Limitations of existing Chemical Biology Public Domain Data sources

[1]Syed Ahsan [2]Muhammad Shahbaz [3]Syed Athar Masood

[1,2]Department of Computer Science
University of Engineering and Technology, Lahore
[3]Department of Engineering Management, College of E&ME Rawalpindi
[1] ahsancs@hotmail.com [2] m.shahbaz@uet.edu.pk [3]atharmasood2000@hotmail.com

Abstract: We have established an HEC funded basic *Analysis Node for Bioinformatics* at, Al-Khwarzmi Institute of Computer Science, U.E.T., Lahore with the objectives of establishing and strengthen bio/chem-informatics infrastructure in Pakistan through linkages with international organizations involved in R&D and manufacturing activities in biotechnology. This paper describes the nature and objectives of collaboration with The Broad Institute of MIT and Harvard (http://www.broad.mit.edu) to create components of a public database that provides transparent access to high-quality compound probe development research data generated as part of ongoing projects at the Broad Institute, enabling the data to be used as effectively by outside researchers as it is by Broad investigators. Limitations of existing ChemBank v2 and need for ChemBank V3 as a new model of public databases is identified. [Dr. Syed Ahsan, Dr. Muhammad Shahbaz,. Dr. Syed Athar Masood. Limitations of existing Chemical Biology Public Domain Data Sources. Journal of American Science 2011;7(10):1-4]. (ISSN: 1545-1003). http://www.americanscience.org.

Keywords: interoperability. Probe development, drug discovery

Introduction:
The Chemical Biology Platform at the Broad developed the original ChemBank and the most recent updated version is ChemBank v2 [1, 2, 3]. ChemBank has been created to provide a platform for sharing data on assays and compounds in what can be termed as "open source" science. ChemBank v2, although pioneering in sharing chemical and biological data to the community at large, is limited in terms of the details and underlying experimental design [1, 4, 5]. As a user you only get to see the result of the experiment and not how we got to these results. The information on the compounds is rather limited with no details on how to synthesize them. Similarly, in the absence of experimental and protocol design information users of the data are rather limited in terms of reproducing the experiment and extend science. As a result, researchers outside of the Broad cannot easily access the wealth of information being generated by the chemical biology platform within the Broad [6,]. Although this data is also publicly available through the PubChem database, PubChem does not provide users with sufficient context to be able to fully interpret the data [6,7,8]. Therefore, it is essential that we create a new version of ChemBank database i.e., ChemBank V3, to provide this additional context and information. Without this, we are making data public without making it useful.
The project execution that will include setting up a facility for cheminformatics development at KICS in collaboration with Broad Institute at MIT and Harvard will throw open an unprecedented market opportunity in the near future. The proposed facility will function as a Technology Driven Company providing a technological platform to provide various types of goods and services in drug development chain and will try to emulate the development of IT sector international outsourcing. This project in collaboration with two of the top universities of the world (Harvard and MIT) and with a world leader in chemical biology research (The Broad Institute) aims to be only a small but very important step in this direction.

2. Chemical Biology at Broad
Before laying out the scope of the project we want to introduce the reader to small molecule compounds, biological screening, high throughput screening and ChemBank v2 to set up the context.

Small Molecules:
The most general definition of Small molecule compounds would be that these are chemical compounds that typically have molecular weight less than 500. Most drugs are small molecules and the initial phases of the Drug Discovery Process involves using small molecule compounds that are either isolated from Natural Products or are Synthesized (Synthetic Compounds) through a series of chemical reactions in the lab otherwise. The use of small molecule compounds includes trying to find compounds that either inhibit or promote certain

biological function in a cell, a protein, an enzyme or another biological entity. Here is a typical small molecule [5,6] :

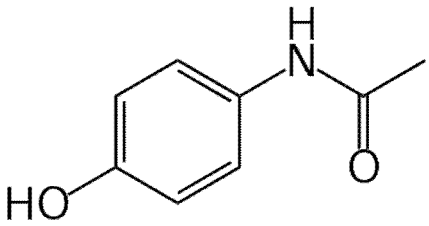

Figure 1 - Molecular structure of acetaminophen aka Paracetamol

Biological Assays and Screening:

Biological Assay is a type of scientific experiment widely used for starting points in the development of new drugs. In the simplest terms a biological assay is set of experiments designed to bring a biological element (such as an enzyme, a protein, a cell or a whole organism) together with a number small molecule compounds and be able to observe the inhibition or promotion of certain biological activity with each of the compounds. Typically a micro titer plate (also called an assay plate) is used to conduct the experiments in parallel. An assay plate has a total of 384 wells where each well can hold about 20-30µL of liquid and hence 384 independent experiments can be run at the same time. Manual or robotic tools are used to transfer biology into each of the wells followed by the compounds. The biology in the entire plate for all wells is kept the same but each well gets a unique small molecule compound. The experiments are designed in such a way that the observation of activity is possible through various means such as the amount of certain wavelength of light being emitted or the florescence or luminescence for each well. Automated instruments are used to read each well of the plate for individual activity reads. Other assays are designed to be read through imaging the wells under an automated microscope and inferences are based on sophisticated image analysis. The data generated from the reader is then analyzed for results and inferences of the experiment [9,10,11].

Ultimately, the compounds that show desired behavior are further analyzed for confirmation, efficacy and safety. Most of the time there would be very few compounds that show the desired activity and hence the term "screening" in the sense of filtering.

High Throughput Screening:

High Throughput Screening (HTS) is the scaled-up and mechanized version of the screening process where the objective is to screen hundreds of thousands of diverse small molecules for desired biological activity. The same protocol or experiment design that was used at a smaller scale of screening is used to run thousands of plates through a series of instruments such as liquid transfer devices (to add biology and to add compounds), mixers, incubation units (to keep the plates at a certain temperature for a given period of time) and plate readers (to read the results). All of this is coordinated by industrial automation robots. [9.10]

3. Limitations of ChemBank v2 and PubChem

PubChem is currently the most comprehensive source of information on biological activity of small molecules [8,11,12]. It contains over 50 million chemical structures and hundreds of biological assays. Significantly, it houses the results of the NIH Molecular Libraries Initiative, which executes over 100 assays per year on a centralized library of 300,000 compounds. This data comprises the bulk of the assay results in PubChem [12]. While it is comprehensive, there are some barriers preventing more effective use of the data within PubChem. First, the assay results are only associated to compound structures, not to specific physical samples. This makes it impossible to verify the quality of the sample used in the screen [13, 14]. Second, the assay descriptions used in PubChem are uncontrolled free text, which leads to inconsistent documentation of important biological details of the assay and prevents effective machine interpretation of the biology. Third, experimental controls and analysis methods are not rigorously tracked, making it difficult to compare runs submitted separately [14.15,16].

ChemBank is a public web site that makes the small molecule compounds and assay results data available to the scientific community. The Chemical Biology Platform at the Broad developed the original ChemBank and the most recent updated version is ChemBank v2.

ChemBank has been created to provide a platform for sharing data on assays and compounds in what can be termed as "open source" science. The entity relationships shown in Figure 2 summarize the information that ChemBank v2 holds:

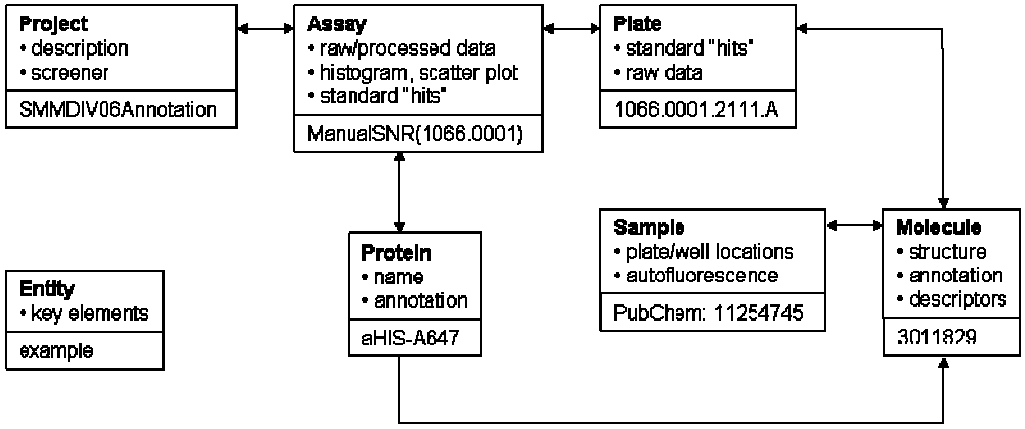

Figure 2: Entity diagram of ChemBank V2

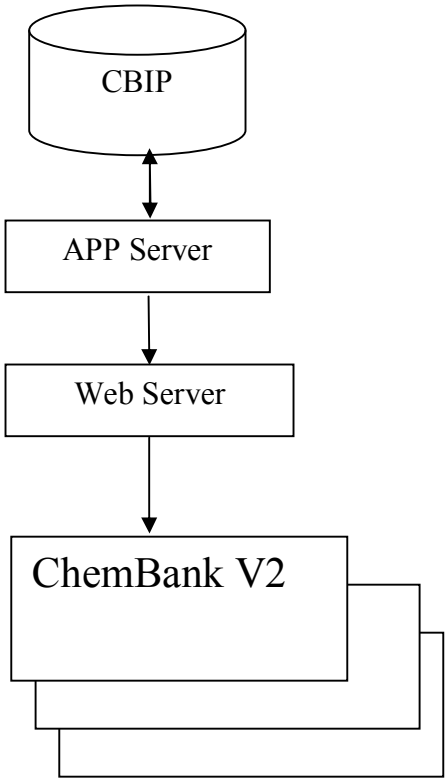

Figure 3: ChemBank v2 Architecture

ChemBank v2, although pioneering in sharing chemical and biological data to the community at large is limited in terms of the details and underlying experimental design. As a user you only get to see the result of the experiment and not how we got to these results [17]. The information on the compounds is rather limited with no details on how to synthesize them. Similarly, in the absence of experimental and protocol design information users of the data are rather limited in terms of reproducing the experiment and extend science. These limitations strongly advocate the development of next generation of ChemBank which may serve as an example for other databases to follow.

References

1.Burks, C., J. W. Fickett, et al. (1985). "The GenBank nucleic acid sequence database." Comput Appl Biosci**1**(4): 225-33.
2.Clamp, M., B. Fry, et al. (2007). "Distinguishing protein-coding and noncoding genes in the human genome." Proc Natl Acad Sci U S A 104(49): 19428-33.
3.Cuff, J. A., G. M. Coates, et al. (2004). "The Ensembl computing architecture." Genome Res 14(5): 971-5.
4.Curwen, V., E. Eyras, et al. (2004). "The Ensembl automatic gene annotation system." Genome Res**14**(5): 942-50.
5.Evan E. Bolton1 et.al.,(2008), PubChem: Integrated Platform of Small Molecules and Biological Activities.
6.Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature 391(6669): 806-11.
7.Howard J. Feldman et.a l(2005) ' CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules' , 2005 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies
8.Hubbard, T., D. Barker, et al. (2002). "The Ensembl genome database project." Nucleic Acids Res 30(1): 38-41.
9.Hur, J. and D. J. Wild (2008). "PubChemSR: A search and retrieval tool for PubChem." Chem Cent J**2**: 11.
10.Kanehisa, M. (2002). "The KEGG database." Novartis Found Symp 247: 91-101; discussion 101-3, 119-28, 244-52.
11.Karp, P. D., M. Riley, et al. (2000). "The EcoCyc and MetaCyc databases." Nucleic Acids Res**28**(1): 56-9.
12.Kathleen Petri Seiler, et al (2008)." ChemBank: a small-molecule screening and cheminformatics resource database, Nucleic Acids Research, 2008, Vol. 36, Database issue D351–D359.
13.Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res**12**(6): 996-1006.
14.Robert L. Strausberg, *et al. (2009) , '* From Knowing to Controlling: A Path from Genomics to Drugs Using Small Molecule Probes , *Science 294* – 300.
15. Shoemaker, D. D., E. E. Schadt, et al. (2001). "Experimental annotation of the human genome using microarray technology." Nature**409**(6822): 922-7.
16 Williams, A. J. (2008). "Public chemical compound databases." Curr Opin Drug Discov Devel**11**(3): 393-404.
17.Yanli Wang, et al (2009), ' PubChem: a public information system for analyzing bioactivities of small molecules '.Nucleic Acids Research, 2009, Vol. 37, Web Server issue W623–W633

2/8/2011