Week 1 (18.10 - 24.10.2019)

**Part 1**

I have summarised and tried my best to capture the key points and essence of the papers that I have read.

**Paper 1: Introduction to Cheminformatics**

This paper gives one a deeper understanding of what Cheminformatics really is and how it is related to or differs from Bioinformatics.

**1.1 How Cheminformatics is related to Bioinformatics:**

* + Both of them need electronically accessible databases and database search tools.
  + They both require structure prediction software, data exchange standards and structure visualization software.

**1.2 How Cheminformatics differs from Bioinformatics:**

* + Bioinformatics focuses on large molecules like DNA, RNA whereas Cheminformatics focuses on molecules whose size is <1000 Da

The linkage between small and large molecules is what connects Bioinformatics and Cheminformatics. Small molecules like metal ions regulate genes and genes (DNA, RNA) in turn are responsible for synthesis and degradation of most small molecules.

**1.3 Five most active areas in Cheminformatics research:**

* 1. Cheminformatic data formats
     + Chemical substructures and subtle chemical variations are not captured efficiently with text representations. In contrast, fingerprint representations serve as a more effective way to quickly match chemical structures.
     + Examples include Structure Data Format (SDF), MOL Format, PDB formats etc
  2. Cheminformatic Utilities
     + This involves software needed for data format conversion, chemical structure reading/writing, chemical structure visualization, chemical property and feature calculation etc.
     + Examples include OpenBabel, JOELib etc.
  3. Cheminformatic Databases
     + These databases can be divided into 5 further categories: Chemical compound databases, spectral databases, small molecule pathway databases, drug or pharmaceutical databases and metabolomics databases.
     + The most popular are PubChem, ChemSpider, ChEBI.
  4. Predictive tools for Cheminformatics
     + The availability of free experimental chemical data has led to the development of a number of high quality tools that help in prediction of chemical and biological properties, metabolism etc.
     + Predictive Cheminformatics falls into 5 categories: Spectral prediction, structure prediction, chemical property prediction, biological property prediction and metabolism prediction.
  5. Analytical tools for Cheminformatics
     + Cheminformatic analysis tools combine different forms of databases along with visualization tools that incorporate machine learning, feature selection and pattern recognition.
     + These tools are used to analyse drug screening, toxicological and natural product screening data.

**Paper 2: Machine Learning in Cheminformatics and drug discovery**

There are many methods that help convert compound structure into chemical information suitable for ML pipeline, such as chemical graph retrieval, descriptor generation, fingerprint construction.

* 1. Chemical Graph Theory
     + It states that chemical structures are fully specified by their graph representations and hence have all the information required to model and provide an insight into a wide range of biological activities.
     + Chemical graphs represent atomic connectivity using a bond adjacency matrix or a topological distance matrix.
  2. Chemical Descriptors
     + They are numerical features extracted from chemical structures for molecular data mining, compound diversity analysis and compound activity prediction.
     + Chemical descriptors can be 0D, 1D, 2D, 3D or 4D, out of which 2D is the most common one.
  3. Chemical Fingerprints
     + They are high dimensional vectors, commonly used to uniquely identify an chemical compound using analytic methods like x-ray spectroscopy.
  4. Chemical Similarity Analysis
     + The objective of chemical similarity analysis is to search a given database and return compounds with similar structures and bioactivities as the query compound.
     + The basis for this search operation is that compounds with similar structures will share similar bioactivities.

**2.1 Supervised Machine Learning**

Labels are assigned to the training data, and trained models can predict labels for given data inputs. They include,

* 1. Naive Bayes
     + This is a classification method that tells us the probability of the test data being assigned to the correct label based on the relative proportions of labels in the training set.
  2. Regression Analysis
     + Given a set of training data points, linear regression analysis tries to find a linear function that fits the given data well, with least loss (L1 or L2 loss).
     + But the linear nature of the model is sometimes not enough to model complex QSAR systems.
  3. K-Nearest Neighbours
     + The labels from the closest nodes in the training set to the query are transferred to the query using a majority vote rule.
     + But the value k is a free parameter and hence has to be set by the user and this influences the model a lot.
  4. Support Vector Machines
     + It is a classification method used to project given training data into a sufficiently high dimensional space where an optimal hyperplane that can easily separate the classes.
  5. Neural Networks and Deep Learning
     + Neural networks is a set of ML algorithms that are inspired by the operations of neurons in the brain. Each neuron receives many inputs, performs a weighted sum of the inputs and generates an activation signal based on an activation function on the weighted sum.
     + Deep learning is an extension of ANN's that use specialized architecture to learn and extract useful features from raw data.

**2.2 Unsupervised Machine Learning**

This includes techniques that learn the pattern of the underlying features directly from unlabelled data. Examples include PCA, ICA, Clustering algorithms.

There are several methods to combat the curse of dimensionality and collinearity for supervised and unsupervised ML, like regularization (L1, L2), dimensionality reduction (PCA) and genetic algorithms.

**2.3 Areas in which ML for Cheminformatic can be improved**

* + Increased use of Big Data in ML to predict wider range of biological phenomena.
  + Incorporating multiple data types and sources, also known as ‘data fusion’ techniques, that aggregate structural, genetic and pharmacological data from the molecular to organism level.
  + Methods that are more interpretable and reduce overfitting.

**Paper 3: QSAR Modelling: Where have you been? Where are you going?**

QSAR models find a broad area of application in for assessing potential impact of chemicals, materials and nanomaterials on human health and ecological systems. The surge in availability of chemical data has led to the rise of research and improved techniques and methods in the field of QSAR modelling.

**3.1 Challenges in QSAR modelling**

* + Failure to take into account data heterogeneity: For example, data from the same species.
  + In QSAR, the unit that has to be used is molar units (mol/kg) and not weight units (mg/kg).
  + Use of highly collinear descriptors: Reduces quality of prediction.
  + Errors in descriptor values
  + Unacknowledged omission of data points like outliers.
  + Replication of chemicals in a dataset, which could happen due to different names for the same chemical.
  + Overfitting of data.
  + Large number of descriptors
  + Lack of descriptor autoscaling.
  + Inadequate amounts of training or test data sets.
  + Inadequate QSAR model validation

**3.2 Current Trends in QSAR Methodology**

3.2.1 Chemical Data Curation:

* + There are two main types of errors in the input data: directly related to chemical structures, and related to the associated experimental measurements.

3.2.2 QSAR in Toxicity Prediction:

3.2.3 QSAR Prediction of Metabolism:

* + Special focus is placed on drug metabolism during drug discovery since a major issue in pharmacotherapy is adverse drug reactions. A drug metabolite might be very different from the parent drug and hence can cause a lot of harm or severe side effects.

3.2.4 Interpretability of QSAR models:

* + Here, the term interpretability refers to the ability of a user to understand and rationalize the chosen descriptors and the predictions of the model.
  + Easily interpretable descriptors and predictions helps user further understand the mechanism(s) of action.

3.2.5 Multitask Modelling:

* + QSAR models are generally developed for individual target properties, but sometimes a single chemical can have many biological functions. Taking this into account in the models could improve predictions drastically.

**3.3 Novel applications of QSAR and future trends**

3.3.1 QSAR modelling of Peptides:

* + Anti-Microbial Peptides (AMP's) form an integral part of the immune system of several living organisms. In recent years, we have seen the rise of resistance of microbes against antibiotics and this is mainly where AMP's have an advantage.
  + They have low toxicity, broad range of activity and minimal level of development of resistance by microbes.
  + But even so, we do not fully understand the Structure-Activity relationship and many techniques like sequence based AMP modelling are being used.

3.3.2 QSAR modelling of Chemical Mixtures

* + It is intuitive that a mixture of chemical compounds can be more powerful and effective than their individual counterparts. But QSAR modelling does not do well for mixtures, mainly because of lack of reliable data.
  + This problem can be overcome by selecting good descriptors for the mixture.

**3.4 QSAR and regulatory decision support**

* + QSAR plays an important role in accessing various properties of a chemical compound and how suitable it is to the environment and other life forms. It is used in the regulation of food, cosmetics, industrial chemicals etc
  + As governments and environmental agencies come to heavily depend on results from QSAR modelling, there is a need to come up with new policies and regulations to increase confidence in QSAR predictions and provide regulatory bodies with enough scientific basis to make decisions from these predictions.

**3.5 Best practices and future of QSAR modelling**

* + Best practices are integral to ensure the overall integrity and validity of any QSAR modelling study and hence we focus on 4 areas: data collection and curation, model building, rigorous external validation and model use.
  + External validation is a very important step as it tells us about the quality of the model.
  + Also, model interpretability is a key issue as it is being used by many people from the non-scientific communities to come up with new regulations and policies.
  + The relative use computational modelling and possibility of carrying out advanced calculations without really knowing the intricacies has led to a significant failure in adhering to best practices standards.
  + But with rigorous editorial and peer review standards, the number of low quality QSAR publications have significantly decreased.

**Part 2**

I went through the tutorial on how to get started in Galaxy and how to use some tools and got some hands-on experience.

I am already familiar with scikit-learn, but not so much with Keras. Hence I went through some tutorials and videos to know more about it.