Towards Biomedical Data Integration for Analyzing the Evolution of Cognition

Amrapali Zaveri*, Katja Nowick**, and Jens Lehmann*

*University of Leipzig, Institute of Computer Science, AKSW Group,, Augustusplatz 10, D-04009 Leipzig, Germany, zaveri@informatik.uni-leipzig.de, lehmann@informatik.uni-leipzig.de

**Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics (IZBI), University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany. nowick@bioinf.uni-leipzig.de

Abstract: Cognition is determined by function and interplay of several hundred if not thousand genes with a considerable overlap in the phenotypes and genes causing different cognitive diseases. We argue that these diseases should not be studied in isolation, but that data allowing to study them should be integrated. Ultimately, this will allow researchers to more easily answer questions, which would otherwise require time-consuming research. Specifically, we propose to use Linked Data publication, data integration and querying methods, which has been successfully used in other life science domains. In this initial effort, we converted 12 different datasets, integrate them and provide a first demonstration of the added value by showing how a set of relevant queries over the integrated data can be answered.

1 Introduction

Cognition refers to a group of mental processes that includes memory, attention, language (production and understanding), reasoning, learning, problem solving as well as decision making. Cognition is determined by function and interplay of several hundred if not even thousand genes. There is a considerable overlap in the phenotypes and genes causing different cognitive diseases. We argue that these diseases should not be studied in isolation.

However, current approaches to study the evolution of cognition diseases involve the querying of independent disparate datasets. This process is not only time consuming, for example, because datasets might be in different formats, but also inefficient when any one of the dataset is updated or changed. Recently, the Linking Open Data (LOD) initiative has made several datasets publicly available[1]. In particular, life science datasets have been converted to a single machine-interpretable format called RDF (Resource Description Format)[2]. These datasets have been interlinked to produce a huge corpus of life science datasets via the Bio2RDF project [BNT+08]. We believe that it is time to take

[1] http://lod-cloud.net
[2] http://www.w3.org/RDF/

advantage of these resources to advance in some exciting research questions related to the evolution of human cognition, for instance we would like to easily answer the following questions:

- Which genes are involved in determining cognition and have changed during primate evolution?
- Which genes have been positively selected in humans but are also implicated in cognitive diseases?
- Which genes differ in expression between humans and chimpanzees during development or aging and have been associated with cognitive decline during aging?
- Do genes involved in cognition and behavior show higher diversity within humans and higher divergence between humans and chimpanzees?

In this paper, we utilize LOD for analyzing the evolution of cognition. In particular, we first identify datasets relevant for the analysis and convert them into the RDF format (Section 2). Thereafter, we interlink the datasets so as to obtain an integrated dataset containing all the relevant information (Section 3). This integrated dataset is then queried to extract information aligned to the research questions above (Section 4). We summarize the obtained results in Section 5.

2 Datasets and RDF Conversion

We identified 12 relevant dataset that could provide information which help analyzing the cognition of evolution. These datasets were either available as CSV (Comma Separated Values), TSV (Tab Separated Values) format, simple text files or even as PDFs. All datasets were converted into a single format – the Resource Description Format (RDF)[3]. The conversion of all datasets to RDF would not only help the different datasets to be easily integrated but also assist in answering the research questions by querying the integrated dataset as opposed to extracting information individually. We converted the data using LODRefine[4] as well as Sparqlify[5]. In general, each row was transformed into a triple (a fact containing a subject, predicate and object) pertaining to each gene. Each gene, in turn, was given a unique identifier based on the gene symbol to create a URI (Uniform Resource Identifier), which identifies it as a single globally re-usable resource. In the following, we describe each dataset and the relevant variables extracted from each.

AutDB. AutDB[6], a modular database for autism research, is the first publicly available genetic database for autism spectrum disorders. The database aimed to collect all gene

[3] http://www.w3.org/RDF/
[4] http://code.zemanta.com/sparkica/
[5] http://aksw.org/Projects/Sparqlify.html
[6] http://autism.mindspec.org/autdb/Welcome.do

information related to autism and was built by integrating data from various areas of autism research obtained from peer-reviewed published scientific literature. The database also contains interactive molecules that illuminate the molecular functions of genes implicated in autism, which allow for cross-modal navigation. These molecules are of (i) human genes (evidence for association of genes with autism), (ii) animal models (characteristics of animal models created from altering expression of these genes), (iii) protein interactions (compiles all known molecular interactions of proteins produced from these genes) and copy number variants (CNV) (which curates all known CNVs linked to autism).

The data is available for download in *CSV* format[7] and contains the gene name and symbol; chromosome number and location; evidence of support for autism; number of positive and negative gene association studies as well as the reference and most cited reference for each gene.

Genes2Cognition. Genes to Cognition (G2C) is a neuroscience research program which aims to discover fundamental biological principles and important insights into brain disease such as finding the basis of neurodegenerative diseases. The project has a publicly available database called G2Cdb [MMP+09] that stores data resources from the research program for basic and clinical neuroscience. G2C uses genome information to understand cognition at the molecular, cellular and systems neuroscience levels.

The data is available for download in *CSV* format (along with text and XLS)[8]. The gene and its symbol, species it belongs to and a description is provided.

Catalogue of Parent of Origin Effects. The Catalog of of Parent of Origin Effects contains a collection of imprinted genes. In contrast to most genes, imprinted genes are only expressed from the paternal or maternal allele. Some of these genes haven been implicated in social behavior. The catalog provides the gene names, a description and genomic location for the genes, as well as cross-species information.

FunDO. FunDO is a project which explores genes using the **Fun**ctional **D**isease **O**ntology annotation [OFH+09]. A list of genes are retrieved and the relevant diseases, based on statistical analysis of the Disease Ontology[9] annotation database, are identified. The Unified Medical Language System (UMLS) MetaMap Transfer tool (MMTx) was utilized to discover the gene-disease relationships from the GeneRIF[10] database. The results were validated against the Homayouni gene collection using recall and precision measurements by comparing them agains the Online Mendelian Inheritance in Man (OMIM) annotations.

The mappings are available in *text* format[11] with the disease, gene symbol and ID.

[7] http://autism.mindspec.org/autdb/search
[8] http://www.genes2cognition.org/db/GeneList/L00000016
[9] http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page
[10] http://www.ncbi.nlm.nih.gov/gene/about-generif
[11] http://django.nubic.northwestern.edu/fundo/media/data/do_lite.txt

Allan Brain Atlas. The ALLAN Human Brain Atlas[12] is a publicly available online resource of gene expression information particularly in the human brain. The dataset contains genome-wide microarray based gene expression profiles in the human brain along with accompanying anatomic and histologic data. In particular, data about 6 brains with a total of 4,000 unique anatomic samples characterized across 60,000 probes per sample is available.

The complete normalized microarray dataset is available for download[13] in *CSV* format. Even though the probe and sample metadata data is available, we only required the microarray expression values.

GWAS. The National Human Genome Research Institute has published a catalog of published genome-wide association studies (GWAS) [LPH+09]. The catalog mainly contains the examination of many common genetic variants in different individuals to analyze if any variant is associated with a trait. The focus of GWAS is typically on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases.

The catalog is available for download in *text* format[14]. The file mainly contains the PubMed ID, disease examined, chromosome position, reported and mapped genes, reported p-value for strongest SNP risk allele and SNP information. The Gene ID was selected as the unique identifier in this case.

Genetic Association DB. The Genetic Association Database [ZDG+10] is an archive of human genetic association studies of complex diseases and disorders. The data is extracted from published articles in peer reviewed journals specifically on candidate genes and GWAS studies. This database allows a user to easily identify medically relevant polymorphism from the large volume of polymorphism and mutational data, in the context of standardized nomenclature.

The data is available for download in *TSV* format[15]. Each record belongs to a particular gene and contains information about the publication, locus number, DNA start and end, disease, phenotype, chromosome band and alleles. The Gene ID was chosen as the unique identifier for each record.

ID-TFs. Transcription factors (TFs) play a major role in regulating the activity of other genes. They are thus key for dynamic and plastic biological processes like cognition and behavior. We collected a list of all human TFs from [JSSN09]. This list contains the gene symbols along with the indication of whether it contains NS-non syndromic or S-syndromic ID -intelligence disorder. Ones with S-ID means having lower IQ and other phenotypes (e.g. smaller brain, bad hearing, maybe even heart problems). Ones with NS-ID, means only having lower IQ and no other phenotype. Thus, one interpretation would

[12] `http://www.brain-map.org/`
[13] `http://human.brain-map.org/static/download`
[14] `http://www.genome.gov/admin/gwascatalog.txt`
[15] `http://geneticassociationdb.nih.gov/cgi-bin/download.cgi`

be that the only function of NS-ID genes is to determine IQ, while S-ID genes have other functions besides determining IQ.

Autistic Train Genes. Autism is a human disorder that affects the behavior of the individuals. Genes implicated in autism can thus provide information on the genes and pathways that are important for controlling behavior. A collection of Autism genes was extracted from [C.11] along with information on the cytoband, disorder, inheritance pattern, ASD/autistic traits and references.

Ensembl. Ensembl[16] is a bioinformatics research project, in collaboration with the Wellcome Trust Sanger Institute and the European Bioinformatics Institute (EBI). Its databases contain information on the genomes of chordates (including primates and mouse), invertebrates as well as yeast and is easily available for download and search.

We retrieved the alternative gene names (available in *text* format), the ortholog information for humans and mouse ((available in *TSV* format) as well as the mappings between the ensembl and gene IDs (also available in *text* format) from Ensembl.

Human Positive Selection Candidates. In [BFAW$^+$05], the authors performed a genome wide scan for regions under positive selection. They calculated a lot of statistics, but the most interesting ones are the dN/dS and Ka/Ks values. if dN/dS or Ka/Ks is larger than one, this means that the gene has changed a lot between humans and chimps.

All these 12 datasets were converted to RDF, which produced a total of $385,786$ triples.

3 Dataset interlinking

The converted RDF datasets are interlinked with each other as well as with other external datasets. The gene symbol is the common element in all the datasets and, thus, the datasets were integrated using this symbol. Therefore, when one queries the integrated dataset for any particular gene symbol, information from all the datasets can be readily obtained.

Additionally, we identified potential candidates of external datasets to which the integrated dataset can be linked to obtain more relevant information. The external datasets that we identified to be useful are: HUGO Gene Nomenclature Committee (HGNC) (for genomic, proteomic and phenotypic information); Gene Ontology Annotation (for annotations to proteins in the UniProt knowledgebase); Online Mendelian Inheritance in Man (for mendelian disorders and relations between genotype and phenotype); PubMed (for literature references); Medical Subject Headings (for using the controlled vocabulary for indexing articles); NCBI taxonomy (for the nucleotide or protein sequence) (homologene) and Reference Sequences (for genomic DNA, transcripts, and proteins). All these datasets have already been converted to RDF and are available via the Bio2RDF [BNT$^+$08] project.

[16] `http://www.ensembl.org/index.html`

4 Dataset Querying and Initial Results

After converting and interlinking the datasets, we obtained a single integrated compendium containing all the relevant information. We loaded the integrated datasets in a Virtuoso triple store[17] (a database for RDF data). The dataset is available at SPARQL endpoint `http://db0.aksw.org:8895/sparql` with the graph name `http://aksw.cogevo.org`. Therefore, as the next step we performed SPARQL[18] (query language for RDF) queries over the integrated dataset to help us answer our research questions (Section 1).

As a preliminary example, we chose our first question:"Which genes are involved in determining cognition and have changed during primate evolution?". First, we started by intersecting our table on ID-TFs with our table on Human Positive Selection Candidates. The first table provides us information about transcription factors (TFs) that have been associated with Intellectual Disability (ID). Patients with this disability display reduced Intelligence Quotients (IQs). TFs are an important class of proteins, as they regulate the activity of other proteins and are thus key for all functions of the individual; in this case for determining cognitive abilities. From the table on Positive selection we retrieved the information on dN/dS ratios for each gene. This ratio represents the ratio of the number of mutations leading to an amino acid sequence change (presumably changing the function of the protein encoded by the gene) vs. the number of mutations that do not lead to an amino acid change (are functionally neutral). The higher this ratio, the faster the protein is evolving. Commonly, genes with dN/dS rations ¿ 1 are assumed to evolve under positive selection.

```
PREFIX cog:<http://aksw.cogevo.org/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bio2rdf:<http://bio2rdf.org/hgnc_vocabulary:>
PREFIX go:<http://bio2rdf.org/goa_vocabulary:>
PREFIX autdb:<http://aksw.cogevo.org/autdb/>

SELECT ?s ?symbol ?dnbydns
FROM <http://aksw.cogevo.org>
WHERE {
       ?s rdf:type cog:gene .
       ?s bio2rdf:approved_name ?symbol .
       ?s cog:dnDs ?dnbydns .
       ?gene go:symbol ?symbols .
       ?gene cog:nsid ?ns .
       FILTER (?symbol = ?symbols)
      }
```

Listing 1: Exemplary SPARQL query querying two different datasets.

Our query (Listing 1) retrieved one gene that is an ID-TF and has a dN/dS ratio of ¿ 1 (dN/dS = 1.33), the FMR2 gene. This gene has thus changed significantly more during primate evolution and might be under positive selection in humans. FMR2 has been linked to non-syndromic intellectual disability (NS-ID) [SSH+11, MNS+13]. For patients with mutations in FMR2 it has been reported that they are mentally retarded in associated with learning difficulties, communication deficits, attention problems, hyperactivity, and autis-

[17] `http://virtuoso.openlinksw.com/`
[18] `http://www.w3.org/TR/rdf-sparql-query/`

tic behavior [MME+09]. Thus, with the result FMR2 we identified a gene that is involved in determining cognition and has significantly changed during primate evolution.

5 Conclusions and Future Work

In this paper, we have describe our preliminary work and ideas to use Linked Data publication, to demonstrate its use in analyzing the evolution of cognition. We identified 12 relevant datasets, converted them to a single machine-readable format, RDF and interlinked them. Moreover, we queried the integrated dataset to show an example in order to illustrate the potential benefits of this project. Results show that this approach is promising and as future work, we plan to perform further different queries over the integrated dataset to help answer the various research questions in analyzing the evolution of cognition in humans as well as compared to different species, for e.g. mouse. With this use case, we hope to illustrate an example that would bridge the gap between biomedical and informatics domains such that they can benefit from each other.

References

[BFAW+05] Carlos D. Bustamante, Adi Fledel-Alon, Scott Williamson, Rasmus Nielsen, Hubisz Melissa T., Stephen Glanowski, David M. Tenanbaum, Thomas J. White, John J. Sninsky, Ryan D. Hernandez, Daniel Civello, Mark D. Adams, Michele Cargill, and Andrew G. Clark. Natural selection on protein-coding genes in the human genome. *Nature*, pages 1153 – 1157, 2005.

[BNT+08] Francois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008.

[C.11] Betancur C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.*, pages 42 – 77, 2011.

[JSSN09] Vaquerizas J.M., Kummerfeld S.K., Teichmann S.A., and Luscombe N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–63, 2009.

[LPH+09] Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., and Manolio T.A. otential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*, May 2009.

[MME+09] Bensaid M, M Melko, Bechara EG, Davidovic L, Berretta A, Catania MV MV, Gecz J, Lalli E, and Bardoni B. FRAXE-associated mental retardation protein (FMR2) is an RNA-binding protein with high affinity for G-quartet RNA forming structure. *Nucleic Acids Research*, 37(4):1269079, 2009.

[MMP+09] Croning M.D., Marshall M.C., McLaren P., Armstrong J.D., and Grant S.G. G2Cdb: the Genes to Cognition database. *Nucleic Acids Research*, 37, 2009.

[MNS+13] M Melko, LS Nguyen, M Shaw, L Jolly, B Bardoni, and J Gecz. Loss of FMR2 further emphasizes the link between deregulation of immediate early response genes FOS and JUN and intellectual disability. *Hum Mol Genet.*, 2013.

[OFH+09] John D. Osborne, Jared Flatow, Michelle Holko, Simon M. Lin, Warren A. Kibbe, Lihua J. Zhu, Maria I. Danila, Gang Feng, and Rex L. Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10(1), 2009.

[SSH+11] G.M. Stettner, M. Shoukier, C. Höger, K. Brockmann, and B. Auber. Familial intellectual disability and autistic behavior caused by a small FMR2 gene deletion. *Am J Med Genet A.*, 155A(8):2003–7, 2011.

[ZDG+10] Yonqing Zhang, Supriyo De, John R. Garner, Kirstin Smith, S. Alex Wang, and Kevin G. Becker. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Medical Genomics*, 3(1), Jan 2010.