Analysis of Biomedical Literature to Identify Protein-Heart Disease Relationships

Ashlyn Jew
Winter 2021 - STATS 199

Introduction

The heart proteome is an extensive list of proteins that are expressed in the heart by an organism. Compiling a proteome requires large experiments and many resources. There already exists various human heart proteomes that we can reference, but the proteomes do not always match and are not complete. In fact, transcriptome analysis shows that 73% of all human proteins are expressed in the heart[1], but the existing heart proteomes do not contain all of the forementioned "73% of all human proteins". We want to analyze published biomedical literature to discover proteins that are missing from existing heart proteomes.

Our goal is to reference previous experimental results to see what has been observed about those cardiac proteins in the past. We want to know which proteins are specific to the heart and what relationships exists between these proteins and heart diseases. We want to demonstrate how text mining of biomedical literature and subsequent analysis can be used to assemble a full cardiac proteome and reveal potential protein-disease relationships.

Methods

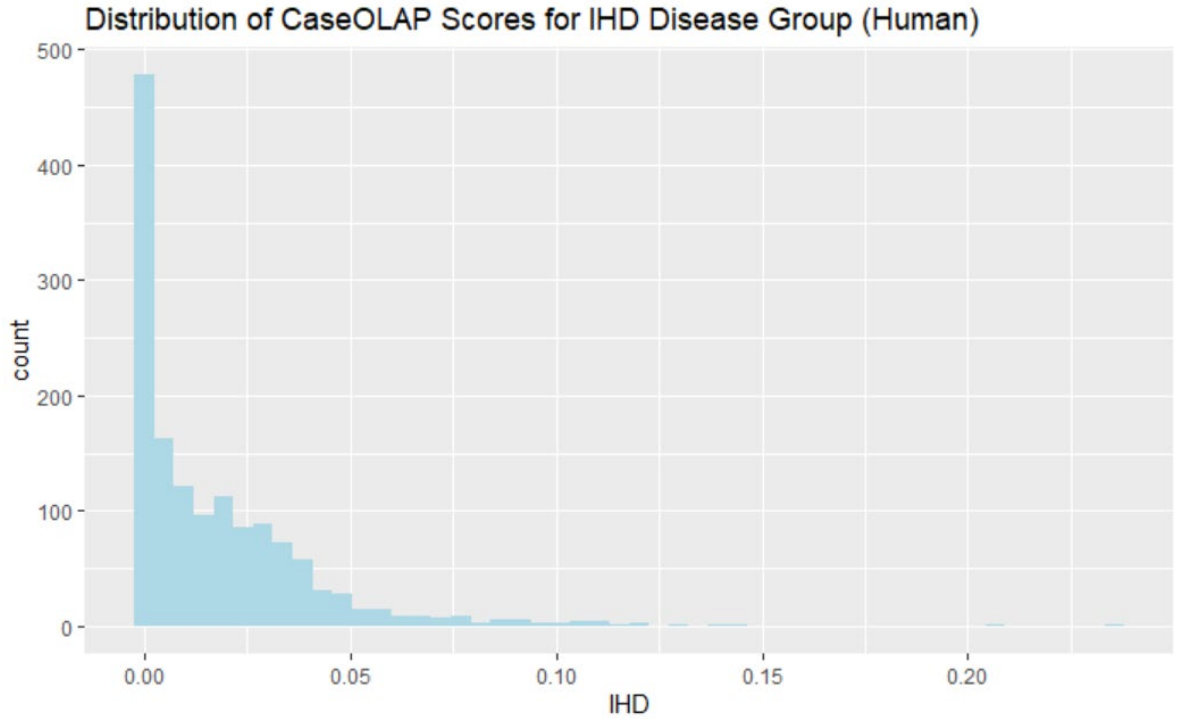
Our lab uses a text mining tool called CaseOLAP[2] to extract biological insights from large collections of literature and quantify associations between text subsets and proteins. We feed a substantial amount of cardiovascular disease-related literature from the literature database PubMed as well as sets of proteins from the UniProtKB database into CaseOLAP to get a score for each protein. CaseOLAP scores are calculated by considering three components[2]: Integrity, Distinctiveness, and Popularity. Integrity measures the quality of searched protein phrases, Distinctiveness measures how distinctive a protein is to a cardiovascular disease group, and Popularity measures the frequency of the protein in the literature collection we input. The CaseOLAP algorithm quantifies these components and generates a score that we use for subsequent analysis. The CaseOLAP score quantifies every protein's relationship with each group of cardiovascular disease that we are interested in. A score close to 0 represents a weak association, and a score close to 1 represents a strong association.

The protein sets we collected from the UniProtKB database are those of the human taxon, the mouse and rat taxon, and the pig taxon. The literature collection was obtained through the PubMed Central API with search terms related to the following eight groups of cardiovascular disease:

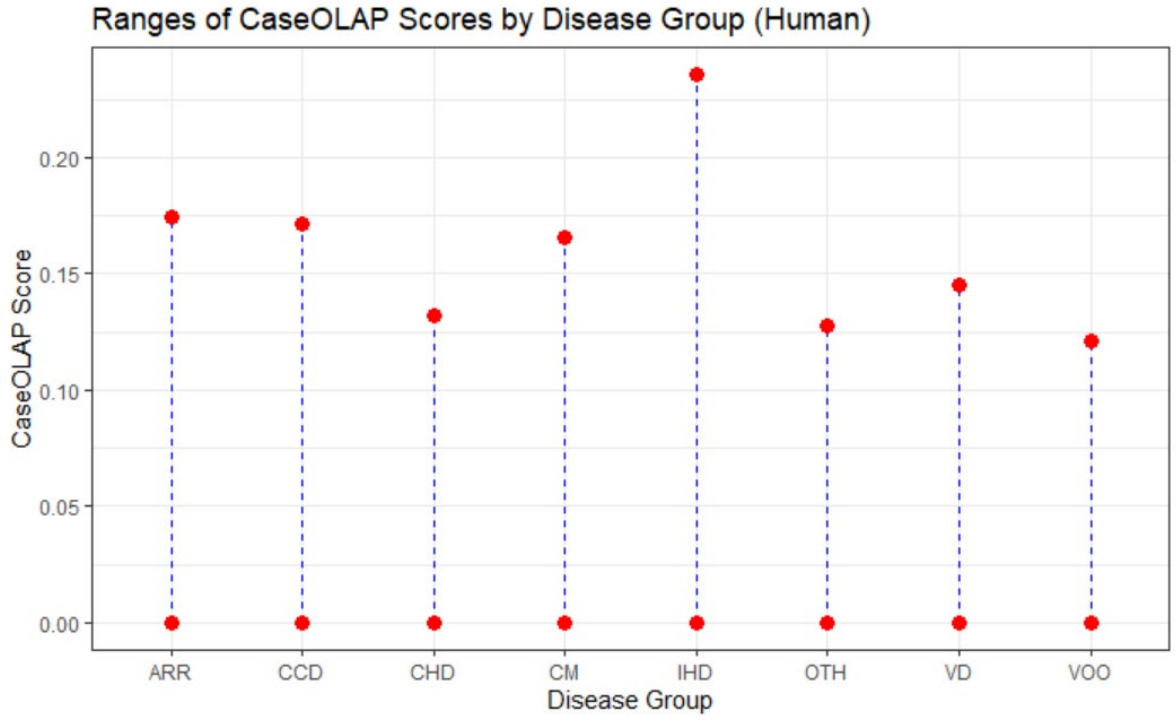
1. IHD: Myocardial Ischemia
2. CM: Cardiomyopathy
3. ARR: Cardiac Arrhythmias
4. VD: Heart Valve Diseases
5. CHD: Congenital Heart Defects
6. CCD: Cardiac Conduction System Disease
7. VOO: Ventricular Outflow Obstruction
8. OTH: Other (Cardiomegaly, Endocarditis, Heart Arrest, Heart Rupture, Ventricular Dysfunction, Pericarditis)

Results

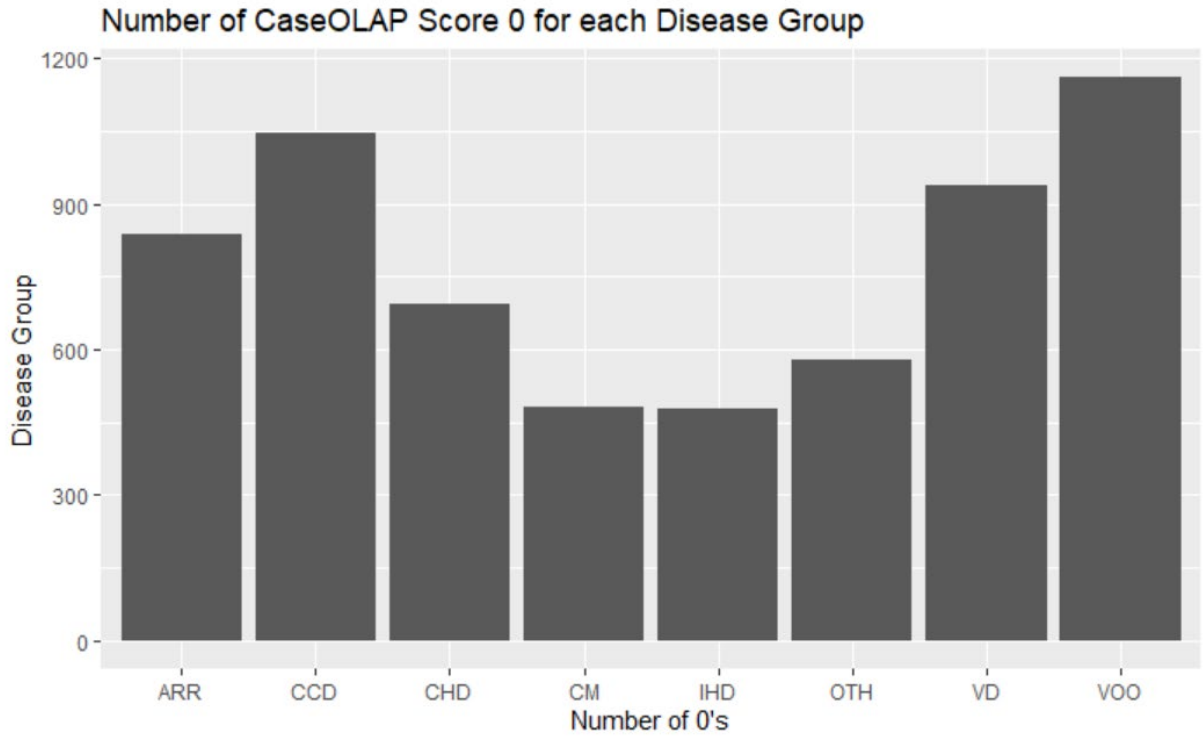
The CaseOLAP scores for each protein are heavily skewed right, with many scores equal to 0 and the highest score for the human protein set being 0.2355156.



We observe a similar distribution in all eight disease groups for the human protein set and the mouse and rat protein set.



The number of CaseOLAP scores of 0 are greater in the VOO and CCD disease groups and lesser in the IHD and CM groups. This may indicate that proteins related to Ventricular Outflow Obstruction or Cardiac Conduction System Disease are not very well-studied or mentioned in biomedical literature.



Discussion

As observed above, the CaseOLAP scores are heavily skewed right, with many scores of 0 and a maximum of less than 0.3. However, this can be expected as the PubMed literature we collected only consisted of article abstracts, so specific proteins may not be mentioned very often. Also, many published literature may be editorials, reviews, or documents that typically do not list specific proteins. Our lab is currently extracting a larger collection of literature that contains full article text, where more details such as protein names and aliases may be listed more often than in abstracts. Thus, we expect to see higher CaseOLAP scores after we input the new literature collection into CaseOLAP.

CaseOLAP scores just tells us meaningful correlation, but do not reveal why certain proteins are more associated with certain disease groups. These scores are just the starting point of discovering relevant cardiac proteins and their relationships with cardiovascular disease. We can identify which proteins and disease groups are lacking in research and which proteins may potentially be important in a disease group. We are looking for hidden proteins that may not have been listed in previous heart proteomes and are awaiting more research.

Text mining and analysis of biomedical literature may prove to be very useful as there are millions of articles that are published, and researchers are expected to be updated with the latest publications in their field. The work completed in this project can be replicated for topics other than heart disease and may prove to be useful in other contexts as well.

References

(1) "The Heart-Specific Proteome." 2021. The Human Proteome in Heart - The Human Protein Atlas. The Human Protein Atlas. Accessed January 6. https://www.proteinatlas.org/humanproteome/tissue/heart.

(2) Liem, D. A.; Murali, S.; Sigdel, D.; Shi, Y.; Wang, X.; Shen, J.; Choi, H.; Caufield, J. H.; Wang, W.; Ping, P.; Han, J. Phrase Mining of Textual Data to Analyze Extracellular Matrix Protein Patterns across Cardiovascular Disease. American Journal of Physiology-Heart and Circulatory Physiology 2018, 315 (4), H910–H924. https://doi.org/10.1152/ajpheart.00175.2018.