



CSHALS Conference on Semantics in Healthcare and Life Sciences
February 22 – 24, 2012
Royal Sonesta Hotel Boston, Cambridge, MA, USA.

Semantic integration to characterize microbial pathogens: Multi-resource enrichment of experimental proteomic and genomic datasets

Erich Gombocz¹⁾, James Candlin²⁾

¹⁾ IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, USA

²⁾ SAGE-N Research, 1525 McCarthy Blvd, Suite 1000, Milpitas, CA 95035, USA

Correspondence: egombocz@io-informatics.com

Topic Area: **Applications to Emerging Health Disciplines**

SUMMARY (385 words)

Bacterial and viral-caused infectious diseases account for major health threats globally, yet the characterization, identification and understanding of them has been scientifically challenging. This is mainly due to the fact that while there is a wealth of information (and even complete genomes) available, its integrated utilization in context of the biological system to better understand causes and similarities in infectious diseases is still in its infancy. This work tries to address some of the many obstacles involved in this endeavor as it attempts to identify peptides from different microorganism with common mechanism of actions causing disease, and to use them as biomarkers to detect pathogenic microbial threats prior to onset of disease symptoms to help in outbreak prevention.

The workflow to accomplish this goal consist of 5 steps. The first step is a thorough peptide analysis of microorganism via mass spectrometry and their identification by sequence scoring (Sorcerer, indexed SEQUEST search, BioWorks). The second step is the annotation of peptides with genes and genomic sequences relevant to protein expression to qualify the accuracy of the identification. Step 3 involves the use of public domain microbial databases (PATRIC, ICTV, VIDA, Viral ORFeome, miRBase) to semantically integrate the experiments with organism taxon-specific functional genomic and pathway information relevant to diseases caused by the pathogens. Based on sequence similarity, sequences are clustered into homologous protein families (HPFs), and those protein families are enriched with annotations including functional classification, related protein structures, taxonomy, protein length, boundaries of conserved regions and bacterial or virus-specific genes. Further enrichment is achieved through addition of disease-related pathways (BioCyc, KEGG). The resulting knowledgebase provides a network with functional annotations to peptides and their relationships to diseases (Sentient Knowledge Explorer). In Step 4, those peptides in the network are identified which have similar disease-causing functions and appear in several pathogens. Interrogating the network via semantic queries (SPARQL) results in discovery of key pathway intersections commonly involved in the disease. The last step is the creation of molecular marker signatures (SPARQL, Applied Semantic Knowledgebases - ASK) and test their validity as decision support in multiplexed assays.

Future applications will apply this technology for rapid detection of biological threats, to characterize origin and type of disease outbreaks and to develop preventive measures (such as broadly applicable drugs or vaccines) effective for entire classes of pathogenic organism.

2 Tables, 3 Figures