

IntegromeDB: semantic integration of Transcriptional Regulation Data

Michael Baitaluk^{1,*} and Julia Ponomarenko^{1,2}

¹San Diego Supercomputer Center and ²Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA.

*Contact: baitaluk@sdsc.edu

Experimental and predicted data on gene transcriptional regulation are distributed throughout many heterogeneous sources. Yet no solution has been proposed aimed to automatically integrate these data, and no resource exists that provides an ‘one-stop shop’ experience for the users seeking information essential for deciphering and modeling gene regulatory networks.

We propose IntegromeDB, a semantic graph-base ‘deep-web’ data integration system that automatically captures, integrates, and manages publicly available data on transcriptional regulation, along with other relevant biological information. The problems of data integration are addressed through ontology driven data mapping, multiple data annotation, and heterogeneous data querying, also enabling integration of the user’s data. IntegromeDB integrates over 100 experimental and computational data sources on genomics, transcriptomics, genetics, functional and interaction data on gene transcriptional regulation in eukaryotes and prokaryotes.

To address the problem of data cleaning and conflict resolution, we developed the reconciliation procedures that identify controversies, or inconsistencies, in the data. The examples of data inconsistencies include, but are not limited to, the following: (i) two different genes are assigned to the same synonym; (ii) two genes with the same name point to different chromosomal locations; (iii) two genes with different names point to the same chromosomal location; and (iv) different objects have the names with a common string, for example, p53, p53(361-393), p53(modified:Thr:212), and pCMX-mutant-p53. Ideally, such inconsistencies should be resolved by a curator. However, because of limited human resources and for the sake of fully automated integration, we currently avoid human intervention and instead provide the details on retrieved properties by data sources that can be viewed for each gene/protein by clicking ‘Details by Data Sources’ on the query result page at <http://www.integromedb.org>. To evaluate quality of the integrated data, we estimated inconsistencies in public databases which were integrated in IntegromeDB. The inconsistencies were found in about 3% of genes/proteins (~12,000); they are reported at <http://www.integromedb.org>.

The web page www.integromedb.org provides keyword/ID, wildcard and multiple word search capabilities, statistics on integrated data by categories and databases, information on retrieved properties by data sources for each gene/protein that can be accessed from the query result page, and inconsistencies in the integrated data. The web site was designed primarily with the purpose to give the user an opportunity to glance at the integrated data rather than to provide complex data mining and analysis capabilities that are implemented in the BiologicalNetworks application.

We explore IntegromeDB web-site search capabilities, using two example queries ‘relb AND diabetes AND Alzheimer’ and ‘rela AND diabetes AND Alzheimer’. RelA (p65) and RelB transcription factors belong to the same family of NF- κ B factors; they can form heterodimers with other NF- κ B factors, p50 (NF- κ B1), p52 (NF- κ B2), and c-Rel, to each other, and RelA can form homodimers.

The first query, ‘relb AND diabetes AND Alzheimer’, returns four genes, Tnf, ESR1, CD40, and AhR. In comparison, EB-eye search (<http://www.ebi.ac.uk>) (Jones, Cote et al. 2008) for the same query returns no entries from any database. While Entrez (Sayers, Barrett et al. 2009) returns three genes, Tnf, ESR1, and CD40. The AhR (aryl hydrocarbon receptor) protein, which was found in IntegromeDB but not in Entrez, was shown to interact with RelB, and AhR:RelB dimers regulate transcription of many genes, functioning as coordinators of inflammatory responses (Vogel, Sciuillo et al. 2007; Vogel and Matsumura 2009), meaning that AhR in fact relates to both RelB, diabetes, and Alzheimer’s disease. One of the reasons why Entrez did not find AhR gene would be that it searches keywords, publications, and gene properties; while IntegromeDB searches relations between publications to the database objects (genes) associated with query words.

The second query, ‘rela AND diabetes AND Alzheimer’, returns six genes, Tnf, K60 (IL-8), PTGS2, BAX, STMY3, and Mlana. While Entrez returns nine genes, PTGS2, Tnf (mouse), IL8, TNF, TP53, SIRT1, PRKCD, ESR1, PRKACA. The query results of Entrez and IntegromeDB have three common genes only; that can be explained by Entrez being more up to date. However, IntegromeDB found three genes that were not found in Entrez, specifically, BAX, STMY3, and Mlana.

We investigated those three genes, searching the query result pages for the query words. The BAX (Bcl2-associated protein X) gene expression is known to be regulated by NF- κ B factors, specifically RelA (Grimm, Schneider et al. 2005). It was demonstrated that patients with diabetes (Varo, Vicent et al. 2003), as well as Alzheimer’s disease (Ait-ghezala, Abdullah et al. 2008) have a pro-inflammatory state as indicated by elevated levels of plasma CD40L. Also, it was shown that BAX mRNA levels are altered in peripheral blood mononuclear cells from both mild cognitive impairment and Alzheimer’s disease patients (Gatta, Cardinale et al. 2009). The STMY3 (Stromelysin-3 precursor) gene expression is associated with expression of p53 in different types of cancers (Sharma, Chattopadhyay et al. 2004). p53 is known to be directly regulated by RelA factor (Jeong, Radonovich et al. 2004). The elevated levels of pro-apoptotic p53 and its oxidative modification by the lipid peroxidation product, HNE, were reported in brain from subjects with amnesic mild cognitive impairment and Alzheimer’s disease (Cenini, Sultana et al. 2008). Also, polymorphisms in p53 are known to be associated with diabetes (Szoke, Molnar et al. 2009). Mlana (Melan-A protein, MART-1) stimulates T-cells to increase secretion of TNF- α (Elluru, van Huyen et al. 2008), which is a direct target of RelA (Shakhov, Kuprash et al. 1990). It was shown that expression of TNF- α increases in both diabetes (Gordin, Forsblom et al. 2008) and Alzheimer’s disease (Baranowska-Bik, Bik et al. 2008).

Thus, all three genes considered above are directly or indirectly associated with RelA and both Alzheimer’s disease and diabetes. The fact that IntegromeDB found those genes while Entrez did not supports the aforementioned statement that IntegromeDB approaches integration of data and the search differently than Entrez; specifically, IntegromeDB integrates data objects (genes) and performs search by object properties rather than searching keywords in publications. The considered examples clearly demonstrate the power of the proposed approach: novel knowledge on gene-disease associations was obtained using IntegromeDB in a matter of minutes; no other system could reveal those associations.