Published in final edited form as:

Pharmacogenomics. 2010 April ; 11(4): 515–518. doi:10.2217/pgs.10.48.

Teaching computers to read the pharmacogenomics literature ... so you don't have to

Yael Garten and

Biomedical Informatics, Stanford University, Stanford, CA 94305, USA

Russ B Altman

Departments of Bioengineering & Genetics, Stanford University, Stanford, CA 94305, USA

In the last decade, pharmacogenomics has emerged as an active and productive discipline, drawing on both pharmacology and genomics. With powerful high-throughput technologies, we are discovering rich connections between genes, drugs, variants and phenotypes. The productivity of pharmacogenomics research has led to a rapid expansion of the associated scientific literature, widely dispersed among numerous journals and difficult to track. An important set of challenges thus arises: how to track new findings? How to recall known connections? How to identify gaps in understanding? Recent advances in the technology of text mining promise to help.

Knowledge reported in the literature is often embedded in dormant PDF documents, disconnected from related information and underlying data [1]. The reality, of course, is that the entities of interest (drugs, genes, model systems, experimental methods, phenotypes and so on) are interconnected in intricate ways, via pathways and networks often too complex to commit to memory. Text mining can connect these entities, and provides us with the opportunity to assess our understanding by viewing a connected map of entities. The benefit is clear; an accurate knowledge map would assist researchers, save their time and allow them to read less, cover more and (most importantly) generate hypotheses to bring to the lab.

In the context of pharmacogenomics, the key entities are drugs, gene variants and phenotypes. Knowledge of these entities and the interactions between them provides answers to associated fundamental questions: 'which gene products metabolize a drug?', 'what drug response phenotypes are affected by a particular gene?' or 'which gene variant decreases response to drug X?' This knowledge also forms the basis for generating candidate gene lists for important phenotypes, and for detecting unexpected drug–drug interactions. The vision of a computer-based system to answer such questions was a far-off fantasy a decade ago. However, text mining and natural language processing techniques, particularly those tailored to biomedical text, have matured in recent years, and we can now contemplate a continually updated and automatically generated comprehensive view of pharmacogenomic knowledge.

© 2010 Future Medicine Ltd

Author for correspondence: Tel.: +1 650 725 3394 Fax: +1 650 723 8544, russ.altman@stanford.edu.

For reprint orders, please contact: reprints@futuremedicine.com

Financial & competing interests disclosure

The authors acknowledge support from NIH LM07033 (Yael Garten) and GM61374 (Yael Garten and Russ B Altman). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

How has text mining advanced? Text mining is the process whereby information is extracted from human-generated natural language text using informatics algorithms. Thus, unstructured text is structured into a machine-readable format. The text mining task consists of several key steps: document retrieval, entity recognition (also known as ‘named entity recognition’), and relationship extraction (either simple term co-occurrence or more advanced semantic understanding). The last decade has shown great progress in each of these steps. The automation of these tasks provides a method that scales with the double-exponential growth of Medline [2], which currently adds an additional two papers every minute [3].

Document retrieval: identifying articles relevant to pharmacogenomics

Most publications containing pharmacogenomic information are published in a set of approximately 20 key journals, as described by Lascar and Barnett [4] and from our experience at the Pharmacogenomics Knowledgebase (PharmGKB [101]). However, important publications are also found in many other journals at a lower frequency, and so methods to identify such publications automatically are critical. MScanner is a classifier of Medline publication abstracts, which learns to distinguish articles relevant to pharmacogenomics versus those that are not, by using the Medical Subject Heading (MeSH) terms [102] assigned to the article by the MeSH curators and learning from a set of examples provided as input to the algorithm [5]. MeSH terms are selected manually by the curators, from a controlled vocabulary used by the National Library of Medicine. Other algorithms use the words in the abstracts [6]. These have been tuned for pharmacogenomics, but there are also general purpose algorithms for finding relevant literature by ranking documents, such as ReleMed® [7], XplorMed [8], GoPubMed® [9,10], GoAnnotator [11], PubFocus [12] and Information Hyperlinked over Proteins (iHOP) [13]. Some search engines allow a semantic search that can include relationships between abstract concepts (e.g., gene’ or ‘drug’ instead of specific genes and drugs) [14,15].

Entity recognition: identifying genes, drugs, variants & phenotypes

There has also been significant progress in recognizing biomedical entities (e.g., genes, drugs and variants) in text. Genes are particularly difficult to recognize automatically owing to the frequent use of common English words as gene names and other such issues [16]. The Critical Assessment of Information Extraction systems in Biology (BioCreAtIvE) challenge [17–20] was a community-wide effort to compare systems for identifying genes in the text, and catalyzed good progress in this task. Standard tools now exist to tag genes found in text, such as ABNER [21], LingPipe [103] and BANNER [22]. In addition, several applications can identify genetic variations in text (e.g., C3435T), including MutationFinder [23], MarkerInfoFinder [24], MuteXT [25] and Pharmspresso [14].

Relationship extraction: identifying relationships between entities

It is critical not only to identify the key entities in text, but also their relations to one another. Several systems can extract relationships using entity co-occurrence within sentences or paragraphs. These have been applied to large bodies of text to build protein–protein interaction networks, gene–drug networks, and drug–disease networks. The FACTA system presents a table summarizing co-occurring concepts for a protein of interest [26]. Other examples include Chang *et al.* [16], AliBaba [27], XplorMed [8], Chilobot [28], Textpresso [15], EBIMed [29], CoPub [30] and iHOP [13]. Integrating these networks with human-curated databases and experimental data can add an additional layer of confidence in the extracted relationships [31,32]. We have reported that text-mined relationships can sometimes perform as well as manually curated ones in the context of candidate gene prediction [33]. Recent work focuses on extracting information from figures, captions and

tables, widespread in biomedical articles and previously inaccessible to automated systems [34].

Finally, new methods have been developed for extracting deeper semantics from sentences [35–37]. These methods rely on four advances:

- Improved statistical sentence parsing technologies that can identify parts of speech (e.g., noun phrases, verb phrases and prepositional phrases) [38–40];
- Huge collections of biomedical text for mining relationships based on recurring patterns or templates of use (e.g., common occurrence of ‘X metabolizes Y’ or ‘Y inhibits X’);
- Improved annotated biomedical corpora [41,104] that provide gold standard datasets with which to assess computer algorithms;
- Ontology standards for representing semantics so that multiple equivalent language forms can be mapped to a common conceptual framework.

These combine to allow us to now recognize the actual meaning or semantics within text [42,43]. For example, knowing that ‘CYP2C9 metabolizes warfarin’ from a gold standard annotation, we can parse other sentences in a large corpus to find other genes and drugs that are related by the verb ‘metabolizes’. We can then map these genes and drugs to a standard ‘metabolize’ concept, which may be associated with more general terms (e.g., ‘interact’) and more specific terms (e.g., ‘oxidize’) as well as synonyms. By building a large number of such interactions, we can answer many questions regarding drug metabolism based on a fully automated analysis of text [44–46]. These analyses can also include the search for negation (e.g., ‘CYP2D6 does not metabolize warfarin’) and contradictory statements.

For the foreseeable future, scientists will publish their findings in papers that are written for other humans. Although structured digital abstracts written with carefully controlled terminologies would simplify the job for computers, they would make the scientific literature unbearable for humans to read. Indeed, recent experiments show that authors are not very good at re-expressing their findings using controlled terminologies [47]. Thus, the challenge of computer analysis of text is likely to remain. Recent progress in extracting useful and semantically accurate information regarding pharmacogenomics is creating an infrastructure that will not only enable powerful queries by users, but will also allow us to combine these observations to generate new hypotheses, uncover cryptic relationships, and fine tune our understanding of the genetic pathways underlying pharmacokinetics and pharmacodynamics.

Bibliography

Papers of special note have been highlighted as:

- of interest
- ■ of considerable interest
- 1 ■ ■. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D. Calling International Rescue: knowledge lost in literature and data landslide! *Biochem J.* 2009; 424(3):317–333. Interesting review about initiatives to prevent the knowledge and discoveries of research from being lost in the overwhelming amount of published literature. [PubMed: 19929850]
- 2 ■ ■. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 2008; 9(Suppl 2):S8. Excellent review of text mining in biology; describes methods and available tools. [PubMed: 18834499]

3. Hull D, Pettifer, Kell DB. Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol*. 2008; 4(10):E1000204. [PubMed: 18974831]
4. Lascar C, Barnett P. Defining and searching pharmacogenetics and pharmacogenomics to identify its core research journals. *Science & Technology Libraries*. 2005; 26(1):69–88.
5. Poulter GL, Rubin DL, Altman RB, Seoighe C. MScanner: a classifier for retrieving Medline citations. *BMC Bioinformatics*. 2008; 9:108. [PubMed: 18284683]
6. Rubin DL, Thorn CF, Klein TE, Altman RB. A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. *J Am Med Inform Assoc*. 2005; 12(2):121–129. [PubMed: 15561790]
7. Siadat MS, Shu J, Knaus WA. Relemed: Sentence-level search engine with relevance score for the medline database of biomedical articles. *BMC Med Inform Decis Mak*. 2007; 7:1. [PubMed: 17214888]
8. Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: A tool for exploring medline abstracts. *Trends Biochem Sci*. 2001; 26(9):573–575. [PubMed: 11551795]
9. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res*. 2005; 33(Web Server issue):W783–W786. [PubMed: 15980585]
10. Vanteru BC, Shaik JS, Yeasin M. Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genomics*. 2008; 9(Suppl 1):S10. [PubMed: 18366599]
11. Couto FM, Silva MJ, Lee V, et al. Goa nnotator: linking protein go annotations to evidence text. *J Biomed Discov Collab*. 2006; 1:19. [PubMed: 17181854]
12. Plikus MV, Zhang Z, Chuong CM. Pubfocus: Semantic medline /pubmed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics*. 2006; 7:424. [PubMed: 17014720]
13. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*. 2005; 21(Suppl 2):II252–II258. [PubMed: 16204114]
14. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*. 2009; 10(Suppl 2):S6. [PubMed: 19208194]
15. Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. 2004; 2(11):E309. [PubMed: 15383839]
16. Chang JT, Altman RB. Extracting and characterizing gene–drug relationships from the literature. *Pharmacogenetics*. 2004; 14(9):577–586. [PubMed: 15475731]
17. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*. 2005; 6(Suppl 1):S1. [PubMed: 15960821]
18. Morgan AA, Lu Z, Wang X, et al. Overview of biocreative II gene normalization. *Genome Biol*. 2008; 9(Suppl 2):S3. [PubMed: 18834494]
19. Smith L, Tanabe LK, Ando RJ, et al. Overview of biocreative II gene mention recognition. *Genome Biol*. 2008; 9(Suppl 2):S2. [PubMed: 18834493]
20. Yeh A, Morgan A, Colosimo M, Hirschman L. biocreative task 1a: gene mention finding evaluation. *BMC Bioinformatics*. 2005; 6(Suppl 1):S2. [PubMed: 15960832]
21. Settles B. Abner : an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*. 2005; 21(14):3191–3192. [PubMed: 15860559]
22. Leaman R, Gonzalez G. Banner : an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput*. 2008:652–663. [PubMed: 18229723]
23. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L. Mutationfinder. a high-performance system for extracting point mutation mentions from text. *Bioinformatics*. 2007; 23(14):1862–1865. [PubMed: 17495998]
24. Xuan W, Wang P, Watson SJ, Meng F. Medline search engine for finding genetic markers with biological significance. *Bioinformatics*. 2007; 23(18):2477–2484. [PubMed: 17823133]
25. Horn F, Lau AL, Cohen FE. Automated extraction of mutation data from the literature: application of mutext to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*. 2004; 20(4):557–568. [PubMed: 14990452]

26. Tsuruoka Y, Tsujii J, Ananiadou S. Facta : a text search engine for finding associated biomedical concepts. *Bioinformatics*. 2008; 24(21):2559–2560. [PubMed: 18772154]
- 27■. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: Pubmed as a graph. *Bioinformatics*. 2006; 22(19):2444–2445. Authors provide an easy-to-use interactive application for browsing biological networks extracted on-the-fly from results of PubMed queries. Edges of the graph display the supporting evidence (sentence from PubMed abstract) linking two entities. [PubMed: 16870931]
28. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*. 2004; 5:147. [PubMed: 15473905]
29. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P. Ebi Med – text crunching to gather facts for proteins from Medline. *Bioinformatics*. 2007; 23(2):E237–E244. [PubMed: 17237098]
30. Alako BT, Veldhoven A, van Baal S, et al. CoPub Mapper: Mining medline based on search term co-publication. *BMC Bioinformatics*. 2005; 6:51. [PubMed: 15760478]
31. Gonzalez G, Uribe JC, Tari L, Brophy C, Baral C. Mining gene–disease relationships from biomedical literature: weighting protein–protein interactions and connectivity measures. *Pac Symp Biocomput*. 2007:28–39. [PubMed: 17992743]
32. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. Stitch : interaction networks of chemicals and proteins. *Nucleic Acids Res*. 2008; 36(Database issue):D684–D688. [PubMed: 18084021]
33. Garten Y, Tatonetti NP, Altman RB. Improving the prediction of pharmacogenes using text-derived drug–gene relationships. *Pac Symp Biocomput*. 2010:305–314. [PubMed: 19908383]
- 34■. Hearst MA, Divoli A, Guturu H, et al. BioText search engine: beyond abstract search. *Bioinformatics*. 2007; 23(16):2196–2197. Authors describe BioText, an application that mines and searches within figures, captions and tables, as well as titles and abstracts. Important advance in mining the scientific literature. [PubMed: 17545178]
- 35■. Hunter L, Lu Z, Firby J, et al. Opendmap : an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*. 2008; 9:78. Important paper; unlike traditional parsers, which extract information as substrings of text, the authors use an ontology to map the information mined from text onto a model of biology, which allows powerful reasoning mechanisms to be used on the extracted information. [PubMed: 18237434]
36. Cheng D, Knox C, Young N, et al. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res*. 2008; 36(Web Server issue):W399–W405. [PubMed: 18487273]
37. Fundel K, Küffner R, Zimmer R. RelEx – relation extraction using dependency parse trees. *Bioinformatics*. 2007; 23(3):365–371. [PubMed: 17142812]
38. Buyko, E.; Faessler, E.; Wermter, J.; Hahn, U. Event extraction from trimmed dependency graphs. *Proceedings of the Workshop on BioNLP: Shared Task; Association for Computation Linguistics, CO, USA. 2009. p. 19-27.*
39. Clegg AB, Shepherd AJ. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*. 2007; 8:24. [PubMed: 17254351]
40. Marneffe, MD.; Manning, CD. The Stanford typed dependencies representation. *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation. Coling 2008 Organizing Committee; Manchester, UK. 2008. p. 1-8.*
- 41■. Kim JD, Ohta T, Tateisi Y, Tsujii J. Genia corpus – semantically annotated corpus for bio-textmining. *Bioinformatics*. 2003; 19(Suppl 1):I180–I182. Important resource of manually annotated literature used to train biological text mining systems. [PubMed: 12855455]
42. Bundschuh M, Dejori M, Stetter M, Tresp V, Kriegel HP. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*. 2008; 9:207. [PubMed: 18433469]
43. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. *Pac Symp Biocomput*. 2007:209–220. [PubMed: 17990493]

44. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.* 2008; 9(6):R96. [PubMed: 18549479]
45. Tari L, Hakenberg J, Gonzalez G, Baral C. Querying parse tree database of Medline text to synthesize user-specific biomolecular networks. *Pac Symp Biocomput.* 2009:87–98. [PubMed: 19209697]
46. Tari L, Anwar S, Liang S, Hakenberg J, Baral C. Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning. *Pac Symp Biocomput.* 2010:465–476. [PubMed: 19908398]
47. Ceol A, Chatr-Aryamontri A, Licata L, Cesareni G. Linking entries in protein interaction database to structured text: the febs Letters experiment. *FEBS Lett.* 2008; 582(8):1171–1177. [PubMed: 18328820]

Websites

101. PharmGKB. www.pharmgkb.org
102. National Library of Medicine. www.nlm.nih.gov/mesh
103. Alias-i: LingPipe 3.9.0. <http://alias-i.com/lingpipe>
- 104 ■. Penn bioie. <http://bioie ldc.upenn.edu> Another important resource of manually annotated literature used to train biological text mining systems