DOPING ANALYSIS: REPORT

TEAM MEMBERS

- Bhavik Agarwal 2018385
- Mrinal 2018398
- Gavish Gupta 2018390
- Prutyay Gautam 2018403
- Ria Gupta 2018405
- Sanskar Sachdeva 2018411

INTRODUCTION

Research Question - Can we prevent the un-intentional banning of a sportsperson due to unknowingly eating or consuming something, later causing a ban also is there a platform where we can get all the information revolving around the banned drugs and compounds by WADA?

We must prevent inadvertent doping caused by the intake of a banned substance. Most sportpersons are not aware of the associated food and medicines which may cause accidental doping and can ban them by WADA. Additionally, there is a need to list some of the health risks and diseases associated with that intake.

- FINDING BANNED DRUGS/FOOD SUBSTANCES
  Used the list of banned compounds/chemicals provided by WADA(World Anti-Doping Agency).
- FINDING TARGET PROTEIN OF THE BANNED DRUGS
  Finding target proteins and their sequences by using different databases and doing molecular pathway analysis.

- PREDICT IF A PROTEIN IS A TARGET OF A BANNED DRUG
  Predicting whether a protein can be targeted by a banned drug using a Machine Learning algorithm.

- MOLECULAR PATHWAY ANALYSIS
  Finding molecular pathways of drugs banned by WADA and using it for further analysis.

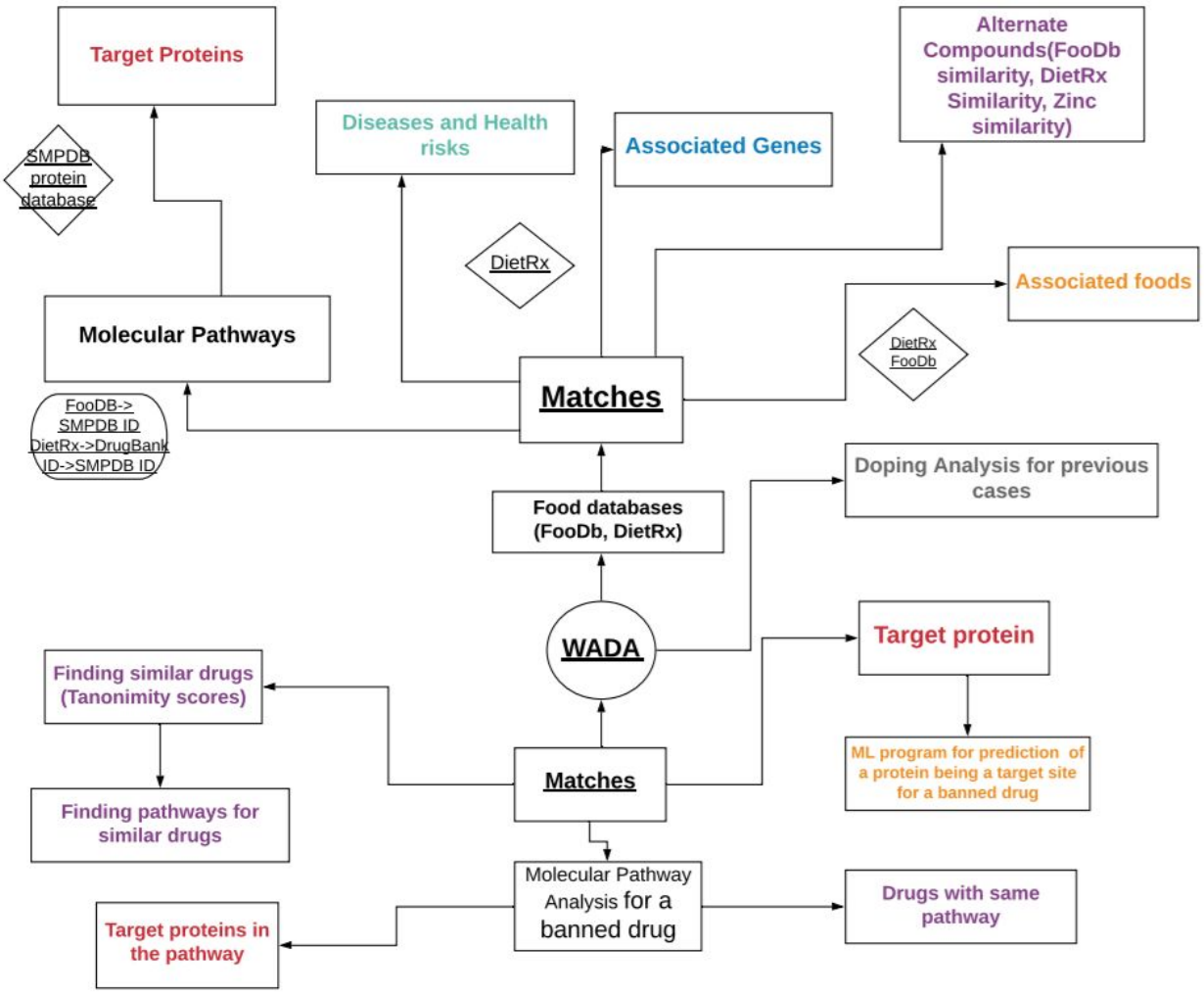
- ALTERNATIVE DRUGS/FOOD SUBSTANCES
  Providing Alternate food items/medicines available for banned food items that reflect the same nutritional value or similar drug effect.

    - Drugs
        - Tanimoto Similarity
        - Pathway Similarity

    - Food Substances
        - DietRx similarity
        - ZINC similarity
        - FooDB similarity

- ANALYSING HEALTH RISKS AND DISEASES ASSOCIATED
  Listing down all potential health risks and diseases associated with consuming banned substances.

WORKFLOW


Target Proteins
SMPDB protein database
Molecular Pathways
FooDB-> SMPDB ID
DietRx->DrugBank ID->SMPDB ID
Diseases and Health risks
DietRx
Associated Genes
Alternate Compounds(FooDb similarity, DietRx Similarity, Zinc similarity)
Associated foods
DietRx FooDb
Matches
Food databases (FooDb, DietRx)
Doping Analysis for previous cases
WADA
Target protein
ML program for prediction of a protein being a target site for a banned drug
Finding similar drugs (Tanonimity scores)
Finding pathways for similar drugs
Matches
Molecular Pathway Analysis for a banned drug
Drugs with same pathway
Target proteins in the pathway

METHODOLOGIES

All the steps and methods which we have used are towards finding/creating a solution for our research question.

DATABASES USED:

1. WADA
2. DrugBank
3. FooDb
4. DietRx
5. SMPdb metabolites
6. SMPdb proteins
7. Foodball
8. FoodData Central
9. FoodComex

TOOLS USED:

1. Python-Pandas
2. Superdrug2
3. Jupyter notebook
4. Scikit-learn
5. Bio.SeqUtils
6. ProteinAnalysis
7. Numpy
8. Machine Learning
9. Web-Scraping
10. ZINC

STEPS:

1. **Finding banned Drugs/Food Substances**
    a) **Drugs**
        - Tools used: Python CSV
        - Description: Extracting the list of drugs and compounds banned by WADA. The IUPAC/common names of compounds banned by WADA were matched by the Drugbank database “Drugbank vocabulary” using a python script FindBannedDrug.py, Readme, Output

    b) **Food**
        - Tools used: Python CSV

- Description: Extracting the list of compounds found in food items banned by WADA. The IUPAC/common names of compounds banned by WADA were matched by offline FooDB and DietRX databases using python scripts.
  FindBannedFood_FooDB.py, Readme, Output
  FindBannedFood_DietRx.py, Readme, Output

2. Analysing drug matches

a) Target proteins of these banned drugs:

1. Finding the target proteins and sequences of these banned drugs:

- Tools used: Python, CSV
- Description: Target proteins and their sequences are extracted by the script FindTargetProteins.py, Readme. This is done via DrugBank Database. Input files include a csv file containing drug bank ids of all banned drugs - DrugBankMatched.csv and a fasta file containing all target protein's names and their protein sequences in a fasta format: protein.fasta . This is the Output obtained.

2. Predicting whether a protein is targeted by a banned drug or not: (ACCURACY=94.9%)

- Tools Used: Python, Jupyter notebook, Scikit-learn, Bio.SeqUtils, ProteinAnalysis, Numpy,Machine Learning, Pandas
- Description:
  First we marked our list of banned drugs and non-banned drugs sequences with 1 and -1. Here '1' means that the target protein is targeted by the drug; and '-1' indicating that the target protein is not targeted by the drug.
  Then we found out certain features like
    - Amino Acid Composition
    - Molecular weight
    - Helix formation
    - etc.

  from various libraries mentioned in tools. These features were found on running against target protein sequences of the protein targeted by that specific drug. After that, we divided our dataset for 70 percent training and 30 percent testing. We trained on the basis of various models like:
    - Random Forest Classifier(accuracy:0.9473449352967426)
    - Logistic Regression(accuracy:0.9494476062939404)
    - DecisionTreeClassifier(accuracy:0.9163039839303649)

So we found out Logistic Regression model to be best with accuracy 94.9 percent.
- Details: input file, Python script or Jupyter Python Notebook, Readme

b) Molecular pathway analysis:

1. Finding pathways followed by banned drugs("Banned Pathways")

- Tools used: SMPDB metabolites database, Python
- Description: For this analysis, the molecular pathways database was used and the drugbank id attribute of this database was matched with the banned drug's drugbank I found in step(1.a). Python script, Readme, Output

2. Finding proteins targeted by "Banned Pathways"

- Tools used: smpdb protein linked database, python
- Description: The smpdb pathway IDs found in the step(3.b.1) were used to get the target proteins from the smpdb protein linked database. Python script, Readme, Output

3. Finding pathways information of tanimoto similar drugs:

- Tools used: Python, CSV
- Description: For this analysis first the drugbank ids of the common names of the tanimoto similar drugs were extracted from the drugbank vocabulary and then these drugbank ids were used to get the pathway information of the drugs. Python script, Readme, Output

3. Analysing food matches

a. Finding Associated food items of a banned chemical(DietRx)

- Tools: DietRx website
- Description: For each banned food compound, we extracted all associated food items i.e those foods which consist of these compounds.These associated foods are considered to be banned because of the presence of banned chemical compounds. Output Folder containing the list of associated food substances for each compound.

b. Finding Associated food items of a banned compound(FooDB)

- Tools: FoodDB website

- Description: For each banned compound it's FoodDB id was used to access the information from the online tool and the associated food items were extracted from the online tool. These food items have a certain amount of this banned compound present in them. Python script, Readme, Output

c. Finding Associated genes of a compound

- Tools: DietRx
- Description: For each banned food compound, we extracted all associated genes present in the DietRx database. It will help us analyse which genes are most affected by banned compounds. Output Folder containing the list of associated genes for each food compound.

d. Molecular Pathway Analysis(DietRx)

- Tools: Drugbank, SMPDB_metabolise, Python, CSV
- Description: DrugBank IDs for the banned chemicals found from the DietRx database were extracted. The DrugBank IDs were then used to find pathway details with the SMPDB IDs using the SMPDB_metabolise database.
  drugid.py, Readme, Output
  dietrx_pathways.py, Readme, Output

e. Finding Target Proteins(DietRx)

- Tools: SMPDB_proteins database, Python CSV
- Description: Target Proteins were found for banned compounds by using their SMPDB IDs from the SMPDB_proteins database.
  dietrx_targetproteins.py, Readme, Output

f. Finding Target Proteins(FooDB)

- Tools:FooDB
- Description:Target Proteins were found for banned compounds by analyzing their SMPDB ids which were further found out by linking various csv files in jupyter using python 3 and pandas.It can be found in this csv file(OutputCsv).Script used to make: .pyfile or Jupyter notebook. TARGET_PROTEIN_NAME.

g. Molecular Pathway Analysis(FooDB)

- Tools:FooDB
- Description: Pathways were found for banned compounds by analyzing their smpDb ids which were further found out by linking various csv files in jupyter using python 3 and pandas. It can be found in this csv file

OutputCsv .Script used to make :.pyfile or Jupyter notebook. Readme. It is in column P_NAME.

4. **Alternate Drugs for the banned drugs**
    a) **Finding alternate drugs on the basis of Tanimoto similarity:**
        - Tools used: Python, chromedriver.exe, SuperDrug2
        - Description:
            - For this analysis, we used the SuperDrug2 tool to get Tanimoto similar drugs for our banned drugs. The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto Similarity Index.

            - The Drugbank ids of banned drugs were given as an input and the output was the Tanimoto similar drug with its common name and its percentage of similarity. Python script, Readme, Output

    b) **Finding alternative drugs on the basis of similar pathways**
        - Tools: smpdb metabolites, python
        - Description: In this analysis, we extracted all the non-banned drugs corresponding to those pathways which are linked to banned drugs. Therefore now we basically have an alternative of the banned drug on the basis of a similar pathway. Python script, Readme, Output

5. **Alternate Compounds present in food substances**
    a) **Finding alternative compounds on the basis of DietRx similarity:**
        - Tools: DietRx
        - Description: Extracted all similar food components present in DietRx database. Output folder(of DietRX)
    b) **Finding alternative compounds on the basis of FooDB similarity:**
        - Tools:FoodDB
        - Description OutputCsv Alter_name columns shows the alternative And whether the alternative compound is valid(i.e. Not banned) can be found in VALID COLUMN.
    c) **Finding alternative compounds on the basis of ZINC similarity:**
        - Tools: DietRx
        - Description: ZINC similarity is a type of Tanimoto Similarity with a threshold of 40% similarity. Output folder

6. **Analyzing Health Risks and Diseases associated**
    a) **Finding associated diseases with banned food substances**
    - Tools: DietRx, FooDB
    - Description:
        - Extracting all associated diseases of banned food items using the DietRX database. Output folder
        - Extracting all health risks of banned food items using the FooDB database. It can be found in this csv file(OutputCsv).Script used to make : .pyfile or Jupyter notebook.Readme:FooDB(Readme). It is in the columns function,effects.

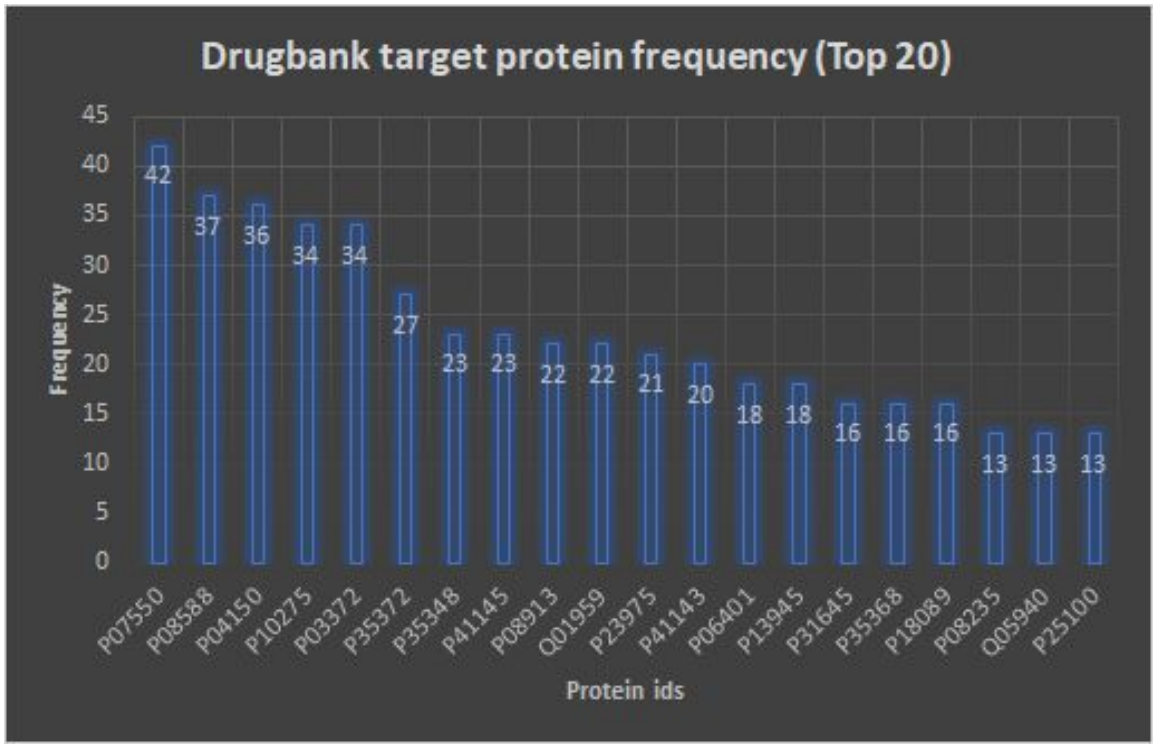
7. **Analysis of recorded Drug Doping cases in various sports**
    Resources used: Drug Study
    a) **Finding the most common drugs used by a sportsperson in a given sport.**
        - Description:
            - Using records of previously recorded doping cases, to analyse the drug most commonly used as a dopant in a sport.
            - Link to Bar graph for the same.

    b) **Finding the total number of cases recorded for a particular sport**
        - Description:
            - Using the recorded data, to analyse the sports and competitions most prone to drug abuse.
            - Link to Pie Chart for the same.

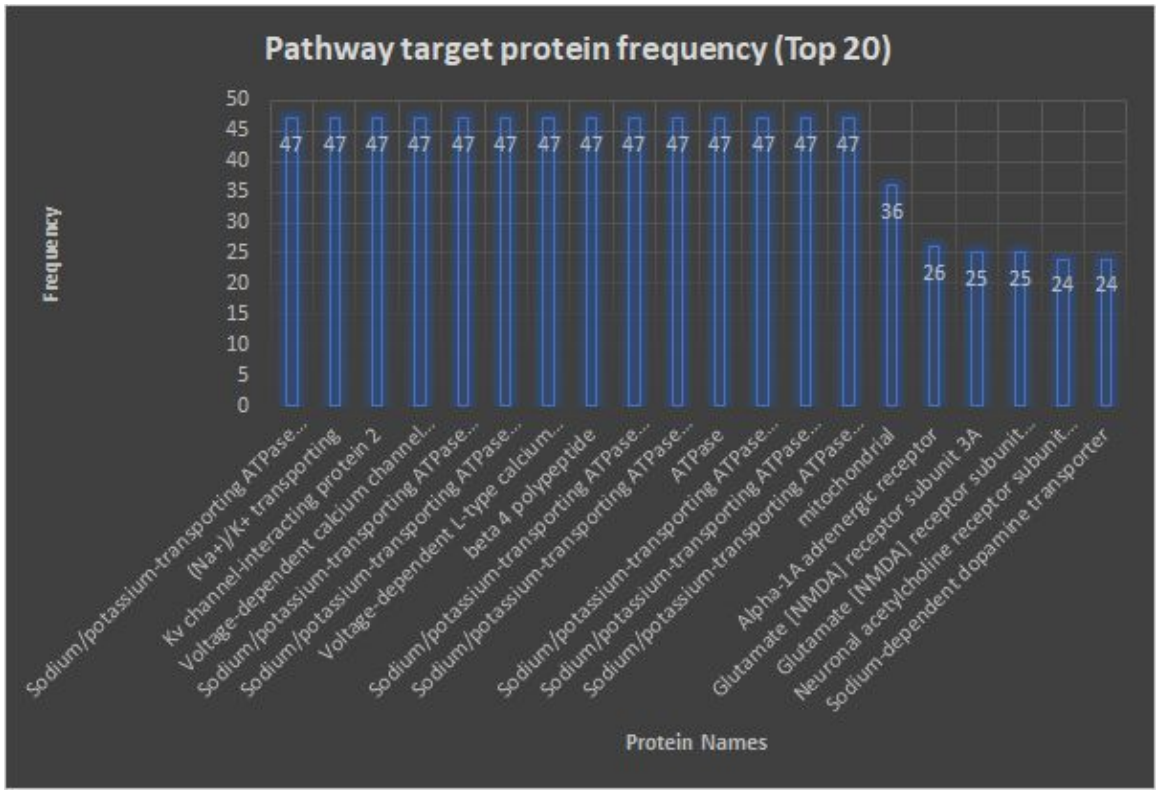
    c) **Finding the total number of cases recorded from a country**
        - Description:
            - Analyzing the countries on the basis of number of drug-doping cases, irrespective of the sport.
            - Link to Pie Chart for the same.

RESULTS

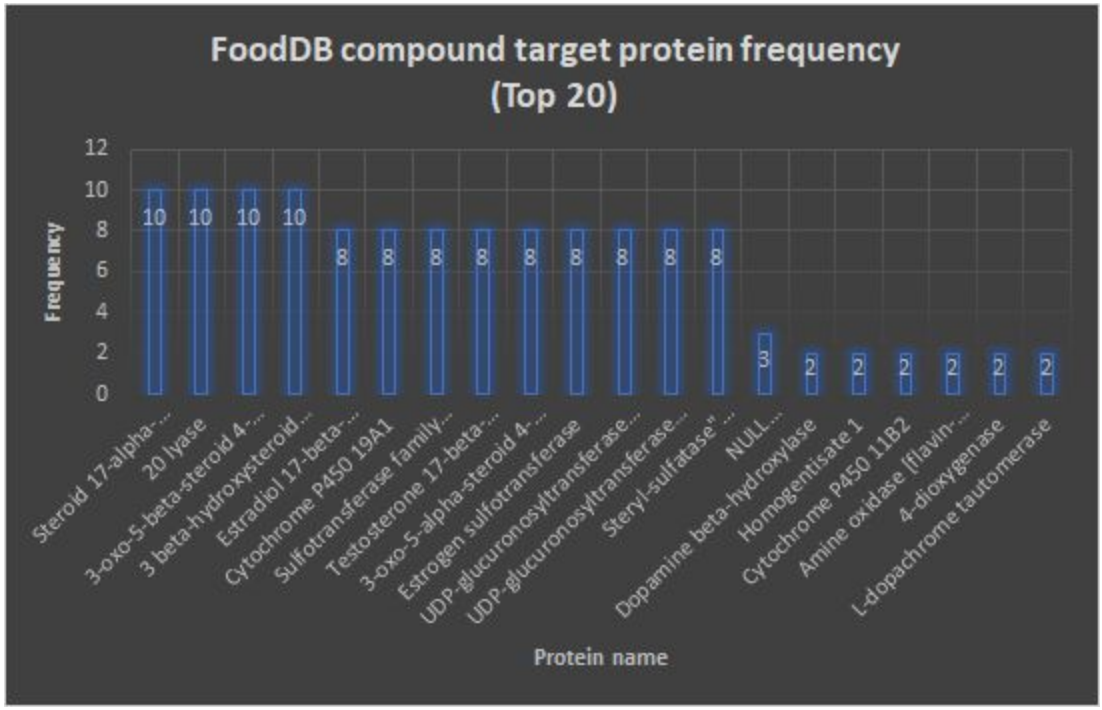
1. We have found alternative drugs on the basis of Tanimoto similarity and similar pathways.
2. Found molecular pathway information of banned/Tanimoto similar drugs.
3. Frequency plot of Target Proteins of drugs banned by WADA. The highest target protein frequency is of P07550 - Beta-2 adrenergic receptor. The beta-2-adrenergic receptor binds epinephrine with an approximately 30-fold greater affinity than it does norepinephrine. So it basically helps to boost the athletes for a short amount of time.



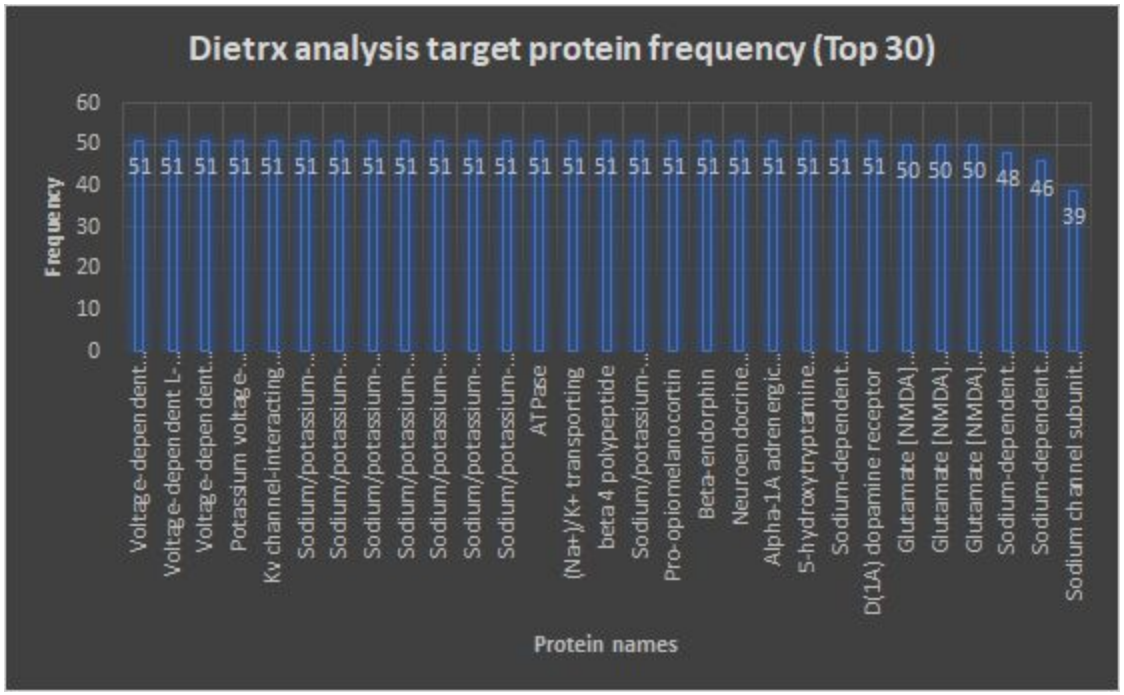
4. Most of the proteins like (Na+)/(K+) transporting ATPase, etc which are affected by doping are directly related to ATP generating mechanism which boosts the doped athletes for a short amount of time.



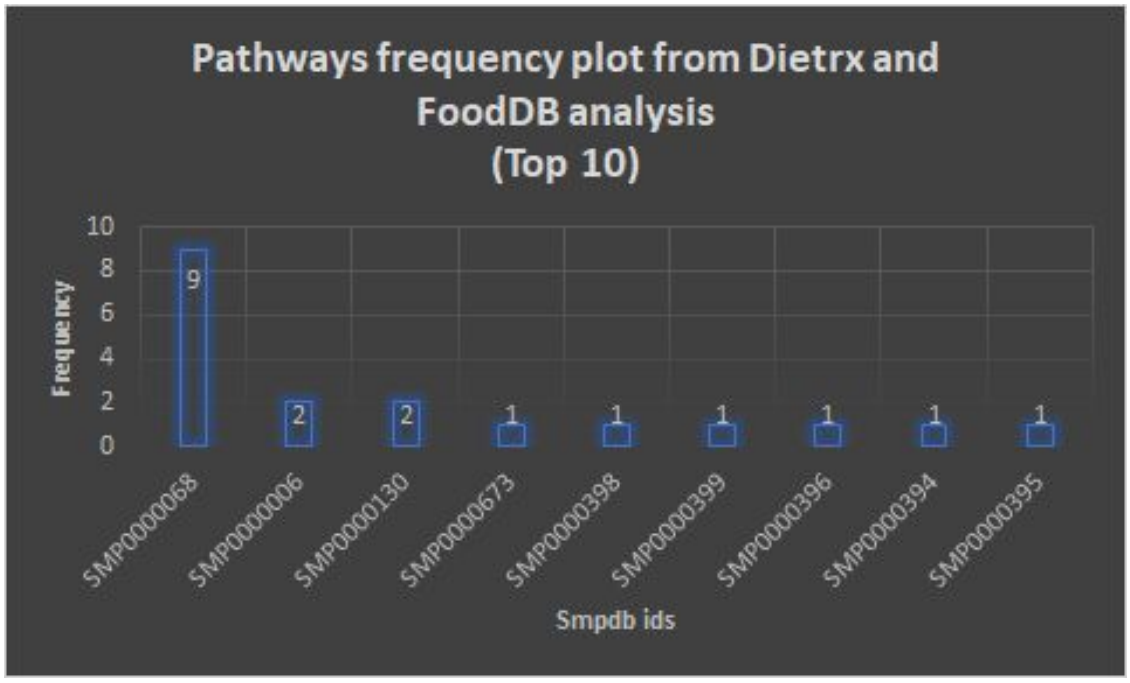
5. Frequency plot of most commonly targeted proteins by the banned compounds found in food items from the FoodDB



6. Frequency plot of most commonly targeted proteins by the banned chemicals found in food items from the DietRx database.



7. This is the frequency plot of the most used pathways by the banned compounds found in the food items extracted from both Dietrx and FooDB databases.



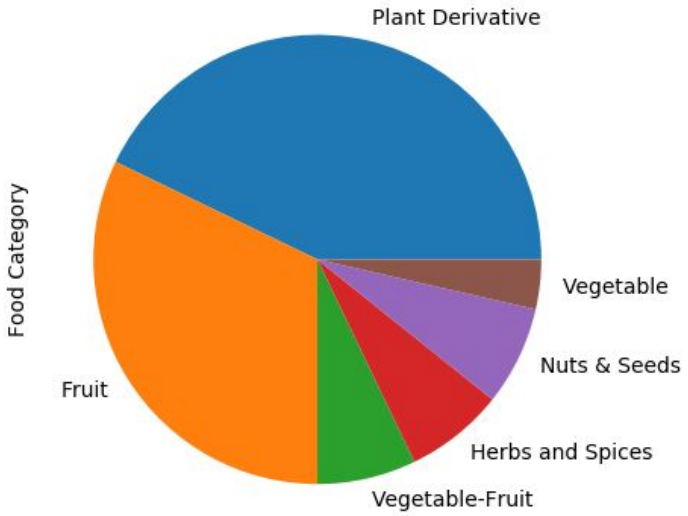
8. Predicting proteins whether they are targeted by a banned drug or not with 94% accuracy.

9. Providing alternate drugs to banned drugs on the basis of Tanimoto similarity.

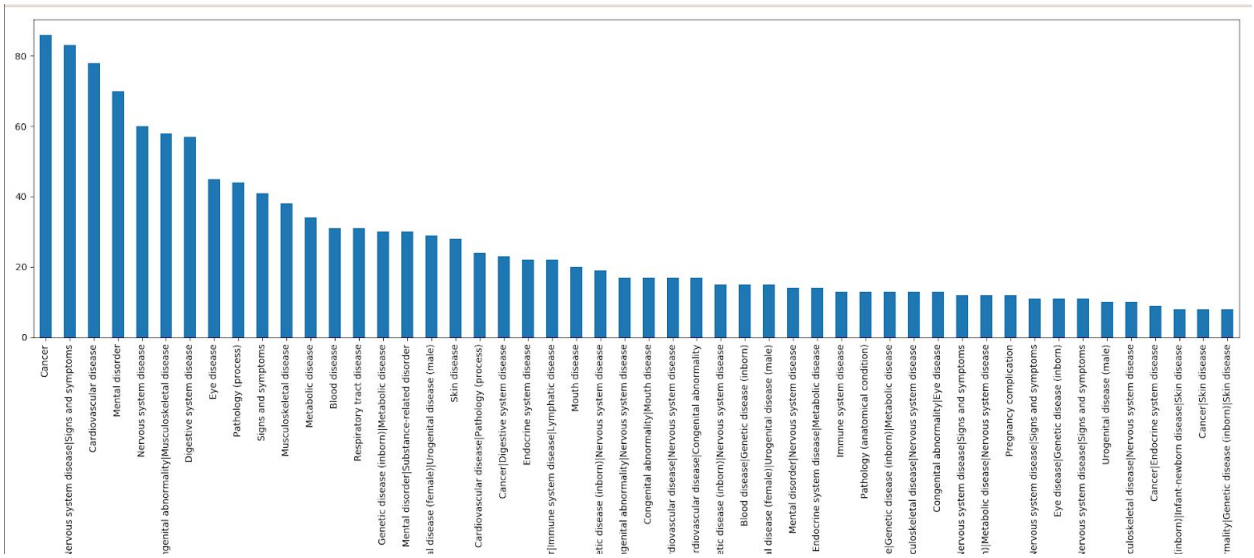
10. Grouping banned drugs on the basis of pathways followed.

11. Sensitivity was given more priority than specificity initially in finding the banned food chemicals and compounds. This is because the food database was not fulfilling our exact search terms to find the banned food chemicals/compounds. There was particularly not a single attribute of the database through which we could compare our banned drug/compund's common name or IUPAC name.

12. Proportion of food categories being affected by banned drugs. We can conclude that a high proportion of Plant derivatives contain substances which are considered banned by WADA.

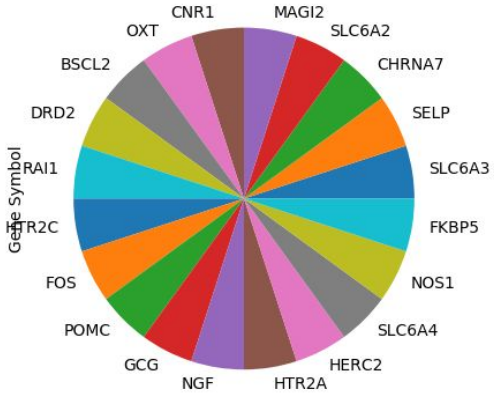


13. Diseases associated with the banned substances. Highest affected diseases as per the results were cancer and nervous system related diseases.



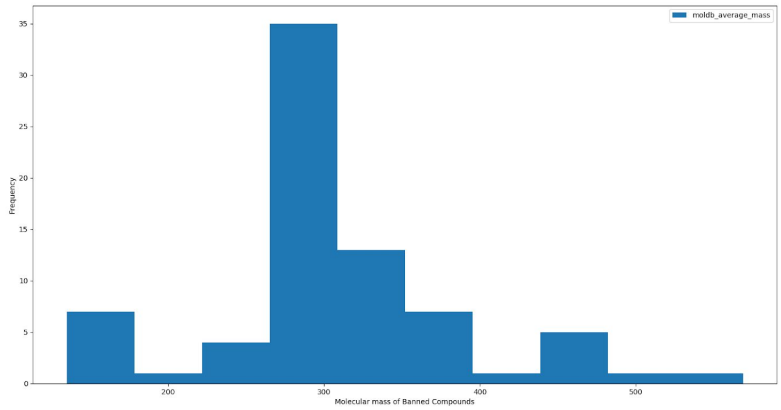
(View full image here.)

14. List of top 20 genes affected by banned food components.

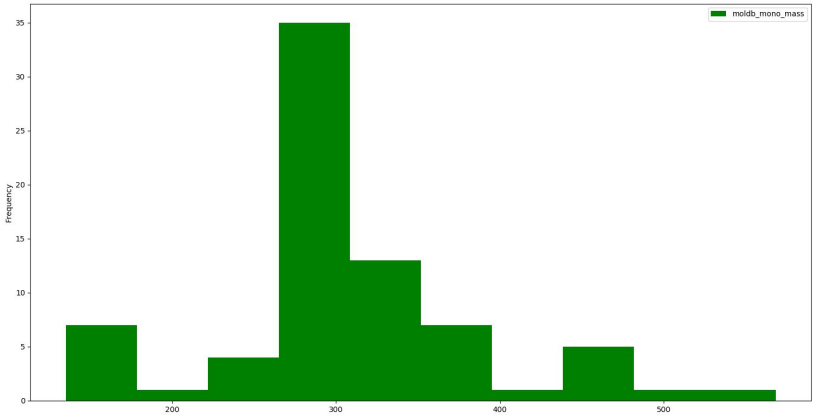


15. We have completed the analysis for drug-doping cases previously recorded, in all sports by different sportspersons.

16. All the below results were found out by analyzing different chemical properties of all the banned compounds in FooDB database.CSV FILE:chemicalproperties SCRIPT :.pyfile or .ipynbfile README:chemreadme.

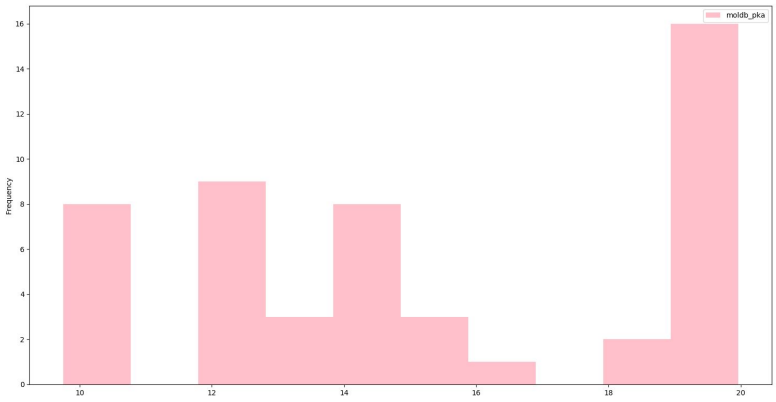


a. Frequency of Average Molecular mass of Banned compounds.



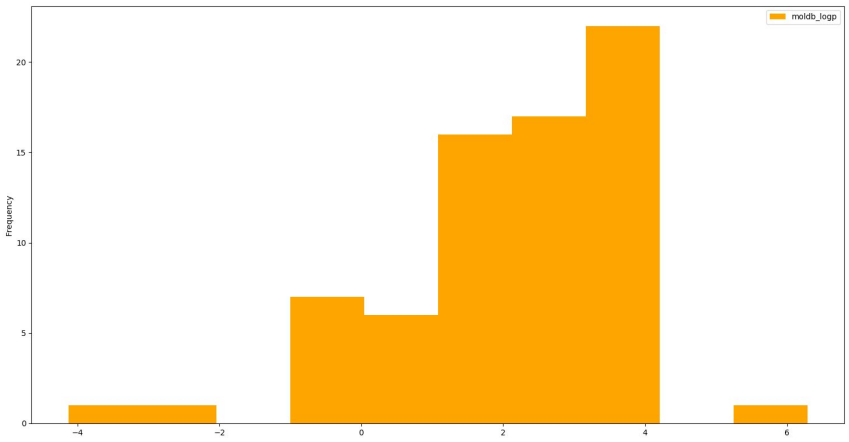
b. Frequency of mono mass of Banned compounds.

Mono mass-Monoisotopic mass (Mmi) is one of several types of molecular masses used in mass spectrometry.



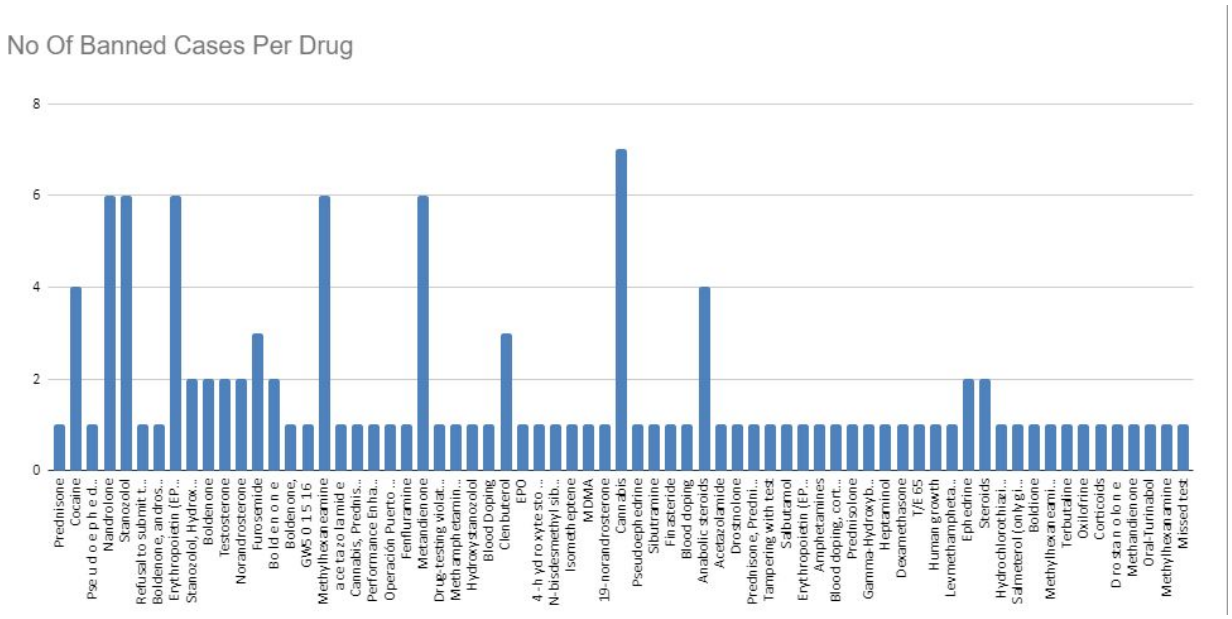
c. Frequency of log pka value of Banned compounds.

Pka-The **pKa value** is one method used to indicate the strength of an acid. **pKa** is the negative **log** of the acid dissociation constant or Ka **value**.

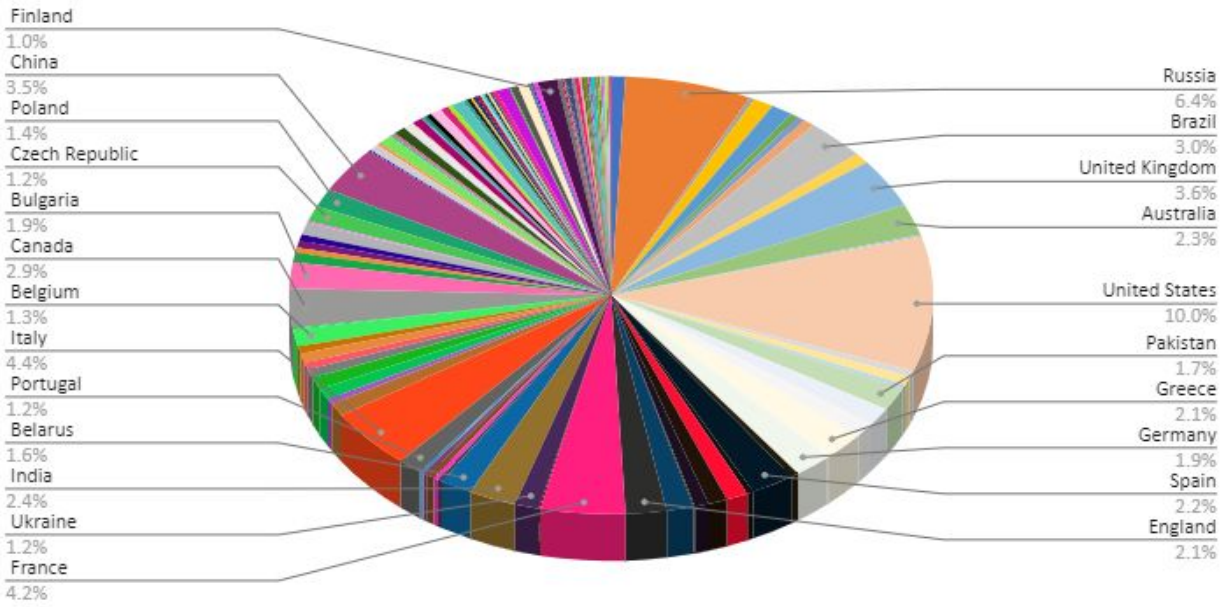


d. <u>Frequency of logp values of molecular weight of Banned compounds.</u>
logP-A negative **value** for **logP** means the compound has a higher affinity for the aqueous phase (it is more hydrophilic); when **logP** = 0 the compound is equally partitioned between the lipid and aqueous phases; a positive **value** for **logP** denotes a higher concentration in the lipid phase (i.e., the compound is more lipophilic).

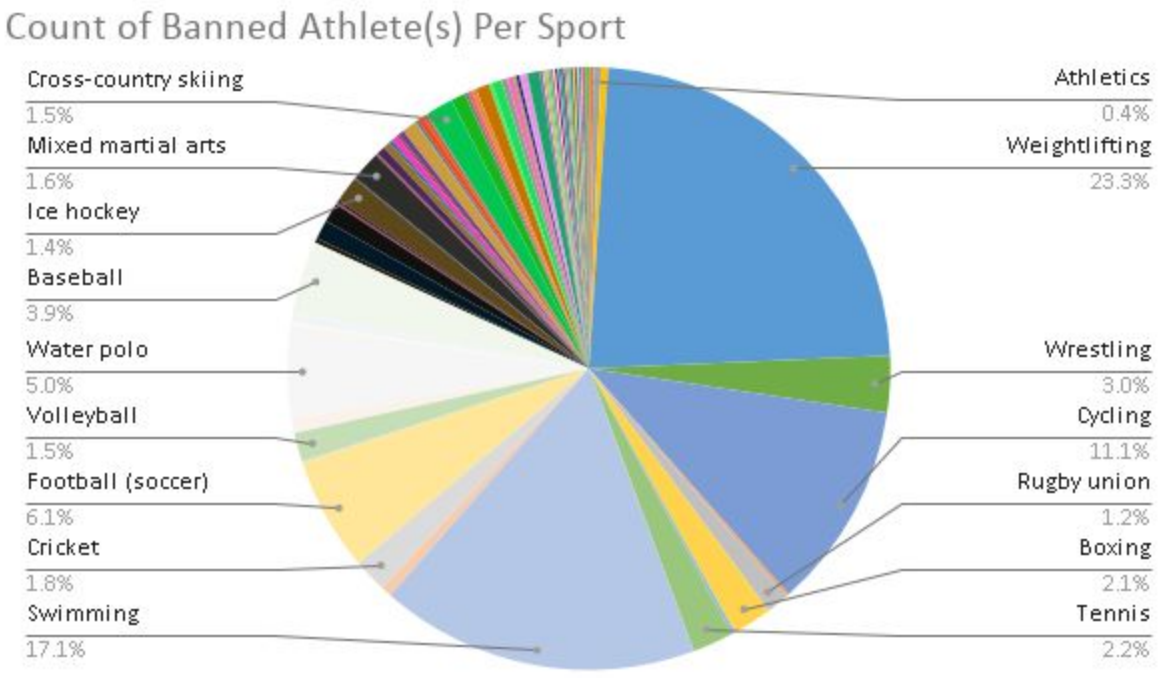
MOST COMMONLY USED DRUGS AS DOPANT:



ALL COUNTRIES WITH THE RESPECTIVE NUMBER OF DRUG ABUSE CASES RECORDED IN DIFFERENT SPORTS:



SPORTS MOST PRONE TO DRUG ABUSE BY THE PARTICIPANTS/ SPORTSPERSONS:



LIMITATIONS

- QUANTITY or CONCENTRATION of banned drug/food compounds is not considered while doing analysis. e.g some compounds are banned only when concentration of substance X is found greater than a specified limit.

- We haven't given special consideration to those compounds which are banned only in a specific sport. Sport specific Analysis is not done.

- As WADA doesn't distinguish between geographical locations, we also have not segregated our data with respect to different geographical locations.

- WADA has a separate list of those substances which are banned only during competition. This special segregation is not done in the study. For better analysis and greater data, We have considered all banned drugs under one category.

- Masking Agents like Diuretics are not considered in the study.

- Since our study revolves around consumption of banned substances, different ways of doping (eg Blood doping by plasma therapy) does not come under the scope of this study.

- We had a setback when the components present in food databases and those banned by WADA were not compatible because of different formats (IUPAC and common names). As a result, the data extracted for food substances was slightly compromised.

- We couldn't provide Alternate food items, rather, we provided alternate compounds which can be changed to convert the food.

- We could not find any database which could relate Health limitations or risks caused by doping through consumption of drugs.

- We couldn’t perform statistical modelling to give target locations of a food. We have provided the target location of a specific drug as the earlier was beyond our scope of study.

- FooDB database does not contain complete information which can relate to SMPDB ids, which could have been helpful for analysing pathways. We tried linking banned compounds to foods but there were no matches as information for food was limited and not present for all the compounds. Similar case was in pathways smpdb ids were not available for all banned compounds.

- The DietRx database did not have many matches for banned chemicals in food items. Moreover, due to less efficient relations between databases such as DrugBank and SMPDB, pathways and target proteins were not found for all the DrugBank IDs.

CONCLUSIONS

At the end of this detailed study analysis, we are able to confidently state the following conclusions:

1) We can predict whether a protein is targeted by a banned drug or not by a high accuracy of 94.9%.
2) We can find alternate (non banned) compounds which can be used as replacements to the banned substances. This will significantly help in prevention of accidental doping.
3) Cancer and Nervous system diseases are the most commonly associated diseases caused by doping.
4) Most of the common dopants banned by WADA affect the ATP generating mechanisms useful for providing short bursts of energy during the competition.
5) Most of the drugs banned by WADA had their target pathway as Androgen and Estrogen Metabolism according to our results which is related to muscle growth and aerobic capacity in males and females.
6) The USA can be called the “Doping Hub” of the world, accounting for almost 10% of total cases in the world. Doping is most common in weightlifting, accounting for nearly 23.3% of the total number of doping cases in all sports.
7) Some of the most affected proteins by banned drugs are Beta-2 adrenergic receptors, Steroid 17-alpha-hydroxylase/17, 3-oxo-5-beta-steroid 4-dehydrogenase etc.

REFERENCES

- Sterling and Irwin, *J. Chem. Inf. Model, 2015* *http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559*
- Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *J Cheminform* **7,** 20 (2015). https://doi.org/10.1186/s13321-015-0069-3
