Research Plan

GADDNET: A Platform for Connecting Researchers via the Genes and Diseases They Studied and Will Study

Emma Yang

Adult Sponsor: Avi Ma'ayan

Category: Computational Biology and Bioinformatics

Rationale

The GADDNET platform aims to bridge the gap between genes that could be important to advancing drug discovery research and the amount of research attention those genes are receiving. Much of the human genome is not being actively studied, while well-researched genes continue to be funded and to be the focus of most investigators.

However, an investigation by Stoeger et al. [ref] on the reasons why potentially important genes are ignored found that the discrepancy is not related to the function of these genes. In fact, they found that biomedical research, especially genetics and genomics research, is guided largely by previous experimentation and the characteristics of genes rather than the physiological significance of individual genes or their connection to diseases. Researchers keep going back to relatively well-understood genes and deepening research in those areas, rather than reaching into new genes and developing an initial understanding of those genes.

This is mainly because access to information about connections between diseases and under-studied genes is not widely available. However, swaths of data are available to make computational predictions about such associations. Providing access to such predictions is one of the goals GADDNET aims to achieve.

GADDNET aggregates connections between genes and investigators that already study the diseases these genes, but also to gene-disease associations that are computationally predicted. The platform gives authors an intuitive way to access information about their own and others research focus, to allow researchers to broaden their attention to consider additional genes and drugs that they may wish to investigate. By putting this data in the hands of investigators, the platform has the potential to accelerate drug discovery research and to enable the identification of understudied drug targets. It will also help to broaden the distribution of genes that are being actively researched and expanding the pool of well-understood genes and their functions in relation to other genes, drugs, and disease mechanisms. With access to this network visualization tool, researchers have the ability to discover lesser-known genes that could be potentially become important to advancing their research and broadening the distribution of research that define gene function.

Engineering Goal & Expected Outcomes

GADDNET aims to combine data from PubMed, Geneshot and ARCHS4 (from the Ma'ayan Lab), and DISEASES (from the Jensen Lab) to create a web-based search engine that visualizes this data as a network. The expected outcome is that investigators will use the platform to discover other lesser-known genes that they may decide to study. In addition, patients and clinicians will be able to find information about a disease, who is studying it, and which genes are associated with it.

The user will be able to generate a network based on five types of queries: investigator, gene, drug, institution, and disease. An investigator query would generate a network centered around a person and will show the genes, diseases and drugs that they published based on abstracts on PubMed. In addition, predicted genes, and other authors relevant to those genes, drugs, and diseases will be added to the network. A gene query would create a network centered around a specific gene, showing the top authors who have published the most papers about those genes, the connected diseases, drugs, and predicted genes. The disease query would create a network surrounding a disease, showing genes related to the disease and the other diseases, authors, drugs, and predicted genes related to those genes.

The platform will also leverage natural language processing (NLP) to allow an investigator or a patient to make a query using a question, rather than a structured query. Integrating NLP into the platform will allow complex questions that query information from multiple sources and multiple types of queries to be combined into a single question. The question format will make it easier and more intuitive to navigate the GADDNET interface. For example, the user could ask, "What investigators are researching the BRCA1 gene?" or "What genes are predicted to be associated with Alzheimer's disease?" The platform would then generate the network from the relevant query that answers the user's question.

Procedures and Data Analysis

GADDNET was developed around the design of a search engine that would allow the user to search the data available through the platform from multiple different perspectives: searching for investigators, genes, drugs, and diseases. The network design had to be interactive so that the user could explore the network to find information pertinent to their interests. Because of the magnitude of the dataset and the potential scale of the results of many queries, the web app had to be scalable and optimized to prevent slowdown when the user interacted with it.

GADDNET is developed with React.js, with a server written in Node.js. The server is used to access a MySQL database containing the different datasets such as the DISEASES dataset from the Jensen Lab, PubMed data from the PubMed API, and a list of investigators who have published papers about genes, their Open Researcher and Contributor ID (ORCiD) IDs using the ORCiD API. React.js supports the front-end interface of the platform. Data on genes predicted to be relevant by co-expression was sourced from the Geneshot API. The network was created using the d3.js library.

Data Collection:

All of the data sourcing was rooted in a list of investigators who had conducted and published genetics and drug discovery research on PubMed.

The back-end server “keeps track” of the investigators using their ORCiD numbers, a unique identification code for scientific and academic authors and contributors.

The names of the investigators were also used to search PubMed for publications that each investigator has published. The PubMed articles found from the PubMed API for each author was mapped to the gene that they primarily discussed using the GeneRIF dataset from the National Center for Biotechnology Information (NCBI). Most authors have researched and published work on a gene multiple times, so the publication count for each gene was also included in the resulting dataset.

Disease-gene associations were sourced from the DISEASES dataset, which was created and is maintained by the Jensen Lab. Geneshot provides the co-expression gene data for the platform.

Constructing the Network:

The data is visualized so that it can be displayed by the d3 network. After a user submits a search term and the relevant SQL and API queries are made through the server, the resulting collection of data is parsed based on where it came from: disease dataset, predicted genes, or genes the author worked on. The color of the node representing each row of data depends on what type of data it is. Gene data, disease data, and author data all have different colors. The data is parsed so that it is organized as a collection of nodes and links, and is displayed in the user interface.

Natural Language Processing:

The natural language processing in GADDNET is powered by the Google Natural Language API (NLP API). When the user submits a query, GADDNET sends the question to the NLP API. The API parses out which words are pertinent to the query that will produce the network that answers the user's question. For example, if the user asked, "Who is working on BRCA1?," the NLP API would return the word "BRCA1" as it was the subject of the question.

GADDNET then searches through each database to find out whether the word the NLP picked out was a person, gene, or disease. Based on what the subject of the question was, the search engine then uses the server to make the corresponding query and to display the network that answers the user's question.

Risk and Safety

There are no safety risks associated with this project.

Bibliography

Facebook Inc. (2018, June 26). Create a New React App. Retrieved November 12, 2019, from React website: https://reactjs.org/docs/create-a-new-react-app.html

Google Cloud. (2019, April 10). Natural Language. Retrieved December 12, 2019, from Google Cloud website: https://cloud.google.com/natural-language/

Kamenzky, N., & Abonil, T. (2019, November 4). Request-Promise. Retrieved December 5, 2019, from NPM website: https://www.npmjs.com/package/request-promise

Lachmann, A., Schilder, B. M., Wojciechowicz, M. L., Torre, D., Kuleshov, M. V., Keenan, A. B., & Ma'ayan, A. (2019). Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research*, *47*(W1), W571-W577. https://doi.org/10.1093/nar/gkz393

Lerner, A. (2016, November). Introduction to Promises. Retrieved December 9, 2019, from FullStack React website: https://www.newline.co/fullstack-react/30-days-of-react/day-15/

National Center for Biotechnology Information. (2011, May 27). Retrieve PMC article identifiers (PMCIDs) from a search. Retrieved December 10, 2019, from PubMed Central website: https://www.ncbi.nlm.nih.gov/pmc/tools/get-pmcids/

Oprea, T. I., Bologa, C. G., & Brunak, S. (2018). Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery*, *17*, 317-332. https://doi.org/10.1038/nrd.2018.14

ORCiD. (2019, October 3). Basic Tutorial: Searching Data Using the ORCiD API (3.0+). Retrieved December 10, 2019, from Member Support Center website: https://members.orcid.org/api/basic-tutorial-searching-data-using-orcid-api-30

NO ADDENDUMS EXIST