Community Platform for Data Wrangling of Gene and Genetic Variant Annotations

SCRIPPS RESEARCH INSTITUTE

PI: WU, CHUNLEI

Grant Number: 1 U01 HG008473-01

Biomedical knowledge is often summarized and structured in the form of annotations of biological entities such as genes, genetic variants, diseases, and pathways. These annotations are fragmented across dozens of data repositories like NCBI Entrez, Ensembl, UniProt, and hundreds (or more) of other specialized databases. While the volume and breadth of annotations is valuable, their fragmentation across many data silos is often frustrating and inefficient. Bioinformaticians everywhere must continuously and repetitively engage in data wrangling in an effort to comprehensively integrate knowledge from all these resources, and these uncoordinated efforts represent an enormous duplication of work. The problem of fragmentation is exacerbated (perhaps even fundamentally caused) by the inability of data providers to efficiently contribute to existing repositories. As a result, annotaion providers must generate new resources in order to host newly-generated annotations that are unavailable in the central repositories. In this proposal, we will create a hybrid solution that combines the high performance of a centralized system with the flexibility and breadth of a federated system. The centralized component will provide high-performance computational infrastructure for the integration, query and access of biological annotations. The technical design of this component will be based on our successful MyGene.info web services (://mygene.info). The federated component builds on our extensive background in crowdsourcing. We will build community infrastructure that allows the small- and medium-scale data wrangling that is already being performed (and repeated) by many scientists to be aggregated into a single big-data resource. Additionally, semantic interoperability will be added to our system to ensure that it will integrate with current and future Linked Data applications. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: A primary challenge in the biomedical Big Data era is that the vast amount of scientific discoveries outpaces the traditional efforts of structuring them in a computable form. Successful completion of this work will result in a platform to harvest structured data from individual researchers directly, and speed up biomedical research with this aggregated community intelligence.