International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 795-802www.ijaem.net ISSN: 2395-5252

Computational Models for Inflammatory Arthritis Protein Activity Based on Primary Screening Datasets

Divyansh Babel[1], Shreyans Babel[2]
[1]*Disha Delphi Public School, Kota, India*
[2]*Department of Computer Science at University of Massachusetts Amherst*

Date of Submission: 21-06-2020 Date of Acceptance: 07-07-2020

ABSTRACT

In this article, we will analyze the agitating protein bioassay causing rheumatoid arthritis. Computational systems biology approach will be applied to the bioassay via primary screening. Proteins play an indispensible role as they are crucial bio molecules. The function of proteins is accomplished by the structure and the regulatory role they play. The experimental investigation of datasets will be conducted in order to build a dynamic model of protein. The knowledge of dynamic behaviour and its skeleton is understood by structure modelling. Endogenous proteins which perform a particular role in sensitivity and infection can protect the immune system from falsely assaulting the body and directing to joint inflammation. This finding will help in detecting essential medication against rheumatoid arthritis.

Protein bioassay will aid in finding an autoimmunity target, thus proposing that an animal model formation can be done using a molecule activator. Though productive-throughput screening of the peptide activity is a time-consuming task i.e. meant to be represented on wide scale. The relevance of this ailment throughout the globe and possible novel medications have revolutionised the requirement to advance the exploration of distinct molecules with arthritis activity. Thus, the estimating technique involving Machine Learning has been globally used to construct classifiers for high-throughput virtual screen to rank molecule for further investigation. The accessibility of datasets based on high productivity yielding bioassay publically forms a base for computing techniques to build predictive models. Moreover, this perspective would decrease the cost, labour as well as time required to run high-throughput screens.

An attempt to use LYP activators, we are able to create foretelling models based on large primary screening datasets of proteins with logical worth. Currently, we have employed the use of a specific protein tyrosine phosphatase activity from original screening bioassay available in public accessibility to raise accurately arguing models, compelling molecules from bigger molecular collection.

LYP (lymphoid tyrosine phospatase) encrypted by PTPN22 gene, takes part in T-cell receptor signalling along with its regulation. LYP portraying single-nucleotide polymorphism (SNP) is related with a variety of disorders. An approximate idea of an animal model design is suggested by the gain of function mutatedallele.The working of the enzymes will be analysed as per the data outlined. Accessing the repercussions of certain inflammatory proteins via high-throughput data is hard enough to conclude as the focus is on system level understanding rather than individual components.Thisresearch work elaborates a mathematical overview using computational methods and primary screening type bioassay to generate a well-established protein model. Here, we have used the machine learning algorithm named as Naïve Bayes which has given an accuracy of around 76% and accurately predicted more than 450 drugs, out of which we have presented 5 drugs which can be potentially accepted for the treatment of Rheumatoid Arthritis.

I. INTRODUCTION

Rheumatoid arthritis is a general inflammatory arthritis which is accountable for disability in an individual. It falls in the category of autoimmune disease featuring inflammatory arthritis and supplementary spinous engagement. As per the study made in 2003, itsubsisted in Native American population decades ago, later started existing in Europe in the 17th century[1], [2]. Previous theories on its pathogenesis are concerned with the autoantibodies as well as resistant compounds. T-cell mediated antigen peculiar feedback; T-cell devoid of cytokine web and antagonistic tumour like reaction of rheumatoid synovial membrane have also been expressed. Recently, the autoantibodies have hinted their comeback by supplying to it. Specific therapeutic interventions can be fabricated to decrease synovial agitation and joint damage in rheumatoid arthritis[3]–[5]. Complicated interactivity between hematopoietic and stromal cells ends in synovial inflammation. Modern researches show that RA Synovial fibroblast plays a destructive part in the disease.

This ailment affects about 1% of the juvenile inhabitants in developed nations. It destroys the peripheral joints, ligaments and tendons which heldin charge of joint deformity as well as physical infirmity[6], [7]. The pathological reasons are deteriorated with rise in the sedentary living style with decreased physical performance. It causes joints swelling, cartilage wrecking and impairment of ossein. Imperfect meals, changing living style, toxic environment can be explained as supplementary grounds for RA. This crippling syndrome involves diverse factors that perform a key role in the pathophysiology in its development[8].

Protein-protein interactivity provides basic rationale behind major biological processes. They play a significant character in signalling, trafficking as well as intra and extra cellular act. Cell communication is a fundamental procedure that happens when a signal is transmitted, employing different receptors, enzymes, catalysts, secondary, messengers, TF[6], [9], [10]. Any obstruction in the signalling may lead to a harmful condition whereas accurate flow of proteins including signal transduction defines an underlying mechanism responsible for causing disease. Anyhow, as all the biological systems are dynamic in nature, constant depiction of molecular network can give important but relatively less understanding. A large-scale study can portray information about the system's etiquettes under diverse situations by in silico stimulation, hypothesis testing and argumentation[11], [12]. Biological modelling is directed via differential equations that give a detailed investigation of a network study. Although improper data can head to alternate models' formation which can be employed as another qualitative system experiment.

Studies conducted in numerous animal models of autoimmune illnesses have exhibited that a slight imbalance in the synchronisation of the excitation or performance of immune system cells can be a huge issue in regards to the health. Function of immune system cells can obstruct the modulation of autoimmunity that consequences in a wide variety of human ailments such as type 1 diabetes (T1 D), rheumatoid arthritis (RA), inflammatory bowel disease (IBD) and multiple sclerosis (MS).A research conducted in 2017 indicates that prior agitating cytokines such as IL-6, IL-1β and TNF-α perform a remarkable part in pathophysiology of RA, these pro-inflammatory peptides itself has been earmarked for the designing of new RA drugs. These cytokines induce agitation through different mechanisms that further leads to origination of angiogenesis, leucocyte linkage and tissue depravity in RA. Non-steroidal anti-inflammatory drugs (NSAIDs) are preferred at first place;alternatively, they have grave injurious effects that further raise the risk of haemorrhage[13], [14]. Diverse record states that there are still no effective anti-arthritis drugswith less negative effect.Other small molecules have also been discovered for cure of RA.

The inefficiency of extensive system magnitude understanding of disease's intricate biological pathway and its control mechanisms, it has been suggested that whole cell phenotypic screen offer a better chance when related to single gene based biological screen.Computational method analysis like Machine Learning techniques are successfully employed as a supportive method for primary screening. The availability of building blocks bioassay datasets in an open domain that can be publically accessed allows us to develop predictive computational models that can be efficiently employed to concentrate on molecules biological essay.

1.1 Linking PTPN22 with autoimmunity

A single nucleotide polymorphism was found in the lymphoid tyrosine phosphatase (LYP), on chromosome lp13 encrypted by PTPN22 gene in 2004. A new report explains a powerful interconnection of the disease linking PTPN22 allele with the existence of anti-deiminated protein antibodies in RA. Thus, it increases the risk of RA manifolds. The PTPN22 SNP may also supply as an anticipating factor. This bond between disease associating allele and the duration of disease have been viewed in RA. More research is required to find if the SNP relates with the clinical variability, course and extremity of disease. As well as detection of PTPN22 genotype will be useful for evaluating the treatment[7], [15].

The protein tyrosine phosphatase non-receptor type22 (VTPN22), also called lymphoid tyrosine phosphatase (LYP), is conveyed particularly in blood cells and is a crucialguard keeper of T and B cell receptor communication. LYP behaves as downstream receptor linking to hindering the performance of pivotal signalling effectors in TCR signalling. Particularly LYP has been proposed in dephosphorylation of +ve modulating tyrosine residue in aimedSrc-family protein tyrosine kinases, involving FynT as well as Lck[16], [17].

1.2 Layout and task performed by PTPN22 encryptedphosphatase

Human PTPN22 is detected on chromosome lp13.3-13.1 and expresses an 807-AA residue protein named as LYP. An experiment revealed that the human and mice genes are the orthologs of same PTPN22 genes. The enzymatic domains are highly identical (~70%), the respective protein differs in their c-termini and all antibodies apprehend either human LYP or mouse PEP, but not both.

The major splice design of man LYP is a 105 kDa protein with an N-terminal catalytic tyrosine phosphatase with a high similarity to other specifically tyrosine classical non-receptor tyrosine phosphatases. The C-closing 2/3rd of the protein is speculated to be highly disordered. It has been viewed that LYP can occur as a substituting altered form with a relative shorter C-end. LYP is exhibited to connect with an adapter protein Grb2 and E3 ligase c-Cbl, but the constraint role of these associations and their relevance are still unknown. To highlight that the C-end 24 amino acids are well conserved among all the members of PEST protein tyrosine phosphatises which is commonly known as C-terminus homology region (CTH domain)[16], [17]. Additionally,the latest screening using complete length LYP has explored the presence existence of SNP in human PTPN22 which causes severe life-long autoimmune disease.

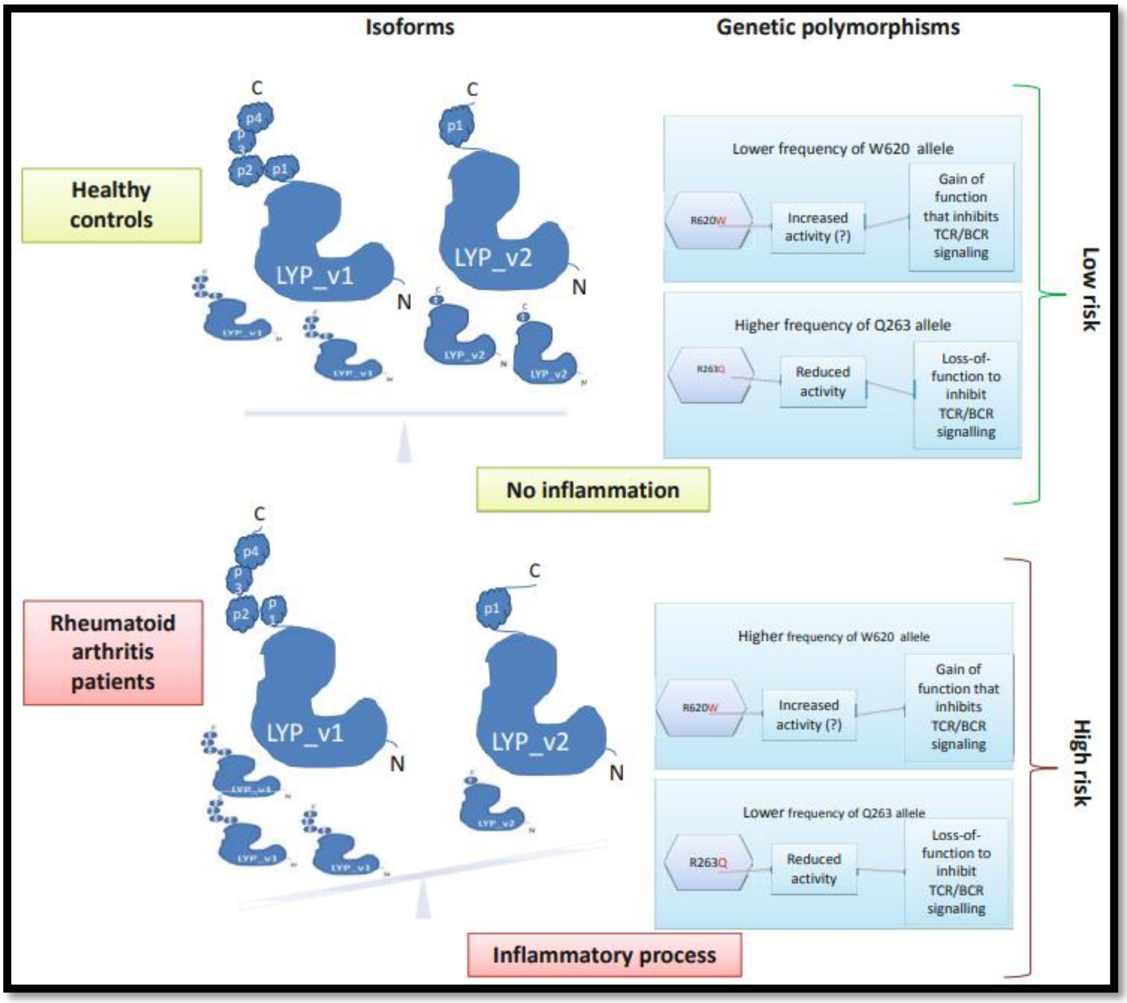
Isoforms
Genetic polymorphisms
Healthy controls
C
p4
p3
p2
p1
LYP_v1
LYP_v2
N
Lower frequency of W620 allele
R620W
Increased activity (?)
Gain of function that inhibits TCR/BCR signaling
Higher frequency of Q263 allele
R263Q
Reduced activity
Loss-of-function to inhibit TCR/BCR signalling
Low risk
No inflammation
Rheumatoid arthritis patients
Higher frequency of W620 allele
R620W
Increased activity (?)
Gain of function that inhibits TCR/BCR signaling
Lower frequency of Q263 allele
R263Q
Reduced activity
Loss-of-function to inhibit TCR/BCR signalling
High risk
Inflammatory process

Figure 1: Effects of isoforms of LYP on RA00000

1.3 Biochemistry of catalysts enciphered by disease-predisposing allele

It is expected that the threshold for TCR signalling would be mutated by SNP as LYP and Csknegatively regulates the TCR signalling pathway. The PTPN22 T1858SNP fascinated the researchers as R620 is a significant residue in the Pro-rich motif in LYP that links to the SH3 domain of Csk, displaying that LYP*W620 cannot associate with Csk. The consequence of the illness linked allele hold various implications for the molecular methods through which LYP proceeds in TCR signalling the -ve modulation T-cell receptor signalling may contrast the restricting ability of LYP, which exhibit that the linking of LYP to Csk is not more considerable for its working in TCR signalling[18]. Studies indicate that usually only a part of LYP has an ability to obstruct TCR signalling. When studied through immunofluorescence staining and confocal microscopy, the inner-cellular position of LYP*W620 is probable to be identical to thatLYP*R620 containing very less plasma membrane in resting T-cells.

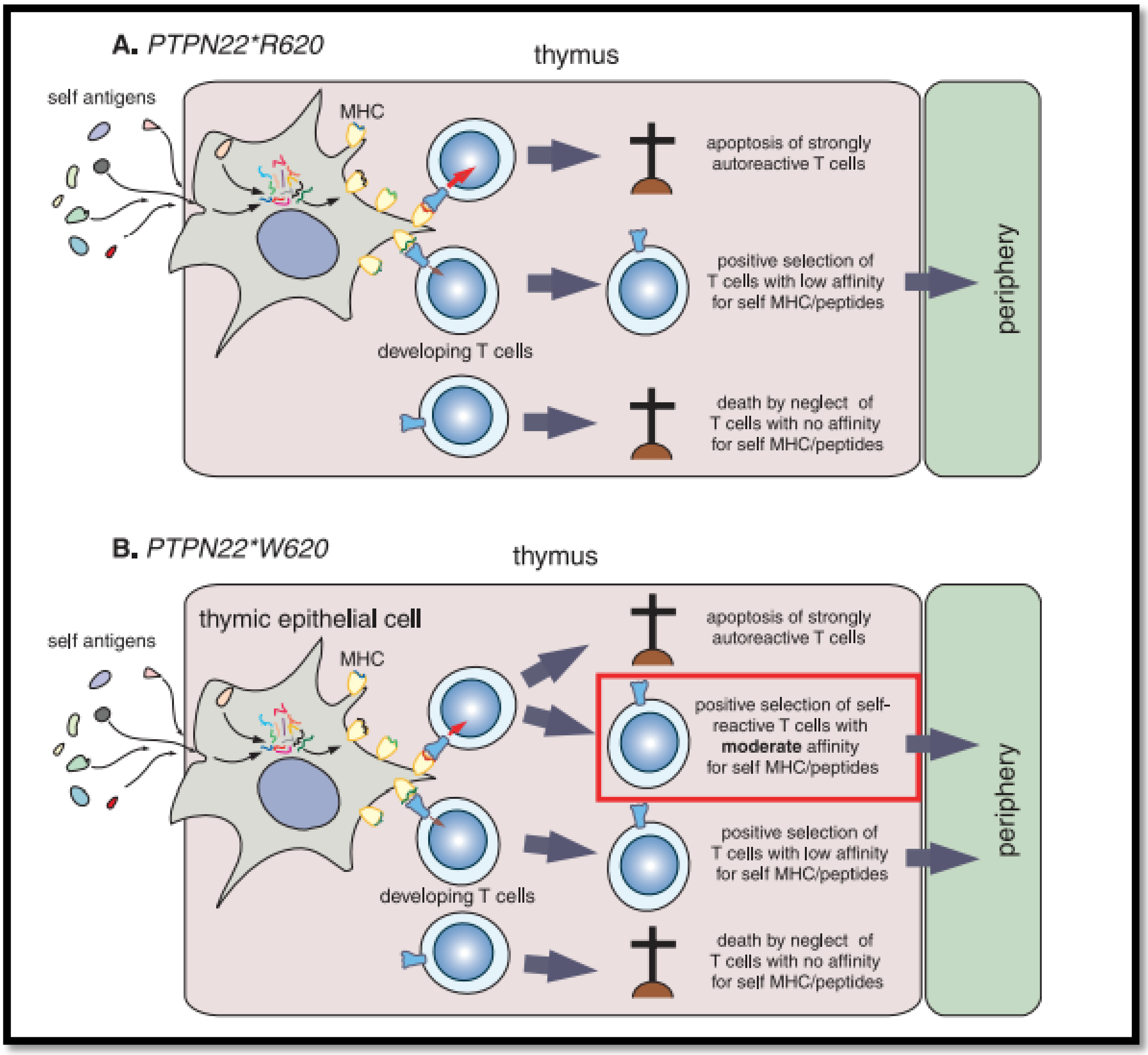
A. PTPN22*R620
thymus
self antigens
MHC
apoptosis of strongly autoreactive T cells
positive selection of T cells with low affinity for self MHC/peptides
developing T cells
death by neglect of T cells with no affinity for self MHC/peptides
periphery
B. PTPN22*W620
thymus
thymic epithelial cell
self antigens
MHC
apoptosis of strongly autoreactive T cells
positive selection of self-reactive T cells with moderate affinity for self MHC/peptides
positive selection of T cells with low affinity for self MHC/peptides
developing T cells
death by neglect of T cells with no affinity for self MHC/peptides
periphery

Figure 2: Thymic selection by LYP variants.

1.4 Molecular mechanisms of PTPN22 spliced form and rheumatoid arthritis directs autoimmunity

Modified Splicing is a method in which distinguished mRNA transcripts are produced from a unitgene so that final splice alteration can be converted into dissimilar protein isoforms. Thus, the substitute isoforms split some of the arrangementbut not whole. Different spliced forms, encrypting LYP_v1 and LYP_v2 have been effectively defined. The former isoform ciphers an amino-terminal PTP area, a middle area of concealed function plus a carboxy-terminus region consisting fourproline-enhanced motifs entitled P1 to P4. The later isoform displays a dwarf carboxyend consequentially in the unavailability of P2, P3 and P4 motifs. Experiment reveals that the presence of PTPN22 sliced types are diverse in RA sufferers. The mRNA of the different splices isoforms was estimated. The T-cell behaviour is administered by the interconnection of

LYP with CD2-joining peptide-1[7], [19]. LYP-v2 is inefficient in consensus sequence that can influence the interactivity with CD2-holding protein-1as in LYP-v1. Next, the phosphorylation of LYP_v1 may hang on the cell regulation process: whose series is recognised by CDK1. The vacation of consensus sequence in LYP_v2 isoform can affect the cell process in the lymphoid cells which are conveying this isoform. Eventually, the 2 isoforms can interconnect with various sets of Src homology 3 domains as isoform contains 4 possible SH3 domain holding locations but isoform2 contains only 1.

Presenting the idea that LYP*W620 is a gain-of-function modification i.e. statistically more agile, it predisposes to autoimmunity because it represses TCR signalling more influentially in the course thymic growth, directing to the survival of autoreactive T-cells that will generally be removed by the -ve choice of individuals. LYP as well as its working is still under a question hence there is a possibility that it affects the immune system in a complicated way[12], [17]. Additionally, some autoimmune ailments that associates PTPN22 are not specified to be fundamentally T-cell mitigated. Since LYP is seen in all leucocytes it also performs a part in antigen processing and depiction in dendritic cells and NKC.

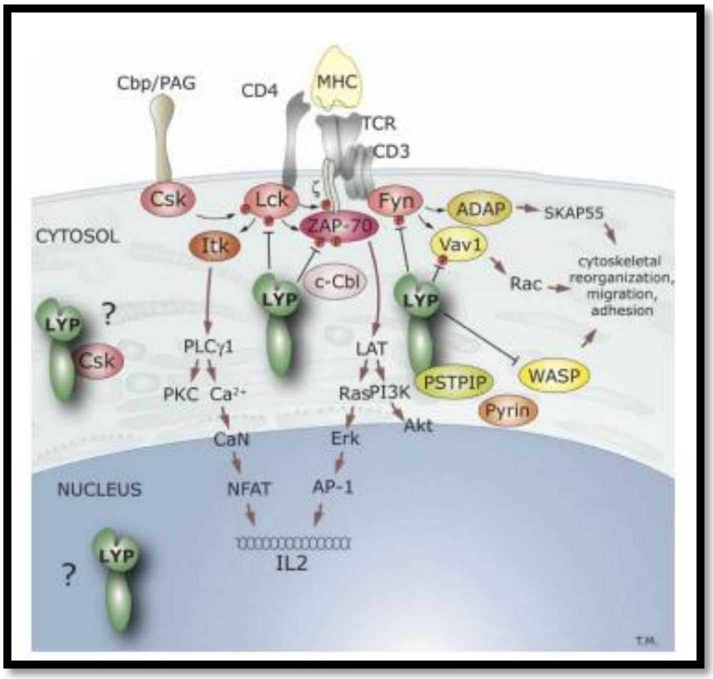

Figure 2: Function of LYP in TCR signalling

II. METHODOLOGY

1) Biological assay data

The primary screening evidence of LYP activators acquired from PubChem data repository maintained by NCBI. The information was established on protein target using two tyrosine phosphatase isoforms. LYP screening was performed at Columbia University as a part of the molecular library screening network. High-throughput sets for LYP inhibitors-- an autoimmunity target[20]. The bioassay was performed using material provided by different scientists, material used are recombinant LYP protein, bis-Tris, NaCl, DTT, PEG, DiFMUP and Na-Ortho vanadate. A range of buffers of batch size of 205 plates, including:

1. Assay bufferwith pH 6, 0.15M bis-Tris

2. Enzyme buffer: an assay buffer with 1% PEG and 5mM DTT which is added immediately before use.

3. LYP working solution composing 7.5nM in enzyme buffer (prepared fresh).

4. DiFMUP working solution, possessing 150uM DiFMUP in assay buffer, 1.5% DMSO (freshly prepared and light sensitive).

5. Ortho vanadate working solution having 8mM Na- Ortho vanadate in assay buffer.

Following the assay principle which says full enzymatic activity is available in a 62 kD N-terminal catalytic domain of the phosphatase which is expressed in and purified from bacteria. The primary screening results obtained from the screening against MLSMR 100K set. The data recorded was normalised on per plate basis. Incubation of the recombinant peptide with the appropriate substrate DiFMUP results in the conversion of substrate to a fluorescent derivative. Determinating position of the equation is evaluatedby calculating fluorescence by 360nm & emission at 460nm. The consequence enzymatic performance under the condition of this assay is set at 100%. The % excitation for every compound was calculated using the following formula:

$$\% \text{ Activation} = 100*[(\text{compound} - \text{MedianBackground})/ (\text{MedianCompound} - \text{MedianBackground})]$$

A threshold hit of 37% activation was reported to recognise potential hits that correspond to the average + 5 standard deviations.

2) Molecular descriptors

Generated by the algo and freely available windows-based descriptor calculation software PowerMV. This software allows an environment for sighting, descriptor production and its ability are only restricted by accessible memory. Due to huge amount of amalgams used in bioassay the dataset files were partitioned to slight SDF portfolio employing a script present onMayaChemTools. Approx 96,409 substances were tested, out of that 103 were active and rest not active. Every descriptor corresponds to a particular molecular feature that was deliberated for every complex in the datasets. The bioactivity estimate was comprehended as the last index tagged as conclusion representing the class trait which helped in model construction.

3) Data pre-processing

The quality of the dataset was enhanced by removing supplementary descriptors, lowered themagnitude of the dataset. The dataset was arranged according to class. At last, a customised script was used to divide the evidence in 80% trainingand legitimate set and 20% test set. The segregation-based modelling was done via training cum validation set. 5-fold CV was employed in all model designing. In each repetition of an n-fold CV, single fold is used for checking and the other (n-1) fold is used for tutoring the classifier[21], [22]. The trial outcomes are gathered and moderate over all folds. This provided a cross validated approximation of the concluding decision figures.

4) Via Machine Learning

ML explores the information utilizing algorithm and customises classifies and its constituents. Set investigation and conceptualisation using a flexible approach using an analysing program. In the following section depiction of four ultra-modern classifiers specifically Naive Bayesthat are instructed to fabricate blueprint. Let's have a quick review about all of them:

4.1 Naive Bayes

The most effective algorithm based on Bayes theorem which shares a common principle, where every set of features is classified independent of oneanother. This classifier is manipulated in sentiment scrutiny, recommendation systems where 'being independent' poses a disadvantage as in almost every example the predictors are dependent and hinders the performance[22], [23]. It computes the conditional probability of every descriptor and then foretells the class with highest possible approximation.

5) Designing classification models

When the datasets are misrepresented with biological assays poses an issue. An imbalanced dataset i.e. altered by the class representation and biased instances. The complication arises due to his imbalance, resulting in highly false negative results. Standard error-based classification methods when implemented on an imbalanced data makes it more cost sensitive and raises the prognostic ability of the classifier. The cost sensitivity can be handled by two ways, first by customising the cost making algorithm and second by designing wrapper class which can interchange the existing algo to levy manageable unit[23].

6) Based on Performance

Several performanceevaluations were utilised to finalise the upshot. True Positive Rate (TPR) is a proportion of predictive true actives to original count of actives (TP/TP+FL), False Positive Rate (FPR) is a proportion of foresee false active to actual count of inactives (FP/FP+TN). The precision applies to the current quantity. It can also be attributed to reactiveness to point +ve outcome and specificity to spot –ve output. A more sensitive and a specific result produce less error. In the screening procedure, the enrichment factor (EF) visualises a prominent factor to be analysed.It uses the enhancement of the hit rate when contrasted to the random selection. BCR i.e. steady segregation proportion describes joint norm for sensitivity and peculiarity producing an accurate result for uneven datasets. ROC i.e. receiver operating characteristic is a statistical chart, PPR v/s FPR for the categorizing system. The ranking of the classifier is based randomly as the selected +ve instance will be graded higher than chosen -ve instance.

III. RESULTS AND DISCUSSIONS

The primary screening of the datasetActivation_Primary_2uM to find the inflammatory proteins causing RA reveals normalised % activation at 2mM activator concentration of the primary acid. The dataset consisted overall 96,409 tested compounds having 103 effectual and 96,306 inert substance. Initial experimentation was performed using quality base programmes only. The model with 20% FP rate which is under the threshold check was obtained, revealing excellent accuracy as the data was highly imbalanced.

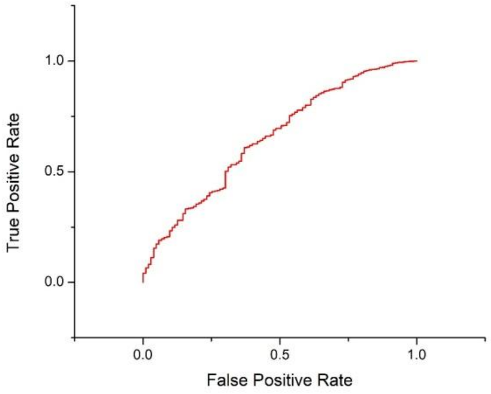

Fig –ROC Curve

A tabular representation of the success predictive model each classifier is depicted in a table.Also, the analytical activity of the best classification model gained from each classifier is represented in another table. Binary classification methods were used to select the best representation of each data. The ROC curve is the most reliable approach for characterising the screening results. All of the above algorithms revealed significant data which can comfortably distinguish the positive and negative tags following the AUC quantities.

Our goal is to reachabsolutesensitivity and specificity. Although, all the four classifiers were highly sensitive with value more than 80% specific. Naïve Bayes seems to be highly sensitive, it emerged to be the finest among all having higher ROC value. Obtaining correct practical solution via primary screening is our goal. In silico techniques used to quantify the enrichment employing EF on different datasets. Model designing using the quality dataset for autoimmunity target protein were successful via different classifiers.

TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area
0.768	0.558	0.803	0.783	0.184	0.653

Table – Representing the Features of the Machine Learning Model

Naïve Bayes which has given an accuracy of around 76% and accurately predicted more than 450 drugs, out of which we have presented 5 drugs which can be potentially accepted for the treatment of Rheumatoid Arthritis.

DrugBank ID	Drug Name	Confidence Level
DB01399	Salsalate	98.8%
DB00861	Diflunisal	97.8%
DB01283	Lumiracoxib	96.7%
DB04552	Niflumic acid	96.5%
DB11323	Glycol salicylate	94.4%

IV. CONCLUSION

The protein bioassay of a publically available dataset, the ML techniques were efficient in model computation. Based on high AUC values and a possible BCR rate, we know that the prognostic models are effective. Understanding the biological process and its need is our major concern with a pinch ofmechanism involved. It is possible to create various models using publically available dataset by amending them. Different properties like virulence, metabolic processes as well as bioaccumulation can also be considered for designing models.

REFERENCES

[1]. P. Isomäki, “Cytokines in rheumatoid arthritis,” in *Scientific Basis of Healthcare: Arthritis*, 2012.

[2]. J. A. Mackintosh, A. Stainer, L. J. De Sadeleer, C. Stock, W. A. Wuyts, and E. A. Renzoni, “Rheumatoid arthritis,” *ERS Monogr.*, 2019, doi: 10.1183/2312508X.10014019.

[3]. L. A. Henderson *et al.*, “On the alert for cytokine storm: Immunopathology in COVID-19,” *Arthritis Rheumatol.*, 2020, doi: 10.1002/art.41285.

[4]. J. S. Smolen *et al.*, “Rheumatoid arthritis,” *Nat. Rev. Dis. Prim.*, 2018, doi: 10.1038/nrdp.2018.1.

[5]. D. L. Scott, F. Wolfe, and T. W. J. Huizinga, “Rheumatoid arthritis,” in *The Lancet*, 2010, doi: 10.1016/S0140-6736(10)60826-4.

[6]. W. M. Mikkelsen, T. W. Bunch, and J. J. Calabro, “Arthritis & rheumatism,” *Arthritis Rheum.*, 1981.

[7]. G. S. Firestein and I. B. McInnes, “Immunopathogenesis of Rheumatoid Arthritis,” *Immunity*. 2017, doi: 10.1016/j.immuni.2017.02.006.

[8]. P. J. O’Connor *et al.*, “Arthritis,” in *Medical Radiology*, 2020.

[9]. G. S. Firestein, “Evolving concepts of rheumatoid arthritis,” *Nature*. 2003, doi: 10.1038/nature01661.

[10]. I. B. McInnes and G. Schett, “Cytokines in the pathogenesis of rheumatoid arthritis,” *Nature Reviews Immunology*. 2007, doi: 10.1038/nri2094.

[11]. C. G. Helmick *et al.*, “Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I,” *Arthritis Rheum.*, 2008, doi: 10.1002/art.23177.

[12]. National Rheumatoid Arthritis Society, “The Economic Burden of Rheumatoid Arthritis,” 2010.

[13]. D. M. Lee and M. E. Weinblatt, “Rheumatoid arthritis,” *Lancet*. 2001, doi: 10.1016/S0140-6736(01)06075-5.

[14]. K. E. Barbour, C. G. Helmick, M. Boring, and T. J. Brady, “Vital signs: Prevalence of Doctor-Diagnosed arthritis and arthritis-attributable activity limitation – United States, 2013-2015,” *Morb. Mortal. Wkly. Rep.*, 2017, doi: 10.15585/mmwr.mm6609e1.

[15]. L. A. Coleman and R. Roubenoff, “Arthritis,” in *Encyclopedia of Human Nutrition*, 2012.

[16]. L. K. Stamp and L. G. Cleland, “Rheumatoid arthritis,” in *Optimizing Women’s Health through Nutrition*, 2007.

[17]. V. Majithia and S. A. Geraci, “Rheumatoid Arthritis: Diagnosis and Management,” *American Journal of Medicine*. 2007, doi: 10.1016/j.amjmed.2007.04.005.

[18]. J. U. Scher and S. B. Abramson, “The microbiome and rheumatoid arthritis,” *Nature Reviews Rheumatology*. 2011, doi: 10.1038/nrrheum.2011.121.

[19]. A. Rutherford, E. Nikiphorou, and J. Galloway, “Rheumatoid arthritis,” in *Comorbidity in Rheumatic Diseases*, 2017.

[20]. J. D. Maccuish and N. E. Maccuish, “Chemoinformatics applications of cluster analysis,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2014, doi: 10.1002/wcms.1152.

[21]. H. Masnadi-Shirazi and N. Vasconcelos, “Cost-sensitive boosting,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, doi: 10.1109/TPAMI.2010.71.

[22]. N. Friedman, D. Geiger, M. Goldszmidt, G. Provan, P. Langley, and P. Smyth, “Bayesian Network Classifiers *,” *Mach. Learn.*, 1997, doi: 10.1023/A:1007465528199.

[23]. A. C. Schierz, “Virtual screening of bioassay data,” *J. Cheminform.*, 2009, doi: 10.1186/1758-2946-1-21.