BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 6 2012, pages 897–899
doi:10.1093/bioinformatics/bts040

Databases and ontologies

Advance Access publication January 27, 2012

SynaptomeDB: an ontology-based knowledgebase for synaptic genes

Mehdi Pirooznia[1,*], Tao Wang[2], Dimitrios Avramopoulos[2], David Valle[2], Gareth Thomas[3], Richard L. Huganir[3], Fernando S. Goes[1], James B. Potash[4] and Peter P. Zandi[1,5,*]

[1]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, [2]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, [3]Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, [4]Department of Psychiatry, Carver College of Medicine, University of Iowa School of Medicine, Iowa City, IA and [5]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The synapse is integral to the function of the brain and may be an important source of dysfunction underlying many neuropsychiatric disorders. Consequently, it is an excellent candidate for large-scale genomic and proteomic study. However, while the tools and databases available for the annotation of high-throughput DNA and protein are generally robust, a comprehensive resource dedicated to the integration of information about the synapse is lacking.

Results: We present an integrated database, called SynaptomeDB, to retrieve and annotate genes comprising the synaptome. These genes encode components of the synapse including neurotransmitters and their receptors, adhesion/cytoskeletal proteins, scaffold proteins, membrane transporters. SynaptomeDB integrates various and complex data sources for synaptic genes and proteins.

Availability: http://psychiatry.igm.jhmi.edu/SynaptomeDB/

Contact: mpirooz1@jhmi.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 30, 2011; revised on January 2, 2012; accepted on January 18, 2012

1 INTRODUCTION

The synapse is fundamental to the structure and function of the brain through its role in connecting neurons into circuits (Chua *et al.,* 2010). As a result, the synapse is an excellent target of large-scale study of neuropsychiatric disorders. Over the past decade, the number of identified synaptic proteins has increased dramatically, creating a need for a comprehensive resource to integrate information about synaptic genes and proteins (the 'synaptome') from multiple heterogeneous sources. These genes encode components of the synapse including neurotransmitters and their receptors, adhesion/cytoskeletal proteins, scaffold proteins, transporters and others (Wu *et al.,* 2010). Here, we report on an integrated database, SynaptomeDB, which provides a detailed and experimentally verified annotation of all known synaptic proteins. The development of SynaptomeDB has been motivated by a desire to have a database that can interoperate with different resources by simply storing the object keys (database identifiers), some important attributes such as symbol and name, and the relationships among the objects. The database does not actually store large objects, like sequence records, but built-in web services can retrieve the subset of objects of interest on demand from other sources such as EBI and NCBI.

*To whom correspondence should be addressed.

2 METHODS

2.1 Data collection

The human synaptome protein list was compiled from a review of all peer-reviewed proteomics studies from 2004 to 2010, as well from as publicly available databases that included proteins in the post- and, pre-synapse, the presynaptic active zone and the synaptic vesicle (Abul-Husn *et al.*, 2009; Zhang *et al.*, 2007). To date, more than 2200 published studies have reported data on post- or pre-synaptic genes and proteins, active zone and vesicles. Synaptome genes were annotated based on RefSeq (GRCh37/hg19) and UCSC (hg19) to identify human orthologs. We further annotated these genes by querying 42 databases covering all aspects of biology, including genes, proteins, pathways and other biological concepts. Both the annotation and curation processes are fully automated and can be executed regularly. The flowchart of the SynaptomeDB construction along with sample queries and explanation of curation process are illustrated in Supplementary Material 1. SynaptomeDB is a gene-centered relational database. It relies primarily on existing database identifiers derived from community databases such as NCBI, GO (Ashburner *et al.*, 2000), EBI (Goujon *et al.*, 2010) and Ensembl (Flicek *et al.*, 2011) as well as the known relationships among those identifiers based on the NCBI Refseq (Mudunuri *et al.*, 2000). The Relational database makes it possible to enhance the SynaptomeDB as an extensible platform for integration with other environments such as variation analysis. Regular updates of the database will be performed to incorporate new information. First, the annotation process will be performed on new genes. A simple update will then be executed to populate the other data in the database. The relational structure of the database allows updates to populate automatically in all related fields. Regular updates of the database will be performed weekly to incorporate new information.

2.2 Pathway enrichment

The Overrepresentation analysis (Fury *et al.*, 2006), detailed in Supplementary Material 2, was performed against a collection of gene sets curated in the Molecular Signatures Database (MSigDB)

(Liberzon *et al.*, 2011) to identify pathways that are enriched for synaptic genes, which can inform subsequent biological analyses. Here, the proportion of genes in a given pathway appearing on the SynaptomeDB list is compared with the proportion of genes not appearing on the list, and a hypergeometric test (Holmans, 2010) is performed to test for differences in these proportions. This analysis is also fully automated and can be updated as new genes and sets are identified.

3 RESULTS

We assembled a list of genes ($n = 1886$) that encode all known proteins of the synapse. This comprises 575 genes encoding proteins in the presynaptic nerve terminal and active zone, 107 from the synaptic vesicles and 1755 from the postsynaptic density (there is some overlap between categories). The list includes strong candidates for a number of neuropsychiatric disorders such as, for example, *ANK3* for bipolar disorder (Ferreira *et al.*, 2008), *GRM7* for major depression (Shyn *et al.*, 2011), *PDE4B* for schizophrenia (Kahler *et al.*, 2010) and *SHANK3* for autism (Gauthier *et al.*, 2009). SynaptomeDB is a database with a web front application resource that integrates the various and complex data sources for these synaptic genes.

3.1 Database design and features

The database is created using MySQL 5.5. The parsers are written in perl and Bioperl (Stajich *et al.*, 2002). The Ensembl BioMart is also used to create some of the tables. A conceptual model of the database is shown in Supplementary Material 3. These tables describe fundamental information about a particular gene: name, description, associated accession numbers, chromosome location, function and comparative map information among other variables. Information from Ensembl also occupies a significant part of the database. It is important to note that no extensive cleaning of the data is performed during the database creation and update process. As detailed in Supplementary Material 1, the major cleaning process involves character screening to make sure the data is compatible for HTML viewing as well as database query. This allows automatic updates and eliminates some well-known problems created by data cleaning.

3.2 Web interface

SynaptomeDB provides a user-friendly web interface. Users can query SynaptomeDB using gene information such as names, gene IDs, synonyms and genomic regions. The output consists of a graphical representation of protein structure from PDB (Berman *et al.*, 2000), protein–protein interactions from STRING (von Mering *et al.*, 2003) and protein domain architecture from HPRD (Keshava *et al.*, 2009). All information was hyperlinked to its original resources. SynaptomeDB allows the user to export multiple samples from different sample sets, in a desired order, to a number of common file formats including Excel, Word, CSV and XML. The web interface of SynaptomeDB provides a rich set of functions for searching the database. In general, search results are initially presented as the summary statements of individual gene records contained in SynaptomeDB, along with additional links to the gene detail page that reveal all details of the gene records returned by the query. A simple text search function is also provided to enable maximum flexibility in searching all records. The advanced search page provides complex searching functions. General database statistics are shown on the home page and reveal a quick summary of genes, as well as the last updates of the system.

4 FUTURE PLANS AND CONCLUSIONS

The database was constructed following guideline described previously (Kirov *et al.*, 2005). It can be used to answer complex queries, such as defining a set of candidate genes based on the genome localization or specific function. The database provides a valuable resource to both experimental and bioinformatics groups by bringing together different sources of information and functional annotation in one place, and in a high-throughput fashion. A synaptome-based strategy for psychiatric genetic sequencing is valuable because there is evidence for synaptic proteins playing a role in psychiatric disorders (Glessner *et al*., 2010), and because these proteins represent the most 'druggable' targets for pursuit of novel therapies. Our application will further research in this area both in its current form and with additional modifications that will include incorporating navigation based on GO and functional pathways and networks among the Synaptome genes in the DB and to also include or link with protein–protein interactions. We intend to extend SynaptomeDB to connect to other psychiatric resources such as SZGene (Allen *et al.*, 2008) and Alzgene (Bertram *et al.*, 2007), and also to integrate variants from several ongoing studies that include synaptome genes, including the 1000 Genomes Project.

Funding: National Institutes of Health (R01-MH087979 to J.B.P.), (K01-MH093809 to M.P.), (R01-MH083738 to P.P.Z.); Johns Hopkins Brain Science Institute.

Conflict of Interest: none declared.

REFERENCES

Abul-Husn,N.S. *et al.* (2009) Systems approach to explore components and interactions in the presynapse. *Proteomics*, **9**, 3303–3315.

Allen,N.C. *et al.* (2008) Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.*, **40**, 827–834.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bertram,L. *et al.* (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.*, **39**, 17–23.

Chua,J.J. *et al.* (2010) The architecture of an excitatory synapse. *J. Cell Sci.*, **123**, 819–823.

Ferreira,M.A. *et al.* (2008) Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.*, **40**, 1056–1058.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Fury,W. *et al.* (2006) Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **1**, 5531–5534.

Gauthier,J. *et al.* (2009) Novel de novo SHANK3 mutation in autistic patients. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **150B**, 421–424.

Glessner,J.T. *et al.* (2010) Strong synaptic transmission impact by copy number variations in schizophrenia. *Proc. Natl Acad. Sci. USA*, **107**, 10584–10589.

Goujon,M *et al.* (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.

Holmans,P. (2010) Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.*, **72**, 141–179.

Kahler,A.K. *et al.* (2010) Association study of PDE4B gene variants in Scandinavian schizophrenia and bipolar disorder multicenter case-control samples. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **153B**, 86–96.

Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Kirov,S.A. *et al.* (2005) GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments. *BMC Bioinformatics*, **6**, 72.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Mudunuri,U. *et al.* (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–556.

Shyn,S.I. *et al.* (2011) Novel loci for major depression identified by genome-wide association study of sequenced treatment alternatives to relieve depression and meta-analysis of three studies. *Mol. Psychiatry*, **16**, 202–215.

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

von Mering,C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.

Wu,H. *et al.* (2010) To build a synapse: signaling pathways in neuromuscular junction assembly. *Development*, **137**, 1017–1033.

Zhang,W. *et al.* (2007) SynDB: a Synapse protein DataBase based on synapse ontology. *Nucleic Acids Res.*, **35**, D737–D741.