Application of natural language processing (NLP) to metabolomic/lipidomic data for new knowledge discovery from existing scientific literature

Aliakbar Panahi[1]; Samuel Henry[1]; Daniel Contaifer[1]; Bridget T McInnes[1]; Dayanjan Wijesinghe[2]
[1]Virginia Commonwealth University, Richmond, VA;[2]Virginia Commonwealth University, Richmond, VA

Introduction

Advances in mass spectrometry allows quick and accurate identification of metabolites that characterize diseases. However, transfer of this information to new and actionable knowledge with respect to the disease is not a trivial matter. This is especially the case considering the current pace of scientific publications. In this regards, the ability to intelligently query all of scientific knowledge with respect to the metabolic findings has the potential to provide new insights to the disease under investigation. Such an approach will also enable the verification of findings and easily identify new metabolic pathways previously undiscovered. Here we demonstrate the use of natural language processing to gain new insights towards cardiac arrest based on the results of metabolomic and lipidomic analysis.

Methods

Plasma samples (n=28) were collected from cardiac arrest patients at time of arrival to the hospital post resuscitation and following therapeutic hypothermia target body temperature. Lipids and water soluble metabolites were extracted from these samples followed by analysis with untargeted lipidomic and metabolomic approaches. The resultant data were normalized and the top most significant lipids and metabolites were identified for further processing using natural language processing. This was undertaken using a literature-based discovery ABC co-occurrence model, where B terms were limited to the metabolites identified, and C terms to diseases. The output diseases were hierarchically clustered and ranked using linking term count. The resultant data were visualized as a graph using Gephi for easy navigation and interpretation.

Preliminary Data

The lipids and metabolites phosphocholine, ceramide, histidine, acylcarnitine, sphingomyelin, lysophosphocholine, phenylalanylphenylalanine, docosapentaenoic acid, glucosylceramide, leucine, docosadienoic acid, choline, pipecolic acid, oleoyl L-carnitine, arginine, docosahexaenoic acid, eicosatrienoic acid, eicosadienoic acid, L-norleucine, were identified as the most significant lipid and metabolic differences between arrival and target body temperature. 17 of the target metabolites were identified as co-occurring with previous studies with cardiac arrest indicating, that two of the metabolites has never before been reported in this context. Several diseases were identified as demonstrating high levels of co-occurring metabolites. Among those, Fish Eye Disease ranked high with 15 of the metabolites of interest also being reported. Investigating this disease identified deficiency in lecithin cholesterol acyltransferase (LCAT) to be the cause of Fish Eye Disease. However, very little

literary evidence was found for a direct relationship between LCAT and cardiac arrest indicating a new finding that can be verified via additional metabolite searches for LCAT abnormalities from our existing data for cardiac arrest. Such new knowledge provides hitherto unknown drug targets for treating diseases, such as modulating LCAT for treating post cardiac arrest syndrome. Similar high ranking diseases (e.g. diabetes, septicemia) were also found to have high levels of relationship to cardiac arrest via their common metabolites, confirming the presence of long suspected, but never before proven, common underlying metabolic circuits between these different diseases. The practical applications of such findings include the ability to generalize insights gained and treatments devised for one disease to others that are closely linked metabolically leading to faster translation of benchside research to clinical treatments. In summary, the findings from our study highlights the great potential for new knowledge discovery by directly coupling the output of metabolomic and lipidomic data for investigated diseases, to the entirety of existing and up-to-date scientific literature via natural language processing.

Novel Aspect

We report for the first time, the use of NLP and LBD techniques for gaining additional insights in metabolomic data.