A Computational Linguistics Approach to the Identification of Biological Factors that Contribute to the Development and Progression of Lung Cancer

C. M. Frenz [1*], C. Luo [2], E. Urgard[2] and A. Metspalu[2]
[1]In Silico Biotechnologies, New York, USA
[2]Department of Biotechnology, University of Tartu, Estonia
[*]Corresponding Author: chris@insilicobiotechnologies.com

Abstract - *The copious volumes of biomedical literature being generated have created a need for the development of text mining algorithms to identify and extract and pertinent biological information. This pilot study demonstrates a computational linguistics approach to identifying genes, proteins, and other biological factors that are associated with the development and progression of lung cancer.*

Keywords: lung cancer, linguistics, bioinformatics, text mining

1 Introduction

Over the course of the last couple of decades the biomedical sciences have undergone an explosion in the amount of biomedical literature that is published, with indexes of biomedical literature, such as Pubmed, housing over 20 million articles (as of March 2011). While the massive amount of information available in such a large corpus of literature is of clear benefit to researchers, the sheer numbers of documents that can match any given query often makes the task of finding needed pieces of information a difficult one. For this reason, researchers in the biomedical sciences are continually turning toward the development of computational tools to perform data mining tasks, ranging from the identification of genes that play a role in certain biological outcome [1-3] and the identification of mutations [4] to the identification of high quality Web resources [5], and many areas in between.

Current methodologies for text mining the biomedical literature include techniques such as regular expression based pattern matching [6], the development and use of biological concept ontologies [7], and the development of specialized parsers designed to perform Natural Language Processing (NLP) of the biomedical literature [3, 8]. This study seeks to expand upon the current methodological approaches by developing a simplistic yet robust methodology for the identification of genes, proteins, and other biological factors that contribute to a disease of interest or other biological state. The developed methodology makes use of commonly used computational linguistics techniques, by first requiring the establishment of a corpus of biological literature pertaining to the disease or biological state of interest and then performing a word frequency analysis on the corpus to identify all of the unique words in the corpus. A pruning technique is then applied to the listing of unique words in order to remove all English language and biomedical specific jargon words, leaving a resultant list which primarily contains the names of genes and proteins that contribute to the disease or biological state of interest. The technique was tested via the identification of biological factors that contribute to or are associated with the development and metastasis of lung cancer.

2 Methods

2.1 Establishment of a Lung Cancer Corpus

A corpus of text pertaining to lung cancer was developed via the modification of the PREP.pl perl script [6, 9], which was designed to retrieve Pubmed abstracts and perform regular expression based pattern matching against them. The script was modified to retrieve all Pubmed abstracts pertaining to the keyword query "lung cancer" and save the title and text of each abstract to the corpus. Only the titles and text were added to the corpus since other abstract data such as journal name abbreviations and author names would add additional "words" to the corpus that would not be valid biological factors and would be very difficult to prune out during later processing of the data. Execution of this script resulted in a corpus consisting of 186 MB of text.

2.2 Processing of the Corpus

A word frequency analysis of the corpus was performed that identified every unique word in the system as well as how often each unique word appears in the system. This analysis was performed via a Perl script which identified words as being unique if they were separated by 1 or more non-alphanumeric characters (i.e. \W+). This analysis

resulted in over 177,000 unique words ranging from the most popular word "of", which occurred 1,338,203 times, to words like "Gp96" (a heat shock protein), which only occurred a single time.

Upon completion of the word identification via the word frequency analysis, the list of unique words was pruned by comparing the identified word to a dictionary of words to remove all words that would not be the names of genes, proteins, or other biological factors that could play a potential role in lung cancer, such as English language words. Initial testing of the technique made use of the words.txt dictionary file that comes standard with any Linux distribution as a basis for identifying English language words, but this lead to a high false positive rate, since the dictionary did not contain much of the biomedical jargon and terminology that appears is the biomedical literature, but is not a gene/protein name. Thus, the dictionary was expanded to include such jargon and biomedical terminologies (e.g. "mesenchyme"), in order to better prune the list of biological factor candidates. Other common false positive candidates, include common non-gene/protein name acronyms, such as NSCLC (non-small cell lung carcinoma) or NK (natural killer), cancer drugs being tested within the published literature, such as cisplatin, and tumor cells types, such as A549. Further improvements to the false positive rate are made possible by incorporation of these words into the dictionary as well. Adding in common misspellings and typographical errors would be a way to further prune the list of potential biological factors.

3 Results and Discussion

The newly established methodology does provide a means of successfully identifying genes, proteins, and other biological factors that can contribute to diseases such as lung cancer, as illustrated by the results in Table 1, which demonstrates the top eighteen most frequent biological factors that are associated with the development and progression of lung cancer. Among the listings in Table 1 are many well characterized tumor suppressors (e.g. p53) and oncogenes (e.g. myc), as well as other biological factors crucial to the progression of cancer such as VEGF. In some cases in Table 1, the factors identified may be very broad (e.g. kinase and cyclin), but more specific instances of these categories are usually identified at lower frequencies, such as PKC (occurred 1068 times), a type of kinase identified by the technique.

Table 1: The top 18 most frequently appearing biological factors associated with lung cancer.

Biological Factor	Number of Appearances	Sample Reference
p53	16028	[10]
EGFR	14055	[11]
kinase	10274	[12]
VEGF	6168	[13]
CEA	5800	[14]
IFN	4287	[15]
TGF	4255	[16]
MMP	4002	[17]
RR	3235	[18]
TNF	3182	[19]
COX	2838	[20]
cyclin	2614	[21]
caspase	2498	[22]
NNK	2464	[23]
IGF	2426	[24]
Bcl	2285	[25]
myc	2279	[26]
EGF	2140	[27]

This ability for specificity is also demonstrated when one considers that the technique has the capacity to pick up biological factors that may only have few or even singular occurrences in the corpus, as illustrated in Table 2, which contains a sampling of factor names that only occurred 1-2 times in the corpus. While the entries in Table 2 serve to illustrate the specificity of the technique, however, they do demonstrate one limitation of the current implementation in that the word frequency analysis does not correlate multiple possible spellings of the same name as being identical. For example, even though "IRS1" only occurred twice (as written), there were alternative matches in the corpus such as "IRS-1". These spelling variants are currently recognized as unique, but future iterations of the word frequency analysis software will be modified to treat them as the same. This spelling issue also applies to ATF6, ELAV3, Dnmt3a, and Gp96 as well. It is notable, however, that relatively few publications currently explore these genes/proteins regardless of spelling. For example, Gp96 is a heat shock protein associated with lung cancer in less than 5 English abstracts [28-30].

Table 2: A sample of biological factor names that appeared as written only 1-2 times.

Biological Factor	Number of Appearances	Sample Reference
IRS1	2	[31]
ELAV3	1	[32]
ATF6	1	[33]
Dnmt3a	1	[34]
Gp96	1	[29]

4 Conclusion

In all, this pilot study demonstrates the potential for using word frequency analysis as a means of identifying the names of genes, proteins, and other biological factors that could play a role in the development of lung cancer or other biological conditions. The utility of the developed methodology, however, is dependent on the use of a robust dictionary of words and terms to be excluded from consideration, although it is hypothesized that the development of such a dictionary is broadly applicable to performing such an analysis across a diversity of biological contexts. It is the contention of the author that the continued expansion of such a dictionary of exclusion terms, would result in a technique that could lead to the rapid identification of candidate genes with a minimal amount of human data curation.

5 References

[1] J. H. Chiang, H. C. Yu, and H. J. Hsu, "GIS: a biomedical text-mining system for gene information discovery," *Bioinformatics*, vol. 20, pp. 120-1, Jan 1 2004.

[2] C. M. Frenz and D. A. Frenz, "The application of regular expression-based pattern matching to profiling the developmental factors that contribute to the development of the inner ear," *Adv Exp Med Biol*, vol. 680, pp. 165-71, 2010.

[3] D. Sahoo, J. Seita, D. Bhattacharya, M. A. Inlay, I. L. Weissman, S. K. Plevritis, and D. L. Dill, "MiDReG: a method of mining developmentally regulated genes using Boolean implications," *Proc Natl Acad Sci U S A*, vol. 107, pp. 5732-7, Mar 30 2011.

[4] F. Horn, A. L. Lau, and F. E. Cohen, "Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors," *Bioinformatics*, vol. 20, pp. 557-68, Mar 1 2004.

[5] J. A. Young and C. M. Frenz, "Automated extraction of health resource URLs from biomedical abstracts," in *2010 Long Island Systems Applications and Technology Conference (LISAT)* 2010, pp. 1-3.

[6] C. M. Frenz, "Deafness mutation mining using regular expression based pattern matching," *BMC Med Inform Decis Mak*, vol. 7, p. 32, 2007.

[7] H. M. Muller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS Biol*, vol. 2, p. e309, Nov 2004.

[8] H. Chen and B. M. Sharp, "Content-rich biological network constructed by mining PubMed abstracts," *BMC Bioinformatics*, vol. 5, p. 147, Oct 8 2004.

[9] C. Frenz, *Pro Perl Parsing*: Apress, 2005.

[10] X. Wang, D. C. Christiani, J. K. Wiencke, M. Fischbein, X. Xu, T. J. Cheng, E. Mark, J. C. Wain, and K. T. Kelsey, "Mutations in the p53 gene in lung cancer are associated with cigarette smoking and asbestos exposure," *Cancer Epidemiol Biomarkers Prev*, vol. 4, pp. 543-8, Jul-Aug 1995.

[11] J. G. Paez, P. A. Janne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon, K. Naoki, H. Sasaki, Y. Fujii, M. J. Eck, W. R. Sellers, B. E. Johnson, and M. Meyerson, "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science*, vol. 304, pp. 1497-500, Jun 4 2004.

[12] M. G. Kris, R. B. Natale, R. S. Herbst, T. J. Lynch, Jr., D. Prager, C. P. Belani, J. H. Schiller, K. Kelly, H. Spiridonidis, A. Sandler, K. S. Albain, D. Cella, M. K. Wolf, S. D. Averbuch, J. J. Ochs, and A. C. Kay, "Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer: a randomized trial," *JAMA*, vol. 290, pp. 2149-58, Oct 22 2003.

[13] P. Salven, T. Ruotsalainen, K. Mattson, and H. Joensuu, "High pre-treatment serum level of vascular endothelial growth factor (VEGF) is associated with poor outcome in small-cell lung cancer," *Int J Cancer*, vol. 79, pp. 144-6, Apr 17 1998.

[14] R. Salgia, D. Harpole, J. E. Herndon, 2nd, E. Pisick, A. Elias, and A. T. Skarin, "Role of serum tumor markers CA 125 and CEA in non-small cell lung cancer," *Anticancer Res*, vol. 21, pp. 1241-6, Mar-Apr 2001.

[15] K. Mattson, A. Niiranen, T. Ruotsalainen, P. Maasilta, M. Halme, S. Pyrhonen, M. Kajanti, M. Mantyla, K. Tamminen, A. Jekunen, S. Sarna, and K. Cantell, "Interferon maintenance therapy for small cell lung cancer: improvement in long-term

survival," *J Interferon Cytokine Res*, vol. 17, pp. 103-5, Feb 1997.

[16] H. J. Baek, S. S. Kim, F. M. da Silva, E. A. Volpe, S. Evans, B. Mishra, L. Mishra, and M. B. Marshall, "Inactivation of TGF-beta signaling in lung cancer results in increased CDK4 activity that can be rescued by ELF," *Biochem Biophys Res Commun*, vol. 346, pp. 1150-7, Aug 11 2006.

[17] X. Zhang, S. Zhu, G. Luo, L. Zheng, J. Wei, J. Zhu, Q. Mu, and N. Xu, "Expression of MMP-10 in lung cancer," *Anticancer Res*, vol. 27, pp. 2791-5, Jul-Aug 2007.

[18] H. Uramoto, K. Sugio, T. Oyama, T. Hanagiri, and K. Yasumoto, "P53R2, p53 inducible ribonucleotide reductase gene, correlated with tumor progression of non-small cell lung cancer," *Anticancer Res*, vol. 26, pp. 983-8, Mar-Apr 2006.

[19] V. Flego, A. Radojcic Badovinac, L. Bulat-Kardum, D. Matanic, M. Crnic-Martinovic, M. Kapovic, and S. Ristic, "Primary lung cancer and TNF-alpha gene polymorphisms: a case-control study in a Croatian population," *Med Sci Monit*, vol. 15, pp. CR361-5, Jul 2009.

[20] R. E. Harris, J. Beebe-Donk, and G. A. Alshafie, "Reduced risk of human lung cancer by selective cyclooxygenase 2 (COX-2) blockade: results of a case control study," *Int J Biol Sci*, vol. 3, pp. 328-34, 2007.

[21] O. Gautschi, D. Ratschiller, M. Gugger, D. C. Betticher, and J. Heighway, "Cyclin D1 in non-small cell lung cancer: a key driver of malignant transformation," *Lung Cancer*, vol. 55, pp. 1-14, Jan 2007.

[22] D. A. Fennell, "Caspase regulation in non-small cell lung cancer and its potential for therapeutic exploitation," *Clin Cancer Res*, vol. 11, pp. 2097-105, Mar 15 2005.

[23] S. Razani-Boroujerdi and M. L. Sopori, "Early manifestations of NNK-induced lung cancer: role of lung immunity in tumor susceptibility," *Am J Respir Cell Mol Biol*, vol. 36, pp. 13-9, Jan 2007.

[24] S. J. London, J. M. Yuan, G. S. Travlos, Y. T. Gao, R. E. Wilson, R. K. Ross, and M. C. Yu, "Insulin-like growth factor I, IGF-binding protein 3, and lung cancer risk in a prospective study of men in China," *J Natl Cancer Inst*, vol. 94, pp. 749-54, May 15 2002.

[25] I. Porebska, E. Wyrodek, M. Kosacka, J. Adamiak, R. Jankowska, and A. Harlozinska-Szmyrka, "Apoptotic markers p53, Bcl-2 and Bax in primary lung cancer," *In Vivo*, vol. 20, pp. 599-604, Sep-Oct 2006.

[26] J. C. Bergh, "Gene amplification in human lung cancer. The myc family genes and other proto-oncogenes and growth factor genes," *Am Rev Respir Dis*, vol. 142, pp. S20-6, Dec 1990.

[27] F. Ciardiello and G. Tortora, "Interactions between the epidermal growth factor receptor and type I protein kinase A: biological significance and therapeutic implications," *Clin Cancer Res*, vol. 4, pp. 821-8, Apr 1998.

[28] T. Kojima, K. Yamazaki, Y. Tamura, S. Ogura, K. Tani, J. Konishi, N. Shinagawa, I. Kinoshita, N. Hizawa, E. Yamaguchi, H. Dosaka-Akita, and M. Nishimura, "Granulocyte-macrophage colony-stimulating factor gene-transduced tumor cells combined with tumor-derived gp96 inhibit tumor growth in mice," *Hum Gene Ther*, vol. 14, pp. 715-28, May 20 2003.

[29] N. Shinagawa, K. Yamazaki, Y. Tamura, A. Imai, E. Kikuchi, H. Yokouchi, F. Hommura, S. Oizumi, and M. Nishimura, "Immunotherapy with dendritic cells pulsed with tumor-derived gp96 against murine lung cancer is effective through immune response of CD8+ cytotoxic T lymphocytes and natural killer cells," *Cancer Immunol Immunother*, vol. 57, pp. 165-74, Feb 2008.

[30] S. Singhal, R. Wiewrodt, L. D. Malden, K. M. Amin, K. Matzie, J. Friedberg, J. C. Kucharczuk, L. A. Litzky, S. W. Johnson, L. R. Kaiser, and S. M. Albelda, "Gene expression profiling of malignant mesothelioma," *Clin Cancer Res*, vol. 9, pp. 3080-97, Aug 1 2003.

[31] Z. Ma, S. L. Gibson, M. A. Byrne, J. Zhang, M. F. White, and L. M. Shaw, "Suppression of insulin receptor substrate 1 (IRS-1) promotes mammary tumor metastasis," *Mol Cell Biol*, vol. 26, pp. 9338-51, Dec 2006.

[32] V. D'Alessandro, L. A. Muscarella, M. Copetti, L. Zelante, M. Carella, and G. Vendemiale, "Molecular detection of neuron-specific ELAV-like-positive cells in the peripheral blood of patients with small-cell lung cancer," *Cell Oncol*, vol. 30, pp. 291-7, 2008.

[33] N. Dioufa, E. Kassi, A. G. Papavassiliou, and H. Kiaris, "Atypical induction of the unfolded protein response by mifepristone," *Endocrine*, vol. 38, pp. 167-73, Oct 2011.

[34] M. Fabbri, R. Garzon, A. Cimmino, Z. Liu, N. Zanesi, E. Callegari, S. Liu, H. Alder, S. Costinean, C. Fernandez-Cymering, S. Volinia, G. Guler, C. D. Morrison, K. K. Chan, G. Marcucci, G. A. Calin, K. Huebner, and C. M. Croce, "MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B," *Proc Natl Acad Sci U S A*, vol. 104, pp. 15805-10, Oct 2 2007.