INFORMATION EXTRACTION ON SALMONELLA PATHOGENESIS AND SALMONELLA-TOMATO INTERACTIONS FROM BIOMEDICAL LITERATURE

Christina Bernard1,2, Nyasha Chambwe2, Tolulola Oyeleye2, Hari H.P. Cohly2, and Raphael D. Isokpehi2

1Science Education Partnership Award (SEPA) Program Jackson State University, Jackson, MS 39217, USA
2Center for Bioinformatics & Computational Biology, Department of Biology, Jackson State University, Jackson, MS 39217, USA

Abstract: *Salmonella* is a group of bacteria that inhabit the intestines of humans and animals, and some species cause diarrhea in humans. The most common types of *Salmonella* bacteria found in the United States are *Salmonella* serotype typhimurium and *Salmonella* serotype enteritidis. Salmonellosis usually lasts between four and seven days causing diarrhea, fever and abdominal cramps in its host. It is very serious in the elderly, infants and people with impaired immune systems. The infection is usually transmitted through contaminated food. *Salmonella* serotype Saintpaul was recently associated with salmonellosis linked to contaminated tomatoes, jalopenos and cilantro. This serotype is not very common, and there is limited information available regarding it. Our objective was to construct a corpus of PubMed abstracts on *Salmonella* and then use information retrieval and extraction techniques to rank abstracts that contain pathogenesis descriptors. As of June 17, 2008, 57,734 abstracts (289,923 sentences) with *Salmonella* as the Medical Subject Heading (MeSH) were retrieved from PubMed. Scores are determined by the frequency of a word descriptor of pathogenesis mentioned in the abstract. In this pilot study, scores for two pathogenesis descriptors; invasion and colonization were determined. In addition, scores were determined for the keyword “tomato” to gain insights into salmonella-tomato interactions. The scores for abstracts ranged from 1 to 14. Comparison of abstracts containing descriptors of early events in pathogenesis revealed 781 and 613 prioritized abstracts for invasion and colonization respectively. Furthermore, a prioritized set of 104 abstracts on studies on *Salmonella* and tomato were identified. At the sentence level, 1661, 1173 and 349 sentences were prioritized for invasion, colonization, and tomato keywords respectively. The sentence count per abstract text ranged from 1 to 54. Future research will include analyses of additional descriptors and evaluation of tokens in the sentence for biological named entities such as gene names and cell types to gain insights into host-pathogen interactions. The *Salmonella* corpus of PubMed abstracts can be searched via a Textpresso-powered text mining engine at http://compbio.jsums.edu/textpresso.

Acknowledgements: Science Education Partnership Award (SEPA) Program at Jackson State University; U.S. Department of Homeland Security under Grant Award Number 2007-ST-104-000007 (Bioinformatics in Biodefense Career Development Program); Mississippi NSF-EPSCoR “Innovations through Computational Sciences” Award (EPS-0556308); Research Centers in Minority Institutions (RCMI) – Center for Environmental Health (NIH-NCRR G12RR13459-09). 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.