab.*, *1*, 19.

[319] Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y. & Friedman, C. (2006). PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput,* 64-75.

[320] Natarajan, J. & Ganapathy, J. (2007). Functional gene clustering via gene annotation sentences., MeSH and GO keywords from biomedical literature. *Bioinformation.*, *2*, 185-193.

[321] Cakmak, A. & Ozsoyoglu, G. (2008). Discovering gene annotations in biomedical text databases. *BMC Bioinformatics.*, *9*, 143.

[322] Jin, B., Muller, B., Zhai, C. & Lu, X. (2008). Multi-label literature classification based on the Gene Ontology graph. *BMC Bioinformatics.*, *9*, 525.

[323] Abulaish, M. & Dey, L. (2007). Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining. *Data & Knowledge Engineering.*, *61*, 228.

[324] Baumgartner WA., Jr.., Lu Z., Johnson HL., Caporaso JG., Paquette J., Lindemann A., et al. (2008). Concept recognition for extracting protein interaction relations from biomedical text. *Genome Biology., 9,* Suppl *2*, S9.

[325] Heinrich, K. E., Berry, M. W. & Homayouni, R. (2008). Gene tree labeling using nonnegative matrix factorization on biomedical literature. *Computational Intelligence and Neuroscience.*, Article ID: 276535.

[326] Spasic, I., Schober, D., Sansone, S. A., Rebholz-Schuhmann, D., Kell, D. B. & Paton, N. W. (2008).. Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*,*a*, Suppl, *5*, S5.