The Hypothesis Web

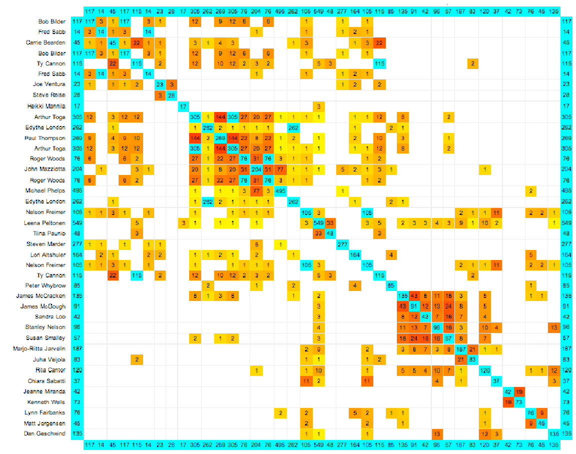
D.S. Parker, PI
Wesley W. Chu, Co-PI
NIH Grant 1 RL1 LM009833-01
http://www.hypothesisweb.org

Project Goals

The central goal of the Hypothesis Web project within the Consortium for Neuropsychiatric Phenomics is to aid in the development of interdisciplinary hypotheses spanning multiple disciplines of neuroscience. The Hypothesis Web itself is to be a software platform that permits collaborative formulation of complex scientific hypotheses, so that:

- hypotheses are the focus of collaboration.
- literature, data, and annotations are combined within a single system.
- interdisciplinary research groups work together to develop a web site for a given hypothesis that organize a group's accumulated evidence.
- resulting hypothesis web sites can be shared, modularized, and published.

This platform is to be used for collaborative development of hypotheses within the other CNP projects, particularly as regards the development of phenomics as a discipline. This is an important part of the effort because understanding genotype-to-phenotype associations in neuroscience can require a more holistic viewpoint. This is particularly true for the complex neuropsychiatric syndromes under study at the CNP.



Sample output from PubAtlas, showing strengths of associations between CNP researchers as manifested in the PubMed literature.

Progress over the initial period November 2007 – May 2008

Our initial progress for the Hypothesis Web effort has been substantial. We have progressed in four areas:

- Development of *PubAtlas*, a new tool (and website) that permits high-level views of the PubMed literature, giving perspective on key concepts and their associations.
- Initial implementation of hypothesis models in *PubAtlas*, comprising multi-level maps of associations among concepts based on co-occurrence statistics from the scientific literature.
- Development and refinement of controlled vocabularies (lexica) for syndromes, symptoms, cognitive concepts, cognitive tasks, and neural systems.
- Initial implementation of the *Phenowiki*, an on-line system for collaborative annotation of scientific literature, enabling representation of cognitive concept and task characteristics, and the associations of the concepts with other levels of knowledge.

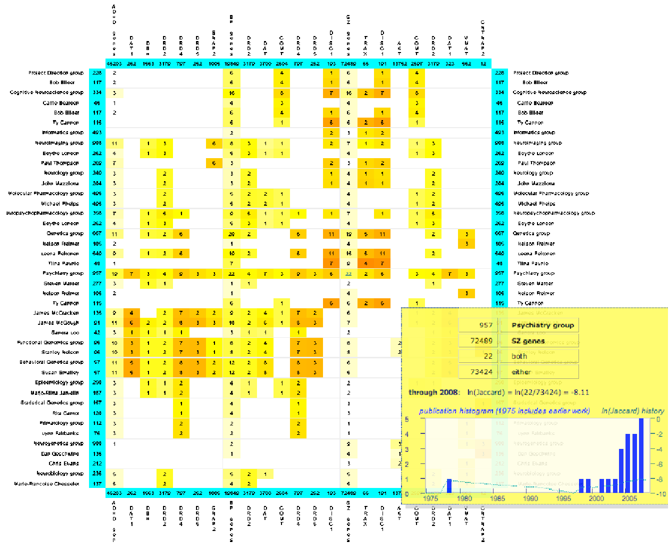
Relative to the goal of the Hypothesis Web in enabling and monitoring interdisciplinary activity and joint development of hypotheses and annotations, it is important to track associations between the many different levels of interdisciplinary science. All these areas of our progress contribute significantly toward this goal.

PubAtlas

Over this period we developed PubAtlas (http://www.pubatlas.org), a new web service that provides a visual interface to PubMed. In some sense PubAtlas is a "BLAST for PubMed", extending the PubMed interface to give a high-level view. The design emphasizes MeSH and can be extended in many ways. It was also designed to live up to the "usability" promise we have made. For example, the figure above shows the result produced by PubAtlas for co-occurrence in the PubMed literature of CNP researchers with one another. The entries in the (symmetric) table contain links that run the appropriate PubMed query needed to obtain the indicated publications. Color intensities reflect the degree of co-occurrence; stronger association is reflected by greater red intensity.

Several unique features of PubAtlas include its ability to find associations between two sets of queries (such as CNP reearchers and CNP genes of interest), handle hierarchies of queries (such as groups of researchers or taxonomies of genes), and present the entire history of some literature association (co-occurrence of queries) over decades. An example, showing the history of publications by the CNP psychiatric team for genes that have been raised as potentially related to Schizophrenia, is shown below. Currently PubAtlas is fully operational and openly available on the Web, but we have not publicized or promoted it, as we wish first to explore possibilities of its adoption by NCBI.

Literature Mining



Associations between CNP researchers and genes, with a histogram showing the history of work by Psychiatry researchers for Sz genes.

We have made progress in representation of semantic knowledge used in CNP hypotheses. The semantic network for the neuropsychiatric domain in the UMLS is inadequate, motivating us to develop the ontology for cognitive phenotypes. With CNP domain experts we constructed a lexicon of 900 relatively concise terms for cognitive phenotypes. We then used statistical models (LSI, PLSA, and HTMM) to extract concepts from PubMed in the CNP focus areas of Memory Mechanisms and Response Inhibition and cluster similar concepts of a given domain. Our domain experts are currently evaluating this concept clustering.

Scientific researchers tend to seek answers to high-level (conceptual) scenario-specific queries, and terms used at this level can differ significantly from those one would need to use in a PubMed query. One of the applications of the semantic network of cognitive phenotypes is to use knowledge to expand a user query into more scenario-specific terms that match the relevant literature. For example, “which cognitive tasks are related to cognitive control?” would arise in scenarios seeking to relate cognitive function and tasks (tests measuring phenotypes in neuropsychiatric research). Clearly, knowing that this is the scenario/context behind a query permits greater precision in query expansion (expanding the query to retrieve documents from PubMed). Jointly with CNP domain experts, we have identified three important types of scenario-specific queries: 1) connecting cognitive functions and cognitive tasks; 2) connecting neuroanatomy and cognitive tasks; and 3) connecting cognitive functions and neuroanatomy. We have implemented preliminary query expansion for these scenarios. For a given set of domain specific documents, based on the grouping of the terms in the semantic network, we are able to generate query expansions based on TFIDF (product of term frequency (TF) and inversed document frequency (IDF) statistics) for these terms from the given set of selected documents. For each scenario-specific query, e.g., related tests of cognitive control, we computed the TFIDF weights of cognitive control with every test defined in the semantic network and then rank the “tasks” based on their corresponding weights with cognitive controls. The top-*k* most similar “tasks” will be expanded, and the documents containing cognitive control and any of these tests will be returned.

PhenoWiki

Toward extending the Hypothesis Web as an interactive web-based tool that enables interdisciplinary interaction not only within our Consortium but by the broader scientific community, we have implemented PhenoWiki, a site for collaborative development of phenomics information. PhenoWiki combines a Wiki with relational database elements to enable annotation of literature for cognitive phenotypes using flexible and structured site. A summary of this work has been recently published.

Publications

Sabb FW, Bearden CE, Glahn DC, Parker DS, Freimer N, Bilder RM, A collaborative knowledge base for cognitive phenomics, *Mol Psychiatry* 2008 Apr; 13(4): 350-60. [Epub Jan 8. PMID: 18180765.]

Parker DS, Poldrack RA, Sabb FW, Bilder RM, Quantologies, under review at *Bioinformatics Journal.*

Plans

Over the next period we plan to work on query expansion, providing more effective means for CNP researchers to access PubMed and other biological information sources. This requires both a more complete ontology and new mechanisms for interpreting and responding to user queries in PubAtlas. PubAtlas will also be extended in many ways, towards living up to its name (a system for producing sets of related literature maps). We will also work more closely with other CNP projects on representation of hypotheses.