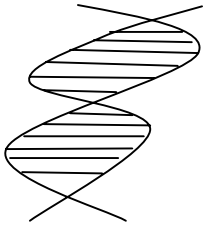

## A report on:

# Molecular Feature Mining in HIV Data

Stefan Kramer, Luc De Raedt and Christoph Helma

Institute for Computer Science, Machine Learning Lab

Albert-Ludwigs University Freiburg

**by Arpit Gattani**

gattani@cc.usu.edu

---

## Problem Statement:

**AIDS** is one of the most perilous thing happened to human community in last couple of decades. Lately the developments in Data Mining field enabled the database research community to suggest and implement some very effective solutions to biomedical field and especially to AIDS.

There are about 43576 compounds in the DTP Database, which were classified as the compounds having capability to defend a HIV-1 attack on human body. The prime problem on which the authors of this paper concentrated is the **reorganization of the molecular substructure those are frequent in the compounds (with minimum support)** which were found active towards the capability to protect human body against HIV-1 and **molecular substructure those are infrequent in the compounds (with maximum support)** which were found inactive for the same. Above problem was more complicated than normal because the distribution of the data was highly skewed as only 1.3% of the molecules were known to be active and 2.7% were moderately active, leaving remaining 96% molecules to be inactive.

Another problem with which the authors were concerned about was the **storage of such a large data and its processing format**. Other already proposed database techniques like connection-tables, multirelational databases and graphical representation would have also proved a bit more time and space consuming. For both of these problems, authors proposed efficient and accurate solutions.

The paper was aimed towards the reorganization of the molecular structure frequent in the anti-HIV-1 compounds and enables the pharmacist to develop more effective and improved drugs.

## Contribution of the paper:

Inductive databases are one of the data mining techniques which authors used to find the particular patterns in the available database. ***"Inductive Databases integrate data with patterns, i.e. generalizations and regularities"***(from paper). Authors proposed a domain specific inductive database called **MOLFEA (Molecular Feature Miner)** to compute the chemical compounds which searches for the fragments in a chemical compound. A specific format to store chemical compounds was used called SMILES (Simplified Molecular Input Line Entry System) and the fragments were formulated in SMARTS language**. Levelwise version space algorithm** was used by MOLFEA to compute the inductive database queries.

**SMILES representation of the Molecules:** SMILES follow some notions to represent the chemical data. First, elements are denoted by their typical symbols in chemical world and the bonds between them is represented by '-', '=' or '#' respectively for single, double or triple bond. Second, the elements of aromatic atoms are represented in lower case while other atoms in uppercase. Third, Hydrogen atoms and single bonds, under certain conditions, are needed not to be represented. They can be calculated by other given information. Fourth, cyclic structures are represented by breaking one bond in each ring. The atoms adjacent to the broken bond obtain the same number. And last side-structures are written between the brackets.

SMILES plays a very significant role in chemical databases representation. ***SMILES is the language of chemists. "SMILES is a language designed, understood and "spoken" by (computational) chemists"*** (from paper). It gives very compact database to store and manipulate. Also there are many efficient and optimized tools available which uses SMILES.

**SMARTS representation of Fragments:** SMARTS is a subset of the SMILES which is proposed to represent the fragments. The subset of SMILES used in the paper is ***M***. The properties of ***M*** described in the paper are,

- fragment ***M*** is ordered by the '*is more general than'* relation. This means is fragment g is more general than fragment s than it can be represented by $g \leq s$.
- in ***M***, two syntactically different fragments, *g* & *s*, can only be same if reversal of *s* is equal to *g*.
- $g \leq s$ if and only if *g* is a subsequence of *s* or the reversal of *s*.

**Approach to the solution domain:** Simple Apriori approach, in general, lacks the user control and focus over the derivation of the association rules. Authors used constraints based mining rules for mining frequent fragments from a data set. The primitive constraints used in the algorithm are which can be imposed on any given fragment *f* were,

$f \leq p$, $p \leq f$, $\neg(f \leq p)$, $\neg(p \leq f)$: where *f* is the unknown target fragment and p is a specific pattern, *freq(f, D)* denotes the relative frequency of a fragment *f* on a set of molecules D. This can be defined as the percentage of molecules in *D* that *f* covers and $freq(f, D_1) \leq t$, $freq(f, D_2) \geq t$ where t is a positive real number and $D_1$ and $D_2$ are sets of molecules. This constraint means the relative frequency of *f* on the database $D_i$ should be longer than or equal to *t*. These primitive constraints were combined to mine the features of interest in the data. A solution space $sol(c_1 \wedge c_2 ..... \wedge c_n)$ was defined. For every constraint c it is defined if it is monotonic or anti-monotonic. A constraint c was defined monotonic if

for every s, g belongs to *M*: $(g \leq s) \wedge$ *(g belongs to sol(c)) implies that (s belongs to sol(c)).*

The monotonic constraints authors dealt with are $(p \leq f)$ and $freq(f, D1) \leq t$ and other two are anti-monotonic. Solution spaces of the Constraints were bounded by the borders. The boundary values was defined as S(c) = min*(sol(c)) and* G(c) = max*(sol(c)).* With the above defined constraints and cases, Levelwise version space algorithm is applied to the dataset. With each incremented level, infrequent fragments were deleted from the dataset and the

process was repeated with next collected dataset. This process continues to run till we have an empty set left. An improved form of the Levelwise version algorithm has been applied which eliminated all the infrequent fragments rather than the frequent fragments from the dataset at a particular level. That helped to compute the frequent fragments at a level without taking into the account the fragments those were not included in the previous level.

**Implementation:** The levelwise version space algorithm was applied in the Prolog and the tests at each level were implemented in SMARTS and SMILES using Daylight Toolkits. Database was picked up from DTP AIDS Antiviral Screen Databases (http://dtp.nic.nih.gov).

Experiment was done by applying the MOLFEA to two sets of database, D1 and D2 and the frequent fragments were find in D1 and infrequent fragments were find in D2. The choices of the constraints were made independently. Minimum relative frequency was kept 3%, arbitrarily, to find the minimum support in active compounds and maximum frequency in inactive compounds, and to find the minimum support in active compounds and maximum support in moderately active compounds. $\chi^2$-test was applied to a 2 x 2 table with class and frequency of fragments as variables which determine the maximum frequency allowed in the non-active compounds and to get the significance of the frequent fragment.

**Results:** Results of the experiments were categorized in two different tables. First shows the frequent fragments found at different levels and time taken by the system to process fragments with minimum support in active compounds and time taken to process fragments with maximum support in moderately active and inactive compounds. Results shows that the majority of the solution fragments were found in the middle levels of the experiments and also the time taken was maximum at those levels. Second table was formed with the ranked fragments from boundary values S and G. The fragments were ranked according to the significance $\chi^2$-test or accuracy (# active compounds / (# active compounds + # inactive compounds). From the experiment, 17 really significant and accurate fragments were selected. The detailed study of those fragments revealed that most of them were close

relative of Azidothymidine which is a strong inhibitor of the HIV-1 replication. The other remaining fragments were found to be the other type of HIV-1 reverse transcriptase inhibitor, thiocarboxanilide derivatives.

## Limitations of the Approach:

Approach adapted in this paper has some limitations. First, despite the fact that SMILES representation of chemical structures is more compact and clearer to understand than normal techniques; **it is not unique and is restricted to specific biochemical data.** Second**,** the algorithm used in this particular approach is the standard Levelwise version space algorithm which is a variant of the Apriori. **Apriori algorithms are lately found inefficient in finding long patterns in data set.** Value of support needs to be high to get some practically important rules. Third, the substructures found at one level do not combine to generate substructures for next level. In this case the fragments generated at the next level are not the fragment with all the variations of the previously found subfragments. Fourth, above algorithm was approached to find linear fragments, i.e. the chains of atoms. However such algorithms have limited use in many other real world applications as **most deals with the frequent structures comprised of rings and common branch points**. Fifth, the algorithm does not keep track of the canonical structures of the frequent fragments found. Hence more time was expended on mining some redundant data. Last but not least, the algorithm starts with set S = {$\perp$) for monotonic constraints. But if we are starting our search from bottom, it's not possible to do it with S = {$\perp$).

## Space for Further Enhancements:

As mentioned earlier, Apriori algorithms are found inefficient in dealing with database with long patterns. Many a times to reduce the difficulty of managing the generating patterns and data set, researcher maximize the value of support which in turn leads to the lost of some important data patterns. Else some kind of pruning strategy has to be applied to make the generated dataset manageable. During the pruning many a times we prune the false

positives and leave false negative unpruned. To overcome this problem, a more efficient and accurate algorithm can be used. Various new algorithms are proposed recently which deals with long data bases more efficiently than Apriori or its variants like **MaxMiner algorithm, pattern decomposition algorithms and gSpan (graph based Substructure pattern mining).** gSpan algorithm process the long chemical database by representing the entire dataset in graph pattern. Then the DFS is applied to the dataset to find frequent patterns without candidate generation and false pruning. Though the gSpan technique will take more memory, work done and hard to implement, it will generate more accurate results in much less time. In the field of biomedical data mining, it is more vital and valuable to get accurate results than fast and convenient. Also the result generated by graphical pattern mining technique would be easily understandable by the people who are ultimately going to use those results like biologists and pharmacists.

Another limitation of the current approach, inability to combine the frequent substructures found at one level with next level, can be solved by such a graphical approach. For any level of DFS, we can choose the lower bound dynamically that includes the frequent subfragments found at the previous level. Canonical structures can also be traced efficiently while the DFS.

## Conclusion:

As a student of Data Mining, paper helps me a lot to understand some of the basic concepts of the field like constraint based mining techniques, inductive database mining techniques and apriori rules. **MOLFEA** allows us to find the sequential fragments in the data. It can be very useful in finding many hidden secrets in protein and DNA databases which can lead to the discovery of treatment for many deadly diseases like cancer and AIDS.