The Hypothesis Web

D.S. Parker, PI
Wesley W. Chu, Co-PI

NIH Grant 1 RL1 LM009833

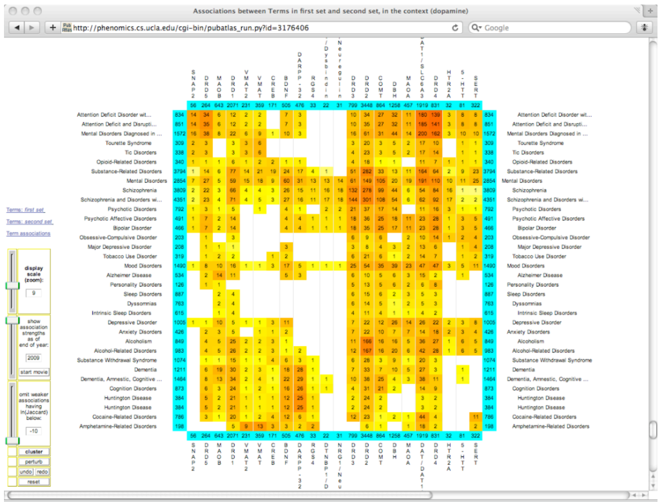
http://www.hypothesisweb.org

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 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.

The Hypothesis Web 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 the neuropsychiatric syndromes under study at the CNP require transdisciplinary collaborations.


Associations between Terms in first set and second set, in the context (dopamine)
http://phenomics.cs.ucla.edu/cgi-bin/pubatlas_run.py?id=3176406
Google
Terms: first set
Terms: second set
Term associations
display scale (zoom):
9
show association strengths as of end of year:
2009
start movie
omit weaker associations having ln(Jaccard) below:
-10
cluster
perturb
undo
redo
reset

		SNAP2	DRD5	MAOB	DRD1	VMAT2	VMAT	CREB	BDNF	DARPP-32	RGS4	DTNBP1/Dysbindin	NRG1/Neuregulin	DRD3	DRD2	COMT	DBH	MAOA	DAT/DAT1/SLC6A3	DRD4	HTR2A	5-HTT	SERT	
		56	264	643	2071	231	359	171	505	476	33	22	31	799	3448	864	1258	457	1919	831	32	81	322	
Attention Deficit Disorder wit...	834	14	34	6	12	2	2		7	3				10	34	27	32	11	180	139	3	8	8	834
Attention Deficit and Disrupti...	851	14	35	6	12	2	2		7	3				10	35	27	32	11	185	141	3	8	8	851
Mental Disorders Diagnosed in ...	1572	16	38	8	22	6	9	1	10	3				16	61	31	44	14	200	162	3	10	10	1572
Tourette Syndrome	309	2	3		3	3	6							3	20	3	5	2	17	10		1	1	309
Tic Disorders	338	2	3		3	3	6							4	23	3	5	2	17	14		1	1	338
Opioid-Related Disorders	340	1	1	1	6	1	2	2	1	1				4	18	1	1		11	7	1	1	1	340
Substance-Related Disorders	3794	1	14	6	77	14	21	19	24	17	4	1		51	282	33	13	11	164	64	2	9	23	3794
Mental Disorders	2854	7	27	5	59	15	18	9	60	31	13	13	14	61	149	105	20	19	191	110	10	11	25	2854
Schizophrenia	3809	2	22	3	66	4	4	3	26	15	11	16	18	132	278	99	44	6	54	84	16	1	1	3809
Schizophrenia and Disorders wi...	4351	2	23	4	71	4	5	3	27	16	11	17	18	144	301	108	54	6	62	92	17	2	2	4351
Psychotic Disorders	792	1	3	1	5		1		4	1		2	2	21	37	17	14		11	16	3	1	1	792
Psychotic Affective Disorders	491	1	7	2	14				4	4	1	1	1	18	36	25	18	11	23	28	1	3	5	491
Bipolar Disorder	466	1	7	2	14				4	4	1	1	1	18	35	25	17	11	23	28	1	3	5	466
Obsessive-Compulsive Disorder	203		1		3									6	9	6		2	10	14	1	2	4	203
Major Depressive Disorder	208		1	1	1				3					3	8	4	3	2	13	6		1	4	208
Tobacco Use Disorder	319		1	2	2			1	2					6	21	6	1	2	14	7		1	2	319
Mood Disorders	1490	1	8	10	16	1	1	3	17	5	1	1	1	25	54	35	39	23	47	47	3	5	11	1490
Alzheimer Disease	534		2	14	11				5	3				6	10	5	6	3	15	2		1		534
Personality Disorders	126		1	1										5	13	5	6	2	6	8				126
Sleep Disorders	887			2	4									6	15	6	2	3	5	3				887
Dyssomnias	763			2	4									6	14	5	1	2	5	3				763
Intrinsic Sleep Disorders	615			1	4									4	11	4	1	1	4	2				615
Depressive Disorder	1005	1	1	10	5	1	1	3	11					7	22	12	26	14	26	22	2	3	8	1005
Anxiety Disorders	426		2	3	5		1	1	2					7	22	10	7	7	14	18	2	3	4	426
Alcoholism	849		4	5	25	2	2	3	1					11	166	16	16	5	36	27	1	6	5	849
Alcohol-Related Disorders	983		4	5	26	2	2	3	1	2				12	167	16	20	6	42	28	1	6	5	983
Substance Withdrawal Syndrome	1074		1	1	15	1	1	4	6	3	1			6	28	3	9	1	20	3				1074
Dementia	1211		6	19	30	2	3	1	18	28	1			7	33	7	10	5	27	3		1		1211
Dementia, Amnestic, Cognitive ...	1464		8	13	34	2	4	1	22	29	1	1		10	38	25	4	3	38	11		1		1464
Cognition Disorders	873		6	3	24	1	2	1	16	26	1	1		4	31	21	2		14	9				873
Huntington Disease	384		5	2	21	1	1	1	12	25	1			2	24	2			3	2				384
Huntington Disease	384		5	2	21	1	1	1	12	25	1			2	24	2			3	2				384
Cocaine-Related Disorders	786		3	1	20	1	2	4	12	6	1			12	23	1	2	1	44	4			11	786
Amphetamine-Related Disorders	198				5	9	13	3	3	2	2				6	1		1	18	2			2	198
		56	264	643	2071	231	359	171	505	476	33	22	31	799	3448	864	1258	457	1919	831	32	81	322	

Output from PubAtlas, showing strengths of associations found by PubMed between mental disorders and certain genes, using the new clustering capabilities. The table shows clustering of disorders by association with neurotransmitter-related genes; e.g., sleep disorders cluster closely with personality disorders and Alzheimer's disease, as these are associated mainly with only the dopamine receptor genes.

Progress over the period June 2008 – May 2009

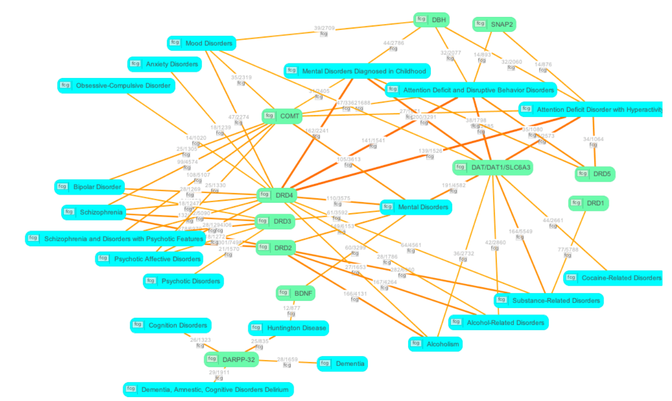
We have made substantial progress for the Hypothesis Web effort in the following areas:

- Expansion of *PubAtlas*, a tool (and website) that permits high-level views of the PubMed literature, giving perspective on concepts and their associations.
- Initial deployment of *HyPoint*, a tool (and website) for constructing multi-level graph representations of hypotheses, with links to the scientific literature, and exploration of hypotheses related to CNP projects.
- Development and refinement of multiple tools for text mining and literature mining, including management of controlled vocabularies (lexica) used in PubAtlas and HyPoint.
- Formalization of roles for informatics to play in phenomics.

These software components are building blocks of the Hypothesis Web, facilitating collaborative development of hypotheses, and in managing associations between multiple levels of transdisciplinary science.

PubAtlas

Over this period we extended PubAtlas (http://www.pubatlas.org), the web service we initiated last year. PubAtlas provides a visual interface to PubMed, and can be viewed as a "BLAST for PubMed", extending the PubMed interface to give a high-level view. The main extension was a new architecture that stores PubMed information in a central knowledge base, permitting many new services to be added over time. We extended PubAtlas to support more forms of association analysis, and in particular included features for client-side interactive biclustering, allowing users to regroup the data in ways that makes it easier to spot patterns of interest. We also included services for downloading results in different formats (including spreadsheet and lexicon), and saving the state of a PubAtlas interaction for later analysis. This allows embedding of links to PubAtlas results to be included directly in other documents; for example http://phenomics.cs.ucla.edu/cgi-bin/pubatlas_run.py?term_collection_1=CNP_people&intersection_threshold=10 gives a dynamic window on collaborations within CNP. This particular link also shows how PubAtlas makes it possible to construct "semi-automated review papers" that summarize the literature relative to a given vocabulary – while remaining both thorough (covering all publications in PubMed) and always up-to-date.


DBH
SNAP2
Mood Disorders
Anxiety Disorders
Obsessive-Compulsive Disorder
Mental Disorders Diagnosed in Childhood
Attention Deficit and Disruptive Behavior Disorders
Attention Deficit Disorder with Hyperactivity
COMT
DAT/DAT1/SLC6A3
DRD5
DRD1
Bipolar Disorder
Schizophrenia
Schizophrenia and Disorders with Psychotic Features
Psychotic Affective Disorders
Psychotic Disorders
DRD4
DRD3
DRD2
Mental Disorders
BDNF
Cocaine-Related Disorders
Substance-Related Disorders
Alcohol-Related Disorders
Alcoholism
Cognition Disorders
Huntington Disease
DARPP-32
Dementia
Dementia, Amnestic, Cognitive Disorders Delirium

PubAtlas now permits export of associations in graph format. This graph has clusters of disorders (with strongly-associated genes) derived from the figure above.

Several services were added to provide alternative kinds of literature map – particularly graph layout maps using the VUE hierarchical graph editor, giving it capabilities of the earlier PubGraph system. For example, the graph figure here shows a graph literature map produced by PubAtlas for the genotype/phenotype associations in the figure above. Edge thickness and color intensities reflect the degree of co-occurrence, as in PubAtlas. Other literature map formats and analytical tools are planned.

One application of PubAtlas has repeatedly generated interest – mapping of associations between researchers and other sets of topics (including other sets of people). PubAtlas now includes lexica for all large neuroscience groups at UCLA, and we plan to extend it with more services specifically for formulating queries about researchers.

We presented PubAtlas at the March 2009 *AMIA Summit on Translational Bioinformatics*, and updated the PubAtlas site. In late April 2009 we will be running on a larger multicore server to handle increasing traffic.

HyPoint

Over this period we developed an initial version of HyPoint (http://www.hypoint.org), a new web service that provides a means for translating a high-level specification into a hypothesis graph. This version is a derivative of PubAtlas – it takes a lexicon of concepts and associations, and derives a VUE graph representing them. A prototype graph developed for the Gpr6 hypothesis developed within the CNP is shown in the figure below.

Literature Mining

We have made progress in representation of semantic knowledge used in CNP hypotheses. Specifically we have results in *lexicon cleaning*, *assertion extraction*, *query extraction*, and *document retrieval*.

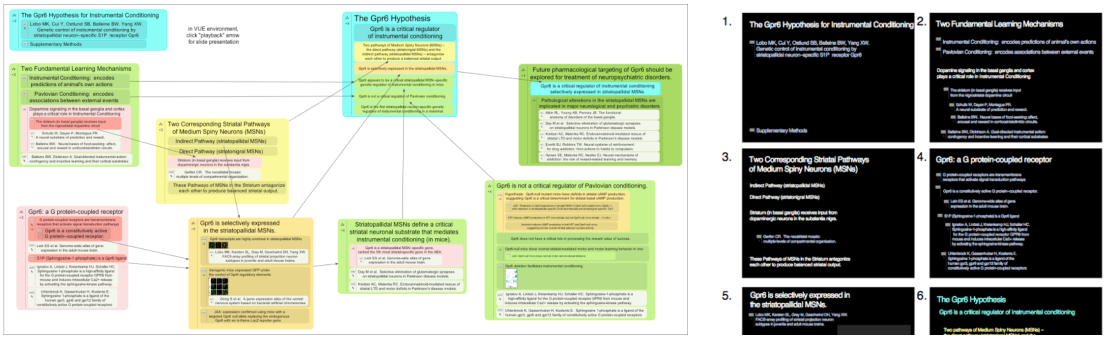
Lexicon Cleaning

We have developed a program that helps clean cognition lexicon, starting from an initial list of over 2,000 *cognition* candidate terms obtained as frequent terms from textbooks and other online resources. To further clean the lexica, we used a stemming tool to group syntactically similar terms together, and then sort them based on their common prefixes or suffixes. For example, family terms can be grouped such as

> attention, auditory attention, central attention, divided attention, focus of attention, focused attention, involuntary attention, passive attention, selective attention, spatial attention, sustained attention

and

> attention, attention capacity, attention span, attention shifting, attentional blink, attentional effort, attentional focusing, attentional resources, attentional state, attentional subsystems.


The Gpr6 Hypothesis for Instrumental Conditioning
in VUE environment, click "playback" arrow for slide presentation
The Gpr6 Hypothesis
Gpr6 is a critical regulator of instrumental conditioning
Two Fundamental Learning Mechanisms
Future pharmacological targeting of Gpr6 should be explored for treatment of neuropsychiatric disorders.
Two Corresponding Striatal Pathways of Medium Spiny Neurons (MSNs)
Gpr6 is not a critical regulator of Pavlovian conditioning.
Gpr6: a G protein-coupled receptor
Gpr6 is selectively expressed in the striatopallidal MSNs.
Striatopallidal MSNs define a critical striatal neuronal substrate that mediates instrumental conditioning (in mice).
1.
The Gpr6 Hypothesis for Instrumental Conditioning
2.
Two Fundamental Learning Mechanisms
3.
Two Corresponding Striatal Pathways of Medium Spiny Neurons (MSNs)
4.
Gpr6: a G protein-coupled receptor
5.
Gpr6 is selectively expressed in the striatopallidal MSNs.
6.
The Gpr6 Hypothesis
Gpr6 is a critical regulator of instrumental conditioning

A HyPoint representation of the Gpr6 hypothesis by CNP investigators Balleine and Yang — arguing the importance of the Gpr6 gene in striatopallidal medium spiny neurons for instrumental learning. The hierarchical graph representation (at left) is hierarchical argument structure, and colors shown here follow the standard CNP 7-level scheme, ranging from proteomic (red) to phenomic (green). The graph can be automatically converted to a PowerPoint-like presentation (at right)), a key feature of HyPoint for facilitating collaborative development, sharing, and publication of hypotheses.

In this way, irrelevant terms can be easily detected and pruned by domain experts. For example in the first group only "attention" and "sustained attention" are kept; furthermore this grouping technique can be used as an aid in constructing concept hierarchies.

Assertion Extraction

We built a knowledge based tool to extract assertions. Given two or more concept terms, the assertion extractor is able to find all sentences (assertions) that mention these concept terms (or their synonyms) in a given knowledge base (e.g. a specified set of scientific documents). For example, for the concept terms "dopamine" and "working memory" and a set of PDF documents by CNP researchers, we obtained the following results from the assertion extractor:

> *Working memory function is strongly dependent upon dopaminergic function in PFC [29, 30]*
> *Dopamine D1 receptors are involved in promoting stability of active representations in working memory [31, 32]*
> *D2 receptors are involved in updating or resetting of working memory neural ensembles [33]*

This tool is particularly useful for researchers who want to construct complex hypotheses.

Query Expansion

Query expansion can be used to relax users' queries so as to retrieve more documents. It can also serve as a tool for constructing a knowledge base of concepts. For a given concept term we are able to find conceptually related concepts by text mining a domain specific corpus. For example, we obtained the following list of the top 5 terms related to the concept term "dopamine" in cognition phenotypes and neural system:

Cognition:

1. [memory consolidation]
2. [insight]
3. [cognition]
4. [motor control]
5. [nondeclarative memory]

Neuroanatomy:

1. [archipallium, striatum]
2. [nucleus accumbens]
3. [frontal cortex, frontal lobe]
4. [prefrontal cortex]
5. [nucleus caudatus, caudate nucleus]

We are now working towards expanding queries containing multiple concepts (e.g., "dopamine" and "working memory"). Clearly, the additional concept term "working memory" will alter the expansion of "dopamine" with different emphasis.

Document Retrieval

We have built a search engine for document retrieval. Given an expanded query, we can retrieve relevant scientific report, and sort them based on their relevance to the query. Document retrieval is critical to the hypothesis web project because the content of the relevant scientific document that supports the hypothesis. Compared with the assertion extractor, document retrieval is more advanced in that it does not require query terms to appear in the same sentence. Instead, the relevance of each document to the query is decided by their term frequency and inverse document frequency (TF-IDF) weights. We have evaluated the precision for complex queries such as "response inhibition AND prefrontal", and obtained good verification of the search results from domain experts.

Plans

Over the next period we plan to consolidate the Hypothesis Web platform. This requires both a more complete ontology and new mechanisms for development of hypotheses. All tools described here (PubAtlas, HyPoint, and the Literature Mining tools) will also be extended in work with other CNP projects on representation of hypotheses.

Publications

Bilder RM, Sabb FW, Cannon TD, London ED, Jentsch JD, Parker DS, Poldrack RA, Evans C, Freimer NB. Phenomics: The systematic study of phenotypes on a genome-wide scale. *Neuroscience*. Epub 2009 Jan 20.

D.S. Parker, W.W. Chu, F.W. Sabb, A.W. Toga, R.M. Bilder, Literature Mapping with PubAtlas -- extending PubMed with a 'BLASTing' interface Proc AMIA Summit on Translational Bioinformatics, San Francisco, March 2009.

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