RATIONALE FOR TEXT MINING RESOURCE FOR LITERATURE ON ARSENIC IN FOODS

Matilda O. Johnson¹, Hari H.P. Cohly², Matthew N. Anyanwu², Raphael D. Isokpehi² and Omotayo R. Awofolu¹

¹Department of Environmental Sciences, University of South Africa, PO Box 392, Pretoria 0003, South Africa
²Center for Bioinformatics & Computational Biology, Department of Biology, Jackson State University, Jackson Mississippi 39217, USA

Abstract: Arsenic is a toxicant derived from the natural environment that has been associated with complex diseases including cancers and diabetes. Despite the importance of dietary sources of inorganic arsenic toxicity, there are currently limited electronic resources that curate arsenic-food relationships. The majority of applications of text mining in biomedical research as focused on gene and protein annotation from diverse biological systems. In the domain of arsenic biomedical research, the Comparative Toxicogenomics Database provides a resource to understand arsenic-gene, arsenic-disease relationships in animal systems. We hereby identify the need for an Internet resource on Arsenic in Foods. A component of the resource would include tools for natural language processing of PubMed and AGRICOLA abstracts as well as Full-Text articles. Additional components proposed are annotated corpora of collection of sentences, full text articles and abstracts on arsenic in foods. We envisage that the resource will be useful for educational and research purposes in the domain of dietary exposure to arsenic, food contamination and risk assessment. As an initial step, in this article, we (i) briefly discuss the sources of arsenic poisoning; (ii) provide a general overview of text mining and its efficacy in mining scientific literature; and (iii) present and demonstrate the importance of mining data for arsenic levels in various foods using text mining applications to elucidate and unravel the emerging knowledge of the importance of arsenic contents of foods. Finally, we make the case for a web resource on Arsenic in Foods to facilitate natural language processing of scientific literature on arsenic in foods.

Keywords: Arsenic, Foods, Information Extraction, PubMed, Text Mining

Acknowledgments: Pittsburgh Supercomputing Center's National Resource for Biomedical Supercomputing (T36 GM008789); U.S. Department of Homeland Security Science & Technology Directorate (2007-ST-104-000007; 2009-ST-062-000014; 2009-ST-104-000021); and Research Centers in Minority Institutions (RCMI) – Center for Environmental Health (NIH-NCRR G12RR13459). Disclaimer: “The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.