# Sage Bionetworks DBP

Every year a wealth of new information technology product launches is fueled by Moore’s Law increases in compute power. In the pharmaceutical industry, however, the exponential growth of biological “omics” data has occurred concurrently with a decline in the number of NMEs approved by the FDA. Sage Bionetworks believes that a fundamental reason biological research productivity does not scale with biological data generation is that the analysis and interpretation of genomic data remains largely an isolated, individual activity. Pharmaceutical R&D pipelines, and even many pre-commercial research programs, consist of a series of linear handoffs between isolated scientists with different areas of expertise. The ability to access, understand, and reuse data, analysis methods, and models at each step is a rate-limiting step in research progress and in developing new cures and treatments for human disease.

Sage Bionetworks’ mission is to catalyze a cultural transition from the traditional single lab, single-company, and single-therapy R&D paradigm to a model with broad precompetitive collaboration on the analysis of large scale data in medical sciences. For this to happen it is critical that: 1) human health data become accessible and reusable by people other than the original data generators allowing multiple parallel approaches to data interpretation; 2) analytical methodologies become fully reproducible and transparent so that results can be vetted and existing analysis techniques quickly applied to new application areas, and; 3) models of biological systems and networks be opened to a variety of users such that theoretical predictions can be rapidly validated experimentally and improve standards of care for patients. Sage Bionetworks is actively engaged with academic and pharmaceutical collaborators in developing technical, policy and strategic solutions to these issues. Part of Sage’s solution is Synapse, a platform for open, reproducible data-driven science, which will provide support for Sage’s research initiatives and serve as a resource for the broader scientific community.

Our collaboration with the NCBO via the Sage Bionetworks DBP will support two specific aims:

**Aim 1: Embed NCBO technology throughout Synapse to facilitate curation, discovery, analysis, and reuse of Sage-hosted global coherent data sets and network models.**

**Aim 2: Use gene set enrichment analysis to dissect relevant substructures in biological networks.**

Synapse will support the reusability of information facilitated by ontology-based services and applications directed at scientific researchers and data curators. Sage Bionetworks is actively pursuing the acquisition, curation, statistical quality control, and hosting of human and mouse global coherent datasets for use by Sage Bionetworks researchers, collaborators, and the broader research community. Global coherent datasets integrate both clinical phenotype and genomic data along with an intermediate molecular layer such as gene expression or proteomic data. Typically studies contain genome-wide genetic variation data (typically SNP and/or CNV) and/or expression profiling (typically mRNA microarray). We expect the release of these sorts of unique, integrative, high value datasets into the public domain will seed a variety of analytical approaches to drive new treatments based on better understanding of disease states and the biological effects of existing drugs. Leveraging semantic technologies in the data curation and hosting processes is needed to provide an environment where researchers can easily find and reuse existing data for new purposes.

Reproducible analysis and disease model reuse require a supporting informatics infrastructure. In the Sage Bionetworks system, users interact with resources via a number of mechanisms depending upon their interests and expertise. The Synapse web portal will be an environment for end user scientists to interact and share data, models, and analysis methods, both in the context of specific research projects, and broadly across otherwise disparate projects. Many other specialized scientific tools can be extended to load data and save results to the Sage Bionetworks platform, or to perform analysis by calling methods executed on a remote service. The Sage Bionetworks Platform is aggressively leveraging and optimizing its architecture to take full advantage of the rapidly maturing cloud computing technologies which will put on-demand supercomputing power in the hands of the average researcher. These more specialized analytical clients would support use cases in data curation and quality control as well as scientific analysis.

Understanding how network structures relate to disease and response to treatments is a core strength of research at Sage Bionetworks. Gene set enrichment (GSE) analysis is an important technique for understanding broader trends in high-throughput gene expression data, such as pathway activity, cellular processes, or disease states. An underlying premise of GSE is that the concerted action of a set of genes is more closely aligned with molecular function and activity than any single gene, and is more likely to result in a stable and interpretable summary of the data. Using NCBO ontological framework, we will have access to a more diverse collection of gene sets, as well as the potential to derive new gene sets that are more relevant to different biological or clinical domains. A second advantage will be more precise filtering of gene sets to those that are relevant to the domain of study. This filtering reduces the number of statistical tests and can help ameliorate the challenges imposed by multiple hypothesis correction. Sage Bionetworks has several active studies and collaborations that will be utilizing GSE techniques. Among these is collaborative project on human aging with Stanford, UCSF, Columbia and UCSD where we collect large human data sets and generate predictive models of aging.

The Sage Bionetworks DBP will employ a variety of NCBO technologies to deliver functionality to our users as part of the DBP:

* NCBO type-ahead and ontology query services to support free-text search of data, tools, and models hosted by Synapse
* Ontology services to support tools for data curation, including collaborative development of new mechanisms for creating feedback loops between Sage Bionetworks data curators and ontology developers.
* Use of NCBO-hosted ontologies to structure data set annotations and clinical trial data to facilitate data discovery and meta-analysis
* Use of the annotator service for indexing free text descriptions of datasets and other Sage Bionetworks platform resources
* Exposure of Sage Bionetworks content to the NCBO resource index
* Use of NCBO services to support GSEA by scientific research teams
* Scale testing NCBO technology in public cloud environment