Creation of next generation of Open Source Science Databases

[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, NUST College of E&ME, Rawalpindi Pakistan
[1] ahsancs@hotmail.com, [2] m.shahbaz@uet.edu.pk, [3] atharmasood2000@hotmail.com

Abstract: From 1990s onwards, biological and chemical research in both the public and private sectors throughout the world has been transformed into industrial scale by the creation of databases with large amounts of high-quality, freely available DNA sequence data. These databases have not only enabled the comprehensive cataloging of human genes but have also accelerated the discovery of new forms of cellular regulation rendering biology and chemistry a discovery science thus providing the basis for novel experimental approaches. We however feel that the potential opportunities, accessibility and power of open source science and publicly available data have not transformed into gains and significant impact on scientific discovery. In this paper we have identified many issues with the existing conventional chemical biology and molecular biology databases and propose the development of ChemBank v3.
[Syed Ahsan, Muhammad Shahbaz, Syed Athar Masood. Creation of Next generation Open Source Science Databases. Journal of American Science 2011;7(6):952-955]. (ISSN: 1545-1003). http://www.americanscience.org.

Keywords: biological research, open source science, databases

1. Introduction

Life Science research in post-genomic era is integrative, collaborative and mainly insilico. Therefore, in our opinion the life science research community should be provided with an integrated, transparent access to data and analytical tools of experimentation. In this perspective the Broad Institute's mission is to develop powerful new tools for genomic medicine and to make those tools available to the world. Currently, researchers outside of the Broad cannot easily access the wealth of information being generated by the chemical biology platform within the Broad. Although this data is all publicly available through the PubChem database, PubChem does not provide users with sufficient context to be able to fully interpret the data. Therefore, it is essential that we create a new database to provide this additional context and information. Without this, we are making data public without making it useful. 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. 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. 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. 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. 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.

2. Next Generation of Chemical Database

In the section above , we discussed the limitations of existing chemical database at Broad. In our opinion, these limitations need to be removed on proiority basis to enable chemical biology and drug discovery as an integrative , collaborative and explorative science. To achieve t his objective, we are proposing the development of ChemBank v3. This database would contain information on small molecules contained in the Broad Institute screening collections along with details of the assays performed on those molecules and any results. Where possible, the database would link to other external databases containing related information, such as ChemSpider and PubChem. Such a database would be transformative in enabling effective use of this research data and in establishing new standards for quality and transparency in public databases. This

proposal articulates about the scope and background, the current state of the art, the objectives & methodology along with milestones and other consideration for the project.

We fully expect that this new version of ChemBank will be used as a model for other new databases and for improvements to PubChem and other existing databases. Our goal with ChemBank is not to supplant these other databases, but to provide an example that it is possible to provide context rich data which they may emulate.

Development of ChemBank V3 will enable us to create a chemical sample repository that links physical compound samples used throughout the probe development process, including primary screening, follow-up chemistry, target identification, and Connectivity Map profiling. Probe-development projects involve ongoing chemical syntheses, including multiple independent syntheses of the same compound. To fully support probe development, *ChemBank* v3 must associate biological data with the physical sample used, not just the compound structure, and with sample-specific analytical chemistry information. *ChemBank* v2, which is the current public sharing site, (and PubChem) associates most information only at the level of abstract structure, not physical sample. We will release chemical sample repository as the first component of *ChemBank* v3 that includes analytical chemistry results on the quality of compounds, and synthesis pathway information when available. This repository will link out to PubChem and ChemSpider to make it simpler to use these resources for probe development. At the outset, the repository will contain both public and DSA versions, as not all samples will be publicly disclosed compounds.

ChemBank v3 will also provide a searchable repository of small molecule assays. The screening center captures all biological assays in an electronic lab notebook. We will provide a read-only version of the public portions of these notebooks with both browse and search capability that will any user to quickly locate assays or protocols of interest. Users can then link from these assays to the small molecule results generated from them.

There are several major challenges associated with this project. There are technical challenges associated with the required integrations. The new ChemBank site must be integrated with the PubChem and ChemSpider sites externally, and additionally the back end must integrate with several commercial products licensed to the Broad Institute, including the CambridgeSoft eNotebook Data Warehouse and the Dotmatics Browser.

Additionally, there is a user interface design challenge since the goal is to make the site approachable for those who are not familiar with the experiments it contains. Our internal interfaces are designed for expert users and so work very differently.

3. Conclusion

The development of ChemBank v3 is a research intensive activity. In our opinion, this will translate into at least two research papers in reputed international journals. Also as Life Science research in post-genomic era is integrative, collaborative and mainly in-silico, the development of ChemBank V3 will help the scientific and reseqrch community achieve the following objectives and advantages. The advantages mentioned below once achieved will help the design and creation of other scientific databases. We feel tat as modern day science is data intensive, lessons learned in one scientific domain can be mapped to other scientific domains.

The following are the objectives that will be achieved by development of ChemBank v3:

i) Support dry lab (In-silico) experiments
ii) Avoid reenactment of experiments
iii) Achieve interoperability of data and applications
iv) Enable reusability of workflows and results
v) Share results through transparent exchange of data
vi) Provide inter-application communication
vii) Create, store and access experimentation procedure/methodology i.e workflows as the workflows are considered the research results in the life science research.
viii) Support the autonomous development and collaborative Research.
ix) Most of the genomics databanks and tools do not yet provide enough standardized computer-readable metadata to facilitate the workflow automation and integration. ChemBank v3 will ease the bottleneck of domain-specific knowledge expert needed to interpret what the data actually represents before using it in the integration experiments.

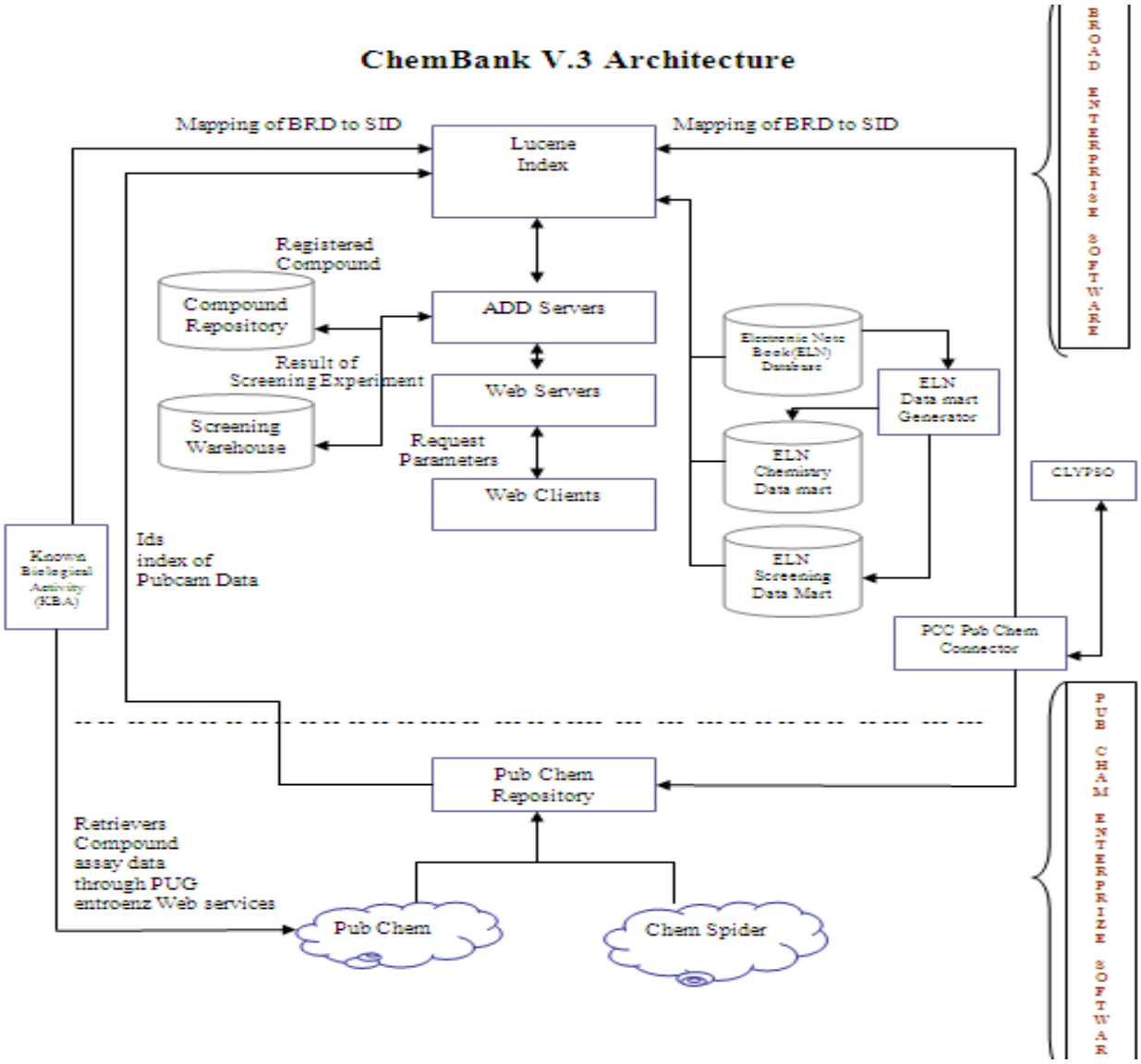

Figure 1: ChemBank V3 Architecture

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. Evan E. Bolton1 et.al.,(2008), PubChem: Integrated Platform of Small Molecules and Biological Activities.
4. Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature 391(6669): 806-11.
5. 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
6. Kathleen Petri Seiler, et al (2008)." ChemBank: a small-molecule screening and
7. cheminformatics resource database, Nucleic Acids Research, 2008, Vol. 36, Database issue D351–D359.
8. Kathleen Petri Seiler , Using ChemBank to Probe Chemical Biology**,** *Current Protocols in Bioinformatics* 14.5.1-14.5.26, June 2008,

Published online June 2008 in Wiley Interscience.
9. Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res**12**(6): 996-1006.
10. Kirill Degtyarenko et al (2008), ' ChEBI: a database and ontology for chemical entities of biological interest', Nucleic Acids Research, 2008, Vol. 36, Database issue
11. Kuhn, R. M., D. Karolchik, et al. (2009). "The UCSC Genome Browser Database: update 2009." Nucleic Acids Res**37**(Database issue): D755-61.
12. Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature**409**(6822): 860-921.
13. Melanie F¨ullbeck,yet .al (2006) ,' Natural products: sources and databases' *Natural Product Reports*
14. Potter, S. C., L. Clarke, et al. (2004). "The Ensembl analysis pipeline." Genome Res**14**(5): 934-41.
15. Robert L. Strausberg, *et al. (2009) , '* From Knowing to Controlling: A Path from Genomics to Drugs Using Small Molecule Probes , *Science 294* – 300.
16. Sachidanandam, R., D. Weissman, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature 409(6822): 928-33.
17. Seiler, K. P., G. A. George, et al. (2008). "ChemBank: a small-molecule screening and cheminformatics resource database." Nucleic Acids Res**36**(Database issue): D351-9.
18. Shoemaker, D. D., E. E. Schadt, et al. (2001). "Experimental annotation of the human genome using microarray technology." Nature**409**(6822): 922-7.
19. Stabenau, A., G. McVicker, et al. (2004). "The Ensembl core software libraries." Genome Res**14**(5): 929-33.
20. Stalker, J., B. Gibbins, et al. (2004). "The Ensembl Web site: mechanics of a genome browser." Genome Res**14**(5): 951-5.
21. Williams, A. J. (2008). "Public chemical compound databases." Curr Opin Drug Discov Devel**11**(3): 393-404.
22. Yanli Wang, et al (2009), ' PubChem: a public information system for analyzing
23. bioactivities of small molecules '.Nucleic Acids Research, 2009, Vol. 37, Web Server issue W623–W633

2/8/2011