**Chris Gessner (crgessne)**

**Big data project**

**Project Description**

This was a combined analytical/parallelization effort whose main aim was to understand the chemical diversity represented in the synthetically feasible chemical space accessible to the automated synthetic laboratory at Eli Lilly. The project spans several disciplines from IT/computer science/big data patterns/abstractions to chemical concepts including chemical/drug space diversity. Practical struggles included manipulation of large data sets with relatively limited system resources. Thus much of the thought/effort went into how to debugging issues related to memory constraints including tuning parameters of Spark and Hadoop jobs. An additional challenge included automating the entire process using ansible.

**Problem statement**

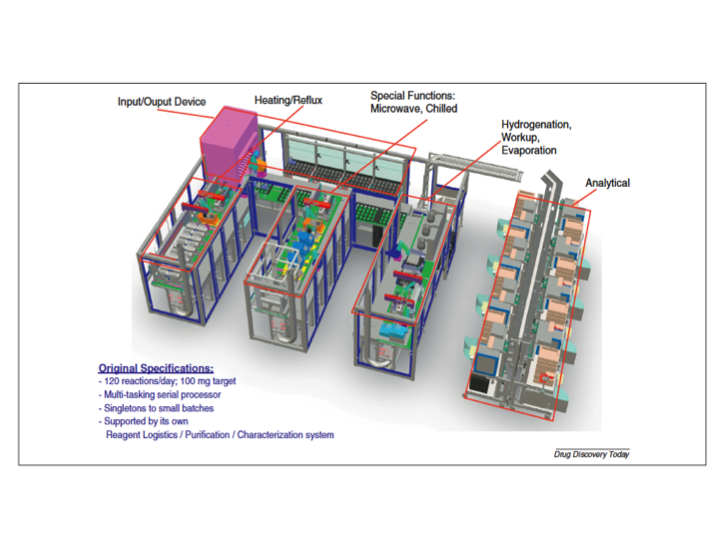
The problem was to leverage big data technologies/techniques combined with FutureSystems infrastructure and class-preferred deployment techniques of ansible to understand the chemical diversity of the automated synthetic laboratory. Furthermore identifying algorithms in the Hadoop ecosystem to apply multidimensional scaling techniques of PCA to reduce 1024-dimensional data to 2-dimensions for visual plotting in x-y coordinate system.

**Purpose and objectives**

**Background/purpose**

For the last several years, Eli Lilly has been involved in creating a new technology involving the automation of certain types of chemical synthesis. This technology is referred to as an automated synthetic laboratory (ASL). It employs advanced robotics and in an automated and even remote fashion, allows the synthesis of chemicals by providing computer program/templates to instruct the robots on exactly how to synthesize the molecule. Additionally sophisticated scientific equipment is used including mass spectrometers that monitor the progression of the chemical synthesis over time, providing electronic feedback in real time. This all allows for someone to remotely synthesize compounds of interest with minimal human intervention {Godfrey, 2013 #1}

The ASL diagram is provided below:

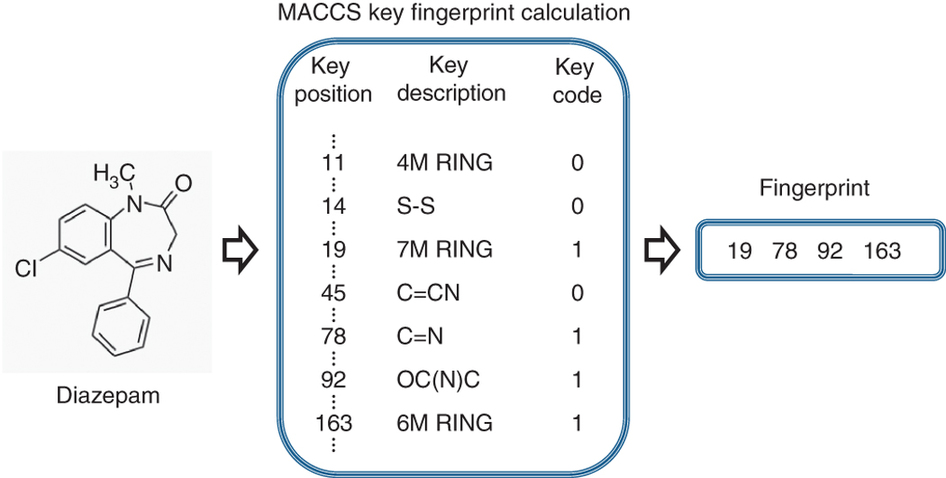


**Objectives**

One of the outstanding questions regarding the ASL is how diverse is the universe of compounds that are synthetically ‘accessible’ to the ASL. In particular what types of drugs can be created from the ASL technology and there any limitations. In order to address this issue, the chemically accessible space of the ASL needs to be compared to the known universe of drugs and optionally ‘bioactives’. Furthermore having the ability to compare the ASL space to all chemical space provides further context for the overall diversity picture represented by the ASL capabilities.

**Molecular Fingerprinting and Scaffold Analysis**

In order to quantitate chemical diversity a compound must first be converted into numbers. These numbers are referred to as a structural key or fingerprint and are comprised of series of bits in a vector where each bit connotes a particular chemical feature or structural aspect. The archetype of these bit vectors is known as the MACCS structural keys and involves 166 keys which represent fixed features of a molecule shown below{Durant, 2002 #2}.

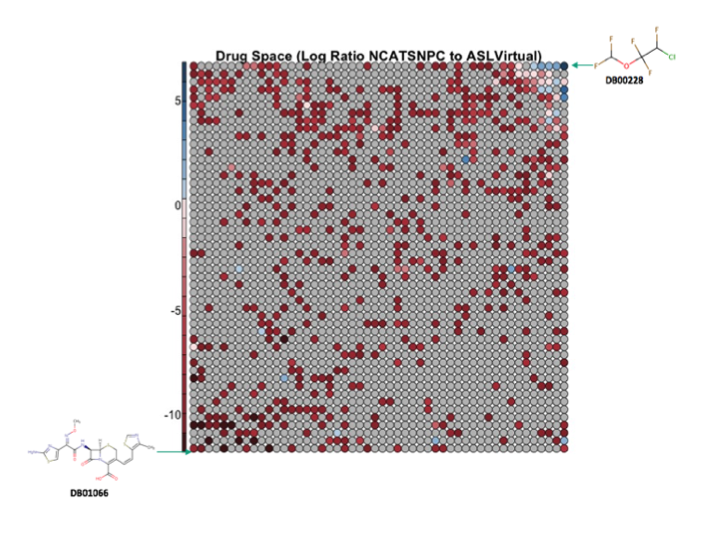


{Nature #3}

A modern modulation of the structural key technique is referred to as a fingerprint and involves treating the molecule as a graph and applies hashing on the chemical graph structure, breaking the structure into fragments, and hashing those fragments into a larger bit vector representing purely structural elements of the molecule. The Chemistry Development Kit (CDK) {Steinbeck, 2003 #3} extended fingerprints are used {Rogers, 2010 #4} here which creates a bit vector of 1024 bits. Each bit represents an arbitrary fragment of the molecule. The CDK is a java based API for chemical manipulations and is the backbone chemical logic for this effort. In addition to generating fingerprints, molecular scaffolds are generated using the CDK. Moreover the fingerprints used in this effort are generated from molecular Murcko scaffolds {Murcko, 1996 #6} of a compound to normalize the structures for more meaningful comparison between molecules.

**Early Efforts to Model Chemical Diversity**

Initial attempts to represent the chemical diversity space of the ASL involved using Kohonen self-organizing maps {Kohonen #7}. Here multidimensional space is reduced to 2 dimensions with a topological ordering that preserves relationships between the higher dimensional data in lower dimensional space (i.e. points close to each other in 2D should also be close to each other in higher dimensions). An example of this effort is shown below.

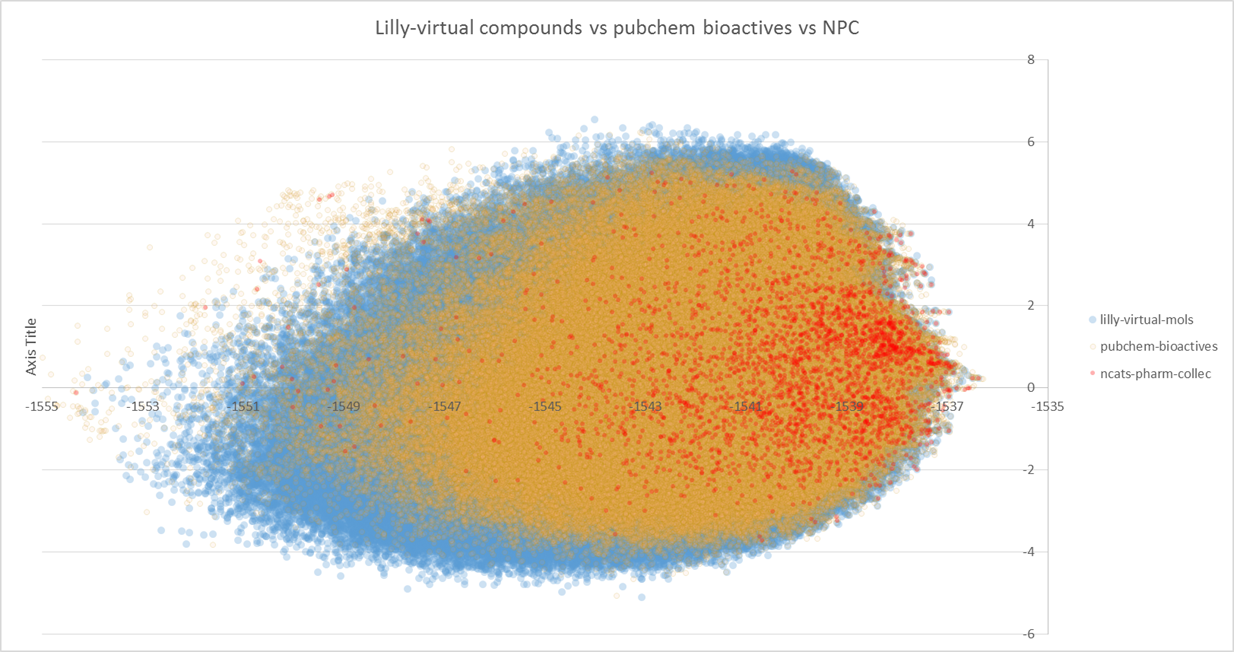


Here red circles represent chemical clusters where the ratio of Lilly virtual compounds is greater than drug space. Blue circles represent the reverse; areas where the ASL could focus on new synthetic strategies. This representation if flawed however because of the large size of the data combined with analysis in R on a single desktop machine. Thus oversimplifications were made including compressing the bit length to 200 instead of 1024. Also the 1.9M Lilly virtual compounds could not be clustered with the drug space compounds; R could not cluster into the millions. So instead only a 100K random sample of molecules from the 1.9M Lilly virtual compounds were clustered with the drug space compounds. The remaining < 1.9M compounds were then ‘mapped’ to the Kohonen SOM, thereby reducing/compressing the true diversity picture.

An alternative strategy to dimension reduce is to use multidimensional scaling. In particular principal component analysis was performed to 2 dimensions. These 2 dimensions were plotted on an x-y coordinate axis

**Results**

Multidimensional scaling techniques of principal component analysis were applied to 1M bioactive compounds from Pubchem, 1.9M Lilly-supplied virtual compounds representing chemically accessible space by the ASL, and 5K drug compounds representing all drug space, publically available from Pubchem sourced from NCATS Pharmaceutical Collection (NPC). The following plot shows all compounds reduced to 2 dimensions from 1024 originally. Blue circles represent Lilly virtual compounds (ASL). Orange circles represent 1M Pubchem bioactive compounds. The red circles represent 5K drug compounds. The x and y axes are the top 2 principal components (which account for the most variation in the data set) over all 3M compounds.



**Findings/Implementation**

Beyond PCA results mentioned above, there are lower level findings related to the overall workflow need to produce the above dimensional reductions over the 3M compounds. Settings/implementation and findings are described below.

Infrastructure: Chameleon, 20 VMs flavor=large, 80 CPU-COREs

Because customization of Hadoop Map-Reduce was needed to read multiline, single record SDF records and because of the large number of records and variability in the computational work required to process uniformly sized chunks of molecular data, customized SDF Mapper/Record reader were implemented in Java:snippets taken from: {Guha #8}. Initially it was thought that Spark could accommodate all map-reduce operations, obviating the need for lower level Hadoop APIs of map-reduce. However after weeks of toil with Spark it became apparent that their textfile and wholetextfiles reader were not sufficient for this data set at scale. Thus the effort became two-part.

Phase-1 Map-Reduce: molecules to fingerprints

Input:

|  |
| --- |
| 4000 SDF files each with 25K molecules therein – pubchem bioactives 100-200MB compressed per SDF file on average. 4X larger uncompressed. |
| 1 SMILES file of 1.9M lines, each is one molecule – Lilly ASL virtual cmpds |
| 1 SDF file of 14K molecular entities (reduces down to 5K drug compounds)- NPC |

Map/reduce using the parallelism of map only (no reduce) to break up the 3M compounds into manageable inputsplits. Then calculate in the mapper the molecular scaffolds and then fingerprints for each chunk of data (set of molecules in SDF format).

Map-reduce Settings:

|  |  |
| --- | --- |
| mapred.max.split.size | 1000000 |
| mapreduce.map.java.opts | -Xmx7500m |
| mapreduce.map.memory.mb | 7600 |
| mapred.reduce.tasks | 0 |
| mapreduce.job.split.metainfo.maxsize | -1 |
| mapred.max.map.failures.percent | 10 |

Phase-2 Principal component analysis in Spark:high dimensional fingerprints to 2D.

Input is 3M CDK-extended fingerprints representing molecules.

Spark PCA settings:

|  |  |
| --- | --- |
| deploy-mode cluster | yes |
| driver-memory | 4G |
| spark.network.timeout | 240s |
| executor-cores | 3 |
| num-executors | 38 |
| executor-memory | 7G |

**Deployment**

Ansible was used for installing/building/compiling and deploying artifacts from local node to remote Hadoop cluster.

Two roles are used.

1. maven
   1. Connects to remotes as user:cc
      1. Needed to escalate sudo permissions as cc to install maven on master0.
2. asl
   1. Connects to remotes as user:hadoop
   2. Handles building deploying src.
   3. Copies data in and out of hdfs
   4. Runs Hadoop/Spark jobs
   5. Copies output from jobs back to local node for successful job verification

Due to the large data sets, the asl ansible play tests/verifies the functionality of

1. creating scaffolds/fingerprints from molecules in SDF format
2. performs PCA analysis on a compound.

**References**

1. Godrey AG et al. Drug Discov. Today 2013, 17, 795-802.

2. Durant JL, et al. J. Chem. Inf. Comput. Sci. 2002, 42, 1273-1280

3. Steinbeck, C. et al. Journal Chem. Inf. Comput. Sci.  2003, 43, 493–500.

4. http://www.nature.com/nprot/journal/v9/n9/fig\_tab/nprot.2014.151\_F2.html

5. Bemis GW, Murcko MA. J. Med. Chem. 1996 19, 2887-93.

6. Rogers D, Hahn M. J. Chem. Inf. Model. 2010, 50, 742-754

7. <https://en.wikipedia.org/wiki/Self-organizing_map>

8. Guha R. <http://blog.rguha.net/?p=293>. Code snippet: Map reduce for SDF RecordReader/Mapper.