Abstract

Preeclampsia is a dangerous condition in pregnant women worldwide, which can lead to various organ damage, eclampsia or even death of the mother and the fetus. This illness has a genetic component. and different genes interact with each other to bring about preeclampsia. Much of the data on interactions can be found in various interaction databases. but most of the time they are not up to date, and most recent data can be found only in literature. However, manual reading of all the literature to find specific information is now becoming impractical due the ever expanding volume of literature.

In the current study, this problem is approached by applying text mining methods on PubMed abstracts. Relevant PubMed abstracts were retrieved and sentence boundary detection was done using GeniaSS tool. Tokenizing and entity recognition was done using GeniaTagger tool. All the data was stored in a MySQL database and Perl was used as the programming language. Gene name normalization was done using a gene name dictionary created for the current work. A sentence which has more than two unique gene symbols and at least one interaction word was selected as candidate sentence for further processing. Information was extracted by manual reading and a web site was created to present the output of the work.

457 sentences were selected as candidate sentences out of 11172 sentences. 42 sentences were found to be describing gene-gene interactions. There were 59 genes participating in 51 interactions.

This approach was evaluated using a sample of 45 abstracts which had 435 sentences. It showed a 61.5% of recall and 28.5% of precision. Current approach has shown promising results, and it has significantly reduced the manual reading work load.

Key words: Pre-eclampsia. Text mining. Gene-Gene interaction