IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Cloud Cheminformatics as a process of pharmaceutical Engineering for Analyzing Chemical Drug Molecules

Deepak Agnihotri[*1], Priyanka Tripathi[2], Dr. Kesari Verma[3]

[*1,2,3] Department of Computer Applications, National Institute of Technology, Raipur, India

agnihotrideepak@hotmail.com

Abstract

This paper presents a strategy for the use of cloud in Cheminformatics as a process of pharmaceutical Engineering for Analyzing Chemical Drug Molecules. The use of computer and informational techniques, applied to a range of problems in the field of Chemical Engineering using internet. These techniques are used in pharmaceutical companies in the process of drug discovery. These methods can also be used in chemical and allied industries in various other forms. The authors have recommended Cheminformatics packages in this paper to analyze drugs which are like small molecule data. These tools can be helpful to obtain the effect of any drug on cells which can be further useful in cure of various diseases like cancer, swine flu etc. In addition these tools can be utilized to create new molecule structure; it can be compared with existing molecules structures and also can show its 2D as well as 3D visualization. One of the Cheminformatics tools discussed in the paper named ChemMine Web Tool can be used on the internet without installing personalized software on their PC.

General Terms Cheminformatics, Cloud computing, ChemMine

Keywords: Clustering, molecules, physicochemical, similarity searching.

Introduction

Cheminformatics (also known as **Chemoinformatics** and chemical informatics) is the use of computer and informational techniques, applied to a range of problems in the field of chemistry. These *in silico* techniques are used in pharmaceutical companies in the process of drug discovery. These methods can also be used in chemical and allied industries in various other forms. The term Cheminformatics was defined by F.K. Brown in 1998[1, 2, 3, 4, and 6]. Cheminformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization [1, 2, 3, 4, and 8]. These latest Cheminformatics packages contains functions for efficient processing of large numbers of molecules, physicochemical/structural property predictions, structural similarity searching, classification and clustering of compound libraries with a wide spectrum of algorithms. In addition, it offers visualization functions for compound clustering results and chemical structures [3, and 4]. This paper is organized in seven sections with its subsections excluding abstract, General Terms, and Keywords. First Section gives the introduction of the authors work in this paper, Second Section is about literature survey, Third Section is about other technologies in the field of Cheminformatics, Fourth section

describes the applications of Cheminformatics, Fifth Section is based on experiments by the authors for this paper, Sixth Section describes various results and their discussions by the authors based on their experiments, and in the Seventh Section the authors have given the conclusion of their work in this paper. All the concerned references of work used in this paper are mentioned at last.

Literature Survey

Cheminformatics and its application in chemical engineering. There are various tools that can be applied in various chemical engineering applications, but the authors have implemented only the tools that can be applied in cloud computing environment only. The authors have used Cheminformatics packages like ChemMine web Tool [3, 4, and 8], ChemmineR [6, and 11], and BioClipse [7, and 9] in this article to analyze drugs which are like small molecule data. Bioclipse is a free and open source workbench for the life sciences. Bioclipse is based on the Eclipse Rich Client Platform (RCP) which means that Bioclipse inherits a state-of-the-art plugin architecture, functionality, and visual interfaces from Eclipse, such as help system, software updates, preferences, cross-

platform deployment etc [7, and 9]. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either directly at the computer or on hardcopy, and a well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities.(Indeed most of the system supplied functions are themselves written in the S language.) The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis [10]. ChemmineR is a Cheminformatics package for analyzing drug-like small molecule data in R. Its latest version contains functions for efficient processing of large numbers of molecules, physicochemical/structural property predictions, structural similarity searching, classification and clustering of compound libraries with a wide spectrum of algorithms. In addition, it offers visualization functions for compound clustering results and chemical structures. The integration of Cheminformatics tools with the R programming environment has many advantages, such as easy access to a wide spectrum of statistical methods, machine learning algorithms and graphic utilities [3, 4, and 6].ChemMine Tools is a free online service for analyzing and clustering small molecules by structural similarities, physicochemical properties or custom data types. This tutorial introduces the functionalities, data formats, methods and algorithms of this web service [3, 4, and 6].

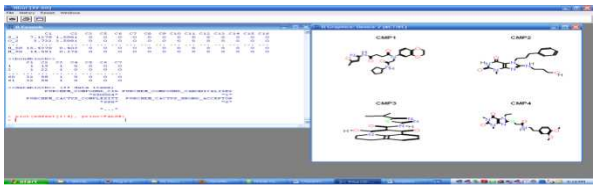


Fig2.1: illustrates ChemmineR Tools functionality

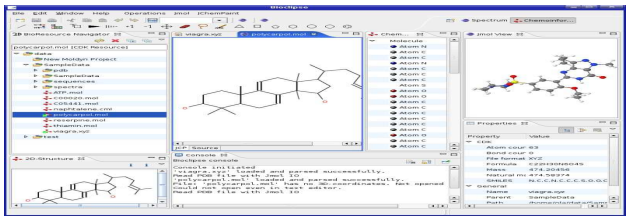


Fig 2.2: illustrate BioClipse functionality

Other Technologies in Cheminformatics

There are several other tools for chemical engineering applications using computers and information technology like, the Cantera using Python for chemical engineering [5], Chemical Engineering using Matlab /Scilab, can also used for Cheminformatics .

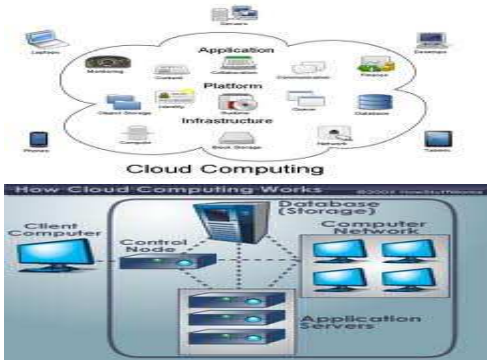
Applications of Cheminformatics

Cheminformatics can be applied for first Storage and retrieval the primary application of Cheminformatics is in the storage, indexing and search of information relating to compounds. The efficient search of such stored information includes topics that are dealt with in computer science as data mining, information retrieval, information extraction and machine learning. Second Chemical File formats .The *in silico* representation of chemical structures uses specialized formats such as the XML-based Chemical Markup Language or SMILES. These representations are often used for storage in large chemical databases. While some formats are suited for visual representations in 2 or 3 dimensions, others are more suited for studying physical interactions, modeling and docking studies [3, and 4].Third Virtual libraries Chemical data can pertain to real or virtual molecules. Virtual libraries of compounds may be generated in various ways to explore chemical space and hypothesize novel compounds with desired properties. Virtual libraries of classes of compounds (drugs, natural products, diversity-oriented synthetic products) were recently generated using the FOG (fragment optimized growth) algorithm. This was done by using Cheminformatics tools to train transition probabilities of a Markov chain on authentic classes of compounds, and then using the Markov chain to generate novel compounds that were similar to the training database [3, and 4]. Fourth Virtual screening In contrast to high-throughput screening, virtual screening involves computationally screening *in silico* libraries of compounds, by means of various methods such as docking, to identify members likely to possess desired properties such as biological activity against a given target. Fifth Quantitative structure-activity relationship (QSAR), this is the calculation of quantitative structure-activity relationship and quantitative structure

property relationship values, used to predict the activity of compounds from their structures. In this context there is also a strong relationship to Chemometrics. Chemical expert systems are also relevant, since they represent parts of chemical knowledge as an *in silico* representation [1, 2, 3, and 4].

About Cloud Computing

You may be familiar with services that involve cloud computing. Some web-based email services are examples of cloud computing implementations. Other examples are web-based document storage, editing and collaboration tools. Cloud computing services are also used for webcommerce. Increasingly web applications are making use of cloud computing, and many contemporary websites use and integrate a a number of cloud computing services. "imagine a world with technology on tap where people can access computing services on demand from any location without worrying about how these services are delivered and where they are hosted. We expect this vision is now becoming a reality." Wikipedia defines cloud computing as:**"the delivery of computing as a service rather than a product, whereby shared resources, software and information are provided to computers and other devices".** Cloud computing provides computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. [12]



Experiments

ChemMine web Tool is a free online service for analyzing and clustering small molecules by structural similarities, physicochemical properties or custom data types. This tutorial introduces the functionalities, data formats, methods and algorithms of this web service. ChemMine tools workbench is shown in fig 4.1, it provides five tool boxes Workbench, Similarity Toolbox, Clustering Toolbox, Search Toolbox, and Property Toolbox for solving chemical engineering problems. Compound Batch Viewing and Format Interconversions Provides utilities to compare compound structures in batches and interconvert between structure formats (e.g. SMILES and SDF). Similarity Comparisons Allows pair wise similarity comparisons among compounds using atom pair and maximum common substructure (MCS) similarity descriptors. Clustering and Data Mining Several clustering and interactive tree/heatmap mining utilities are provided. The clustering utilities include hierarchical, multidimensional scaling and binning clustering. The required distance measures can be calculated by the web service or uploaded by the user. Molecular Property Predictions 38 different physicochemical property descriptors can be calculated for compounds. These include MW, logP, rotatabale bonds, functional groups, etc. Interacting with ChemMine Web Tools from R Power users can interact with this web service from the statistical analysis environment R using the ChemmineR library [3, 4, 6, and 8].

General Functionality

Users can import their custom compounds into the workbench of ChemMine Tools by copy and paste, from local files, or from a PubChem search which includes an online molecular editor. Subsequently, the imported compounds can be submitted from the workbench to the different analysis services .

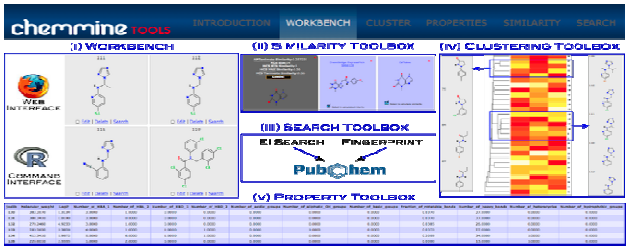


Fig 5.1: Organization of ChemMine Tools

Workbench Overview

All compound structures and other custom data sets are organized in the compound workbench of ChemMine Tools. The workbench interface allows users to add, edit and remove compounds, and to view compound structure images in batches and to submit them to the other online services.**Compound Import:** To import compounds into the workbench, users can choose from the following three options:(1) **SDF/SMILES Import :** Users can import their custom compound structures in the standard SDF and SMILES formats. A brief overview of these structure formats is available on these pages: SMILES and SDF.

(2) Import from PubChem: Alternatively, users can search the PubChem database with text or structure similarity searches and upload the identified

compounds interactively to the workbench by clicking the "Add to Workbench" menus. **Viewing of Compounds in Batches:** Once compounds are imported into the workbench, the user can view them in a list or grid view. The latter allows viewing of all imported compound structures in large batches. Additional functionalities of the workspace interface are: add/edit/delete options and interactive structure similarity searches against PubChem. A starring utility is available in the list view to highlight specific compounds of interest so that they can be easily found in the subsequent analysis outputs of the different clustering tools. (3) **Format Interconversions:** For reformatting purposes, all compounds imported into ChemMine Tools can be saved in SMILES or SDF formats. **Searching for Compounds Similar to Those in Workbench:** After uploading compounds to the Workbench, browse to "My Compounds" and click "Search" next to your compound of choice. This will send the SMILES string to the search tool where you can select either EI or PubChem Fingerprint search. Please note: the similarity search provided by PubChem uses substructure-based fingerprints, while the similarity tools in the clustering and similarity services of ChemMine Tools are based on atom pairs and maximum common substructures[]. Similarity scores between compound pairs can be computed with the Similarity Toolbox. The interface calculates atom pair and maximum common substructure (MCS) similarities with the Tanimoto coefficient, Dice's coefficient, and Tversky index as similarity measures. The MCS tool allows to identify the best substructures that two compounds have in common. The similarity measures provided by the Similarity Toolbox (MCS and AP) differ from PubChem fingerprint similarity, as used in the Search Toolbox. Clustering of compounds by structural and physicochemical similarities is a powerful approach for correlating structural features of compounds with their activities. ChemMine Tools provides facilities for binning clustering, hierarchical clustering and multidimensional scaling (MDS). The required distance matrices for hierarchical and MDS clustering are calculated by all-against-all comparisons of compounds using atom pair similarity measures and transforming the generated similarity scores into distance values. The resulting trees and scatter plots are presented on the web interface in interactive mode using an internally developed cluster mining program that is based on the Google Maps API. After zooming into a tree the tree leaves and internal nodes become click-able to view the corresponding compound structures of a given subtree or cluster. The starring (flagging) utility in the list view of the workspace can be used to highlight compounds of interest in the tree. The tree viewing tool also accepts the upload of custom tables for the generation of heatmaps. This is useful for showing custom data like bioactivity information from HT screens in form of heatmaps next to the hierarchical clustering results. Molecular descriptors provide quantitative information about chemical properties of compounds. They can be very useful for prioritizing lead compounds, property clustering and basic QSAR analyses. 38 different molecular descriptors are currently provided by the ChemMine interface either for custom compounds or those contained in the database. The JOELib package is used for their calculation. After calculating molecular descriptors users are given the option to cluster the workbench compounds based on these data by clicking "Send Table to Workbench and Cluster" [3, 4, 6, 8].

Interacting with ChemMine Tools from R: *Power users can interact with this web service from the statistical analysis environment R using the ChemmineR library.* ***Descriptors, Similarity Measures and Clustering Schemes:*** *This section provides a brief overview of the Cheminformatics and clustering algorithms used by ChemMine Tools.****Structure Similarity Comparisons and Searching of Small Molecules:*** *To compare, cluster and search small molecules with respect to their structural similarities, a common approach is to enumerate their structural features, which are often referred to as structural descriptors. The numbers of common and unique features are then used to calculate a similarity measure among two compounds. The descriptor types and similarity coefficients used by ChemMine Tools are (1) Structural Descriptors, (a) Atom Pairs. Atom pairs are a structural descriptor type that is defined by the shortest paths among the non-hydrogen atoms in a molecule. Each path is described by the types of atoms in a pair, the length of their shortest bond path, the number of their pi electrons and the non-hydrogen atoms bonded to them. The number of atom pairs describing a molecule grows with its number of atoms. To use atom pairs for similarity comparisons, one can simply enumerate their common and unique atom pairs, and then use these numbers to compute a similarity coefficient .*

(b) PubChem Fingerprints Similarity Search: The fingerprints provided by PubChem are a binary representation of the presence and absence of a library of 881 substructure features (see here for details). In this system every molecular structure is described by 881 bits where 1 indicates the presence and 0 the absence of a feature. Compared to atom pairs, the PubChem fingerprints are a knowledge-based system that stores less information than the

much more complex and unbiased atom pair concept. For database searching fingerprints are often much more time and memory efficient, but they are less sensitive than atom pair descriptors. (c) Maximum Common Substructure The maximum common substructure (MCS) problem is a graph-based similarity concept that is defined as the largest substructure (sub-graph) shared among two compounds. It is a pair-wise concept that is not directly related to the above structural descriptors, but its results (e.g. size of MCS relative to source structures) can be used for the computation of the same similarity coefficients. Compared to descriptor-based similarity concepts, the MCS method provides the most accurate and sensitive similarity measure, especially for compounds with large size differences.(2)Similarity Coefficients (a) Tanimoto coefficient: The Tanimoto coefficient is defined as c/(a+b+c), which is the proportion of the features shared among two compounds divided by their union. The variable c is the number of features (or on-bits in binary fingerprint) common in both compounds, while a and b are the number of features that are unique in one or the other compound, respectively [3, 4, 6]. The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. It is important to emphasize that a Tanimoto coefficient of 1 does not necessarily mean that two compounds are identical. It only means that they have identical structural descriptors or identical on-bits in a binary fingerprint. To determine identity among two compounds, InChI strings usually provide a much more reliable solution. This latter feature will become available in ChemMine Tools soon.(b) Tversky Index: The Tversky index is defined as c/(α*a + β*b + c). It extends the Tanimoto index by two weighting variables α and β. If α and β are set to 1 then the index returns the same result as the Tanimoto coefficient. (c) Dice Index: Setting α and β in the Tversky index to 0.5 returns the Dice index. Similarity Measures for Property and Activity Profiles Sets of numeric property values of compounds, such as physiochemical properties or bioactivity values, can also be used to compute a similarity measure among compounds. For instance, ChemMine Tools uses the physiochemical descriptors of compounds - or any numeric custom data set - for the computation of Pearson correlation coefficients as similarity measure for the calculation of an item-to-item similarity matrix that can be converted into a distance matrix for downstream clustering. Clustering based on property values is performed follows: Scaling and centering the property table row-wise by subtracting from each value the row mean and then dividing by the standard deviation of each row. Calculation of an all-against-all Pearson correlation matrix this is transformed into a distance matrix by subtracting each value from 1. Hierarchical clustering of above distance matrix and average linkage as a cluster joining method. Clustering Methods and Schemes (a) Hierarchical Clustering: This service uses the hclust function implemented in R to perform hierarchical clustering. It requires as input a distance matrix of all-against-all compound distances that is generated by subtracting the similarity measure (e.g. Tanimoto coefficient Tc) from one (1 - Tc). The resulting distance matrix is then passed on to the actual clustering program that hierarchically joins the most to least similar items in an agglomerative manner using as cluster joining rule either single, average or complete linkage. The latter parameters are definable by the user. (b) Multidimensional Scaling: Similar to hierarchical clustering, multidimensional scaling (MDS) starts with a matrix of item-item distances and then assign coordinates for each item in a low-dimensional space to represent the distances graphically in a scatter plot. The cmdscale function implemented in R is used for this service. (c) Binning Clustering Binning clustering assigns compounds to similarity groups based on a user-definable similarity cutoff. For instance, if a Tanimoto coefficient of 0.6 is chosen then compounds will be joined into groups that share a similarity of this value or greater using a single linkage rule for cluster joining. This method is based on an internally developed C++ implementation that is very memory efficient since it does not require a distance matrix as input. It calculates the required compound-to-compound distance information [3, 4, 6 and 8].

Browser Recommendation

This site is designed for use with Firefox 3.6 or newer, Internet Explorer 9 or newer, or Apple Safari 5 or newer as it makes use of special features available only in these browsers. Some tools may be unavailable or not function as expected with other browsers [3, 4, and 8].

Results & Discussion

ChemMine Tools provides two powerful structural similarity search algorithms: EI and PubChem Fingerpint. EI Search is an ultra-fast search tool connects directly to the PubChem database, and therefore can return compounds only recently added to PubChem. Both tools accept five different types of input: SMILES strings, structural drawings, SDF, and similarity to existing compounds in the Workbench [3, 4, 6, 8**].**

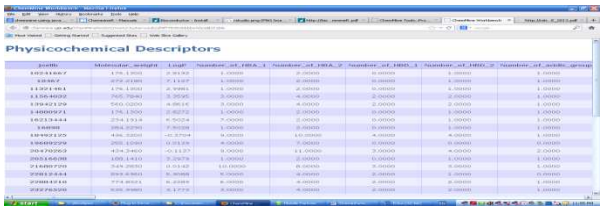


Fig 6.1: Molecular Property Descriptors

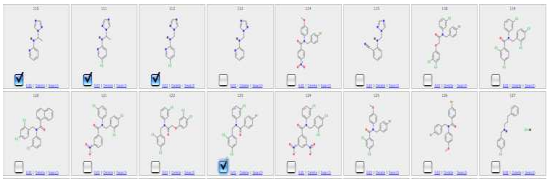


Fig6-2: Compound View in Workbench

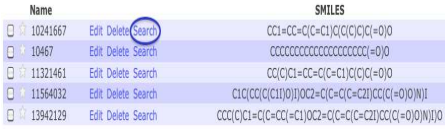


Fig6.3: Search PubChem for Similar Compounds

Fig6.4: Calculation of Compound Similarities

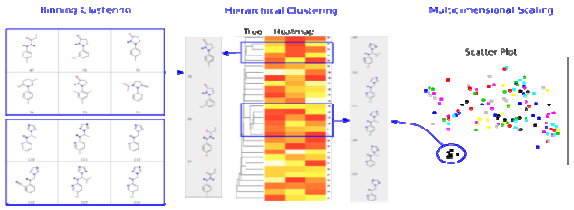


Fig6.5: Clustering tool box

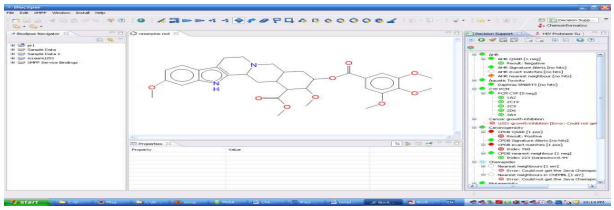


Fig 6.6: BioClipse to analyze the effect of drug on cell DNA/RNA for desease like cancer with its online Decision support system

Conclusions

ChemMine Web Tool and BioClipse with its online decision support system can be used on the internet without installing personalized software on their PC might be helpful for the concerned research persons like pharmaceutical Engineer for identification of various chemical drug molecules. These latest Cheminformatics packages contains functions for efficient processing of large numbers of molecules, physicochemical/structural property predictions, structural similarity searching, classification and clustering of compound libraries with a wide spectrum of algorithms. In addition, it offers visualization functions for compound clustering results and chemical structures. We are working on this tool which is currently limited to a maximum of 1000 compounds, but future updates are planned to expand this limit to at least 10,000. Cloud computing requires not just high speed, but also high quality broadband connections, that are always connected. While many websites are usable on non-broadband connections or slow broadband connections; cloud-based applications are often not usable.

References

[1] All about Cheminformatics at www.cheminformatics.org/ and http://www.en.wikipedia.org/wiki/Cheminformatics

[2] About ChemmineR at http://bioconductor.org /packages /devel /bioc /vignettes / ChemmineR / inst / doc/ ChemmineR.pdf

[3] http://chemmine.ucr.edu/iframe/work/intro/tutorial

[4] http://chemmine.ucr.edu/

[5] All about python project Cantera for Chemical engineerin http://cran.at.r-project.org/

[6] BioClipse can be downloaded from http://sourceforge.net/projects/bioclipse/

[7] ChemMine Web Tools Tutorial http:// chemmine.ucr.edu / ChemMineToolsV2 /work /intro /tutorial # theory

[8] All about BioClipse at http://www.bioclipse.net

[9] All about R project for statistical model at http://www.**r**- project.org/

[10] All about ChemmineR at http://www.bioconductor.org › Bioconductor 2.11 › Software Packages

[11] About cloud computing at http://www.en.wikipedia.org/wiki/Cloud_computing