**Scientific Review 2.0: Let the data tell!**

--Mining for associations from complex biomedical literature using deep learning

**Project description:**

Nowadays, more and more genes were discovered and studied with the encouraging research results blooming every second in the world. When we type any interest gene, we can easily find a lot of related papers online, especially some famous genes, like BCL2. Because of its important potential role in the cancer therapy, BCL2 and its family genes were studied deeply in every aspect. A lot of investigations have been performed on BCL2 family and tremendous results have been published. Even the scientific review papers, which should give the summary of the most research results regarding BCL2 family, are not in a small amount. Furthermore, indicated from the various reviews of BCL2 family, surprisingly, some research results are against each other.

How to cope with this fast-paced changing scientific world and dig into this increasing publication data pool to explore the gene functions and their relationships is a big challenge to all the biomedical researchers. As data scientists, we should shoulder this mission and develop a deep learning model for mining for associations from complex biomedical literature. This model has four goals: 1) to summarize all the past and current discoveries for the interest gene regardless of the research area. For example, some scientist studied BCL2’s signal pathway, some studied the cancer therapy potential of the BCL2, some reported its clinical trial result, etc. This model should be able to summarize all the data from past to the current moment from bench to clinical. 2) to logically develop the relationship with other genes. For example, we can explore the relationship within the BCL2 family or BCL2 gene with other oncogenes, like p53. This model should be smart enough to develop some logical relationship chart. 3) to automatically update. Researchers worldwide produced a lot of results every second. Knowledge about the specific genes grows in an incredible rate. This model should be able to present the most updated information. 4) to deal with the controversial result. For example, BCL2-xL activated BAX genes to promote MOMP was reported in one paper and BCL2-xL and BAX took effect separately in MOMP in another paper. This model should be able to pick these controversial conclusions out and reflect it in the report.

Our final goal is to manage the machine to search and summarize the related data and analyze and produce a review report for our interest automatically. It will save scientists a lot of time and become a strong helper to keep them updated with the newest development.

To achieve this final goal, Ming and I formed a team and tried to develop a trail model using BCL2 family genes as a starting point.

**Approach**:

**First step** is to train the machine to pick up the related functions for the BCL2 gene. We use the BCL2 gene itself as a key word. It will go through all the publications and pick up the BCL2 as text and sort the related phrase, for example, the top 50 words with BCL2 in one sentence. After the first step, we should have the top-50-words list.

**Second step** is to manually pick up the function word and phrase from the top-50-words list. After the second step, we should have a few function phrase linked with BCL2, for example, apoptosis or MOMP.

**Third step** is to train the machine to pick up the related gene names using the BCL2 and the specific function like apoptosis as the key word. After the third step, we should have the top-50-genes with BCL2 and apoptosis.

**Fourth step** is to pick up the BCL2 family genes from the top-50-genes list. After the fourth step, we should have the BCL2 gene, BCL2 family genes and their relationship in some specific function.

**Fifth step** is when several BCL2 family genes have same function linked to each other, we need train the machine to classify them. After the fifth step, the machine should give us some classes within same function.

**Sixth step** is when we found BCL2 and some other family genes are both present in the two controversial functions, we need to pick it up and present this question. After the sixed step, the machine will present us some doubt in the controversial discovery.