Information retrieval

[1] Wilbur, W. J. & Coffee, L. The effectiveness of document neighboring in search enhancement. *Inf. Process. Manage.* **30**, 253–266 (1994).

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

[3] Usuzaka, S., Sim, K. L., Tanaka, M., Matsuno, H. & Miyano, S. A machine learning approach to reducing the work of experts in article selection from database: A case study for regulatory relations of *S. cerevisiae* genes in MEDLINE. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 91–101 (1998).

[4] Manning, C. D. & Schütze, H. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, (1999).

[5] Tanabe, L. *et al.* MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* **27**, 1210–1217 (1999).

[6] Aronson, A. R. *et al.* Methods for accurate retrieval of MEDLINE citations in functional genomics. in *Proc. AMIA Symp.*, volume 12, (2003).

[7] Renner, A. & Aszodi, A. High-throughput functional annotation of novel gene products using document clustering. in *Pac. Symp. Biocomput.*, volume 5, 50–68 (World Scientific, Hawaii, 2000).

[8] Iliopoulos, I. Enright, A. J. & Ouzounis, C. A. Textquest: document clustering of medline abstracts for concept discovery in molecular biology. in *Pac. Symp. Biocomput.*, volume 6, 384–395 (World Scientific, Hawaii, 2001).

[9] Marcotte, E. M., Xenarios, I. & Eisenberg, D. Mining literature for protein–protein interactions. *Bioinformatics* **17**, 359–363 (2001).

[10] Perez-Iratxeta, C., Bork, P. & A., A. M. XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.* **26**, 573–575 (2001).

[11] Bhalotia, G., Nakov, P. I., Schwartz, A. S. & Hearst, M. A. BioText team report for the TREC 2003 genomics track. in *Proceedings of TREC 2003*, volume 12, (2003).

[12] Crangle, C., Zbyslaw, A., Cherry, J. M. & Hong, E. L. Concept extractino and synonymy management for biomedical information retrieval. in *Proceedings of TREC 2004*, volume 13, (2004).

[13] deBruin, B. & Martin, J. Finding gene function using LitMiner. in *Proceedings of TREC 2003*, volume 12, (2003).

[14] Glenisson, P., Antal, P., Mathys, J., Moreau, Y. & De Moor, B. Evaluation of the vector space representation in text-based gene clustering. in *Pac. Symp. Biocomput.*, volume 8, 391–402 (World Scientific, Hawaii, 2003).

[15] Hersh, W. & Bhuptiraju, R. T. TREC genomics track overview. in *Proceedings of TREC 2003*, volume 12, (2003).

[16] Kayaalp, M. *et al.* Methods for accurate retrieval of MEDLINE citations in functional genomics. in *Proceedings of TREC 2003*, volume 12, (2003).

[17] Perez-Iratxeta, C., Perez, A. J., Bork, P. & A., A. M. Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Res.* **31**, 3866–3868 (2003).

[18] Yeh, A. S., Hirschman, L. & Morgan, A. A. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* **19**, i331–i339 (2003).

[19] Aronson, A. R. *et al.* Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. in *Proceedings of TREC 2004*, volume 13, (2004).

[20] Büttcher, S., Clarke, C. L. A. & Cormack, G. V. Domain-specific synonym expansion and validation for biomedical information retrieval. in *Proceedings of TREC 2004*, volume 13, (2004).

[21] Fujita, S. Revisiting again document length hypotheses: TREC-2004 genomics track experiments at Patolis. in *Proceedings of TREC 2004*, volume 13, (2004).

[22] Hersh, W., Bhupatiraju, R. T. & Corley, S. Enhancing access to the bibliome: the TREC genomics track. *Medinfo.* **11**, 773–777 (2004).

[23] Hersh, W. R. *et al.* TREC 2004 genomics track overview. in *Proceedings of TREC 2004*, volume 13, (2004).

[24] Kraaij, W., Weeber, M., Raaijmakers, S. & Jelier, R. MeSH based feedback, concept recognition and stacked classificatino for curation tasks. in *Proceedings of TREC 2004*, volume 13, (2004).

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

[26] Doms, A. & Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.* **33**, W783–W786 (2005).

[27] Goetz, T. & von der Lieth, C.-W. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Res.* **33**, W774–W778 (2005).

[28] Hoffmann, R. *et al.* Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* **283**, pe21 (2005).

[29] Shatkay, H. Hairpins in bookstacks: Information retrieval from biomedical text. *Brief Bioinform.* **6**, 222–238 (2005).

[30] Suomela, B. P. & Andrade, M. A. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics* **6**, 75 (2005).

[31] Xuan, W., Watson, S. J. & Meng, F. GeneInfoMiner—a web server for exploring biomedical literature using batch sequence ID. *Bioinformatics* **21**, 3452–3455 (2005).

Entity recognition

[1] Fukuda, K., Tamura, A., Tsunoda, T. & Takagi, T. Toward information extraction: identifying protein names from biological papers. in *Pac. Symp. Biocomput.*, volume 3, 707–718 (World Scientific, Hawaii, 1998).

[2] Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V. & Jacq, B. Detecting gene symbols and names in biological texts: A first step towards pertinent information extraction. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 72–80 (1998).

[3] Coller, N., Nobata, C. & Tsujii, J. Extracting the names of genes and gene products with a hidden Markov model. in *Int. Conf. Comput. Linguistics*, volume 18, 201–207, (2000).

[4] Krauthammer, M., Rzhetsky, A., Morozov, P. & C., F. Using blast for identifying gene and protein names in journal articles. *Gene* **259**, 245–252 (2000).

[6] Hatzivassiloglou, V., Duboue, P. A. & Rzhetsky, A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* **17 Suppl. 1**, S97–S106 (2001).

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

[] Collier, N., Nobata, C. & Tsujii, J. Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology* **7**, 239–257 (2002).

[9] Franzen, K. *et al.* Protein names and how to find them. *Int. J. Med. Inform.* **67**, 49–61 (2002).

[10] Leonard, J. E., Colombe, J. B. & Levy, J. L. Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics* **18**, 1515–1522 (2002).

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

[12] Bussey, K. J. *et al.* MatchMiner: a tool for batch navigation oamong gene and gene product identifiers. *Genome Biol.* **4**, R27 (2003).

[13] Hanisch, D., Fluck, J., Mevissen, H. T. & Zimmer, R. Playing biology's name game: identifying protein names in scientific text. in *Pac. Symp. Biocomput.*, volume 8, 403–414 (World Scientific, Hawaii, 2003).

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

[15] Narayanaswamy, M., Ravikumar, K. E. & Vijay-Shanker, K. A biological named entity recognizer. in *Pac. Symp. Biocomput.*, volume 8, 427–438 (World Scientific, Hawaii, 2003).

[16] Seki, K. & Mostafa, J. An approach to protein name extraction using heuristics and a dictionary. *Proceedings of the American Society for Information Science and Technology* **40**, 71–77 (2005).

[17] Tsuruoka, Y. & Tsujii, J. Boosting precision and recall of dictionary-based protein name recognition. in *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 41–48, (2003).

[18] Yu, H. & Agichtein, E. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics* **19**, i340–i349 (2003).

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

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

[21] Mika, S. & Rost, B. Protein names precisely peeled off free text. *Bioinformatics* **20**, i241–i247 (2004).

[22] Zhou, G., Zhang, J., Su, J., Shen, D. & Tan, C. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* **20**, 1178–1190 (2004).

[23] Chen, L., Liu, H. & Friedman, C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* **21**, 248–256 (2005).

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

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

[26] Finkel, J. *et al.* Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics* **6**, S5 (2005).

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

[28] Gaudan, S., Kirsch, H. & Rebholz-Schuhmann, D. Resolving abbreviations to their senses in Medline. *Bioinformatics* **21** (2005).

[29] Hakenberg, J. *et al.* Systematic feature evaluation for gene name recognition. *BMC Bioinformatics* **6**, S9 (2005).

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

[31] Hirschman, L., Colosimo, M., Morgan, A. & Yeh, A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics* **6**, S11 (2005).

[32] Kinoshita, S., Cohen, K. B., Ogren, P. V. & Hunter, L. BioCreAtIvE task 1A: entity identification with a stochastic tagger. *BMC Bioinformatics* **6**, S4 (2005).

[33] Kou, Z., Cohen, W. W. & Murphy, R. F. High-recall protein entity recognition using a dictionary. *Bioinformatics* **21**, i266–i273 (2005).

[34] Mani, I. *et al.* Protein name tagging guidelines: lessons learned. *Comp. Func. Genomics* **6**, 72–76 (2005).

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

[36] Mitsumori, T., Fation, S., Murata, M., Doi, K. & Doi, H. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics* **6**, S8 (2005).

[37] Pillet, V., Zehnder, M., Seewald, A. K., Veuthey, A. L. & Petrak, J. GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics* **21**, 1743–1744 (2005).

[38] Schijvenaars, B. J. A. *et al.* Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics* **6**, 149 (2005).

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

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

[41] Tamames, J. Text Detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics* **6**, S10 (2005).

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

[43] Yeh, A., Morgan, A., Colosimo, M. & Hirschman, L. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics* **6**, S2 (2005).

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

[45] Reyle, U. Understanding chemical terminology. *Terminology* **12** (2006). in press.

Information extraction

[1] Brill, E. A simple rule-based part of speech tagger. in *Proc. Conf. Appl. NLP*, volume 3, (1992).

[2] Schmid, H. Probabilistic part-of-speech tagging using decision trees. in *International Conference on New Methods in Language Processing* (, Manchester, UK, 1994).

[3] Hishiki, T. *et al.* Developing NLP tools for genome informatics: An information extraction perspective. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 81–90 (1998).

[4] Sekimizu, T., Park, H. S. & Tsujii, J. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 62–71 (1998).

[5] Blaschke, C., Andrade, M. A., Ouzounis, C. & Valencia, A. Automatic extraction of biological information from scientific text: protein–protein interactions. in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 7, 60–67 (AAAI Press, Menlo Park, CA, 1999).

[6] Craven, M. Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 7, 77–86 (AAAI Press, Menlo Park, CA, 1999).

[7] Ng, S. K. & Wong, M. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform. Ser. Workshop Genome Inform.* **10**, 104–112 (1999).

[8] Andrade, M. A. & Bork, P. Automated extraction of information in molecular biology. *FEBS Letters* **476**, 12–17 (2000).

[9] Brants, T. TnT—a statistical Part-of-Speech tagger. in *Proceedings of the Applied Natural Language Processing Conference*, volume 6, 224–231, (2000).

[10] Humphreys, K., Demetriou, G. & Gaizauskas, R. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. in *Pac. Symp. Biocomput.*, volume 5, 505–516 (World Scientific, Hawaii, 2000).

[11] Proux, D., Rechenmann, F. & Julliard, L. A pragmatic information extraction strategy for gathering data on genetic interactions. in *Proc. Int. Conf.*

Intell. Syst. Mol. Biol., volume 8, 179–285 (AAAI Press, Menlo Park, CA, 2000).

[12] Rindflesch, T. C., Tanabe, L., Weinstein, J. N. & Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. in *Pac. Symp. Biocomput.*, volume 1, 517–528 (World Scientific, Hawaii, 2000).

[13] Stapley, B. J. & Benoit, G. Biobibliometrics: Information retrieval and visualization from co-occurrence of gene names in Medline abstracts. in *Pac. Symp. Biocomput.*, volume 5, 529–540 (World Scientific, Hawaii, 2000).

[14] Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. Automatic extraction of protein interactions from scientific abstracts. in *Pac. Symp. Biocomput.*, volume 5, 707–709 (World Scientific, Hawaii, 2000).

[15] Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17 Suppl. 1**, S74–S82 (2001).

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

[17] Ono, T., Hishigaki, H., Tanigami, A. & Takagi, T. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* **17**, 155–161 (2001).

[18] Park, J. C., Kim, H. S. & Kim, J. J. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. in *Pac. Symp. Biocomput.*, volume 6, 396–407 (World Scientific, Hawaii, 2001).

[19] Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. & Mostafa, J. Detecting gene relations from Medline abstracts. in *Pac. Symp. Biocomput.*, volume 6, 483–495 (World Scientific, Hawaii, 2001).

[20] Wong, L. PIES, a protein interaction extraction system. in *Pac. Symp. Biocomput.*, volume 6, 520–531 (World Scientific, Hawaii, 2001).

[21] Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J. Event extraction from biomedical papers using a full parser. in *Pac. Symp. Biocomput.*, volume 6, 408–419 (World Scientific, Hawaii, 2001).

[22] Blaschke, C. & Valencia, A. The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems* **17**, 14–20 (2002).

[23] Ding, J., Berleant, d., Nettleton, D. & Wurtelle, E. Mining Medline: Abstracts, sentences, or phrases? in *Pac. Symp. Biocomput.*, volume 7, 326–337 (World Scientific, Hawaii, 2002).

[24] Hahn, U., Romacker, M. & Schulz, S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. in *Pac. Symp. Biocomput.*, volume 7, 338–349 (World Scientific, Hawaii, 2002).

[25] Leroy, G. & Chen, H. Filling preposition-based templates to capture information from medical abstracts. in *Pac. Symp. Biocomput.*, volume 7, 350–361 (World Scientific, Hawaii, 2002).

[26] Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M. & Cochran, B. Robust relational parsing over biomedical literature: Extracting inhibit relations. in *Pac. Symp. Biocomput.*, volume 7, 362–373 (World Scientific, Hawaii, 2002).

[27] Raychaudhuri, S., Chang, J. T., Sutphin, P. D. & Altman, R. B. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* **12**, 203–214 (2002).

[28] Regev, Y., Finkelstein-Landau, M. & Feldman, R. Rule-based extraction of experimental evidence in the biomedical domain. *SIGKDD Explorations* **4**, 90–92 (2002).

[29] Becker, K. G. *et al.* PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* **4**, 61 (2003).

[30] Donaldson, I. *et al.* PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics* **4**, art. 11 (2003).

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

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

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

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

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

[36] Chiang, J. & Yu, H. Extracting functional annotations of proteins based on hybrid text mining approaches. in *Proc. BioCreAtIvE Challenge Evaluation Workshop*, (2004).

[37] Chiang, J.-H., Yu, H.-C. & Hsu, H.-J. GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics* **20**, 120–121 (2004).

[38] Daraselia, N. *et al.* Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**, 604–611 (2004).

[39] Huang, M. *et al.* Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics* **20**, 3604–3612 (2004).

[40] Koike, A. & Takagi, T. PRIME: automatically extracted PRotein Interactions and Moolecular Information databasE. *In silico Biology* **5**, 0004 (2004).

[41] Rzhetsky, A. *et al.* GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* **37**, 43–53 (2004).

[42] Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. Extracting regulatory gene expression networks from pubmed. in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, (2004).

[43] Saric, J., Jensen, L. J. & Rojas, I. Large-scale extraction of gene regulation for model organisms in an ontological context. *In silico Biology* **5**, 0003 (2004).

[44] Wattarujeekrit, T., Shah, P. K. & Collier, N. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics* **5**, 155 (2004).

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

[46] Alako, B. T. *et al.* CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics* **6**, 51 (2005).

[47] Bajdik, C. D., Kuo, B., Rusaw, S., Jones, S. & Brooks-Wilson, A. CGMIM: Automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. *BMC Bioinformatics* **6**, 78 (2005).

[48] Blaschke, C., Leon, E. A., Krallinger, M. & Valencia, A. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* **6**, S16 (2005).

[49] Camon, E. B. *et al.* An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* **6**, S17 (2005).

[50] Cohen, A. M., Hersh, W. R. & Spackman, K. Using co-occurrence network structure to extract synonymous gene and prtoein names from MEDLINE abstracts. *BMC Bioinformatics* **6**, 103 (2005).

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

[52] Couto, F. M., Silva, M. J. & Coutino, P. M. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics* **6**, S21 (2005).

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

[54] Domedel-Puig, N. & Wernisch, L. Applying GIFT, a Gene Interactions Finder in Text, to fly literature. *Bioinformatics* **21** (2005).

[55] Ehrler, F., Geissbühler, A., Jimeno, A. & Ruch, P. Data-poor categorization and passage retrieval for gene ontology annotation in swiss-prot. *BMC Bioinformatics* **6**, S23 (2005).

[56] Hao, Y., Zhu, X., Huang, M. & M., L. Discovering patterns to extract protein-protein interactions from the literature: part II. *Bioinformatics* **21** (2005).

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

[58] Jelier, R. *et al.* Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics* **21**, 2049–2058 (2005).

[59] Krallinger, M., Padron, M. & Valencia, A. A sliding window approach to extract protein annotations from biomedical articles. *BMC Bioinformatics* **6**, S19 (2005).

[60] Maier, H. *et al.* LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. *Nucleic Acids Res.* **33**, W779–W782 (2005).

[61] Narayanaswamy, M., Ravikumar, K. E. & Vijay-Shanker, K. Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics* **21**, i319–i327 (2005).

[62] Ramani, A. K., Bunescu, R. C., Mooney, R. J. & Marcotte, E. M. Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology* **6**, R40 (2005).

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

[64] Rice, S. B., Nenadic, G. & Stapley, B. J. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics* **6**, S22 (2005).

[65] Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. Extraction of regulatory gene/protein networks from Medline, (2005). doi:10.1093/bioinformatics/bti597.

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

[67] Skusa, A., Ruegg, A. & Kohler, J. Extraction of biological interaction networks from scientific literature. *Brief Bioinform.* **6**, 263–276 (2005).

[68] Verspoor, K. *et al.* Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics* **6**, S20 (2005).

Text mining

[1] Swanson, D. R. Undiscovered public knowledge. *Library Quarterly* **56**, 103–118 (1986).

[2] Swanson, D. R. Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7–18 (1986).

[3] Swanson, D. R. Migrane and magnesium: Eleven neglected connections. *Perspect. Biol. Med.* **31**, 526–557 (1988).

[4] Swanson, D. R. Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect. Biol. Med.* **33**, 157–186 (1990).

[5] Swanson, D. R. Intervening in the life cycle of scientific knowledge. *Library Trends* **41**, 606–631 (1988).

[6] Smalheiser, N. R. & Swanson, D. R. Assessing a gap in the biomedical literature: Magnesium deficiency and neurological disease. *Neuroscience Research Communications* **15**, 1–9 (1994).

[7] Gordon, M. D. & Lindsay, R. K. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *J. Am. Soc. Inf. Sci.* **47**, 116–128 (1996).

[8] Smalheiser, N. R. & Swanson, D. R. Linking estrogen to Alzheimer's disease: An informatics approach. *Neurology* **47**, 809–810 (1996).

[9] Smalheiser, N. R. & Swanson, D. R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.* **57**, 149–153 (1998).

[10] Hearst, M. A. Untangling text data mining. in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, (1999).

[11] Lindsay, R. K. & Gordon, M. D. Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci.* **50**, 574–587 (1999).

[12] Swanson, D. R. & Smalheiser, N. R. Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. *Library Trends* **48**, 48–59 (1999).

[13] Weeber, M. *et al.* Text-based discovery in biomedicine: The architecture of the DAD-system. *Proc. AMIA Symp.* **20 Suppl.**, 903–907 (2000).

[14] Blaschke, C. & Valencia, A. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform. Ser. Workshop Genome Inform.* **12**, 123–134 (2001).

[15] Hristovski, D., Stare, J., Peterlin, B. & Dzeroski, S. Supporting discovery in medicine by association rule mining in MEDLINE and UMLS. *Medinfo.* **10**, 1344–1348 (2001).

[16] Smalheiser, N. R. Predicting emerging technologies with the aid of text-based data mining: the micro approach. *Comput. Methods Programs Biomed.* **21**, 689–693 (2001).

[17] Blagosklonny, M. V. & Pardee, A. B. Unearthing the gems. *Nature* **416**, 373 (2002).

[18] Srinivasan, P. & Libbus, B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* **20**, i290–i296 (2004).

[19] Wren, J. D., Bekeredijan, R., Stewart, J. A., Shohet, R. V. & Garner, H. R. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* **20**, 389–398 (2004).

[20] Wren, J. D. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics* **5**, 145 (2004).

[21] Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79–81 (2003).

[22] Homayouni, R., Heinrich, K., Wei, L. & Berry, M. W. Gene clustering by Latent Semantic Indexing of MEDLINE abstracts. *Bioinformatics* **21**, 104–115 (2005).

[23] Matsunaga, T. & Muramatsu, M. Knowledge-based computational search for genes associated with the metabolic syndrome. *Bioinformatics* **21**, 3146–3154 (2005).

Integration

[1] Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I. & Takagi, T. Automatic construction of knowledge base from biological papers. in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 5, 218–225 (AAAI Press, Menlo Park, CA, 1997).

[2] Andrade, M. A. & Valencia, A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **14**, 600–607 (1998).

[3] Liu, J. & Rost, B. SAWTED: Structure Assignment With Text Description—enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics* **16**, 125–129 (2000).

[4] Shatkay, H., Edwards, S., Wilbur, W. J. & Boguski, M. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 8, 317–328 (AAAI Press, Menlo Park, CA, 2000).

[5] Blaschke, C., Oliveros, J. C. & Valencia, A. Mining functional information associated with expression arrays. *Funct. Integr. Genomics* **1**, 256–268 (2001).

[6] Chang, J. T., Raychaudhuri, S. & Altman, R. B. Including biological literature improves homology search. in *Pac. Symp. Biocomput.*, volume 6, 374–383 (World Scientific, Hawaii, 2001).

[7] Masys, D. R. *et al.* Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* **17**, 319–326 (2001).

[8] Masys, D. R. Linking microarray data to the literature. *Nature Genetics* **28**, 9–10 (2001).

[9] Chaussabel, D. & Sher, A. Mining microarray expression data by literature profiling. *Genome Biol.* **3**, RESEARCH0055 (2002).

[10] Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Association of genes to genetically inherited diseases using text mining. *Nature Genetics* **31**, 316–319 (2002).

[11] Raychaudhuri, S., Schutze, H. & Altman, R. B. Using text analysis to identify functionally coherent gene groups. *Genome Res.* **12**, 1582–1590 (2002).

[12] Raychaudhuri, S. & Altman, R. B. A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics* **19**, 396–401 (2003).

[13] Raychaudhuri, S., Chang, J. T., Imam, F. & Altman, R. B. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.* **31**, 4553–4560 (2003).

[14] Schlitt, T. *et al.* From gene networks to gene function. *Genome Res.* **13**, 2568–2576 (2003).

[15] Bowers, P. M. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Nucleic Acids Res.* **5**, R35 (2003).

[16] Dieterich, G., Kärst, U., Wehland, J. & Jänsch, L. MineBlast: a literature presentation service supporting protein annotation by data mining of BLAST results. *Bioinformatics* **21**, 3450–3451 (2005).

[17] Fattore, M. & Arrigo, P. Knowledge discovery and system biology in molecular medicine: an application on neurodegenerative diseases. *In silico Biology* **5**, 0019 (2004).

[18] Glenisson, P. *et al.* TXTGate: profiling gene groups with text-based information. *Genome Biol.* **5**, R43 (2004).

[19] Iossifov, I. *et al.* Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics* **20**, 1205–1213 (2004).

[20] Krauthammer, M., Kaufmann, C. A., Gilliam, T. C. & Rzhetsky, A. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15148–15153 (2004).

[21] Djebbari, A., Karamycheva, S., Howe, E. & Quackenbush, J. MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. *Bioinformatics* **21**, 3324–3326 (2005).

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

[23] Korbel, J. O. *et al.* Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* **3**, e134 (2005).

[24] von Mering, C. *et al.* STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2005).

[25] Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. G2D: A tool for mining genes associated to disease. *BMC Genetics* **6**, 45 (2005).

[26] Scherf, M., Epple, A. & Werner, T. The next generation of literature analysis: Integration of genomic analysis into text mining. *Brief Bioinform.* **6**, 287–297 (2005).

[27] Seifert, M., Scherf, M., Epple, A. & Werner, T. Multievidence microarray mining. *Trends Genet.* **21**, 553–558 (2005).

[28] Tiffin, N. *et al.* Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* **33**, 1544–1552 (2005).

Miscellaneous

[1] Nenadic, G., Spasic, I. & S., A. Terminology-driven mining of biomedical literature. *Bioinformatics* **19**, 938–943 (2003).

[2] Shah, P. K., Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics* **4**, 20 (2003).

[3] Schuemie, M. J. *et al.* Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics* **20**, 2597–2604 (2004).

[4] Srinivasan, P. & Hristovski, D. Distilling conceptual connections from MeSH co-occurrences. *Medinfo.* **11**, 808–812 (2004).

[5] Mons, B. Which gene did you mean? *BMC Bioinformatics* **6**, 142 (2005).

[6] Spasic, I. & Ananiadou, S. A flexible measure of contextual similarity for biomedical terms. in *Pac. Symp. Biocomput.*, volume 10, 197–208 (World Scientific, Hawaii, 2005).

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

[8] Yandell, M. D. & Majoros, W. H. Genomics and natural language processing. *Nat. Rev. Genet.* **3**, 601–610 (2002).

[9] Krallinger, M. & Valencia, A. Text-mining and information-retrieval services for molecular biology. *Genome Biology* **6**, 224 (2005).

[10] Dickman, S. Tough mining. *PLoS Biology* **1**, 144–147 (2005).

[11] Rebholz-Schuhmann, D. Facts from text—is text mining ready to deliver. *PLoS Biology* **3**, e65 (2005).

[12] Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6**, S1 (2005).