I571 Fall '08 Main Project

Implementation of Workflows in Pipeline Pilot

Sashikiran Challa

December 10,2008

Pipeline Pilot is a powerful software used in Pharmaceutical Industries like Novartis to streamline data integration, analysis and reporting. It is based around a client-server platform on which one can construct workflows by graphically combining components for data retrieving, filtering, analysis, and reporting. One can retrieve, merge and organize data without performing complex programming or scripting, but just use components that are in-built in the Sci-Tegic Pipeline Pilot. It has the ability to process all common types of file formats with numeric, textual, mathematical, chemical, biological data. It integrates databases, algorithms and different applications on open service oriented architecture. In pharmaceutical industries it can be used to deal with large combinatorial libraries of molecules, build QSAR models, predict activity of molecules, filter out inactive ones, and thus arrive at closest possible drug.

Antidepressant molecules were retrieved from Pubchem, an online database of small organic molecules and their biological activities. 319 molecules were retrieved and were downloaded as a SD file. This SD file containing the Structure, heavy atom count, Canonical Smiles, Isomeric Smiles, etc was read in using SD Reader Component. PubChemCANsmiles property was renamed to 'smiles' and using keep property and Smiles Writer components, a .smi file is generated. In the second pipeline this .smi file was read in using smiles reader component. Then the Lipinski filter component which is based on Lipinski's Rule of Five, that calculates molecular weight, number of H bond donors, H-bond acceptors, ClogP values was used to filter those that satisfied the rule of five. Thus 295 compounds were obtained. These compounds were then piped into Mutagenicity calculator component, which predicted whether a molecule is mutagenic or not. Then Using a Custom filter, with a pilot script in it, all the ones that tested false for mutagenicity were obtained. 285 compounds were predicted to be non-mutagenic.

A web-service can be used in pipeline pilot by loading the .wsdl file onto the 'SOAP with WSDL Support' component available under the integration components in the Pipeline pilot. The Mutagenicity prediction could even be done using the Ames Mutagenicity prediction Web Service available on the Chembiobgrid.org website. Since it took very long nearly 75 minutes to calculate for 285 compounds, the in-built mutagenicity calculator component was used. Molecular Descriptor Webservice was used to calculate the TPSA (Total Polar Surface Area) for those molecules that proved non-mutagenic. TPSA value signifies the bioavailability of the drug molecule. TPSA value should usually range between 60.00 A^2 and 140.00A^2 for a drug molecule to be available to act on the target. If it is below 60.00A^2 the molecule is fully adsorbed and thus

the bioavailability would be less. Thus by using a simple pilot script all those compounds with TPSA value in between $60A^2$ and $140A^2$ were filtered. Then Maximal Common Substructure search was done on all the Molecules obtained. A single Substructure was obtained. Maximal Common Substructure search is normally used in the analysis of set of hits from screening. Thus it can be used to identify common cores with the hits and thereby organize the hits into families.

Thus Pipeline pilot can be affectively used in pharmaceutical industry to handle scientific data, filter out compounds based on different molecular descriptors and thus hasten the drug discovery process.

References:

1) Sci-Tegic Pipeline Pilot Student Edition, Fundamental Training. http://accelrys.com/services/training/scitegic/student.html
2) Sci-Tegic Examples available with the Pipeline Pilot Software
3) Moises Hassan, Robert D.Brown, Shikha Varma-O'Brein and David Rogers. *Cheminformatics analysis and learning in a data pipelining environment.* Review. Molecular Diversity (2006) 10: 283–299, DOI: 10.1007/s11030-006-9041-5.