Vision Statement

In response to the NIH FOA OTA-19009 “Biomedical Translator: Development” we propose to build an Autonomous Relay Agent (ARA) that can characterize and rate the quality of information returned from multiple multiscale heterogeneous knowledge providers (KPs).

Biomedical researchers develop a trust relationship with a knowledge provider (KP) through frequent and continued use. Over time a familiarity develops that drives their understanding and insight on 1) how to structure and invoke more effective queries, 2) the quality of the results they may expect in response to different query parameters and feature values, and 3) how to assess the relevancy of a specific query’s results.

Although this information retrieval paradigm has served the research community moderately well in the past it is not scalable and the number, scope and complexity of KPs is increasing at a dramatic pace (1,613 molecular biology databases reported as of Jan. 2019). Within this ever changing information landscape, a biomedical researcher now has two choices -- either continue using the few KPs they have learned to trust but remain limited in the actionable information they will receive, or invest the time and accept the risk of using a range of new information resources with little or no familiarity and thus uncertain effectiveness. If researchers are to benefit from the vast array of NIH and industry sponsored information assets now available and expanding new information retrieval and quality assessment technologies will be required.

We propose to build an Explanatory Autonomous Relay Agent (xARA) that can characterize query results by rating the quality of information returned from multi-scale heterogeneous KPs. The xARA will utilize multiple information retrieval and explainable Artificial Intelligence (xAI) strategies to perform queries across multiple heterogeneous KPs and rank their results by quality and relevancy while also identifying and explaining any inconsistencies among databases for the same query response. To deliver on this promise, we will utilize case-based reasoning and language models trained with biomedical data (i.e., BioBERT and custom annotation embeddings through Reactome and UniProt) permitting a new level of query profiling and assessment.

Our strategies will permit 1) information gaps to be filled by testing alternative query patterns that produce different surface syntax yet possess semantically related and actionable concepts, 2) inconsistencies to be identified for a given query feature value, and 3) the identification and elimination or merging of semantically redundant query results via similarity metrics enriched by case-based reasoning strategies employed in the explainable AI (xAI) community to identify machine learning model behavior and performance.

The xARA capabilities proposed herein will be based on strategies developed in Dr. Weber’s lab for information retrieval where the desire for greater transparency when reasoning over experimental data is our primary aim. Our multi-institutional team is comprised of senior researchers and software engineers formally trained and experienced in the computer and data sciences, cheminformatics, bioinformatics, molecular biology, and biochemistry.

Inherent risks in querying heterogeneous KPs include the presence of inconsistent labeling of the same biomedical concept within unique KP data structures. Manual engineering may be necessary to overcome such hurdles, but will not be a significant challenge for the initial prototype, since only two well documented KPs are being evaluated. Another noteworthy risk is that the quality of word embeddings generated from UniProt and Reactome may not be sufficient, requiring further textual analysis of biomedical text like PubMed, which is feasible within the timeframe of our project plan.