STYX : A CML based chem.-informatics facility

[1]Syed Ahsan, [2]Muhammad Shahbaz

[1,2]Department of Computer Science
University of Engineering and Technology, Lahore
[1] ahsancs@hotmail.com, [2] muhammad.shahbaz@yahoo.com

Abstract: Developing countries such as Pakistan lose out the benefits of global cheminformatics research unless immediate actions are taken to strengthen the infrastructure for their participation. They rely on dry labs because they don't have the wet labs. They need infrastructure so that cure can be found of the diseases and infections which attack the local human, animal and plant population as international pharmaceutical companies are less interested in these "poor man " diseases for lack of financial turnover. We feel that instead of supplying disease information and data to the foreign companies, we would then use this information to discover new drugs.
[Syed Ahsan, Muhammad Shahbaz. STYX: A CML Based Chem-Informatics Facility. Journal of American Science 2011;7(8):11-12]. (ISSN: 1545-1003). http://www.americanscience.org.

Keywords: STYX; CML based chem.-informatics facility

1. Introduction

The research community has not been able to benefit from the computational power of modern computational machines because of the ad hoc manner in which the data has been recorded over the years [1,2,3]. Only recently, some efforts have been made to mitigate the problems of disparate, heterogeneous, non standardized and distributed data resources. The major impediment to these efforts is the tremendously huge amount of legacy data. We feel that a dependency on these primary biological data resources which are based in the developed countries will result in inheriting these problem s[4,5,6]. Computational power can help Research work but we lack of

- Specialized DBMS for data that constitutes chemical research.
- Semantic access to already done research work.
- Ontological descriptions for already existing research work.
- Interoperability between various data resources.

We are designing a Database through which researchers and scientists will be able to easily retrieve and use the research work done by other scientists, furthermore they will be able to access the data semantically and ontologically [7].
Although there are few databases available in which chemical data is available but it has only developed country data and there is no data available of local diseases of developing countries [8].
Furthermore the data in these databases are in the form of html and xml, but there are not semantic or ontological search methods available in them, so they are not of much use to scientists and researchers.
We will design a database in which chemical data is represented in a standard format CML (chemical markup language), and will make RDF (resource description framework) graphs of such data [9,10]. After making these changes to the available data formats of such research work we will be able to make semantic and ontological search on it and also by representing this data in CML format we will have an advantage that there are some readymade CML enabled applications available, which makes it easy to manipulate and analyze the data[11]. So not only the scientists and researchers can get the data in more appropriate format, in addition if interested they will also be able to extend this work. Furthermore this data would be available worldwide, so other scientists in the world would aware of it and they might be helpful to find the cure of that disease.

2. Semantic Searches for Chemical Biology

Drug discovery is a vast field, and in this fast going era we daily met across some new diseases that originate from different places of the world, and some time scientists get aware of them, when these diseases exist in developed countries, and mostly don't get aware of them when such diseases originate in under developed countries because of those very good pharmaceutical companies discourage research work to take place in such under developed countries because of negligible profits, and diseases in that area

are left untreated unless they don't make their way to such developed countries. For this reason these diseases are called 'Orphan Diseases'[10,11,].

Now how, what we are doing will make good to this entire imbalance which is mentioned above? This will remove the distances between researchers and scientists present in different areas of world, they will be merged in to single unit, and it will be like every new research or experiment results will be available to you anywhere in the world the very next minute it is done. This will make the all the diseases to be cured a global problem because of its being available instantly in useful form and they will start working together in order to cure diseases [2,11].
Although there is research being done in making web searches semantic, but what if we make CHEM informatics semantically searchable? This will be very helpful for scientists to discover cure of diseases being problem of today, and it should not be the case that someone is wasting time in trying to cure what have already been cured somewhere else and giving time to what have to be cured yet, so combining their efforts and strengthening the research work to take it to a whole new level of collaborating their research work.

3. Cheminformatics tools: IT outsource model

As drug discovery gets more technology-focused, many software companies are already functioning as lab less pharmaceutical companies, cashing in on the market for drug discovery informatics (cheminformatics) and bioinformatics. Also, the increase in external licensing deals for drugs by pharmaceutical companies has resulted in more support for smaller, corporate or university drug development research units, and thus a demand for affordable bioinformatics and chemoinformatics tools for these groups, and a corresponding need for education in bioinformatics and chemoinformatics techniques. The financial return for such IT investments is quantifiable and significant. Successful completion of this project will help local IT companies to grow as Technology Driven Companies providing a technological platform to provide various types of goods and services in molecular biology and drug development chain and will try to emulate the development of IT sector international outsourcing. With the bioinformatics market size expected to hit US$ 3 billion by the year 2011 and a phenomenal growth in its offshoots like cheminformatics and pharmacogenomics, we feel that it is the right time for Pakistan to build bioinformatics/cheminformatics infrastructure and capture a large pie of the lucrative market [5.7].

As major pharmaceutical and genome-based biotech companies invest heavily in software, Pakistan's IT companies have a great business opportunity to offer complete solutions to major pharmaceutical and genome-based biotech companies in the world. This project aims to be only a small step in this direction.

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 cheminformatics resource database, Nucleic Acids Research, 2008, Vol. 36, Database issue D351–D359.
7. 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.
8.Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res**12**(6): 996-1006.
9. Kirill Degtyarenko et al (2008), ' ChEBI: a database and ontology for chemical entities of biological interest', Nucleic Acids Research, 2008, Vol. 36, Database issue
10. Kuhn, R. M., D. Karolchik, et al. (2009). "The UCSC Genome Browser Database: update 2009." Nucleic Acids Res**37**(Database issue): D755-61.
11. Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature**409**(6822): 860-921.

5/10/2011