OFFICIAL ABSTRACT and CERTIFICATION

Abstract text mining to create an exhaustive disease-disease correlation database

Suchir Misra
Jericho Senior High School, Jericho, NY, USA

Craniosynostosis, the second most common craniofacial abnormality, shares genetic mutations implicated in cancer progression, yet a correlation between the two has not been elucidated. Disease-disease correlations can assist in developing improved disease treatments, yet finding genetic loci used to establish such correlations is expensive and time-consuming. Databases enhance visualization of disease-disease correlations and disease-gene associations; however, current databases overlook rare diseases and important connections by limiting the pool of diseases studied.

A computational approach was designed to create a database of disease-disease correlations such that correlations with rare diseases could be elucidated. Python programs were written to collect a list of abstract IDs for all genetic papers related to an extensive list of diseases (N = 1857), to sort the abstract IDs numerically and remove duplicates, and to extract gene names from the abstracts. A PostgreSQL database was used to store the data for efficient querying. Disease-disease correlations were determined based on gene overlaps.

The top ten disease-disease connections overall have been previously elucidated, validating the effectiveness of the method used to create the database. Of the top ten disease-disease connections for craniosynostosis, four were newly elucidated.

In the future, publications should denote mutation percentages of genes in their abstracts so the importance of genes mutated in a disease can be considered in future iterations of the program. This study provides a tool to find genetic loci and design improved disease treatments for both rare and common diseases.

Category
Pick one only — mark an "X" in box at right

Animal Sciences	
Behavioral & Social Sciences	
Biochemistry	
Biomedical & Health Sciences	
Biomedical Engineering	
Cellular & Molecular Biology	
Chemistry	
Computational Biology & Bioinformatics	■
Earth & Environmental Sciences	
Embedded Systems	
Energy: Sustainable Materials and Design	
Engineering Mechanics	
Environmental Engineering	
Materials Science	
Mathematics	
Microbiology	
Physics & Astronomy	
Plant Sciences	
Robotics & Intelligent Machines	
Systems Software	
Translational Medical Sciences	

1. As a part of this research project, the student directly handled, manipulated, or interacted with (check ALL that apply):

   □ human participants □ potentially hazardous biological agents

   □ vertebrate animals □ microorganisms □ rDNA □ tissue

2. I/we worked or used equipment in a regulated research institution or industrial setting: ■ Yes □ No

3. This project is a continuation of previous research. ■ Yes □ No

4. My display board includes non-published photographs/visual depictions of humans (other than myself): □ Yes ■ No

5. This abstract describes only procedures performed by me/us, reflects my/our own independent research, and represents one year's work only ■ Yes □ No

6. I/we hereby certify that the abstract and responses to the above statements are correct and properly reflect my/our own work. ■ Yes □ No

This stamp or embossed seal attests that this project is in compliance with all federal and state laws and regulations and that all appropriate reviews and approvals have been obtained including the final clearance by the Scientific Review Committee.